BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 07/08 pp306–318
Volume 2, Number 4, December 2019
DOI: 10.26599/BDMA.2019.9020008

Selective Ensemble Learning Method for Belief-Rule-Base Classification System Based on PAES

Wanling Liu, Weikun Wu, Yingming Wang, Yanggeng Fu*, and Yanqing Lin

Abstract: Traditional Belief-Rule-Based (BRB) ensemble learning methods integrate all of the trained sub-BRB systems to obtain better results than a single belief-rule-based system. However, as the number of BRB systems participating in ensemble learning increases, a large amount of redundant sub-BRB systems are generated because of the diminishing difference between subsystems. This drastically decreases the prediction speed and increases the storage requirements for BRB systems. In order to solve these problems, this paper proposes BRBCS-PAES: a selective ensemble learning approach for BRB Classification Systems (BRBCS) based on Pareto-Archived Evolutionary Strategy (PAES) multi-objective optimization. This system employs the improved Bagging algorithm to train the base classifier. For the purpose of increasing the degree of difference in the integration of the base classifier, the training set is constructed by the repeated sampling of data. In the base classifier selection stage, the trained base classifier is binary coded, and the number of base classifiers participating in integration and generalization error of the base classifier is used as the objective function for multi-objective optimization. Finally, the elite retention strategy and the adaptive mesh algorithm are adopted to produce the PAES optimal solution set. Three experimental studies on classification problems are performed to verify the effectiveness of the proposed method. The comparison results demonstrate that the proposed method can effectively reduce the number of base classifiers participating in the integration and improve the accuracy of BRBCS.

Key words: belief-rule-base; pareto-archived evolutionary strategy; selective ensemble; classification

1 Introduction

In 2006, for the modelling of data characterized by incompleteness, fuzzy uncertainty, probability uncertainty, and non-linearity, Yang et al.[1] extended the evidence-based reasoning algorithm to propose their belief Rule-base Inference Methodology using the Evidential Reasoning approach (RIMER). RIMER is composed of a knowledge base and a reasoning machine, and was developed on the basis of fuzzy logic theory[2], the Dempster-Shafer theory[3,4], and traditional If-then rules[1]. A Belief-Rule-Base (BRB) system is an expert system that adds a confidence distribution to the If-then rule. After the construction of the BRB, quantitative information or qualitative knowledge is input for reasoning and analysis, ultimately to provide an informative basis for decision making.

Present research on BRB systems is mainly focused on the use of a single BRB system. However, the use of a single BRB system has some limitations. Its

• Wanling Liu, Weikun Wu, Yanggeng Fu, and Yanqing Lin are with the School of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China. E-mail: 380509981@qq.com; wwk91@qq.com; ygfu@qq.com; 765305442@qq.com.
• Yingming Wang is with the Institute of Decision Sciences, Fuzhou University, Fuzhou 350116, China. E-mail: ymwang@fzu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2019-03-24; revised: 2019-04-09; accepted: 2019-04-11

reasoning performance is affected by the parameter values, and where the training set is unevenly distributed or the amount of data is small, parameter training can be insufficient. Therefore, the decision information provided by the reasoning results suffers from locality. In 2016, Wu et al.[5] introduced the Bagging and AdaBoost algorithms. Their approach uses the accelerated gradient method[6] to train the parameters of a single BRB system, and then integrates multiple sub-BRB systems with ensemble learning methods to improve the reasoning ability. In ensemble learning, a common approach is to integrate all of the trained learning machines in order to obtain better results; ensemble learning produces better results than using a single BRB system in isolation. However, as the number of individuals participating in ensemble learning increases, the sub-BRB system begins to produce a large number of redundant base learning machines because of the decrease in individual differences. This results in a noticeable decrease in prediction speed and a dramatic increase in storage overhead, ultimately reducing the effective generalization ability of the system.

In response to these deficiencies, this paper proposes BRBCS-PAES, using selective ensemble learning methods for Belief-Rule-Base Classification Systems (BRBCS) based on the Pareto-Archiving Evolution Strategy (PAES). The improved Bagger algorithm is used to train the base classifier, and the training set is constructed by repeated sampling of the data, thereby increasing the degree of difference when the base classifier is integrated. In the base classifier selection stage, the trained base classifier is binary coded (with 1 meaning participation integration, 0 meaning no participation), and the number of base classifiers participating in integration and generalization error of the base classifier is used as the objective function for multi-objective optimization. Employing an elitist retention strategy and an adaptive grid archiving strategy to iteratively arrive at the PAES optimal solution set, three sets of classification data from UCI (University of California, Irvine) are used to verify the effectiveness of the proposed method.

The rest of the paper is organized as follows: Section 2 briefly reviews the basics of BRB, multi-objective optimization, and selective ensemble learning, and reviews some related works; Section 3 introduces the belief rule-base classification system of selective ensemble learning; Section 4 discusses three case studies to demonstrate the efficiency of the proposed method; and Section 5 concludes the paper.

2 Preliminaries

2.1 Belief rule-base and RIMER method

The belief rules in the BRB are extensions of If-then rules[7], adding the distributed confidence frame, the antecedent attribute weight, and the rule weight. The k-th belief rule[8] can be written as follows:

$$R_k : \text{if } A_1^k \wedge A_2^k \wedge \cdots \wedge A_{T_k}^k, \\ \text{then } \{(D_1, \overline{\beta}_{1,k}), (D_2, \overline{\beta}_{2,k}), \ldots, (D_N, \overline{\beta}_{N,k})\} \quad (1)$$

where R_k $(k = 1, 2, \ldots, L)$ represents the k-th rules, L represents the total number of rules; A_i^k $(i = 1, 2, \ldots, T_k)$ represents the antecedent attribute reference of the i-th attribute of the k-th rule, T_k represents the number of attributes in the k-th rule; D_j $(j = 1, 2, \ldots, N)$ represents a set of rule result evaluation levels, N is the set size; and $\overline{\beta}_{j,k}$ $(j = 1, 2, \ldots, N, \quad k = 1, 2, \ldots, L)$ represents the belief degree of the result of the k-th rule on the j-th evaluation level D_j. If $\sum_{j=1}^{N} \overline{\beta}_{j,k} = 1$, then the k-th rules contain complete information, otherwise, the information in the k-th rule is incomplete. θ_k $(k = 1, 2, \ldots, L)$ is the rule weight of the k-th rules, the antecedent attribute weight is $\delta_{k,i}$ $(k = 1, 2, \ldots, L, i = 1, 2, \ldots, T_k)$, and "$\wedge$" expresses the logical conjunction (And operator).

