

A Brief Survey of Genetic-Fuzzy Data Mining Techniques

Chun-Hao Chen[†], Tzung-Pei Hong[‡], and Vincent S. Tseng[†]

[†]Department of Computer Science and Information Engineering,
National Cheng-Kung University, Tainan, 701, Taiwan, R.O.C.

[‡] Department of Computer Science and Information Engineering,
National University of Kaohsiung, Kaohsiung, 811, Taiwan, R.O.C.

[‡] Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, 804, Taiwan, R.O.C.
chchen@idb.csie.ncku.edu.tw, tphong@nuk.edu.tw, tsengsm@mail.ncku.edu.tw

Abstract

This paper surveys some genetic-fuzzy data mining techniques for mining both membership functions and fuzzy association rules. The motivation from crisp mining to fuzzy mining will be first described. Three types of genetic-fuzzy data mining approaches are then described according to the utilized methods and different mining problems, including Integrated Genetic-Fuzzy approaches for items with Single Minimum Support (IGFSMS), Integrated Genetic-Fuzzy approaches for items with Multiple Minimum Supports (IGFMMS) and Divide-and-Conquer Genetic-Fuzzy approaches for items with Single Minimum Support (DGFSMS).

Keywords: Data mining, Fuzzy set, Genetic algorithms, Genetic-fuzzy data mining.

1. 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 is a set of items and Y is a single item. 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. For example, assume whenever customers in a supermarket buy bread and butter, they will also buy milk. From the transactions kept in the supermarkets, an association rule such as "Bread and Butter $\rightarrow$ Milk" will be mined out. Most

previous studies focused on binary valued transaction data. Transaction data in real-world applications, however, usually consist of quantitative values. Many sophisticated data-mining approaches have also been proposed to deal with various types of data presents a challenge to workers in this research field [3, 23, 27].

As to fuzzy data mining, many approaches have also been proposed for mining fuzzy association rules [4, 13, 20, 22, 25]. Most of the previous approaches, however, set a single minimum support threshold for all the items or itemsets and identify the relationships among transactions. In real applications, different items may have different criteria to judge their importance and quantitative data may exist. We thus divide the fuzzy data mining approaches into two types, namely single-minimum-support fuzzy-mining (SSFM) [4, 13, 20, 25] and multiple-minimum-support fuzzy-mining (MSFM) problems [22].

In fuzzy data mining, the first thing that it needs to be done is to define appropriate membership functions because they have a critical influence on the final mining results. Generally speaking, membership functions are defined by experts. That is the best approach, absolutely. However, experts may not always do this since the customers' favorites change all the time. Most of the previous fuzzy data mining algorithms thus assume the membership functions are already known. Of course, it is not suitable when we try to apply it to real applications. The algorithms

that can derive both the appropriate membership functions and fuzzy rules automatically are thus developed. In this paper, we made a survey of several algorithms that can mine both appropriate membership functions and fuzzy association rules [5, 6, 7, 8, 10, 11, 12, 17, 18, 19]. We divide them into four different genetic-fuzzy data mining problems according to the utilized approaches and two types of problems in fuzzy data mining, namely Integrated Genetic-Fuzzy approach for items with Single Minimum Supports (IGFSMS) [5, 8, 11, 17, 18, 19], Integrated Genetic-Fuzzy approaches for items with Multiple Minimum Supports (IGFMMS) [7], Divide-and-Conquer Genetic-Fuzzy approaches for items with Single Minimum Supports (DGFSMS) [6, 10, 12] and Divide-and-Conquer Genetic-Fuzzy approaches for items with Multiple Minimum Supports (DGFMMMS) problems as shown in Table 1.

In the Table 1, in the integrated genetic-fuzzy approaches, they encoded all membership functions of all items (attributes) into a chromosome (also called an individual). The genetic algorithms are then used to derive a set of appropriate membership functions according to the designed fitness function. Finally, the best set of membership functions are then used to mine fuzzy association rules. On the other hand, the divide-and-conquer genetic-fuzzy approaches go in different direction. They encoded membership functions of each item into a chromosome. In other words, chromosomes in a population were maintained just for only one item. We will introduce them more in the following sections.

