Soft Comput
DOI 10.1007/s00500-016-2266-z

FOCUS

Genetic algorithm with a structure-based representation for genetic-fuzzy data mining

Chuan-Kang Ting[1] · Ting-Chen Wang[1] · Rung-Tzuo Liaw[1] · Tzung-Pei Hong[2,3]

Abstract Mining association rules is an important data mining technology aiming to find the relationship among items in the databases. Genetic-fuzzy data mining uses evolutionary algorithm, such as genetic algorithm (GA), to optimize the membership functions for mining fuzzy association rules, and has received considerable success. The increase in data, especially in big data analytics, poses serious challenges to GA in the effectiveness and efficiency of finding appropriate membership functions. This study proposes a GA for enhancing genetic-fuzzy mining of association rules. First, we design a novel chromosome representation considering the structures of membership functions. The representation facilitates arrangement of membership functions. Second, this study presents two heuristics in the light of overlap and coverage for removing inappropriate arrangement. A series of experiments is conducted to examine the proposed GA on different amounts of transactions. The experimental results show that GA benefits from the proposed representation. The two heuristics help to explore the structures of membership functions and achieve significant improvement on GA in terms of solution quality and convergence speed. The satisfactory outcomes validate the high capability of the proposed GA in genetic-fuzzy mining of association rules.

Keywords Genetic algorithm · Chromosome representation · Membership function · Fuzzy association rules · Genetic-fuzzy data mining

1 Introduction

With the rapid increase in data amount, data analytics emerges to leverage and transform the information hidden in the data into explicit knowledge. Data mining aims to explore the data for valuable information and plays a key role in data analytics (Fayyad et al. 1996a, b; Piatesky-Shapiro et al. 1996). Several data mining technologies have been proposed to discover knowledge for different purposes, such as classification (Chang and Lin 2011), clustering (Wagstaff et al. 2001), and association rules (Hong et al. 2008).

Mining association rules is a significant technology of data mining. It attempts to find the relationship among items, also known as homogeneous group or affinity group, from the database and has achieved many successful stories, e.g., prediction of customer's behavior in Walmart. The Apriori algorithm is well known for mining association rules (Agrawal and Srikant 1994). This method establishes association rules based on frequent itemsets according to a user-defined minimum confidence. Srikant and Agrawal (1996) extended the rules into quantitative association rules, which can deal with data with quantitative values or categories. They presented a method similar to the Apriori

Communicated by C.-H. Chen.

✉ Chuan-Kang Ting
ckting@cs.ccu.edu.tw

Ting-Chen Wang
wtc102p@cs.ccu.edu.tw

Rung-Tzuo Liaw
lrt101p@cs.ccu.edu.tw

Tzung-Pei Hong
tphong@nuk.edu.tw

[1] Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan

[2] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan

[3] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan

Published online: 18 July 2016

algorithm but requiring an additional preprocess of data discretization. Recently, fuzzy set theory is introduced to different aspects of association rules (Chan and Au 1997; Hong and Lee 1996; Kuok et al. 1998). In particular, Hong et al. (1999, 2006, 2008) proposed the fuzzy transaction data mining algorithm (FTDA) by adopting fuzzy sets to analogue the values of original data. In the FTDA, quantitative values of data are transformed into fuzzy values according to membership functions. The results are known as fuzzy association rules. The FTDA holds the advantages in extension, tolerance, and suitability for nonlinear systems.

In view of its recognized capability in global search, genetic algorithm (GA) is commonly used to find the optimal setting for membership functions. Although GA has obtained some promising results, there exist two key issues at the design of GA for fuzzy association rule mining. First, the common chromosome representation for genetic-fuzzy data mining comprises the vertex positions but omits the structure information of membership function. Second, the relationship between membership functions involves overlap and coverage. However, it is ordinarily not considered in the design of GA.

This study aims to address the above two issues and improve GA on mining fuzzy association rules. Specifically, we propose a chromosome representation considering the structure of membership functions and their relationship. Using the new representation, two heuristics are developed to eliminate inappropriate arrangement of membership functions and reduce the search space. These enhancements are helpful for the genetic-fuzzy system to mine association rules among transactions. This study carries out a series of experiments to examine the performance of the proposed GA in different data scales.

The remainder of this paper is organized as follows. Section 2 introduces association rules and fuzzy association rules. Section 3 sheds light on the proposed GA. The experimental results are presented and discussed in Sect. 4. Finally, we draw conclusions and recommend the directions for future work in Sect. 5.

2 Mining fuzzy association rules

Association rules infer the coexistence of items and play an important role in data analytics (Agrawal et al. 1993). An association rule can be generally expressed by the following mathematical form:

$$X \rightarrow Y, \tag{1}$$

where X and Y are sets of items. In the famous example of Walmart, the association rule $\{bread, cheese\} \rightarrow \{milk\}$ infers "If buying *bread* and *cheese*, then buying *milk*".

Given an itemset $I = \{I_1, \ldots, I_m\}$ with m items and a database $D = \{T_1, \ldots, T_n\}$ with n transactions $T_k \subseteq I$, an association rule $X \rightarrow Y$ represents the possibility: if $X \subseteq T_i$ then $Y \subseteq T_i$. The importance of an association rule is ordinarily measured by two metrics: support and confidence.

Definition 1 (*Support*) The support of association rule $X \rightarrow Y$ is defined by the probability that X and Y coexist, i.e.,

$$\text{Support}\,(X \rightarrow Y) = \mathcal{P}\,(X \cup Y). \tag{2}$$

Definition 2 (*Confidence*) The confidence of association rule $X \rightarrow Y$ is defined by the probability that Y exists given X exists, i.e.,

$$\text{Confidence}\,(X \rightarrow Y) = \mathcal{P}\,(Y|X) = \frac{\mathcal{P}\,(X \cup Y)}{\mathcal{P}\,(X)}. \tag{3}$$

The Apriori algorithm (Agrawal and Srikant 1994) is well known for mining association rules. Given the minimum support $support_{\min}$, the Apriori algorithm progressively selects the large itemsets $L = \{L_1, \ldots, L_m\}$ from all candidate itemsets $C = \{C_1 \;\ldots, C_m\}$. The items in the large itemsets obtained are then arranged as association rules according to a predetermined minimum confidence $confidence_{\min}$. The time complexity of the Apriori algorithm is $O\,(2^m)$ in that the size of candidate itemsets C is exponential to the number of items m in the itemset I.