The RIMER method is at the core of a BRB system[9], and consists of three main steps[10]: (1) calculate the activation weight; (2) amend with the belief degree; and (3) use an Evidence Reasoning (ER) algorithm to synthesize activation rules[11].

The calculation of the activation weight depends on the input data, the antecedent attribute weight, and the rule weight[12]. Before calculating the activation weight, we need to calculate the individual match of the antecedent attribute for each reference. Assuming that the BRB input x_i $(i = 1, 2, \ldots, M)$ is in numeric form, then according to utility information conversion[7], the matching degree α_i^j of the i-th input relative to the reference value in the k-th rule is calculated from x_i and A_i^k $(i = 1, 2, \ldots, T_k)$ as follows:

$$\begin{cases} \alpha_i^j = \frac{A_i^{k+1} - x_i}{A_i^{k+1} - A_i^k}, \ \alpha_i^{j+1} = 1 - \alpha_i^k, A_i^k \leqslant x_i \leqslant \\ \qquad A_i^{k+1} \text{and } j = k; \\ \alpha_i^s = 0, s \neq k \text{ or } k+1 \end{cases} \quad (2)$$

Then the activation weight of the k-th rule is

calculated as

$$\omega_k = \frac{\theta_k \prod_{i=1}^{T_k} (\alpha_i^k)^{\bar{\delta}_{k,i}}}{\sum_{l=1}^{L} \theta_l \prod_{i=1}^{T_k} (\alpha_i^l)^{\bar{\delta}_{l,i}}},$$
$$\bar{\delta}_{k,i} = \frac{\delta_{k,i}}{\max_{i=1,\ldots,T_k}\{\delta_{k,i}\}} \tag{3}$$

while $\omega_k \in [0,1], k = 1,2,\ldots,L$.

When the input data contains fuzzy, uncertain data, we need to amend the belief degree of each evaluation level of the result portion. The correction formula for the belief degree $\overline{\beta}_{i,k}$ of the i-th evaluation grade D_i of the k-th rule is

$$\beta_{i,k} = \overline{\beta}_{i,k} \frac{\sum_{t=1}^{T_k} \tau(t,k) \sum_{j=1}^{|A_t|} \alpha_{t,j}}{\sum_{t=1}^{T_k} \tau(t,k)},$$
$$\tau(t,k) = \begin{cases} 1, & A_t \in R_k \ (t = 1,2,\ldots,T_k); \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $|A_t|$ indicates the number of candidate values. If the input data is complete, then $\beta_{i,k} = \overline{\beta}_{i,k}$. In the ER algorithm, Wang et al.[13] proposed the ER analysis algorithm to combine all the rules in a BRB; the output $f(x)$ of the BRB can be expressed as

$$f(x) = (D_j, \beta_j), j = 1,2,\ldots,N \tag{5}$$

where β_j represents the belief degree relative to the evaluation result, calculated as

$$\beta_j = \frac{\mu \times [\prod_{k=1}^{L}(\omega_k \beta_{j,k} + 1 - \omega_k \sum_i^N \beta_{j,k})]}{1 - \mu \times [\prod_{k=1}^{L}(1-\omega_k)]} - \frac{\mu \times (\prod_k^L (1 - \omega_k \Sigma_i^N \beta_{j,k}))]}{1 - \mu \times [\prod_{k=1}^{L}(1-\omega_k)]} \tag{6}$$

$$\mu = \left[\sum_{j=1}^{N} \prod_{k=1}^{L} (\omega_k \beta_{j,k} + 1 - \omega_k \sum_{j=1}^{N} \beta_{j,k}) - (N-1) \prod_{k=1}^{L} (1-\omega_k) \sum_{j=1}^{N} \beta_{j,k}) \right]^{-1} \tag{7}$$

Assuming $\mu(D_n)$ is the utility value of the n-th evaluation level D_n, the final numerical output of the BRB system is expressed as

$$y = \sum_{n=1}^{N} \mu(D_n)\beta_n + \frac{\mu(D_1)+\mu(D_N)}{2} \left(1 - \sum_{n=1}^{N} \beta_n\right) \tag{8}$$

2.2 Multi-objective optimization

Multi-Objective Optimization problems (MOPs) have two or more objective functions[14, 15]; they can be stated as follows:

$$\begin{aligned} &\text{Min/Max} y = F(x) = (f_1(x), f_2(x), \ldots, f_n(x)), \\ &\text{s.t.} : g_i(X) \leqslant 0, i = 1,2,\ldots,K_g, \\ &h_j(X) = 0, j = 1,2,\ldots,k_h, \\ &\text{where } x = (x_1, x_2, \ldots, x_m) \in X \subseteq \mathbf{R}, \\ &y = (y_1, y_2, \ldots, y_n) \in Y \subseteq \mathbf{R} \end{aligned} \tag{9}$$

In the above formula, $x = (x_1, x_2, \ldots, x_m)$ is the m-dimensional decision parameters, X is the decision space, $y = (y_1, y_2, \ldots, y_n)$ is n-dimensional target variable, and $F(x)$ is the objective function of mapping m decision spaces to n target spaces. $g_i(X) \leqslant 0$ contains k_g inequality constraints and $h_j(X) \leqslant 0$ contains k_h equality constraints; let X_f denote the set of decision parameters x that satisfy all the constraints.

Definition 1: Pareto dominant

Suppose $x_A, x_B \in X_f$ are the two solutions of $F(x) = (f_1(x), f_2(x), \ldots, f_n(x))$. For x_B relative to x_A to be Pareto dominant ($x_B \succ x_A$), two equations need to be satisfied:

(1) $\forall i = 1,2,\ldots,n, f_i(x_B) \geqslant f_i(x_A)$, that is, in all objective functions, x_B is not worse than x_A;

(2) $\exists i = 1,2,\ldots,n, f_i(x_B) > f_i(x_A)$, that is, x_B is better than x_A in at least one objective function.

Definition 2: Pareto optimal solution

If a solution $x^* \in X_f$ satisfies $\neg\exists x \in X_f : x \succ x^*$, it is called a Pareto optimal solution. In an MOP, a solution is actually an approximation set of candidate solutions which offer trade-offs between the multiple objectives, where an improvement in one objective value will result in a decline in one or more of the others[15].

