2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

Genetic-Fuzzy Mining with Type-2 Membership Functions

Yu Li, Chun-Hao Chen, Tzung-Pei Hong and Yeong-Chyi Lee

***Abstract*—In this paper, a type-2 genetic-fuzzy mining algorithm is proposed for mining a set of type-2 membership functions for mining fuzzy association rules. It first encodes the type-2 membership functions of each item into a chromosome. The quantitative transactions are then transformed into fuzzy values according to the type-2 membership functions. Each chromosome is then evaluated by the number of large 1-itemsets and the suitability factor. The suitability factor consists of three sub-factors - coverage, overlap and difference which are used to avoid three bad types of membership functions. Experiments on a simulated dataset are also conducted to show the effectiveness of the proposed approach.**

I. Introduction

Association rule mining is most commonly used in attempts to induce relationship between items from transaction data [1]. Based on the well-known approach, Apriori algorithm, a variety of mining approaches have also been proposed [4, 10, 23, 24]. Since transaction usually consists of quantitative values in real-world applications, fuzzy data mining algorithms have thus been proposed for handling quantitative transactions and mining fuzzy association rules [4, 13, 14, 17, 18, 27]. In those approaches, they first transforms quantitative transactions into fuzzy values. Then, an Apriori-like approach has been used to finding large itemsets and rules.

However, fuzzy data mining approaches assumed the membership functions to be known in advance. The given membership functions may have a critical influence on the final mining results. Recently, various genetic-fuzzy mining (GFM) approaches have been proposed to derive appropriate membership functions and mining fuzzy association rules [2, 5, 12, 15, 16]. Kaya et al. proposed a GA-based approach to derive a predefined number of membership functions to obtain the maximum profit within the user specified interval of minimum supports [15]. Hong et al. also proposed a genetic-fuzzy data-mining algorithm for extracting both association rules and membership functions from quantitative transactions [12]. Alcalá-Fdez et al. then modified this approach to propose an enhanced approach based on the 2-tuples linguistic representation model [2]. In addition, multi-objective genetic-fuzzy mining approaches have also been proposed [5, 16].

In type-1 fuzzy logical systems, four sources of uncertainties are described in [20]. Thus, type-2 fuzzy sets are able to deal with such uncertainties. Some approaches have been proposed for various applications by utilizing type-2 fuzzy sets, e.g., decision making [3, 21], classification [19], and type-2 fuzzy logic systems [8]. Since finding appropriate membership functions and rules is an optimization task, many genetic fuzzy systems have also been proposed [3], and more information can be found in [6, 9].

In GFM, it also has uncertainties that mentioned in [20]. Thus, in this paper, we propose a type-2 genetic-fuzzy mining (T2GFM) approach for deriving type-2 membership functions for items and mining fuzzy association rules. It first encodes type-2 membership functions of each item into a chromosome. The quantitative transactions are then transformed into fuzzy values according to type-2 membership functions. Each chromosome is evaluated by number of large 1-itemsets and suitability factor. The suitability factor consists of coverage, overlap, and difference factors that are used to avoid deriving bad types of membership functions. Experiments on a simulated dataset are made to show the effectiveness of the proposed approach.

II. The Proposed Type-2 GFM Framework

This section proposes the type-2 genetic-fuzzy mining framework for mining type-2 membership functions that are suitable for items to derive type-2 fuzzy association rules. The proposed framework is shown in Figure 1.

The proposed framework consists of two phases, namely mining type-2 membership functions (MF) and mining type-2 fuzzy association rules phases. The first phase transforms the type-2 membership functions of items into a fixed-length string, which is known as a chromosome (individual). The number of large 1-itemsets and suitability factor are calculates for evaluating fitness values. Genetic operations are then executing for generating more suitable solutions. In the second phase, the derived final type-2 membership functions are utilized for mining type-2 fuzzy association rules.

Yu Li is with the Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan (e-mail: m023040059@student.nsysu.edu.tw).

Chun-Hao Chen is with the Department of Computer Science and Information Engineering, Tamkang University, Taipei, Taiwan (e-mail: chchen@mail.tku.edu.tw).

Tzung-Pei Hong is with the Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, and the Department of Computer Science and Engineering, National Sun Yat-sen University, Taiwan (corresponding author; e-mail: tphong@nuk.edu.tw).