Hong et al. (1999, 2006) extended association rules by considering quantity of items and introducing the notion of fuzzy sets. They proposed constructing fuzzy association rules using the fuzzy support and fuzzy confidence based on the membership functions for all m items in the database D.

Let $\Upsilon_{j,k}$ denote the fuzzy region of k-th membership function for item I_j. The fuzzy membership value $f_{j,k}^{(i)}$ of region $\Upsilon_{j,k}$ is determined by the quantity $v_j^{(i)}$ of the j-th item in the i-th transaction T_i.

Definition 3 The fuzzy support of region $\Upsilon_{j,k}$ is calculated by

$$\text{FuzzySupport}\left(\Upsilon_{j,k}\right) = \frac{1}{n}\sum_{i=1}^{n} f_{j,k}^{(i)}. \tag{4}$$

If a fuzzy region $\Upsilon_{j,k}$ has FuzzySupport $\left(\Upsilon_{j,k}\right) \geq support_{\min}$, it is then included to the large 1-itemset L_1 like the Apriori algorithm.

For a set of fuzzy regions $\Upsilon = \{\Upsilon_1, \ldots, \Upsilon_p\}$, its fuzzy value in transaction T_i is given by the intersection of membership values $f_{\Upsilon_k}^{(i)}$, i.e.,

$$f_{\Upsilon}^{(i)} = \bigcap_{k=1}^{p} f_{\Upsilon_k}^{(i)}, \tag{5}$$

which is commonly implemented by taking the minimum function as the intersection operator. Hence, the fuzzy value of Υ can be computed by

$$f_{\Upsilon}^{(i)} = \min_{1 \le k \le p} f_{\Upsilon_k}^{(i)}. \tag{6}$$

Based on the above equation, the fuzzy support of Υ is defined as follows.

Definition 4 (*Fuzzy support*) The fuzzy support of Υ is defined by

$$\text{FuzzySupport}\,(\Upsilon) = \frac{1}{n} \sum_{i=1}^{n} f_{\Upsilon}^{(i)}. \tag{7}$$

If the fuzzy support of Υ is greater than the minimum support $support_{\min}$, Υ is then added to the large p-itemset L_p. The collection of large itemsets continues until $L_p = \emptyset$.

The fuzzy association rules are built upon the large itemsets and their fuzzy confidence. Given a large p-itemset $L_p = \{\Upsilon_1, \ldots, \Upsilon_p\}$, the candidate rules have the following form:

$$X \to Y, \quad X, Y \subset L_p \text{ and } X \cap Y = \emptyset, \tag{8}$$

where X and Y are two disjoint subsets of L_p representing the antecedent and consequent, respectively. The candidate rules need to be further examined to see if they satisfy the minimum confidence.

Definition 5 (*Fuzzy confidence*) The fuzzy confidence of a candidate rule $R : X \to Y$ associated with L_p is defined by

$$\text{FuzzyConfidence}\,(R) = \frac{\text{FuzzySupport}(X \cup Y)}{\text{FuzzySupport}(X)}. \tag{9}$$

If a candidate rule has fuzzy confidence greater than or equal to the minimum confidence, then it is qualified as a fuzzy association rule. Note that all possible candidate rules for the large itemset L_p need to be examined.

As above indicated, the membership functions exert a significant influence over fuzzy association rules. Optimization of membership functions serves as a key task in fuzzy association rule mining. Several studies propose using evolutionary algorithms to optimize the parameters of membership functions. Hong et al. (1999, 2006) developed GAs for fuzzy transaction data mining. Experimental results show that GAs can find appropriate settings for membership functions. In addition, they presented the divide-and-conquer strategy to improve the efficiency of genetic-fuzzy data mining (Hong et al. 2008). Cai et al. (2010) adopted nonlinear particle swarm optimization (PSO) for mining fuzzy association rules. In the PSO algorithm, a particle represents all the parameters of membership functions. Mishra et al. (2011) applied PSO to mine fuzzy frequent patterns from gene expression data. The initial population is generated by the frequent pattern growth method, and the fitness is defined as the mean-squared residue score.

Aside from membership functions, the minimum support and minimum confidence are two important parameters to be determined. Asadollahpoor-Chamazi et al. (2013) devised an adaptive strategy for setting the minimum support threshold in a cluster-based GA. Chen et al. (2013) presented a fuzzy coherent rule mining algorithm, in which the fuzzy coherent rules are defined by four conditions that are used to replace the minimum support. Instead of generating rules, Alcalá-Fdez et al. (2011) used GA to select fuzzy association rules in the fuzzy rule-based classification system. Lee et al. (2014, 2016) devised a radio frequency identification-based recursive process mining system, where GA evolves a population of fuzzy rules rather than membership functions to find the relations of process parameters and product quality. The cost of calculating the support values is another issue at fuzzy association rule mining. To reduce the computational cost, Chen et al. (2008) proposed dividing the population into several clusters and calculating only the support value of representative chromosome of each cluster.

Some studies consider multiple objectives in mining of fuzzy association rules and formulate it as a multi-objective optimization problem (Fazzolari et al. 2013). For example, Qodmanan et al. (2011) took account of support and confidence, and Meng and Pei (2012) included linguistic quantifier and truth in the fitness function. The proposed methods remove the need to specify the minimum support and minimum confidence. Minaei-Bidgoli et al. (2013) considered more objectives in fuzzy association rule mining: support, confidence, comprehensibility (Wakabi-Waiswa and Baryamureeba 2008), and interestingness (Liu et al. 2000). Their multi-objective evolutionary algorithm adopts Michigan approach, which encodes an association rule as a single chromosome. Antonelli et al. (2014) developed a multi-objective evolutionary learning scheme for fuzzy rule-based classifiers with two objectives, namely accuracy and rule complexity. In the course of evolution, the association rules and membership function parameters are optimized concurrently. Rudziňski (2016) considered the objectives concerning the root-mean-square error and interpretability, and adopted Pittsburgh approach for representation of association rules.

3 Optimization of membership functions

The membership functions play an essential role in mining fuzzy association rules since they determine the fuzzy value of a quantity. The setting for membership functions is key to the expressiveness and interpretability of fuzzy sets and therefore affects the validity of fuzzy association rules. Finding appropriate membership functions for fuzzy association rule mining has been formulated as an optimization problem (Hong et al. 2008).

This study develops a GA using a novel chromosome representation for fuzzy association rule mining. First, the parameters of membership functions are represented as a chromosome. In this study, we propose a novel chromosome representation that additionally considers the structure of membership functions. Then, the GA generates a population of chromosomes as the basis of evolution. By mimicking the nature evolution, the evolutionary process of GA involves parent selection, crossover, mutation, and survival selection. More details about the proposed GA are described below.