MOPs were previously solved by being treated as single-objective problems using the weighted sum approach, but recent years have witnessed significant progress in the development of Evolutionary Algorithms (EAs) for MOPs[15, 16]. The majority of existing Multi-Objective Evolutionary Algorithms (MOEAs) are based on Pareto dominance. In MOEAs, the utility of each individual solution is mainly determined by its Pareto dominance relations with other solutions visited in the previous search. Since using Pareto dominance alone can reduce the diversity of a search, certain techniques such as fitness sharing and crowding have often been used to compensate[15, 17]. Arguably, PAES is one of the most popular Pareto dominance based MOEAs, proposed by Corne et al.[18], in 2000.

2.2.1 Pareto Archiving Evolution Strategy (PAES)

PAES[14,19] is a classical method for the evolutionary multi-objective optimization algorithm. PAES uses (1+1) evolution strategy, in which a population of solutions is used to create offspring solutions using a mutation operator. The dominance relationship of offspring and parental solutions is compared, with the elite retention strategy used to retain the better solution and establish a Non-Dominated Solutions (NDS) file to retain the solution of the previous generation.

The PAES algorithm consists of three parts: (1) Generation of candidate solutions; (2) Selection of candidate solutions; and (3) Construction of the NDS. The algorithm (shown in Fig. 1) randomly generates an initial solution, calculates the target value corresponding to the initial solution, and adds it to the NDS. A candidate solution is obtained by the mutation of a parent solution or the recombination of multiple parent solutions, and the target value of the candidate solution is calculated. If the candidate solution is dominated by the parent solution, then, according to a certain probability of performing a mutation or reorganization operation, a new candidate solution is generated; otherwise, the dominance of the offspring solution is compared with other solutions in the NDS. The NDS is updated by the adaptive grid archiving strategy, and a mutated or recombined solution in the NDS is selected as the new parent solution. The process iterates until it reaches the end condition.

2.2.2 Adaptive grid archiving strategy

The PAES algorithm uses the adaptive grid archiving strategy to maintain diversity in the Pareto-optimal set. The main purpose of the adaptive grid archiving strategy is to make a choice between the parent solution and offspring when updating the NDS. If the candidate solution is dominated by any solution in the NDS, the candidate solution is deleted. The basic idea is to divide the target space into many grids and assign a grid to each individual. The crowded comparison operator is used in various stages of PAES to guide the selection[15].

Reference [20] points out that when appending candidate solutions to the NDS, three things need to be considered: (1) The size of the NDS is limited; (2) The algorithm produces a new non-dominated sub-solution after each iteration; and (3) The distribution of solutions in the NDS must be more uniform than the distribution of solutions in the target space. Based on the above three aspects, for a candidate solution to join the NDS,

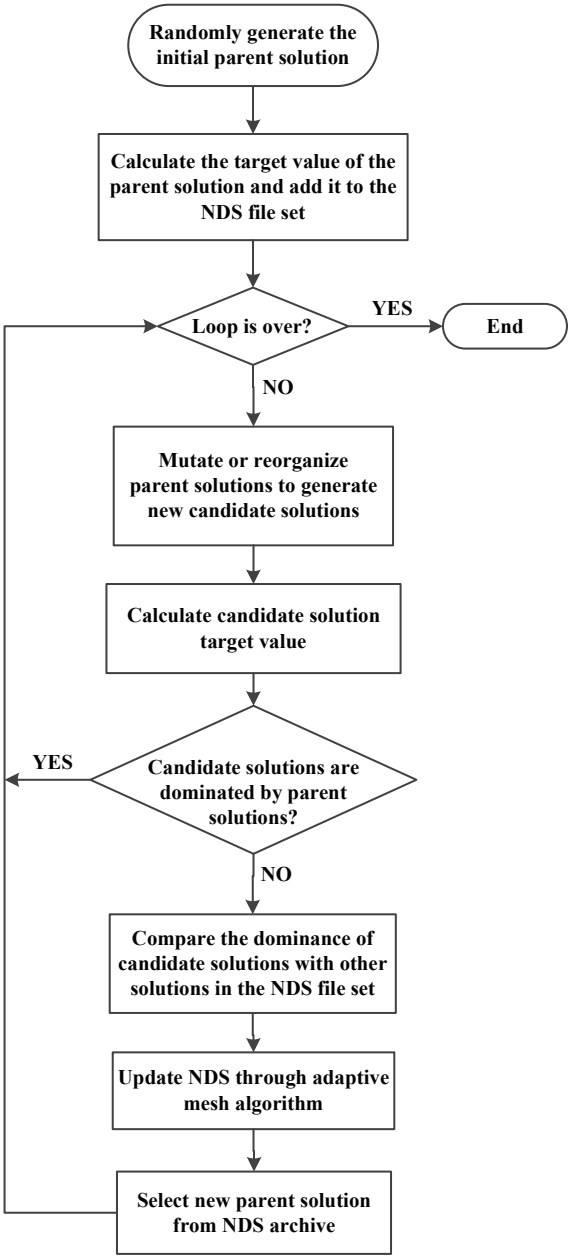

Fig. 1 Process of PAES.

at least one of the following conditions must be met: (1) The NDS is empty; (2) The candidate solution is not dominated by, or the same as, any solution in the NDS; (3) The candidate solution dominates a solution in the NDS; or (4) There is at least one solution in the NDS that non-dominates the candidate solution and has a larger congestion coefficient than the candidate solution. The adaptive grid archiving strategy is shown in Fig. 2.

As shown in Fig. 2, if the candidate solution dominates any solutions in the NDS, all solutions in the NDS that are subject to the candidate solution are deleted, the candidate solution is added into the

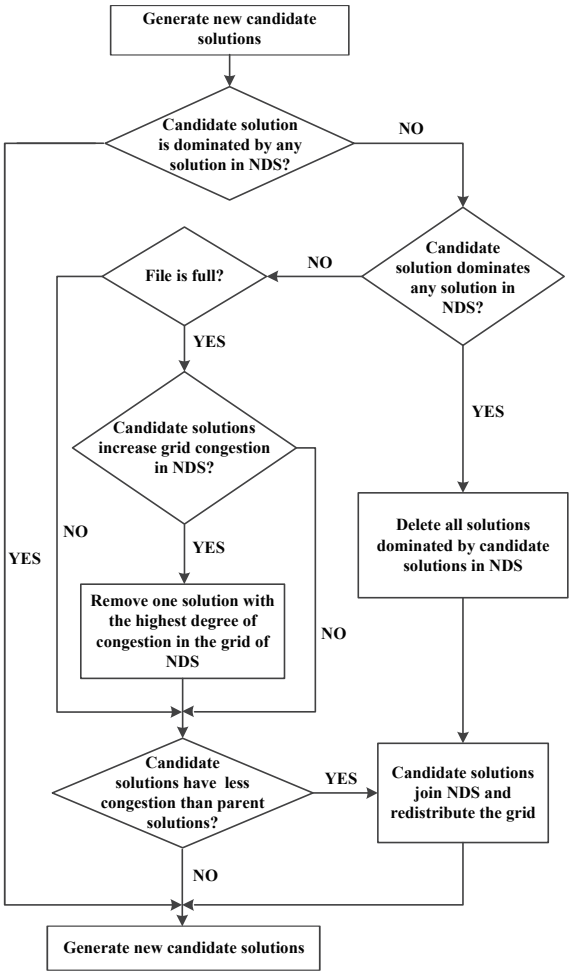

Fig. 2 Process of adaptive grid archiving strategy.