Yeong-Chyi Lee is with the Department of Information Management, Cheng Shiu University, Kaohsiung, Taiwan (e-mail: yeongchyi@csu.edu.tw).

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

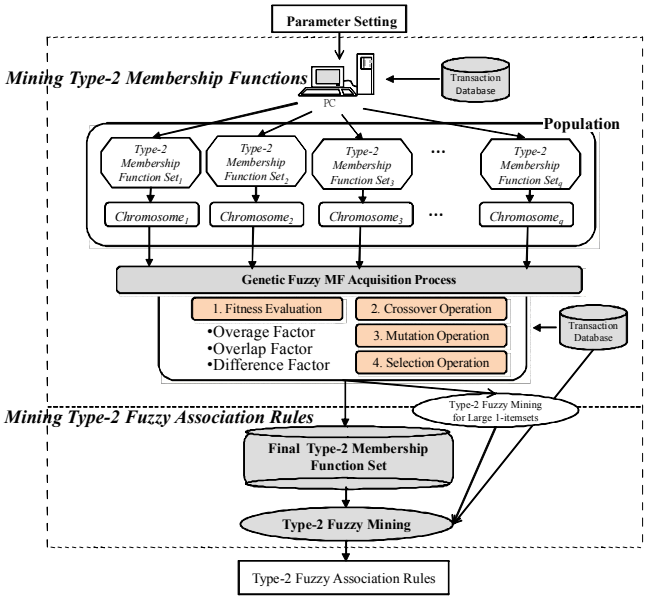

Figure 1: Type-2 genetic-fuzzy mining framework.

III. Components of Proposed Approach

In this section, the details of the components in the proposed approach are described, including chromosome representation, initial population, fitness evaluation and selection, and genetic operations.

A. Chromosome Representation

It is important to encode type-2 membership functions as string representation for GAs to be applied. Several possible encoding approaches have been described in [7, 25]. In order to effectively encode the associated membership functions, two parameters, namely the center abscissa (c_{jk}) and half the spread (w_{jk}) of fuzzy region R_{jk}, are used to represent a type-1 membership function in Parodi and Bonelli [22]. Since type-2 membership functions have upper and lower membership functions, two parameters are not enough, obviously. Thus, an extra parameter, namely d_{jk}, that indicates the difference of half the spread of upper and lower membership functions. In fuzzy rule mining, type-2 membership functions applied to an item are assumed to be isosceles-triangle functions, as shown in Figure 2.

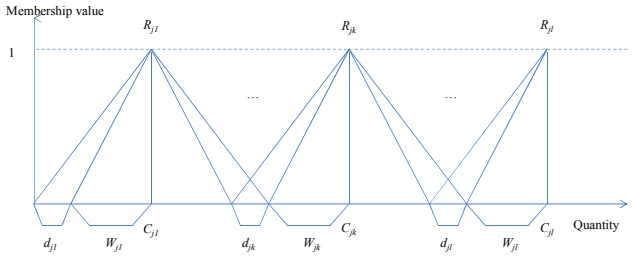

Figure 2: Type-2 membership functions of item I_j.

In Figure 2, R_{jk} denotes the type-2 membership function of the k-th linguistic term of item I_j, c_{jk} indicates the center abscissa of fuzzy region R_{jk}, w_{jk} represents half the spread of the fuzzy region R_{jk}, and d_{jk} means the difference of half the spread of upper and lower membership functions. Based on Parodi and Bonelli, we then represent each type-2 membership function as a tuple (c, w, d), and thus all such tuples for a certain item are concatenated to represent its type-2 membership functions. The set of type-2 membership functions MF_1 for the first item I_1 is then represented as a substring of $c_{11}w_{11}d_{11}\ldots c_{1|I_1|}w_{1|I_1|}d_{1|I_1|}$, where $|I_1|$ is the number of linguistic terms of I_1. The entire set of membership functions is then encoded by concatenating substrings of MF_1, MF_2, ..., MF_j. Since c, w and d are both numeric values, a chromosome is thus encoded as a fixed-length real-number string rather than a bit string.

B. Initial Population

A genetic algorithm requires a population of feasible solutions to be initialized and updated during the evolution process. As mentioned above, each individual within the population is a set of isosceles-triangular membership functions. Each membership function corresponds to a linguistic term in a certain item class. The initial set of chromosomes is randomly generated according to the given dataset with some constraints of forming feasible membership functions.