3.1 Representation

In genetic-fuzzy mining of association rules, GA is usually used to find appropriate membership functions for a given item. The common membership functions include triangular, trapezoidal, Gaussian, and bell functions; in particular, the triangular function is most widely used due to its simplicity and effectiveness. This study thus adopts the triangular function for membership functions, while other shape functions are also applicable. The fuzzy region associated with a triangular membership function is parameterized by the three vertices of a triangle. Accordingly, for an item with ℓ linguistic terms, a chromosome can be represented by 3ℓ real-valued genes to determine the ℓ membership functions. Let $c_{i,j}$ be the j-th parameter of i-th membership function for a given item. A membership function must satisfy the following two constraints:

$$c_{i,1} \leq c_{i,2} \leq c_{i,3} \tag{10}$$

$$c_{1,2} \leq c_{2,2} \leq \cdots \leq c_{l,2}. \tag{11}$$

The first constraint maintains the triangular shape, and the second constraint ensures the order of linguistic terms. Figure 1 illustrates a legal membership function that satisfies the above constraints. Note that the genetic operators may bring about chromosomes violating the constraints. The genes of these illegal chromosomes will then be reordered to fix the problem.

This study proposes a novel chromosome representation by considering the structure, in addition to the parameters, of membership functions. That is, a chromosome is composed of two parts: (1) parameters and (2) structure type.

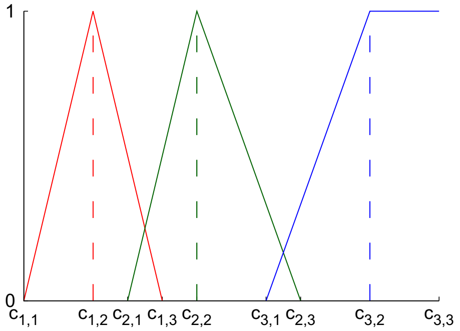

Fig. 1 Example of three membership functions

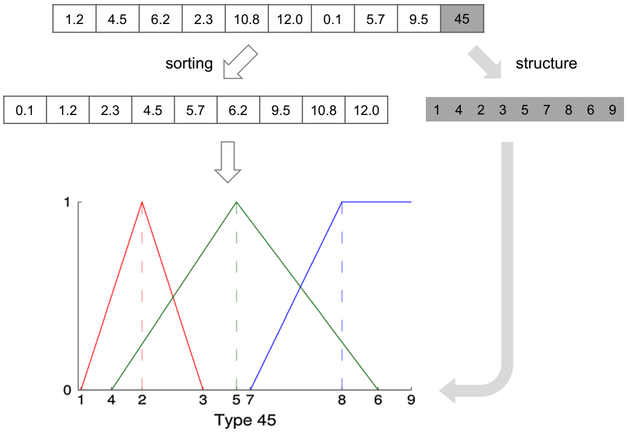

Fig. 2 Example of chromosome representation

The structure type is indexed according to the deployment of membership functions. Figure 2 illustrates a chromosome and its corresponding membership functions. The first 3ℓ real-valued genes represent the parameters of ℓ membership functions, and the last integer-valued gene indicates the structure type. When forming the membership functions, the 3ℓ parameters are first sorted and then arranged according to the structure index; for example, in Fig. 2 the lowest value 0.1 corresponds to the first parameter, the second lowest value 1.2 corresponds to the fourth parameter, and so forth.

The number of structure types is dependent upon the formation of membership functions. For three triangular membership functions, there exist 93 structure types that satisfy constraints (10) and (11) on shape and order. In addition to shape and order, two factors are crucial to the membership functions for fuzzy association rules: coverage and overlap. The former represents the range covered by all membership functions for an item, while the latter is measured by the area covered by two membership functions. In fuzzy systems, the interpretability is associated with coverage and overlap of membership functions. According to coverage and overlap, we classify the 93 structure types into four categories in Table 1. The first category consists of 12 structure types. As Fig. 3a illustrates, the structure types in this category have full coverage and appropriate overlap between membership functions. The second category includes 8 structure types, which have appropriate overlap but incomplete coverage. The third category has the largest number 69 of structure types. Although these structure types can achieve full coverage, the overlap between membership functions is inappropriate. For example, the four structure types in Fig. 3c incur needless overlap of the first and third membership functions. The fourth category comprises 4 structure types with partial coverage and inappropriate overlap.

Introducing structure type to chromosome representation facilitates development of heuristics for filtering out improper membership functions. This study proposes two heuristics considering coverage and overlap in membership functions:

- Coverage:

$$c_{i-1,1} \leq c_{i,1} \leq c_{i-1,3} \tag{12a}$$

Table 1 Classification of membership function structures

	Coverage	
	Full	Partial
Overlap		
Appropriate	12	8
Inappropriate	69	4

$$c_{i+1,1} \leq c_{i,3} \leq c_{i+1,3} \tag{12b}$$

- Overlap:

$$c_{i,3} \leq c_{i+2,1} \tag{13}$$

The above inequalities secure the full coverage and appropriate overlap. Among the 93 structure types, only 12 structure types of the first category are selected according to the two heuristics, which guarantee proper arrangement of membership functions.