Table 1: The four different genetic-fuzzy data mining problems

	<i>Integrated Approach</i>	<i>Divide-and-Conquer Approach</i>
<i>Single Minimum Support</i>	IGFSMS Problem	DGFSMS Problem
<i>Multiple Minimum Support</i>	IGFMMS Problem	DGFMMMS Problem

The remaining part of this paper is organized as follows. Definition of the genetic-fuzzy mining algorithm and its importance are first described in

Section 2. Some integrated genetic-fuzzy approaches for items with single minimum support are then described in Sections 3. Some integrated genetic-fuzzy approaches for items with multiple minimum supports are then described in Sections 4. Some divide-and-conquer genetic-fuzzy approaches for items with single minimum support are then described in Sections 5. At last, the conclusions are given in Section 6.

2. Definition of the Subject and Its Importance

Data mining techniques have been used in different fields to discover interesting information from databases in recent years. Association rules mining is one of the popular data mining techniques since it was proposed by Agrawal et al. [2]. The basic concept of association rules is that it tries to find a rule format likes “If *milk* was bought, Then *bread* was bought” with a support value and confidence value from transaction dataset. The proposed approach was used for binary data. However, in real applications, quantitative data exists and should also be considered. Srikant *et al.* thus proposed a method for mining quantitative association rules from those with quantitative attributes [23]. Fuzzy sets which was proposed by Zadeh and Goguen in 1965 [26] are also used to mine fuzzy association rules since the derived rules are easy to understand for human. Many algorithms are thus proposed for mining fuzzy association rules [4, 13, 20, 22, 25]. Most of these fuzzy data mining algorithms assume the membership functions are already known. In mining problems, the given membership functions may, however, have a critical influence on the final mining results. The developing of mining algorithms that can mine both appropriate membership functions and fuzzy association rules automatically is thus an important task. There at least have two reasons. The first one is that companies may not always ask experts to define the appropriate membership function because it needs to spend lots of money and time. The second reason is that the favorite things of customers change all the time. Some mechanisms are thus needed to adapt the

membership functions to these changes automatically.

3. Integrated Genetic-Fuzzy approaches for items with Single Minimum Support

In IGFSMS problem, many approaches have been published [5, 8, 11, 17, 18]. Hong et al. proposed a genetic-fuzzy data-mining algorithm for extracting both association rules and membership functions from quantitative transactions [11]. They propose a GA-based framework for searching membership functions suitable for mining problems and then use the final best set of membership functions to mine association rules. The proposed framework is shown in Figure 1.

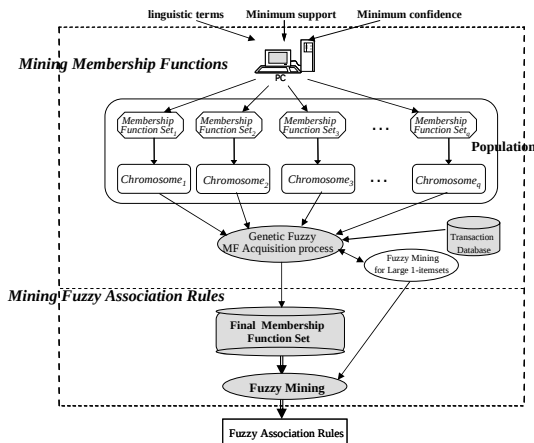


Figure 1: GA-based framework for searching membership functions

The proposed framework consists of two phases, namely mining membership function and mining fuzzy association rules phases. In the first phase, the proposed framework maintains a population of sets of membership functions, and uses the genetic algorithm to automatically derive the resulting one. It first transforms each set of membership functions into a fixed-length string. The chromosome was then evaluated by the number of large 1-itemsets and the suitability of membership functions. The fitness value of a chromosome C_q is then defined as:

$$f(C_q) = \frac{|L_1|}{suitability(C_q)},$$

where $|L_1|$ is the number of large 1-itemsets obtained by using the set of membership functions in C_q . The suitability measure was used to reduce the occurrence of bad types of membership functions. The two bad types of membership function are shown in Figure 2, where the first one is too redundant, and the second one is too separate.

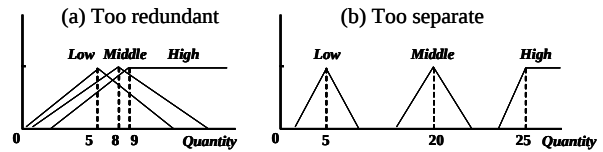


Figure 2: The two bad types of membership functions

Two factors, overlap factor and coverage factor, are used to avoid those shapes. The overlap factor is designed for avoiding the first bad case (too redundant), and the coverage factor is for the second one (too separate). After fitness evaluation, it then chooses appropriate chromosomes 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 the second phase, the final best membership functions are gathered to mine fuzzy association rules. The fuzzy mining algorithm proposed in [14] is adopted to achieve this purpose.