C. Fitness Evaluation and Selection

In order to derive a good set of membership functions from an initial population, the genetic algorithm selects *parent* type-2 membership function sets with high fitness values for mating. An evaluation function is then used to qualify the derived type-2 membership function sets. The performance of the 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 type-2 membership functions. In this paper, we thus propose an evaluation function according to the one used in our previous paper [12], as shown in Equation (1).

$$f(C_q) = \frac{|L_1^{C_q}|}{suitability(C_q)}, \qquad (1)$$

where the $|L_1^{C_q}|$ is the number of large 1-itemsets of chromosome C_q, and the $suitability(C_q)$ can reduce the occurrence of the three bad kinds of membership functions. It consists of three factors, namely overlap, coverage, and difference factors. The first one is used to reduce too redundant, and the second one is used to avoid too separate. Since type-2 membership functions have upper and lower membership functions, the third one is designed to avoid them too close and separate. The suitability of type-2 membership functions in a chromosome C_q is thus defined as Equation (2):

$$suitability(C_q) = \sum_{j=1}^{m}[\, overlapFactor(C_{qj}) + coverageFactor(C_{qj}) + differenceFactor(C_{qj})], \quad (2)$$

where the overlap factor of the type-2 membership functions for an item I_j in the chromosome C_q is defined as Equation (3):

$$overlapFactor(C_{qj}) = \sum_{k \neq i}[max((\frac{overlap(R_{jk}, R_{ji})}{min(w_{jk}, w_{ji})}), 1) - 1], \quad (3)$$

where the overlap ratio $overlap(R_{jk}, R_{ji})$ of two upper

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 upper membership functions for an item I_j is defined as the coverage range of the functions divided by the average value of maximum quantity of that item class in the transactions. The larger the coverage ratio is, the better the derived type-2 membership functions are. Thus, the coverage factor of the upper membership functions for an item I_j in the chromosome C_q is defined as Equation (4):

$$coverageFactor(C_{qj}) = \frac{1}{\frac{range(R_{j1}, ..., R_{jl})}{max(I_j)}} \quad (4)$$

where $range(R_{j1}, R_{j2}, \ldots, R_{jl})$ is the coverage range of the membership functions, l is the number of membership functions for I_j. Note that the overlap factor is designed to avoiding the first bad case of being "too redundant", while the coverage factor is to avoid the second one of being "too separate".

The difference factor of upper and lower membership functions of an item I_j is calculated according to their difference of half the spans. Thus, the difference factor of the upper and lower membership functions for an item I_j in the chromosome C_q is defined as Equation (5):

$$differenceFactor(C_{qj}) = \begin{cases} d_{jk} / w_{jk}, if\ d_{jk} > w_{jk}; \\ 1, if\ d_{jk} = w_{jk}; \\ w_{jk} / d_{jk}, if\ d_{jk} < w_{jk}; \end{cases} \quad (5)$$

where w_{jk} represents half the spread of the fuzzy region R_{jk}, and d_{jk} means the difference of half the spread of upper and lower membership functions.

D. Genetic Operators

Genetic operators are important to the success of specific GA applications. Two genetic operators, the *max-min-arithmetical (MMA) crossover* proposed in [11] and the *one-point mutation*, are used in the proposed approach. The *one-point* mutation operator will create a new type-2 membership function by adding a random value ε to the center, to the spread of an existing linguistic term, say R_{jk}, or to the difference of half the spans of upper and lower membership functions. Assume that c, w and d represent the center, the spread and difference of R_{jk}. The center, the spread, or the difference of the newly derived membership function will be changed to $c + \varepsilon$, $w + \varepsilon$ or $d + \varepsilon$ by the mutation operation. The selection strategy used in the proposed approach can be the elitist or the roulette-wheel strategy.

IV. Proposed Algorithm

In this section, the proposed *T2GFM* is described.

Proposed T2GFM:

INPUT: The input data includes n quantitative transaction data, a set of m items, each with a number of linguistic terms. The parameters consist of a support threshold α, a confidence threshold λ, the population size P, and a split number of point of a type-2 membership function s.

OUTPUT: A set of type-2 membership functions and type-2 fuzzy association rules.