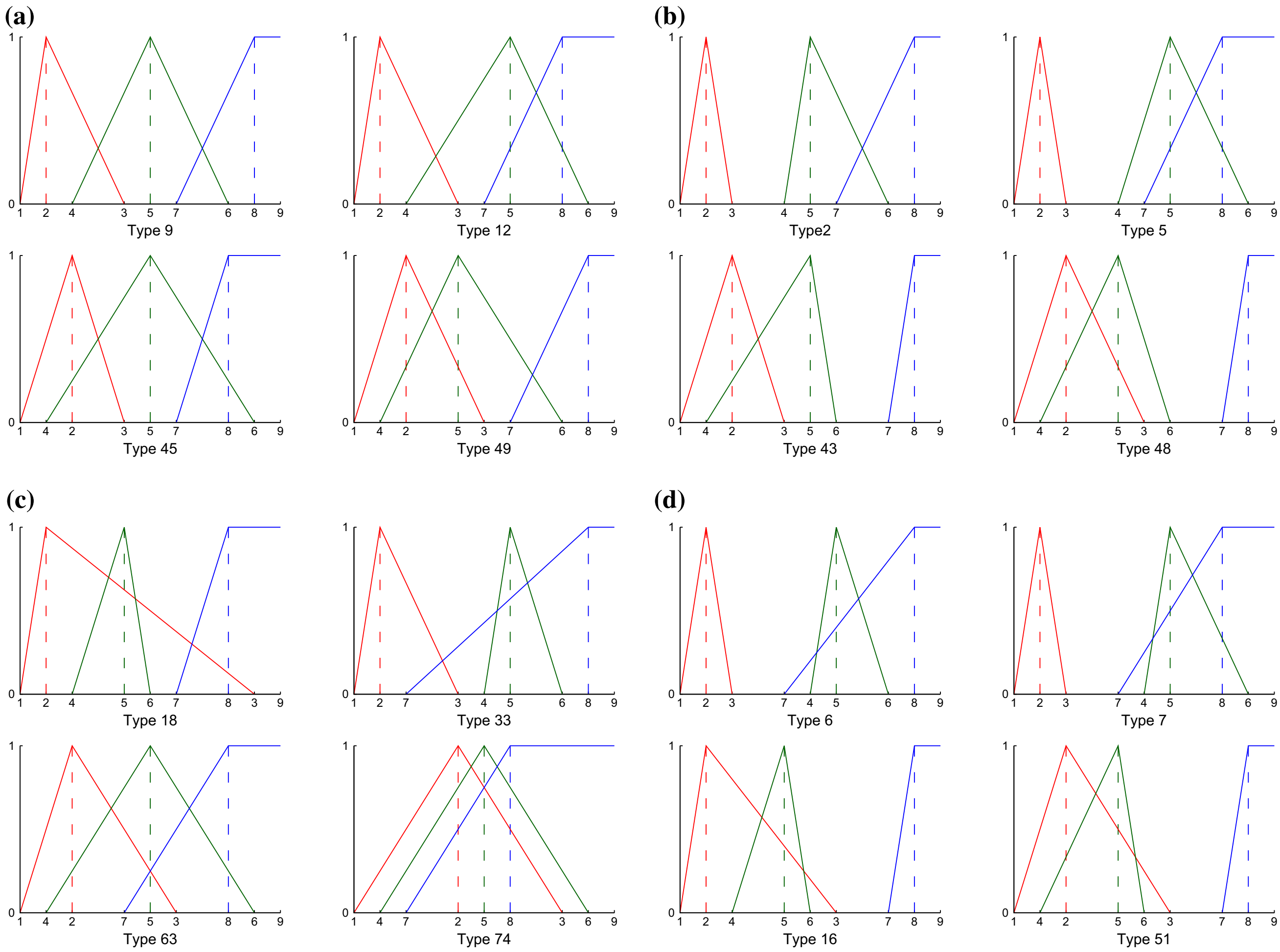

Fig. 3 Example structure types of the four categories. **a** Category I, **b** category II, **c** category III, **d** category IV

The proposed representation using the above heuristics holds two significant advantages. First, it can effectively filter out the structures inappropriate for coverage and overlap of membership functions. Second, the two heuristics decrease the number of structure types and thus reduce the search space of GA. These advantages are beneficial for the convergence speed and solution quality of the GA.

3.2 Genetic operators

The genetic operators of GA include parent selection, crossover, mutation, and survival selection. The parent selection operator picks two chromosomes from the population as the parents for subsequent reproduction. In this study, we adopt the well-known k-tournament selection, which picks as a parent the best of k chromosomes randomly selected from the population. Performing the tournament selection twice yields a pair of parents.

The crossover and mutation operators generate new candidate solutions, namely offspring, by exchanging parental information and slightly altering the composition, respectively. This study employs the max-min arithmetical (MMA) crossover and creep mutation that are widely used in genetic-fuzzy data mining (Herrera et al. 1997; Hong et al. 2008). The MMA crossover generates four candidates in different ways: The first two are the maximum and the minimum of the two parents, respectively; and the other two are produced by the whole arithmetic crossover. The best two of the candidates are selected as offspring. The creep mutation probabilistically changes some parameters to random values. As for structure, the type number is mutated by random resetting.

The survival selection determines the chromosomes surviving into the next generation. The principle "survival of the fittest" is generally used in the survival selection. In this study, we utilize the $(\mu + \lambda)$ survival selection, which considers both parent and offspring populations in the competition for survival.

3.3 Fitness evaluation

The fitness function guides the search direction and strongly affects the performance of GA. For mining fuzzy association rules, the fitness function usually uses the fuzzy support according to the largest itemset. However, this method requires iterative calculations and is hard to be paralleled. Hong et al. (2008) proposed using the divide-and-conquer strategy to address this issue. Specifically, they used separate populations for the items and considered the fuzzy support of large 1-itemset L_1 in the fitness evaluation. In addition, the fitness function takes the coverage and overlap of membership functions into account.

In view of the above advantages, we utilize the fitness function proposed by Hong et al. (2008). For a given item, the fitness of a chromosome is evaluated by the fitness function based on fuzzy support, overlap, and coverage.

Definition 6 (*Overlap factor*) The overlap factor of chromosome C_k is defined by

$$\text{Overlap}\,(C_k) = \sum_{i<j} \left(\max\left(\text{ovlratio}\left(\Upsilon_i, \Upsilon_j\right), 1\right) - 1\right) \tag{14}$$

with

$$\text{ovlratio}\left(\Upsilon_i, \Upsilon_j\right) = \frac{\text{The area covered by both } \Upsilon_i \text{ and } \Upsilon_j}{\min\left(c_{i,3} - c_{i,2}, c_{j,2} - c_{j,1}\right)}\,. \tag{15}$$

The ratio of overlap ovlratio $\left(\Upsilon_i, \Upsilon_j\right)$ is determined by the proportion of the area covered by both membership functions to the smaller of $(c_{i,3} - c_{i,2})$ and $(c_{j,2} - c_{j,1})$, which stand for the right half of the left membership function and the left half of the right membership function, respectively. The overlap factor is nonnegative, and its best value is zero. Note that zero overlap factor indicates adequate overlap, instead of nonoverlap, of all pairs of membership functions.

Definition 7 (*Coverage factor*) The coverage factor of chromosome C_k is defined by

$$\text{Coverage}\,(C_k) = \frac{\max\,(I)}{\text{range}\,(\Upsilon_1, \ldots, \Upsilon_l)}\,. \tag{16}$$

Coverage factor is inversely proportional to the range covered by all membership functions. Therefore, a small coverage factor is preferred, and its best value is 1 for full coverage of the item's quantity.