NDS and the grid is updated. Otherwise, if the NDS is full and adding the candidate solution increases the crowding coefficient of the grid in the NDS, a solution in the grid with the largest crowding coefficient will be deleted. When making the judgement in this step, the crowding coefficient of the grid where the candidate solution is located is compared with the parent solution. If the candidate solution is less crowded, the candidate solution is added to the NDS and the grid is updated; otherwise, the candidate solution is discarded and the process moves to the next iteration.

The size of the NDS will increase or decrease in the iterative process, with the size of the grid automatically adjusted through the adaptive algorithm. Figures 3 and 4 show two common scenarios for adding a candidate solution to the NDS and re-dividing the grid. In Fig. 3, the NDS file is full, so the points of NDS which have a large grid congestion coefficient are randomly removed, and the new solution is then added to the NDS. In Fig. 4, when the NDA file set is not full, we need to repartition

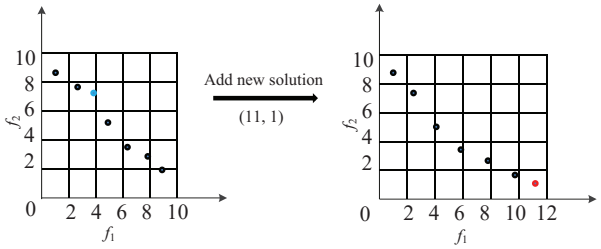

Fig. 3 When the archived set is full, the candidate solutions are added and the grid is updated. (Blue indicates the point to be deleted, red indicates the newly added solution.)

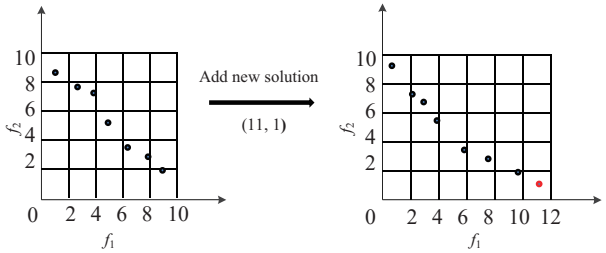

Fig. 4 When the archived set is not full, the mesh is newly meshed and the new solution is added to the NDS.

the grid and add the new solution to the NDS.

2.3 Selective ensemble learning

Selective ensemble learning is a learning algorithm that trains a number of base classifiers and selects some of them to form an ensemble[17]. As shown in Fig. 5, a certain measure is used to select a number of pre-trained base learning machines to form an ensemble base learning machine, with the base learning machine to be processed being equivalent to a different solution to a problem.

2.4 Related works and challenges

A common approach in ensemble learning is to combine all of a set of trained learning machines to obtain better results. Wu et al.[5] put forward the ensemble rule-based learning method, which is combined with AdaBoost and Bagging. It produces better results than a single rule-based system, but as the number of rule-based systems in the ensemble increases, individual differences become increasingly

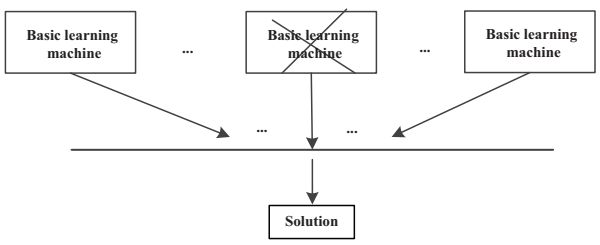

Fig. 5 Basic idea of selective ensemble learning.

difficult to obtain. Furthermore, as a large number of redundant basic learning machines are generated, the prediction rate will significantly reduce, and the storage space requirement will significantly increase, thereby reducing the effective generalization ability. Therefore, Zhou et al.[21] proposed the concept of the selective ensemble in 2002, by which some selection criterion is applied so that only selected basic learning machines are involved in the ensemble.

In order to solve the existing problems, this paper introduces the multi-objective optimization algorithm PAES to select the base classifier for ensemble learning. The improved Bagging algorithm is used as the training strategy to construct the training set, thereby increasing diversity of the base classifier. In the base classifier selection phase, the trained base classifier is binary-coded, an elitist retention strategy is adopted to obtain the PAES optimal solution set, and the solution set is updated using the adaptive grid archiving strategy.

3 Belief Rule-Base Classification System of Selective Ensemble Learning

3.1 Belief rule-base classification system construction methods and training methods

Because belief-rules are traditionally constructed by traversal combination, the RIMER method suffers a "combinatorial explosion" problem during rule building. Taking the categorical Breast Cancer dataset on UCI as an example, containing 30 antecedent attributes, if we assume that each candidate value of the antecedent attributes is set to 3, then the number of constructed BRB rules constructed by traversing the combination is $3^{30} = 205\,891\,132\,094\,649$. The method of traversal combination exponentially increases the combinations with the increase of antecedent attributes, and most real classification problems are multi-attribute. For this reason, Chang et al.[22] proposed the linear combination method of BRBCS, as shown in Fig. 6.

The linear combination method proposed by Chang et al.[22] overcomes the problem of "combination explosion" in the rule construction process, but does not provide a specific solution as to how many rules need to be generated. When the number of rules is too small, the classification performance of the BRB classifier will be reduced; when the number of rules is too high, storage space requirements will soar. Therefore, Ye et al.[23] proposed correlating the number of categories of results

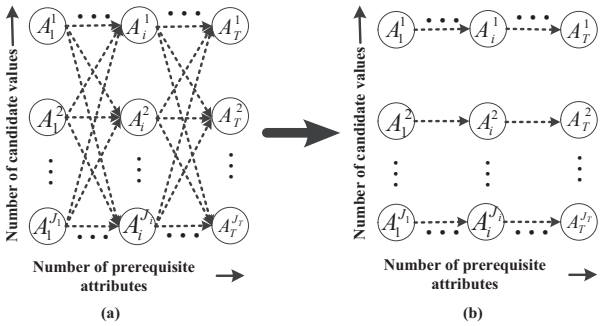

Fig. 6 Different ways of building the rules. (a) Traverse combination and (b) linear combination.

in the classification problem with the number of BRB rules, setting the number of rules equal to the number of categories, and performing a rationality analysis. At the same time, the BRB result evaluation level is mapped to the classification result.