The calculation for large 1-itemsets, however, would still take a lot of time, especially when the database to be scanned could not totally fit into main memory. An enhanced approach, called the cluster-based fuzzy-genetic mining algorithm were thus proposed [8] to speed up the evaluation process and keep nearly the same quality of solutions as that in [11]. The proposed approach also maintains a population of sets of membership functions, and uses the genetic algorithm to automatically derive the resulting one. Before fitness evaluation, the clustering technique was first used to cluster chromosomes. In other words, it used the k -means clustering approach to gather similar chromosomes into groups. The two factors, overlap factor and coverage factor, were then used as attributes for clustering. For example, assume there have ten chromosomes with its coverage and overlap factor and shown in Table 2, where the

column “suitability” represents the pair (coverage factor, overlap factor).

The k -means clustering approach is executed to divide the ten chromosomes into k clusters. In this example, assume the parameter k is set at 3. The three clusters found are shown in Table 3. The representative chromosomes in the three clusters are C_5 (4.37, 0.33), C_4 (4.66, 0) and C_9 (4.09, 8.33).

Table 2: The coverage and the overlap factor of ten chromosomes

C_q	Suitability	C_q	Suitability
C_1	(4, 0)	C_6	(4.5, 0)
C_2	(4.24, 0.5)	C_7	(4.45, 0)
C_3	(4.37, 0)	C_8	(4.37, 0.53)
C_4	(4.66, 0)	C_9	(4.09, 8.33)
C_5	(4.37, 0.33)	C_{10}	(4.87, 0)

Table 3: The three clusters found from the ten chromosomes

Cluster _{i}	Chromosome	Representative chromosome
Cluster ₁	C_1, C_2, C_5, C_8	C_5
Cluster ₂	$C_3, C_4, C_6, C_7, C_{10}$	C_4
Cluster ₃	C_9	C_9

All the chromosomes in a cluster use the number of large 1-itemsets derived from the representative chromosome in the cluster and their own suitability of membership functions to calculate their fitness values. Since the number for scanning a database decreases, the evaluation cost can thus be reduced. In this example, the representative chromosomes are chromosomes C_4 , C_5 , C_9 and it only needs to calculate the number of large 1-itemsets three times. The evaluation results are utilized to choose appropriate chromosomes for mating in the next generation. The offspring membership function sets then undergo recursive evolution until a good set of membership functions has been obtained. Finally, the derived membership functions are used to mine fuzzy association rules.

Besides, Kaya et al. also proposed several genetic-fuzzy data mining approaches to derive membership functions and fuzzy association rules [17, 18, 19]. In [17], the proposed approach tries

to derive membership functions, which can get a maximum profit within an interval of user specified minimum support values and then using the derived membership functions to mine fuzzy association rules. The concept of their idea is shown in Figure 3.

As shown in Figure 3(a), they first used genetic-fuzzy mining approach to derived membership functions from the given quantity transaction database. The final membership functions are then used to mining fuzzy association rules. The concept of maximizing the large itemsets of the given minimum support interval is used to design the fitness function as shown in Figure 3(b). They also extended the approach to mine fuzzy weighted association rules [18]. Furthermore, since the fitness functions are not easily to be defined for optimizing problems, the multi-objective genetic algorithm has thus been developed for this situation [15, 16]. In other words, there always needs more than one criterion to do things well in real applications. A set of solutions, namely non-dominate points (also called Pareto-Optimal Surface) is derived and given to user instead the best one solution by genetic algorithm. Kaya et al. thus proposed a multi-objective genetic algorithm based approaches for mining membership functions, which can generate interesting fuzzy association rules [19]. Three objective functions, namely strongness, interestingness and comprehensibility, are used to find the Pareto-Optimal Surface.