The fitness evaluation considers the fuzzy support as well as the suitability. The suitability is defined as the sum of overlap and coverage factors. Formally, the fitness value of chromosome C_k is computed by

$$f\,(C_k) = \frac{\sum_{X \in L_1} \text{FuzzySupport}\,(X)}{\text{Overlap}\,(C_k) + \text{Coverage}\,(C_k)}. \tag{17}$$

Note that the fitness function considers only the large 1-itemset L_1 obtained from the membership functions set C_k. The exclusion of $L_{>1}$ is beneficial for balance of computation time and quality in calculating fuzzy support (Hong et al. 2008).

As an example, the fitness value of the chromosome in Fig. 2 is calculated as follows. First, the overlap ratios for the three membership functions are computed using (15):

$$\text{ovlratio}\,(\Upsilon_1, \Upsilon_2) = \frac{4.5 - 1.2}{\min(4.5 - 2.3, 5.7 - 1.2)} = 1.5$$

Table 2 Parameter setting

Parameter	Value
Representation	Real numbers (parameters) + integer (structure)
Parent selection	2-tournament
Crossover	MMA ($d = 0.35$)
Crossover rate	$p_c = 0.8$
Mutation	Creep ($\varepsilon = 3$)
Mutation rate	$p_m = 0.01$
Survival selection	$\mu + \lambda$
Population size	50
#Generations	500

$$\text{ovlratio}\,(\Upsilon_1, \Upsilon_3) = \frac{0}{\min(4.5 - 2.3, 9.5 - 6.2)} = 0$$

$$\text{ovlratio}\,(\Upsilon_2, \Upsilon_3) = \frac{10.8 - 6.2}{\min(10.8 - 5.7, 9.5 - 6.2)} = 1.39$$

Using (14), we can obtain the overlap factor

$$\begin{aligned}\text{Overlap}\,(C_k) &= (\max(1.5, 1) - 1) + (\max(0, 1) - 1)\\ &\quad + (\max(1.39, 1) - 1)\\ &= 0.89\end{aligned}$$

The coverage factor is calculated by (16):

$$\text{Coverage}\,(C_k) = \frac{12.0}{12.0 - 0.1} = 1.0084$$

Suppose the item exists in five transactions with quantities 1, 5, 8, 11, and 6. We use (4) to calculate the fuzzy support of each region:

$$\begin{aligned}\text{FuzzySupport}\,(\Upsilon_1) &= \frac{1}{5}\left(\frac{1 - 0.1}{2.3 - 0.1} + 0 + 0 + 0 + 0\right)\\ &= 0.08\\ \text{FuzzySupport}\,(\Upsilon_2) &= \frac{1}{5}\left(0 + \frac{5 - 1.2}{5.7 - 1.2} + \frac{10.8 - 8}{10.8 - 5.7}\right.\\ &\quad \left. + 0 + \frac{10.8 - 6}{10.8 - 5.7}\right)\end{aligned}$$

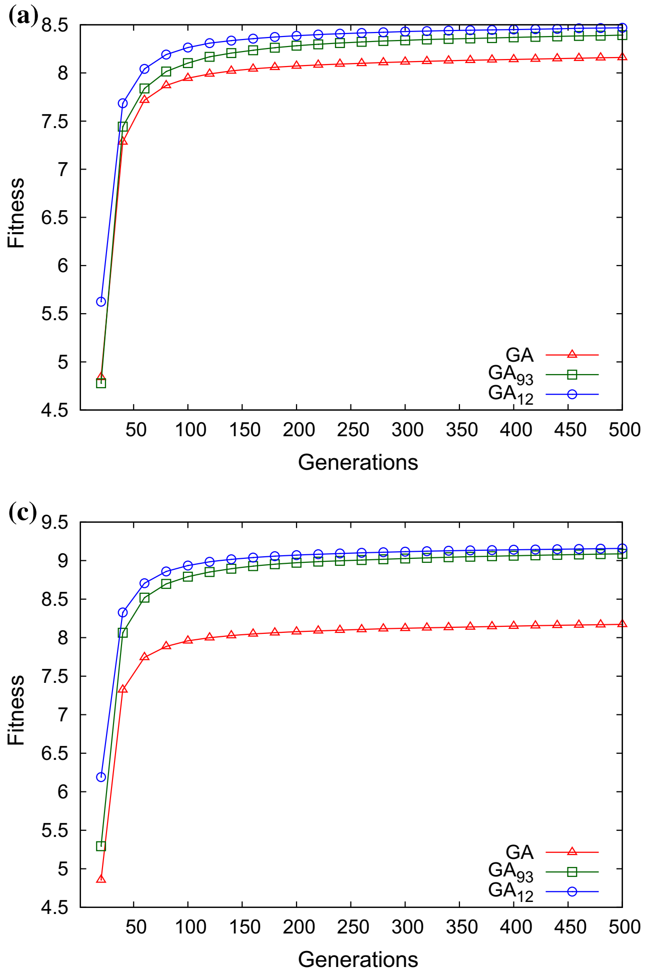

Fig. 4 Progress of MBF against generations for GA, GA_{93}, and GA_{12} on different datasets. **a** 10 k transactions, **b** 30 k transactions, **c** 50 k transactions, **d** 90 k transactions

$$= 0.47$$

$$\text{FuzzySupport}\ (\Upsilon_3) = \frac{1}{5}\left(0 + 0 + \frac{8 - 6.2}{9.5 - 6.2} + 1 + 0\right)$$
$$= 0.31$$

Given a minimum support 0.25, the large 1-itemset is $L_1 = \{\Upsilon_2, \Upsilon_3\}$ in that Υ_2 and Υ_3 have fuzzy support greater than 0.25. Summing their fuzzy support values yields

$$\text{FuzzySupport}\ (C_k) = 0.47 + 0.31 = 0.78$$

Hence, the fitness value

$$f\ (C_k) = \frac{0.78}{0.89 + 1.0084} = 0.41$$

As aforementioned, this study proposes two heuristics for the structure types considering overlap and coverage. The two heuristics guarantee that the coverage factor is smaller than $\frac{\max(I)}{c_{l,3} - c_{1,1}}$; in addition, they help to eliminate the gap between membership functions and avoid strong overlap.

4 Experimental results

This study conducts a series of experiments to examine the performance of the proposed GA on optimization of membership functions for fuzzy association rule mining. In the experiments, we investigate the effects of the proposed chromosome representation and the two heuristics about coverage and overlap. The test algorithms include GA (Hong et al. 2008), GA_{93} (GA using the novel representation), and GA_{12} (GA_{93} applying the two heuristics), where the subscripts 93 and 12 account for the numbers of structure types. Table 2 summarizes the parameter setting for the test algorithms. The minimum support is set to 0.04. Different data sizes are tested in the experiments, including 10, 30, 50, 70, and 90 k transactions, each of which consists of 64 items (Hong et al. 2008). Each experiment includes 30 independent runs of each algorithm.