In order to overcome the problem of "zero activation" in the process of rule activation, Ref. [23] improved the calculation method for the individual matching degree when seeking the activation weight, by returning the normalized value of the inverse of the distance from the input parameter to the candidate value in the rule as the individual match degree. Assuming that x_i represents the i-th attribute value of the input data, the formula for calculating the individual matching degree of the k-th rule is as follows:

$$\alpha_i^k = \frac{1/|A_i^k - x_i|}{\sum_{j=1}^{L} 1/|A_j^k - x_j|} \tag{10}$$

The new rule weight is calculated as

$$\omega_k = \frac{\theta_k \sum_{i=1}^{T} \alpha_i^k}{\sum_{l=1}^{L} (\sum_{i=1}^{T} \alpha_i^l)} \tag{11}$$

The constructed belief rule-base is trained by the Differential Evolution (DE) algorithm. This algorithm first initializes the size NP and the number of iterations T of the population $P^t = \{p_1^t, p_2^t, \ldots, p_{\mathrm{NP}}^t\}$. With $t = 0$, NP individuals in the initial population P^0 are initialized randomly. Each individual in the current population x_i^t $(i = 1, 2, \ldots, \mathrm{NP})$ is then mutated to produce a variant individual V_i^{t+1}; the formula is

$$V_i^{t+1} = x_{r1}^t + F \times (x_{r2}^t - x_{r3}^t) \tag{12}$$

where $x_{r1}^t, x_{r2}^t, x_{r3}^t$ are random and satisfy $r_1, r_2, r_3 \in \{1, 2, \ldots, \mathrm{NP}\} \wedge r_1 \neq r_2 \neq r_3 \neq i$, and F is the scaling factor. The above formula cross-reorganizes $V_i^{t=1}$ and X_i^t to produce cross-members U_i^{t+1}; the cross formula is

$$U_{i,j}^{t+1} = \begin{cases} V_{i,j}^{t+1}, & \text{if (rand()} \leqslant \mathrm{CR}) \text{ or } (j = \mathrm{rb}); \\ x_{i,j}^t, & \text{otherwise} \end{cases} \tag{13}$$

where $U_{i,j}^{t+1}$ represents the j-th dimensional element of the individual U_i^{t+1} after crossover, and $j = 1, 2, \ldots, \text{Dim}$, with Dim representing the dimensions of the optimization problem, rand() represents a random number between $[0, 1]$ and CR is a crossover factor in the range $[0, 1]$. rb is a random integer between $\{1, 2, \ldots, \text{Dim}\}$. The fitness value function is used to calculate the fitness value of an individual U_i^{t+1} and x_i^t. Based on the greedy strategy, the individual with a better fitness value is selected as the individual of the new population. The formula for selecting the individual is

$$x_i^{t+1} = \begin{cases} U_i^{t+1}, & \text{if } f(U_i^{t+1}) < f(x_i); \\ x_i^t, & \text{otherwise} \end{cases} \tag{14}$$

where $f(\cdot)$ is the fitness function, the Mean-Square Error (MSE) or Cross Entropy (CE)[24] can be selected.

3.2 Multi-objective optimization based on PAES for BRBCS selective ensemble learning

The selective ensemble learning process of the BRBCS consists of two steps: (1) Base classifier generation; and (2) Base classifier selection.

In order to generate different base classifiers, several training datasets are obtained by using the Bagging data re-sampling technique. Based on the training datasets, a BRB classifier is constructed by linear combination. The DE algorithm is used to train the parameters of the base classifier, then the trained BRB base classifier is binary-coded (with 1 meaning that the base classifier is involved in the ensemble, and 0 meaning that it is not), and PAES multi-objective optimization[19] is used to find the optimal solution.

For the multi-class classification problem, when there are multiple base classifiers involved in the ensemble, the calculation of the integrated generalization error can be deduced as follows.

Assume that C is the number of categories, in which case the actual class label of the j-th sample d_j satisfies $d_j \in \{1, 2, \ldots, C\}$ and the actual class label f_{ij} of the i-th base classifier on the j-th sample satisfies $f_{ij} \in \{1, 2, \ldots, C\}$, then the generalization error of the i-th base classifier on m samples of the training set is

$$E_i = \frac{1}{m} \sum_{j=1}^{m} \text{Error}(K(f_{ij} d_j)) \tag{15}$$

where

$$\text{Error}(x) = \begin{cases} 1, & \text{if } x = -1; \\ 0.5, & \text{if } x = 0; \\ 0, & \text{if } x = +1 \end{cases} \tag{16}$$

$$K(f_{ij}, d_j) = \begin{cases} +1, & f_{ij} = d_j; \\ -1, & f_{ij} \neq d_j \end{cases} \tag{17}$$

The general meaning of sum_j can be expressed as for the j-th sample, the classifier with the highest number of votes is obtained from the voting results of all base classifiers, that is, the mode of the class flag. The output of all base classifiers on the j-th sample can be represented as

$$\hat{f}_j = \text{sum}_j = \text{mode}_{i=1}^{m}(f_{ij}) \tag{18}$$

Therefore, the generalization error of the integrated base classifier for multi-class tag classification problems is

$$\hat{E} = \frac{1}{m} \sum_{j=1}^{m} \text{Error}(G(\hat{f}_j d_j)) \tag{19}$$

where

$$G(\hat{f}_j d_j) = \begin{cases} +1, & \text{if } \hat{f}_j \neq d_j; \\ 1/\text{mode}, & \text{if } \hat{f}_j \doteq d_j, \text{ mode} > 1; \\ 0, & \text{if } \hat{f}_j \doteq d_j, \text{mode} \doteq 1 \end{cases} \tag{20}$$

In Eq. (20), mode denotes the number of modes in the class tag. For example, if there were 10 base classifiers participating in the integration, and the output class tag set is {1, 1, 1, 2, 7, 7, 7, 8, 8}, then $\hat{f}_j' = 1, 7$, mode $= 2$.