STEP 1: Randomly generate a population of P individuals; each individual is a set of type-2 membership functions for items in transactions.

STEP 2: Encode each set of type-2 membership functions into a string representation.

STEP 3: Calculate the fitness value of each chromosome by the following substeps:

SUBSTEP 3.1: Transform the quantitative value v_{ij} of each transaction datum D_i (i =1 to n) for each encoded group name I_j into fuzzy values f_{ijl}^{lower} and f_{ijl}^{upper} represented as $\left(\frac{[f_{ij1}^{lower}, f_{ij1}^{upper}]}{R_{j1}} + \frac{[f_{ij2}^{lower}, f_{ij2}^{upper}]}{R_{j2}} + ... + \frac{[f_{ijh}^{lower}, f_{ijh}^{upper}]}{R_{jh}} \right)$, using the corresponding type-2 membership functions represented by the chromosome, where R_{jl} is the l-th fuzzy region of item I_j, $1 \le l \le h$, f_{ijl}^{lower} and f_{ijl}^{upper} is v_{ij}'s fuzzy membership value in region R_{jl}, and l is the number of linguistic terms for I_j.

SUBSTEP 3.2: Reduce type-2 fuzzy values into type-1 fuzzy values by centroid type-reduction method with the given split number of point of a type-2 membership function s. Thus, the f_{ijl}^{lower} and f_{ijl}^{upper} are reduced to a fuzzy value f_{ijl}.

SUBSTEP 3.3: Calculate the scalar cardinality of each fuzzy region R_{jl} in the transactions as:

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

SUBSTEP 3.4: Check whether the value of each fuzzy region R_{jl} is larger than or equal to the predefined minimum support value α. If the value of a fuzzy region R_{jl} is equal to or greater than the minimum support value, put it in the large 1-itemsets (L_1). That is: $L_1 = \{ R_{jl} \mid count_{jl} \ge \alpha, 1 \le j \le m, 1 \le l \le h \}$.

SUBSTEP 3.5: Calculate the suitability of each chromosome by Equation (2).

STEP 4: Set the fitness value of each chromosome as its number of large 1-itemsets ($|L_{1q}|$) divided by its suitability ($suitability(C_q)$) as defined in Equation (1).

STEP 5 to 7: Execute crossover and mutation operations on the population. Then, use the *selection* criteria to choose individuals for the next generation.

STEP 8: If the termination criterion is not satisfied, go to Step 3; otherwise, output the set of membership functions with the highest fitness value for mining fuzzy rules.

V. Experimental Results

In this section, the results of the experiments to show the

performance of the T2GFM are described. The experiments were implemented in Java on a personal computer with Intel Core i5 CUP 661 @ 3.33GHz and 1.8GB RAM. A simulation dataset contains 64 items and 10000 transactions was used in the experiments. In the data set, the number of purchased items in transactions was first randomly generated, and the purchased items and their quantities in each transaction were then generated. An item could not be generated twice in a transaction. 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 Herrera *et al.* [11], and the minimum support was set at 0.04 (4%).

Firstly, experiments were made to show the convergence of the proposed approach. The average results of five times are show in Figure 3.

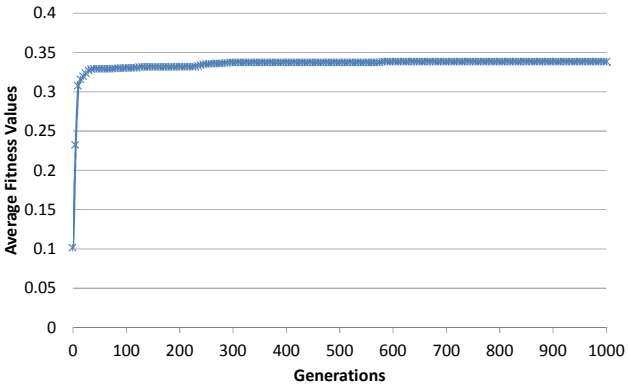

Figure 3: The convergence of T2GFM

Figure 3 shows that the average fitness values are increasing along with the increasing of generations. After 600 generations, it converges to a certain value. The final type-2 membership functions of two items are shown in Figure 4.