Figure 4 shows the progress of mean best fitness (MBF) over generations for the test algorithms on different datasets. According to the results, GA using the proposed representa-

Table 3 MBF and p-values for GA, GA_{93}, and GA_{12} on the data of different sizes

#Tr. (k)	MBF			p-value		
	GA	GA_{93}	GA_{12}	$\text{GA}:\text{GA}_{93}$	$\text{GA}:\text{GA}_{12}$	$\text{GA}_{93}:\text{GA}_{12}$
10	8.16	8.40	**8.47**	**+2.64E−14**	**+6.16E−23**	**+1.56E−03**
30	8.14	8.80	**8.89**	**+4.86E−34**	**+1.69E−47**	**+1.51E−05**
50	8.18	9.09	**9.16**	**+7.27E−47**	**+6.88E−49**	**+1.78E−04**
70	8.15	8.73	**8.82**	**+2.15E−30**	**+4.22E−45**	**+1.27E−04**
90	8.15	8.75	**8.88**	**+9.69E−27**	**+2.72E−46**	**+1.10E−06**

The p-values account for the results of t-test on the MBF obtained from X and Y algorithms (denoted by $X{:}Y$), where positive p-values indicate that Y is superior to X. Boldfaced MBF marks the best result among the three algorithms; boldfaced p-values denote the statistical significance with confidence level $\alpha = 0.01$

Table 4 Overlap, coverage, suitability, and fuzzy support of membership functions obtained from GA, GA_{93}, and GA_{12} on the data of different sizes

#Tr. (k)	Algorithm	Overlap	Coverage	Suitability	Fuzzy support
10	GA	23.23	64.20	87.43	**11.02**
	GA_{93}	6.92	64.13	71.05	9.27
	GA_{12}	**0.17**	**64.10**	**64.27**	8.51
30	GA	23.01	64.16	87.17	**10.96**
	GA_{93}	5.76	64.12	69.88	9.57
	GA_{12}	**0.18**	**64.11**	**64.28**	8.93
50	GA	22.74	64.18	86.92	**10.98**
	GA_{93}	5.93	64.13	70.06	9.89
	GA_{12}	**0.13**	**64.10**	**64.23**	9.19
70	GA	23.54	64.20	87.74	**11.05**
	GA_{93}	6.48	64.13	70.60	9.58
	GA_{12}	**0.14**	**64.10**	**64.24**	8.85
90	GA	23.60	64.18	87.78	**11.05**
	GA_{93}	7.10	64.13	71.23	9.67
	GA_{12}	**0.13**	**64.10**	**64.23**	8.91

Boldface marks the best result among the three algorithms

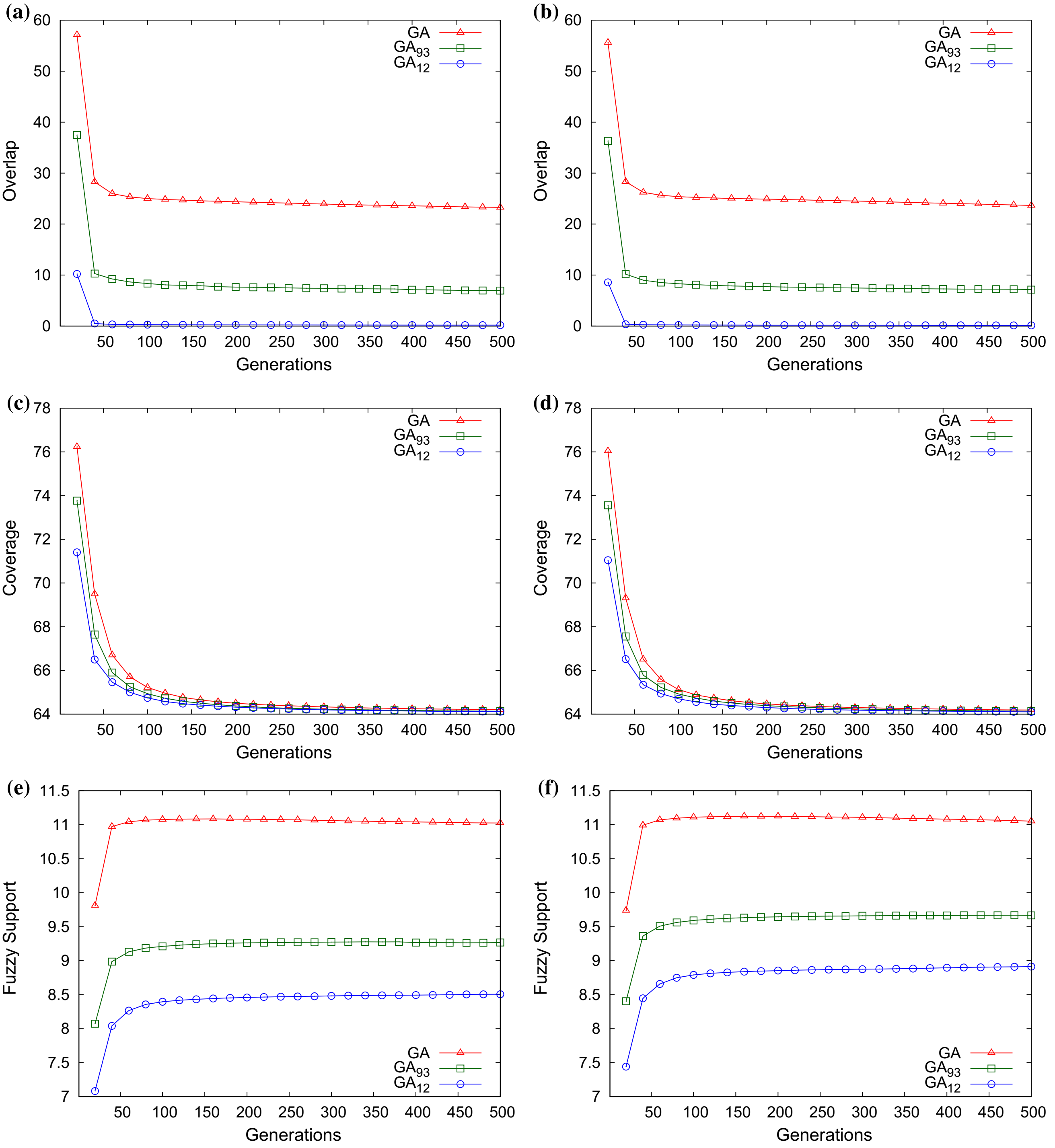

Fig. 5 Variation of overlap (*top*), coverage (*middle*), and fuzzy support (*bottom*) of membership functions obtained from GA, GA_{93}, and GA_{12} among 64 items on the data of 10 k (*left*) and 90 k (*right*) transactions. **a** Overlap (10 k transactions), **b** overlap (90 k transactions), **c** coverage (10 k transactions), **d** coverage (90 k transactions), **e** fuzzy support (10 k transactions), **f** fuzzy support (90 k transactions)