Assuming that the k-th base classifier does not participate in the integration and is therefore removed, the output of the new integration base classifier on the j-th sample is

$$\hat{f}_j' = \text{sum}_j = \text{mode}_{i=1}^{m}(f_{ij} - f_{kj}) \tag{21}$$

And the generalization error of the new ensemble base classifier is expressed as

$$\hat{E}' = \frac{1}{m} \sum_{j=1}^{m} \text{Error}(G(\hat{f}_j' d_j)) \tag{22}$$

Algorithm 1 and Fig. 7 show the steps of the selective ensemble learning methods for the BRBCS based on the PAES algorithm. The initial dataset is S, the test dataset is SI, NUM is the number of constructed BRBCSs, and the parameter training method is the DE algorithm, the classifier selection algorithm is PAES multi-objective optimization, and M is the number of BRBs participating in the ensemble.

4 Experimental Evaluation

In order to evaluate the performance of the proposed BRBCS selective ensemble learning method based on PAES, three sets of classification problems were studied. This section analyzes the classification

Algorithm 1 BRBCS-PAES

Input: initial dataset S, test datasets SI, NUM, M.
Output: Prediction results f.

1: **for** $t = 1$ to NUM **do**
2: $S(t) =$ bootstrap sample from S.
3: Constructing initial BRBCS(t) by linear combination.
4: Use DE algorithm to train BRBCS(t).
5: return BRBCS (t').
6: **end for**
7: Load test set SI, calculate the classification Result(t) of BRBCS(t') ($t = 1, 2, \ldots,$ NUM) on SI.
8: Binary encoding of BRBCS-based classifiers.
9: Set the NDS size, number of iterations, and initial adaptive mesh in PAES to obtain the Pareto optimal solution set.
10: Selecting an appropriate Pareto optimal solution according to the generalization error and the number of selected base classifiers.

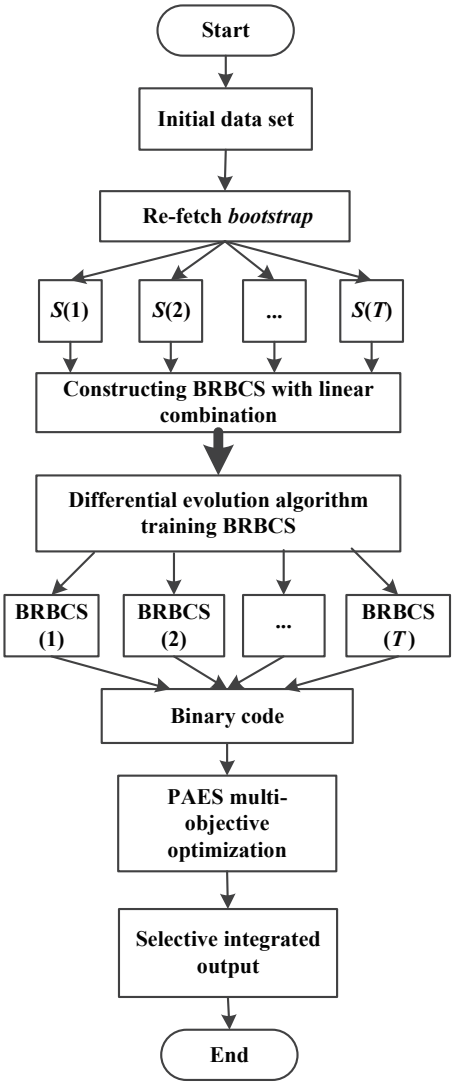

Fig. 7 Process of selective ensemble learning methods of belief rule base classification system based PAES.

accuracy and spatial complexity of the proposed method and compares it with a single expert system, contrasting ensemble learning with the simple voting method with the data-driven ensemble learning of BRBCS, and then reports on the results of tests run on classification datasets. The experimental environment is an Intel Core i5-4570 CPU@3.20 GHz with 8 GB memory running the Windows 10 operating system, and the algorithm is written in Visual Studio 2013.

4.1 Experimental design

The three test datasets used for the experiment was selected from the UCI public test dataset. The three datasets are made up of Breast Cancer data, Iris trait data, and Glass type data. Table 1 lists the number of antecedent attributes and classification categories and the data size of the three datasets.

In the experiment, assuming that the number of categories in the classification problem is C and the number of antecedent attributes is T_k, it can be known from the linear combination of BRBs that each antecedent attribute in the BRB contains C candidate values, the result evaluation level is C levels, the number of rules is L, and the number of training datasets is ND.

The initial settings for each parameter in the BRB were set as follows.

(1) θ_k is the weight of k-th rule, the initial value of θ_k is

$$\theta_k = \text{rand}^k(), k = 1, 2, \ldots, L \tag{23}$$

(2) $\delta_{k,i}$ is the weight of the i-th antecedent attribute in BRB, the initial value of $\delta_{k,i}$ is

$$\delta_{k,i} = \text{rand}^i(), i = 1, 2, \ldots, T_K \tag{24}$$

(3) A_k^i is the referential set of values for the i-th antecedent attributes in k-th rule, the value of A_k^i is

$$A_k^i = \begin{cases} \min\{x_{d,i}\}, d = 1, 2, \ldots, N_D, & \text{if } k = 1; \\ A_i^1 + \frac{(A_i^c - A_i^1)}{(C-1)\times(k-1)}, & \text{if } k \neq 1 \wedge k \neq C; \\ \max\{x_{d,i}\}, d = 1, 2, \ldots, N_D, & \text{if } k = C \end{cases} \tag{25}$$

(4) Q is the evaluation rating for classification results, the value of Q is

$$Q = c\ (c = i = 1, 2, \ldots, C) \tag{26}$$

Table 1 UCI datasets.

Dataset	Attribute number	Category number	Data size
Breast Cancer	30	2	569
Iris	4	3	150
Glass	9	7	214

(5) $\beta_{c,k}$ is the belief degree of the c-th result in the k-th rule, the initial value of $\beta_{c,k}$ is

$$\beta_{c,k} = \text{rand}^c()/\sum_{i=1}^{C}\text{rand}^i() \tag{27}$$

In the DE algorithm, population size NP $= 100$, scale factor $F = 0.5$, crossover factor CR $= 0.9$, and the NDS size in the PAES algorithm was set to 20.

4.2 Selective ensemble for classification problem

To verify the effectiveness of the proposed method, experiments were run three times for each of the three datasets, with a different number of base classifiers generated for each run: 25, 100, and 200. The PAES-BRB method was then used for selective ensemble experiments. Tables 2 – 4 respectively show the average generalization error and the average classification accuracy on the Breast Cancer, Iris, and Glass datasets after applying the PAES selective ensemble with the three different numbers of base classifiers.