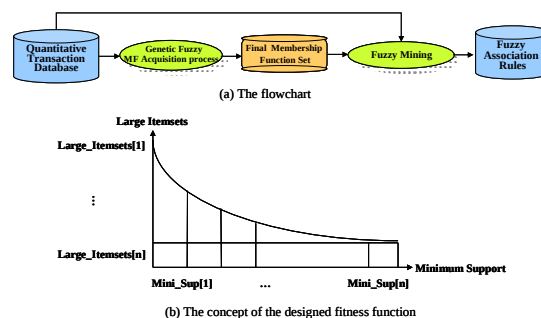


Figure 3: The concept of Kaya's approach

4. Integrated Genetic-Fuzzy approaches for Items with Multiple Minimum Supports

In above subsection, we can know that lots of researches are focused on integrated genetic-fuzzy approaches for items with single minimum support. However, different items may have different criteria to judge their importance. For example, if there have some items and they are expensive, then people usually may not buy it. It is easy to know that the support value of that item is low. Even people may not buy it, manager may also interest on those produces because of their high profit. In such cases, the IGFSMS approaches may not suitable for this problem. In this subsection, we then introduced another genetic-fuzzy data mining approach [7] which extends the approach proposed in [11] to solving this problem. The proposed algorithm combines the clustering, fuzzy and genetic concepts to derive minimum support values and membership functions for items. The final minimum support values and membership functions are then used to mine fuzzy association rules. The genetic-fuzzy mining framework is shown in Figure 4. It is also the first proposed approach for achieving the purpose.

As shown in Figure 4, it can be divided into two phases. The first phase searches for suitable minimum support values and membership functions of items and the second phase uses the final best set of minimum support values and membership functions to mine fuzzy association rules. The proposed framework maintains a population of sets of minimum support values and membership functions, and uses the genetic algorithm to automatically derive the resulting one. It first uses the k -means clustering approach to gather similar items into groups. All items in the same cluster are considered to have similar characteristics and are assigned similar values when a population is initialized. The values (or initialization information) include an appropriate number of linguistic terms for each item, its reasonable membership functions, and a range of its possible minimum support values. It then generates and encodes each set of minimum support values and membership functions into a fixed-length string according to the initialization information.

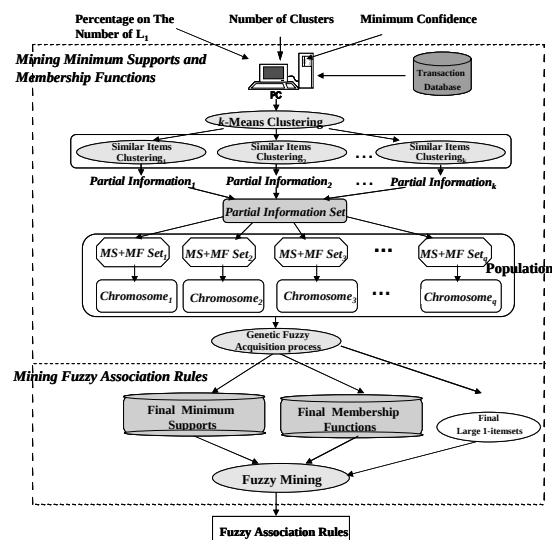


Figure 4: The proposed genetic-fuzzy mining framework for items with multiple minimum supports

In this approach, the minimum support values of the items may be different. It is hard to assign the values. As an alternative, the values can be determined according to the required number of rules. It is, however, very time-consuming to obtain the rules for each chromosome. Usually, a larger number of 1-itemsets will result in a larger number of all itemsets with a higher probability, which will thus usually imply more interesting association rules. The evaluation by 1-itemsets is faster than that by all itemsets or interesting association rules. Using the number of large 1-itemsets can thus achieve a trade-off between execution time and rule interestingness [11].

A criterion should thus be specified to reflect the user preference on the derived knowledge. In this approach, the required number of large 1-itemsets (RNL) is proposed for this purpose. It is the number of linguistic large 1-itemsets that a user wants to get from an item. It can be defined as the number of linguistic terms of an item multiplied by the predefined percentage which reflects users' preference on the number of large 1-itemsets. It is used to reflect the closeness degree between the number of derived large 1-itemsets and the required number of large 1-itemset. For example, assume there are three linguistic terms for an item and the predefined percentage p is set at 80%. The RNL value is then set as $\lfloor 3 \times 0.8 \rfloor$, which is 2. RNL is thus used in the

fitness function to evaluate the goodness of a chromosome. For example, assume there are three linguistic terms for an item and the predefined percentage p is set at 80%. The RNL value is then set as $\lfloor 3 \times 0.8 \rfloor$, which is 2. The fitness function is then composed of the suitability of membership functions and the closeness to the RNL value. The minimum support values and membership functions can thus be derived by GA and are then used to mine fuzzy association rules by a fuzzy mining approach for multiple minimum supports such as the one in [22].