tion, i.e., GA_{93} and GA_{12}, converges faster than the original GA does, validating the effectiveness of the new representation on improving the search efficiency of GA. In addition, the faster convergence of GA_{12} than GA_{93} indicates the advantages of the two heuristics.

Table 3 presents the MBF obtained from the three test algorithms for all 64 items. The table also lists the p-values of one-tailed t-test on the MBF values. With confidence level $\alpha = 0.01$, the t-test results indicate that GA using the new representation, namely GA_{93} and GA_{12}, achieves signifi-

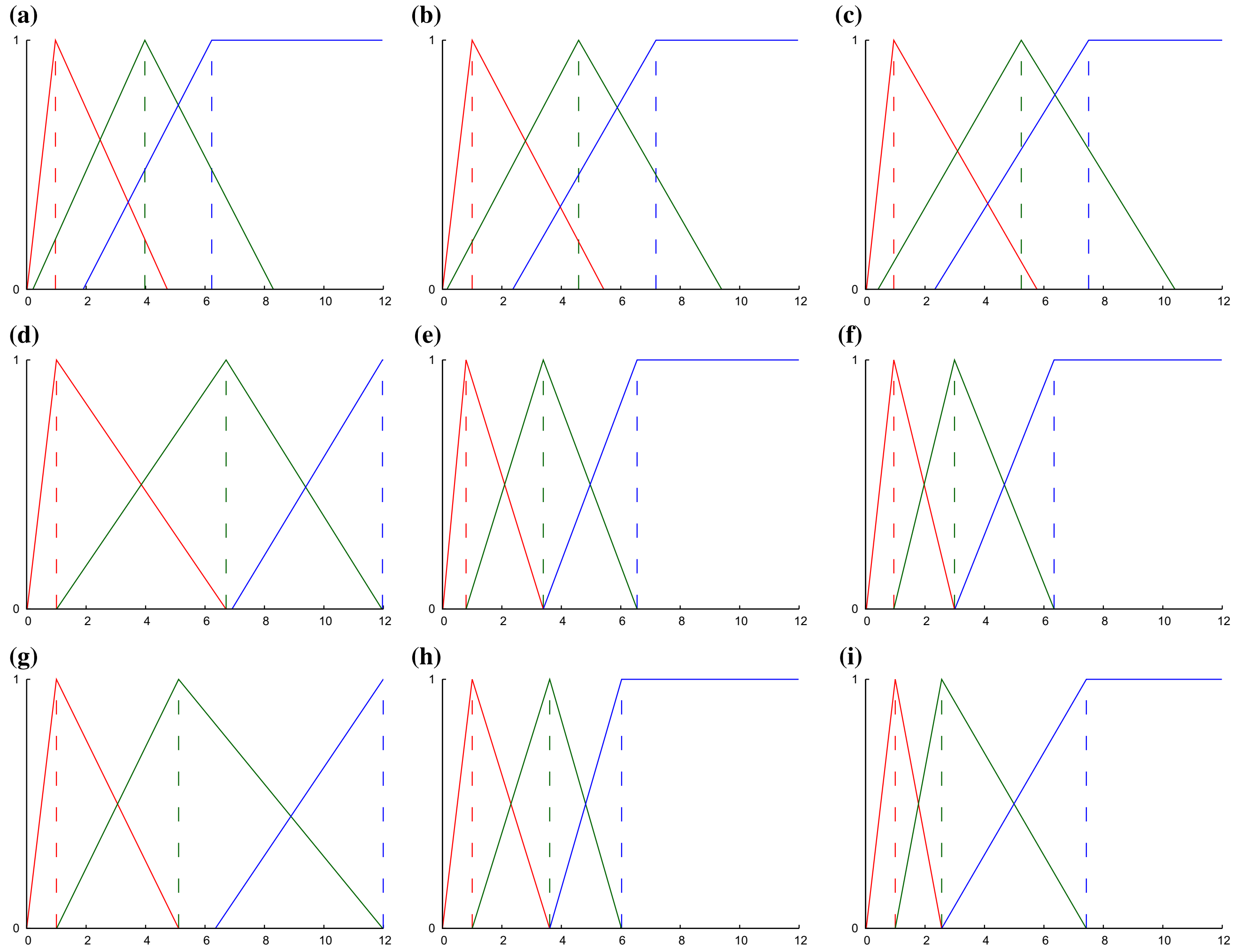

Fig. 6 Comparison of membership functions obtained from GA, GA_{93} and GA_{12} on the data of 10 k (*left*), 50 k (*center*), and 90 k (*right*) transactions. **a** GA (10 k transactions), **b** GA (50 k transactions), **c** GA (90 k transactions), **d** GA_{93} (10 k transactions), **e** GA_{93} (50 k transactions), **f** GA_{93} (90 k transactions), **g** GA_{12} (10 k transactions), **h** GA_{12} (50 k transactions), **i** GA_{12} (90 k transactions)

cantly higher fitness than the original GA does; meanwhile, GA_{12} significantly outperforms GA_{93} on all datasets. These superior outcomes validate the benefits of the representation and two heuristics.

Next, we investigate the influences of the proposed representation and heuristics on the overlap, coverage, and fuzzy support of membership functions obtained. As Table 4 indicates, GA benefits from the two heuristics in coverage and overlap: GA_{93} and GA_{12} obtain better coverage and overlap than GA does. In addition, GA_{12} gains the best coverage and overlap among the three test algorithms. Figure 5 further shows that GA_{12} can find membership functions with good overlap and converge faster than other approaches do. In general, GA_{12} and GA_{93} achieve membership functions with better suitability; nonetheless, they lead to lower fuzzy support than GA. The results on Table 4 and Fig. 5 reflect the trade-off between suitability and fuzzy support. The two heuristics impose constraints on structure types for high suitability; on the other hand, they discourage sacrificing suitability for fuzzy support. As Fig. 6 shows, GA gains the highest fuzzy support at the cost of suitability; in particular, the overlap of membership functions obtained from GA is so high that the fuzzy regions become trivial. By contrast, GA_{93} and GA_{12} maintain adequate overlap and full coverage while pursuing high fuzzy support, which results in more reasonable membership functions and better fitness.