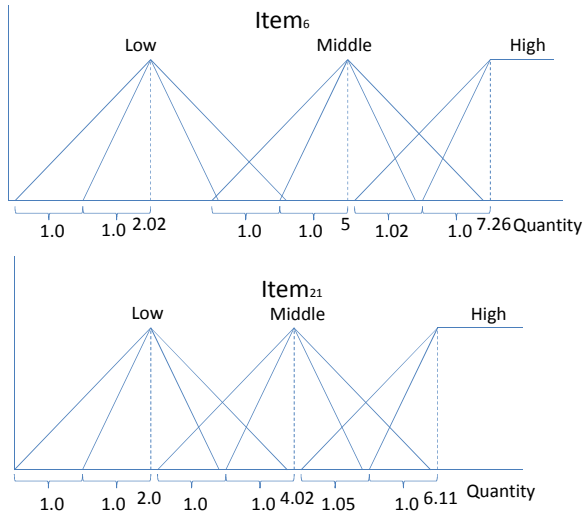

Figure 4: The derived type-2 membership functions

Figure 4 shows that the three bad kinds of membership functions don't appear in the final results. It thus show that the proposed approach can actually get more appropriate membership functions.

VI. Conclusion and Future Work

In this study, we have proposed a type-2 genetic-fuzzy mining algorithm for finding appropriate type-2 membership functions and fuzzy association rules. It first encodes type-2 membership functions for all items into chromosomes. Then, in order to derive good type-2 membership functions, the suitability of a chromosome that consists of three sub-factors is designed. The fitness value of an individual is then evaluated by the number of large 1-itemsets and its suitability value. Experimental results on a simulation dataset show the merits of the proposed T2GFM. However, the proposed type-2 GFM is just a beginning. In the future, we will enhance it to more complex problems, such as T2GFM with taxonomy, T2GFM with multiple minimum supports, and so on.

Acknowledgment

This research was supported by the National Science Council of the Republic of China under grant NSC 102-2221-E-032 -056.

References

[1] R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," *The International Conference on Very Large Databases*, pp. 487-499, 1994.

[2] J. Alcala-Fdez, R. Alcala, 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.

[3] D. Bernardo, H. Hagras, and E. Tsang, " A genetic type-2 fuzzy logic based system for the generation of summarised linguistic predictive models for financial applications," Soft Computing, Vol. 17, No 12, pp. 2185-2201, 2013.

[4] 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.

[5] C. H. Chen, T. P. Hong, Vincent S. Tseng, "A SPEA2-based genetic-fuzzy mining algorithm," *The IEEE International Conference on Fuzzy Systems*, 2010.

[6] F. Herrera, "Genetic Fuzzy Systems: Taxonomy, Current Research Trends and Prospects," *Evolutionary Intelligence*, Vol. 1, pp. 27-46, 2008.

[7] O. Cordón, 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] M. H. Fazel Zarandi and R. Gamasaee, "A type-2 fuzzy system model for reducing bullwhip effects in supply chains and its application in steel manufacturing," *Scientia Iranica*, Vol. 20, No 3, pp. 879–899, 2013.

[9] M. Fazzolari, R. Alcalá, Y. Nojima, H. Ishibuchi, F. Herrera, "A review of the application of Multi-Objective Evolutionary Systems: Current status and further directions," *IEEE Transactions on Fuzzy Systems*, Vol. 21, No. 1, pp. 45-65, 2013.

[10] 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.

[11] 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.

[12] 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.

[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, 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] P. Melin and O. Castillo, " A review on the applications of type-2 fuzzy logic in classification and pattern recognition," *Expert Systems with Applications*, Vol. 40, No 13, pp. 5413–5423, 2013.

[20] J. M. Mendel and R. I. Bob John, "Type-2 fuzzy sets made simple," *IEEE Transactions on Fuzzy Systems*, Vol. 10, No. 2, pp. 117-127, 2002.

[21] Y. Nojima and H. Ishibuchi, "Incorporation of user preference into multiobjective genetic fuzzy rule selection for pattern classification problems," *Artificial Life and Robotics*, Vol. 14, pp. 418-421, 2009.

[22] 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.

[23] 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.

[24] R. Srikant and R. Agrawal, "Mining generalized association rules," The International Conference on Very Large Data Bases, pp. 407-419, 1995.

[25] 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.

[26] R. R. Yager, "Fuzzy subsets of type II in decisions," *J. Cybern.*, vol. 10, pp. 137–159, 1980.

[27] 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.