5. Divide-and-Conquer Genetic-Fuzzy approaches for items with Single Minimum Support

The advantages of the integrated genetic-fuzzy approaches are that they are simple and easy to use. However, if there has lots of items, the integrated genetic-fuzzy approaches maybe need lots of time to find the near optimal solution because the length of a chromosome is too long. Recently, the divide-and-conquer strategy has been used in the evolutionary computation community to very good effect. Many algorithms based on it have also been proposed in different applications [1, 9, 21, 24]. The divide-and-conquer strategy thus becomes a good choice to deal with this situation. Hong et al. thus propose an enhanced algorithm that can divide-and-conquer the derivation process of the membership functions for different items [10]. The GA-based framework with the divide-and-conquer strategy is thus proposed for searching for membership functions suitable for the mining problems. The proposed framework is shown in Figure 5.

The proposed framework in Figure 5 is divided into two phases: mining membership functions and mining fuzzy association rules. Assume the number of items is m . In the phase of mining membership functions, it maintains m populations of membership functions, with each population for an item I_j ($1 \leq j \leq m$). Each chromosome in a population represents a possible set of membership functions for that item. The chromosomes in the same population are of the same length. The fitness of each set of

membership functions is evaluated by the fuzzy-supports of the linguistic terms only in large 1-itemsets and by the suitability of the derived membership functions. The offspring sets of membership functions undergo recursive evolution until a good set of membership functions has been obtained. Next, in the phase of mining fuzzy association rules, the sets of membership function for all the items are gathered together and used to mine the fuzzy association rules from the given quantitative database.

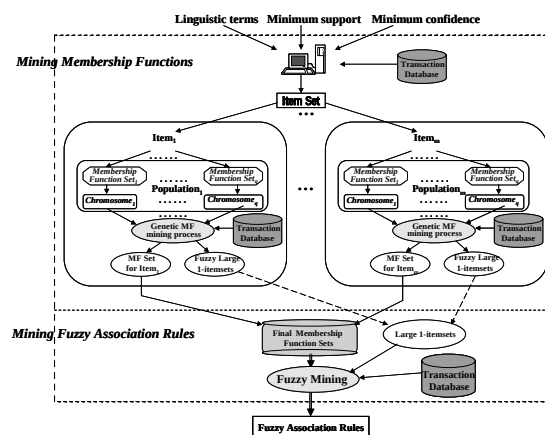


Figure 5: The proposed GA-based framework for fuzzy data mining

Besides, an enhanced approach [6], which combines the clustering and the divide-and-conquer techniques, were also proposed to speed up the evaluation process. The clustering idea is similar to that in *IGFSMS* [8] except that the center value of each membership function is also used as an attribute to cluster chromosomes. The clustering process is thus executed according to the coverage factors, the overlap factors and the center values of chromosomes. For example, assume each item has three membership functions. Totally five attributes including one coverage factor, one overlap factor and three center values, are used to form appropriate clusters. Note that the number of linguistic terms for each item is predefined in the mentioned approaches. It may also be automatically and dynamically adjusted [12].

6. Conclusions

Since association rules was proposed Agrawal et al., lots of researchers devoted themselves to it, no matter what they just improve the existing algorithms, create new ideas or combine the existing one with other techniques. In fact, association rules mining approaches become an important topic. In this article, we thus focus on genetic-fuzzy data mining techniques and lots of literatures are introduced. According to the two skills used to solve the problem and two types of fuzzy mining problems, we divide all algorithms into four classes and three of them are discussed. The remaining one is the divide-and-conquer genetic-fuzzy approaches for items with multiple minimum supports because it still doesn't have any papers yet. Thus, developing algorithms for this part is the first future direction. Besides, how to design algorithms that can derive membership functions for items with taxonomy and fuzzy association rules is also an interesting direction. In general, the future directions on this topic can be very diversity since the rules mining is commonly used in every fields.

7. Acknowledgement

This research was supported by the National Science Council of the Republic of China under contract NSC 96-2213-E-390-003.