5 Conclusions

This study proposes a GA for optimization of membership functions for fuzzy association rule mining. For the GA, we design a novel chromosome representation that considers structure types in addition to parameters of membership functions. Based on the new representation, two heuristics are developed for securing the coverage and moderating the overlap of membership functions. The heuristics can filter out inappropriate arrangement of membership functions; moreover, they help to reduce the search space. For example, the number of structure types is reduced from 93 to 12 for three membership functions; that is, 81 inappropriate structure types are filtered out in light of coverage and overlap.

A series of experiments is carried out to examine the performance of the proposed GA. The experimental results on 10–90 k transactions show that the proposed GA achieves significant performance improvement. The new representation and two heuristics enhance the coverage and overlap of membership functions; in addition, they improve the fitness value and convergence speed of GA. These preferable outcomes show the utility of the proposed chromosome representation and heuristics for the GA. They also validate the effectiveness and efficiency of the GA in finding appropriate membership functions for mining fuzzy association rules.

Future work may further expand the use of the structure types and design of local search operator. The structure type reflects the relationship between membership functions and facilitates designing strategies for their arrangement. In addition to overlap and coverage, advanced heuristics are promising for enhancing the performance of GA on fuzzy association rule mining.

Acknowledgments The authors would like to thank the editor and reviewers for their valuable comments and suggestions.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the international conference on very large data bases, pp 487–499

Alcalá-Fdez J, Alcalá R, Herrera F (2011) A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Trans Fuzzy Syst 19(5):857–872

Antonelli M, Ducange P, Marcelloni F (2014) A fast and efficient multi-objective evolutionary learning scheme for fuzzy rule-based classifiers. Inf Sci 283(1):36–54

Asadollahpoor-Chamazi M, Minaei-Bidgoli B, Nasiri M (2013) Deriving support threshold values and membership functions using the multiple-level cluster-based master-slave IFG approach. Soft Comput 17(7):1227–1239

Cai G-R, Li S-Z, Chen S-L (2010) Mining fuzzy association rules by using nonlinear particle swarm optimization. Quant Log Soft Comput 82:621–630

Chan K, Au WH (1997) Mining fuzzy association rules. In: Proceedings of the international conference on information and knowledge management, pp 209–215

Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(27):1–27

Chen C-H, Li A-F, Lee Y-C (2013) A fuzzy coherent rule mining algorithm. Appl Soft Comput 13(7):3422–3428

Chen C-H, Tseng V-S, Hong T-P (2008) Cluster-based evaluation in fuzzy-genetic data mining. IEEE Trans Fuzzy Syst 16(1):249–262

Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996a) From data mining to knowledge discovery in databases. AI Mag 17:37–54

Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996b) Advances in knowledge discovery and data mining. In: From data mining to knowledge discovery: an overview, AAAI, pp 1–34

Fazzolari M, Alcala R, Nojima Y, Ishibuchi H, Herrera F (2013) A review of the application of multiobjective evolutionary fuzzy systems: current status and further directions. IEEE Trans Fuzzy Syst 21(1):45–65

Herrera F, Lozano M, Verdegay JL (1997) Fuzzy connectives based crossover operators to model genetic algorithms population diversity. Fuzzy Sets Syst 92(1):21–30

Hong T-P, Chen C-H, Lee Y-C, Wu Y-L (2008) Genetic-fuzzy data mining with divide-and-conquer strategy. IEEE Trans Evol Comput 12(2):252–265

Hong T-P, Chen C-H, Wu Y-L, Lee Y-C (2006) A GA-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions. Soft Comput 10(11):1091–1101

Hong T-P, Kuo C-S, Chi S-C (1999) Mining association rules from quantitative data. Intell Data Anal 3(5):363–376

Hong T-P, Lee C-Y (1996) Induction of fuzzy rules and membership functions from training examples. Fuzzy Sets Syst 84(1):33–47

Kuok CM, Fu A, Wong MH (1998) Mining fuzzy association rules in databases. ACM SIGMOD Rec 27(1):41–46

Lee CK-H, Choy K-L, Ho GT-S, Lam CH-Y (2016) A slippery genetic algorithm-based process mining system for achieving better quality assurance in the garment industry. Expert Syst Appl 46:236–248

Lee CK-H, Ho GT-S, Choy K-L, Pang GK-H (2014) A RFID-based recursive process mining system for quality assurance in the garment industry. Intern J Prod Res 52(14):4216–4238

Liu B, Hsu W, Chen S, Ma Y (2000) Analyzing the subjective interestingness of association rules. IEEE Intell Syst Their Appl 15(5):47–55

Meng D, Pei Z (2012) Extracting linguistic rules from data sets using fuzzy logic and genetic algorithms. Neurocomputing 78(1):45–54

Minaei-Bidgoli B, Barmaki R, Nasiri M (2013) Mining numerical association rules via multi-objective genetic algorithms. Inf Sci 233(1):15–24

Mishra S, Mishra D, Satapathy SK (2011) Particle swarm optimization based fuzzy frequent pattern mining from gene expression data. In: Proceedings of the international conference on computer & communication technology, pp 15–20

Piatesky-Shapiro G, Brachman R, Klösgen W, Simoudis E (1996) An overview of issues in developing industrial data mining and knowledge discovery applications. In: Proceedings of knowledge discovering and data mining, pp 89–95

Qodmanan HR, Nasiri M, Minaei-Bidgoli B (2011) Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. Expert Syst Appl 38(1):288–298

Rudzinski F (2016) A multi-objective genetic optimization of interpretability-oriented fuzzy rule-based classifiers. Appl Soft Comput 38:118–133

Srikant R, Agrawal R (1996) Mining quantitative association rules in large relational tables. ACM SIGMOD Rec 25(2):1–12

Wagstaff K, Cardie C, Rogers S, Schroedl S (2001) Constrained K-means clustering with background knowledge. In: Proceedings of the international conference on machine learning, pp 577–584

Wakabi-Waiswa P, Baryamureeba V (2008) Extraction of interesting association rules using genetic algorithms. Int J Comput ICT Res 2:1139–1818