Table 2 shows that the average classification accuracy rates on the Breast Cancer dataset after PAES selective ensemble for the three different numbers of base classifiers were higher than 97%, and that the number of base classifiers had little effect on the classification accuracy. Table 3 shows that the average classification accuracy rate on the Iris dataset after PAES selective ensemble for the three different numbers of base classifiers were higher than 98%. In this case, the classification accuracy increased as the number of base classifiers increased, reaching 99.38% with 200 base classifiers. Table 4 shows that the average classification accuracy on the Glass dataset after PAES selective ensemble for the three different numbers of base classifiers were about 70%, with the classification accuracy again increasing as the number of base classifiers increased. From these experimental results, we can see that the proposed method obtains a lower generalization error and a higher classification accuracy when solving classification problems on the Breast Cancer, Iris, and Glass datasets.

To verify that the number of base classifiers participating in integration using this method in fact is reduced, Tables 5 – 7 show the number of classifiers actually participating in the integration when generating different numbers of total classifiers on the three datasets. Indeed, these results show that with an increase of the number of classifiers involved in the ensemble, this method selects a smaller percentage of classifiers for the ensemble.

In order to more intuitively display the classification accuracy and the number of base classifiers that actually participate in integration, some selected experimental results were plotted as shown in Figs. 8 – 10. Figure 8 shows a scatter plot of the classification accuracy of the experiments on the Breast Cancer dataset along with

Table 2 Average generalization errors and accuracy of classification in Breast Cancer experiments.

Method	Generalization error (%)	Average accuracy (%)
BRBCS-PAES(25)	2.20	97.80
BRBCS-PAES(100)	2.01	97.99
BRBCS-PAES(200)	2.31	97.69

Table 3 Average generalization errors and accuracy of classification in Iris experiments.

Method	Generalization error (%)	Average accuracy (%)
BRBCS-PAES(25)	1.18	98.82
BRBCS-PAES(100)	1.15	98.85
BRBCS-PAES(200)	0.62	99.38

Table 4 Average generalization errors and accuracy of classification in Glass experiments.

Method	Generalization error (%)	Average accuracy (%)
BRBCS- PAES(25)	30.70	69.30
BRBCS- PAES(100)	29.82	70.18
BRBCS- PAES(200)	28.36	71.64

Table 5 Involved classifiers of two ensemble learning methods in Breast Cancer experiment.

Total classifiers	BRBCS-Vote	BRBCS-PAES	Percentage (%)
25	25	3.18	12.72
100	100	5.9	5.90
200	200	4.27	2.14

Table 6 Involved classifiers of two ensemble learning methods in Iris experiment.

Total classifiers	BRBCS-Vote	BRBCS-PAES	Percentage (%)
25	25	2.6	10.40
100	100	3.3	3.30
200	200	6.38	3.19

Table 7 Involved classifiers of two ensemble learning methods in Glass experiment.

Total classifiers	BRBCS-Vote	BRBCS-PAES	Percentage (%)
25	25	6	24.00
100	100	8.6	8.60
200	200	9.92	4.96

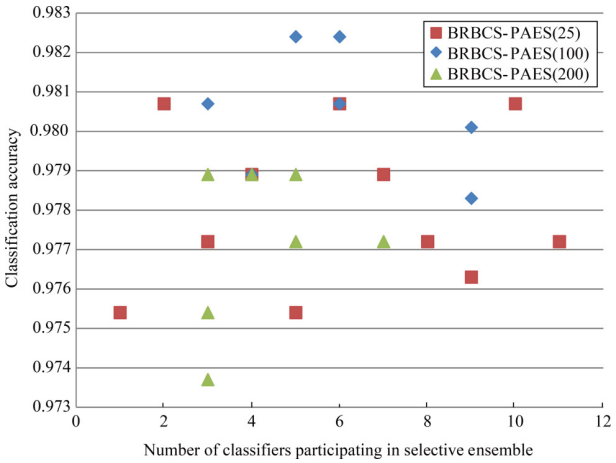

Fig. 8 Breast Cancer's classification accuracy rate of multiple tests and the number of base classifiers participating in the integration.

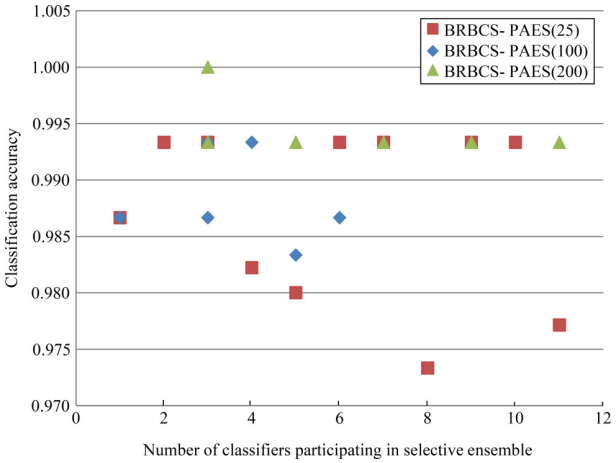

Fig. 9 Iris's classification accuracy rate of multiple tests and the number of base classifiers participating in the integration.

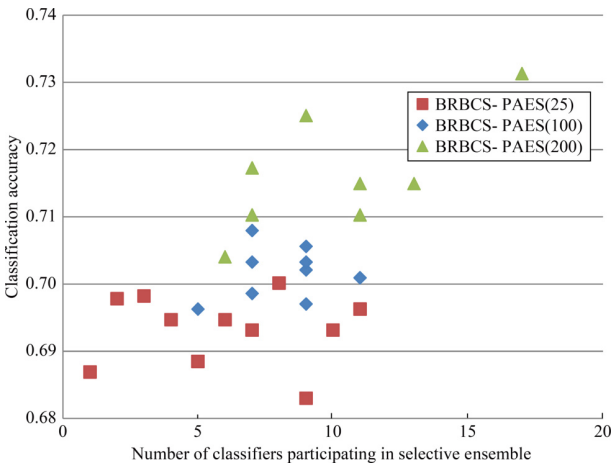

Fig. 10 Glass's classification accuracy rate of multiple tests and the number of base classifiers participating in the integration.

the number of base classifiers participating in ensemble learning. Figures 9 and 10 show the same results for the Iris and Glass datasets, respectively.