8. References

- [1] W. H. Au, Keith C. C. Chan and X. Yao, "A novel evolutionary data mining algorithm with applications to churn prediction," *IEEE Transactions on Evolutionary Computation*, Vol. 7, No. 6, pp. 532-545, 2003.
- [2] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large database," *The 1993 ACM SIGMOD Conference*, Washington DC, USA, 1993.
- [3] R. Agrawal, T. Imielinski and A. Swami, "Database mining: a performance perspective," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, No. 6, 1993, pp. 914-925.
- [4] C. C. Chan and W. H. Au, "Mining fuzzy association rules," *The Conference on Information and Knowledge Management*, Las Vegas, pp. 209-215, 1997.
- [5] C. H. Chen, T. P. Hong and Vincent S. Tseng, "A comparison of different fitness functions for extracting membership functions used in fuzzy data mining," *IEEE Symposium on Foundations of Computational Intelligence*, pp. 550-555, 2007.
- [6] C. H. Chen, T. P. Hong, Vincent S. Tseng, "A modified approach to speed up genetic-fuzzy data mining with divide-and-conquer strategy," Accepted and to appear in *The IEEE Congress on Evolutionary Computation (CEC)*, 2007.
- [7] C. H. Chen, T. P. Hong, Vincent S. Tseng and C. S. Lee, "A genetic-fuzzy mining approach for items with multiple minimum supports," *The IEEE International Conference on Fuzzy Systems*, pp.1-6, 2007.
- [8] C. H. Chen, Vincent S. Tseng and T. P. Hong, "Cluster-based evaluation in fuzzy-genetic data mining," Accepted and to appear in *The IEEE Transactions on Fuzzy Systems*, 2007.
- [9] P. J. Darwen and X. Yao, "Speciation as automatic categorical modularization," *IEEE Transactions on Evolutionary Computation*, Vol. 1, No 2, pp. 101-108, 1997.
- [10] T. P. Hong, C. H. Chen, Y. L. Wu and Y. C. Lee, "Genetic-Fuzzy Data Mining with Divide-and-Conquer Strategy", Accepted and to appear in *IEEE Transactions on Evolutionary Computation*, 2006.
- [11] 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.
- [12] T. P. Hong, C. H. Chen, Y. L. Wu and Vincent S. Tseng, "Finding active membership functions in fuzzy data mining", *The Workshop on Foundations of Data Mining in The Fourth IEEE International Conference on Data Mining*, 2004.
- [13] 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.
- [14] T. P. Hong, C. S. Kuo and S. C. Chi, "Trade-off between time complexity and number of rules for fuzzy mining from quantitative data," *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, Vol. 9, No. 5, pp. 587-604, 2001.
- [15] H. Ishibuchi and T. Yamamoto, "Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining," *Fuzzy Sets and Systems*, Vol. 141, pp. 59-88, 2004.
- [16] Y. Jin, *Multi-objective Machine Learning*, Springer, 2006.
- [17] M. Kaya and R. Alhajj, "A clustering algorithm with genetically optimized membership functions for fuzzy association rules mining," *The IEEE International Conference on Fuzzy Systems*, pp. 881-886, 2003.
- [18] M. Kaya, R. Alhajj, "Genetic algorithms based optimization of membership functions for fuzzy weighted association rules mining," *The International Symposium on Computers and Communications*, Vol. 1, pp. 110-115, 2004.
- [19] M. Kaya, R. Alhajj, "Integrating multi-objective genetic algorithms into clustering for fuzzy association rules mining," *The Fourth IEEE International Conference on Data Mining*, pp. 431-434, 2004.
- [20] C. Kuok, A. Fu and M. Wong, "Mining fuzzy association rules in databases," *SIGMOD Record*, Vol. 27, No. 1, pp. 41-46, 1998.
- [21] V. R. Khare, X. Yao, B. Sendhoff, Y. Jin and H. Wersing, "Co-evolutionary modular neural networks for automatic problem decomposition," *The 2005 IEEE Congress on Evolutionary Computation*, Vol. 3, pp. 2691-2698, 2005.
- [22] 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.
- [23] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," *The 1996 ACM SIGMOD International Conference on Management of Data*, Monreal, Canada, June 1996, pp. 1-12.
- [24] X. Yao, "Adaptive divide-and-conquer using populations and ensembles," *The 2003 International Conference on Machine Learning and Application*, pp. 13-20, 2003.
- [25] 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.
- [26] L. A. Zadeh, "Fuzzy sets," *Information and Control*, Vol. 8, No. 3, 1965, pp. 338-353.
- [27] Z. Zhang, Y. Lu and B. Zhang, "An effective partitioning-combining algorithm for discovering quantitative association rules," *The Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 261-270, 1997.