Figures 8 – 10 show that BRBCS-PAES can effectively reduce the number of classifiers participating in the integration while ensuring the classification accuracy of the integrated system and reducing its space complexity. Some notable results are as follows. When the base classifier number was set to 25, the highest classification accuracy on the Breast Cancer dataset was 98.07%, with 5 classifiers participating; the highest classification accuracy on Iris was 99.33%, with a single classifier; and the highest classification accuracy on Glass was 70.16%, with 7 classifiers participating in the integration. When the number of base classifiers was set to 100, the highest classification accuracy on the Breast Cancer dataset was 98.24%, with 5 classifiers participating; the highest classification accuracy on Iris was 99.33%, with 4 classifiers; and the highest classification accuracy on Glass was 70.79%, with 7 classifiers participating in the integration. When the number of base classifiers was set to 200, the highest classification accuracy on the Breast Cancer dataset was 97.89%, with 3 classifiers participating; the highest classification accuracy on Iris reached 100%, with 3 classifiers; and the highest classification accuracy on Glass was 73.13%, with 17 classifiers participating in the integration. These experimental results show that the proposed method can reduce the number of classifiers involved in the ensemble and ensure the generalization ability of the ensemble system.

4.3 Performance comparison

In order to further verify the effectiveness of this method, Table 8 provides a comparison between this method and alternative approaches[23], with various numbers of base classifiers. Of these alternatives, Naive Bayes, C4.5, SMO, Fuzzy gain measure, Fallahnezhad, and YE-BRBCS use the average effect achieved by a single classifier, and EBRB-Vote is based on the data-driven EBRB using the AdaBoost algorithm and integrated using the simple voting method. BRBCS-Vote is based on the linear approach and uses a simple voting method to ensemble. BRBCS-PAES adopts the selective ensemble learning methods based on PAES. The classification accuracies of BRBCS-Vote and BRBCS-PAES were obtained by averaging multiple experiments.

From Table 8, we can see that BRBCS-PAES can achieve a higher classification accuracy on the three test datasets with the number of base classifiers set to 25, 100, and 200. Its classification accuracy on the Iris dataset is the highest among the comparison methods, and on the Breast Cancer dataset it is bettered only by the Fuzzy gain measure method. BRBCS-PAES has a lower classification accuracy on the Glass dataset than the EBRB-Vote method, but the ERBB-Vote depends on the data for reasoning. As a result, when dealing with large-scale data the number of rules in the EBRB will become very large, leading to higher storage costs and time consumption while searching for rules. It can be seen from Table 8 that the accuracy of BRBCS-PAES is significantly higher than that of the BRBCS-Vote method. When the number of participating classifiers increases, the BRBCS-Vote classification accuracy decreases. In particular, in the Iris dataset experiment, the BRBCS-Vote classification accuracy reduced by 11.33% when the number of classifiers was increased from 25 to 200. In contrast, BRBCS-PAES has a higher classification accuracy on the three datasets and is less affected by the number of base classifiers.

5 Conclusion

In the present study, we proposed BRBCS-PAES: using selective ensemble learning methods for BRBCS based on PAES. The proposed method is effective in solving the problem of a system's prediction speed decreasing and storage space usage increasing rapidly when the number of base classifiers participating in an ensemble is increased. Experimental studies on classification problems demonstrated that the proposed method can effectively promote the performance of BRBCS. Two main conclusions can be drawn from the study, as follows. (1) BRBCS lacks methods to improve its effective generalization ability as the number of sub-BRBs participating in ensemble learning is increased. The use of selective ensemble learning methods is a good approach to dealing with this problem. (2) The proposed selective ensemble learning method for BRBCS using the multi-objective optimization model as the selective strategy is effective. In this method, sub-BRBs are binary coded in the base classifier selection stage. The number of base classifiers involved in the ensemble and the generalization error of the base classifier is taken as a multi-objective optimization function. An elite retention strategy and adaptive grid archiving strategy are used to produce the Pareto optimal solution set. Comparing the proposed method with existing methods on three classification datasets shows that it improves the accuracy of BBRCS and reduces the number of classifiers participating in the ensemble. In future work, we will investigate how to obtain the optimal number of classifiers to participate in the ensemble.

Table 8 Classification accuracy of the proposed method compared to other methods in different classifiers.

Method	Accuracy (%)		
	Breast Cancer	Glass	Iris
Naive Bayes	95.90	42.90	96.00
C4.5	94.71	67.90	95.13
SMO	97.51	58.85	96.69
Fuzzy gain measure	98.14	69.14	96.88
Fallahnezhad	97.17	56.50	97.46
YE-BRBCS	97.60	61.86	96.93
EBRB-Vote(25)	93.58	86.21	98.60
BRBCS-Vote(25)	95.46	66.35	96.00
BRBCS-Vote(100)	95.43	63.56	94.10
BRBCS-Vote(200)	93.84	61.86	84.67
BRBCS-PAES(25)	96.99	69.30	98.82
BRBCS-PAES(100)	97.99	70.18	98.85
BRBCS-PAES(200)	97.69	71.64	99.38

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Nos. 71501047 and 61773123) and the Natural Science Foundation of Fujian Province (No. 2019J01647).

References

[1] J. B. Yang, J. Liu, J. Wang, H. S Sii, and H. W. Wang, Belief rule-base inference methodology using the evidential reasoning approach-RIMER, *IEEE Trans. Syst. Man Cybern. A Syst. Hum.*, vol. 36, no. 2, pp. 266–285, 2006.

[2] C. L. Hwang and K. Yoon, Methods for multiple attribute decision making, in *Multiple Attribute Decision Making: Methods and Applications A State-of-the-Art Survey*, C. L. Hwang and K. Yoon, eds. Springer, 1981, pp. 58–191.

[3] A. P. Dempster, A generalization of Bayesian inference, *J. Roy. Stat. Soc.*, vol. 30, no. 2, pp. 205–232, 1968.

[4] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, UJ, USA: Princeton University Press, 1976.

[5] W. K. Wu, Y. G. Fu, Q. Su, Y. J. Wu, and X. T. Gong, GDA based ensemble learning methods for parameter training in belief rule base, (in Chinese), *J. Front. Comput. Sci. Technol.*, vol. 10, no. 12, pp. 1651–1661, 2016.

[6] W. K. Wu, L. H. Yang, Y. G. Fu, L. Q. Zhang, and X. T. Gong, Parameter training approach for belief rule base using the accelerating of gradient algorithm, (in Chinese), *J. Front. Comput. Sci. Technol.*, vol. 8, no. 8, pp. 989–1001, 2014.

[7] J. B. Yang, Rule and utility based evidential reasoning approach for multi-attribute decision analysis under uncertainties, *Eur. J. Oper. Res.*, vol. 131, no. 1, pp. 31–61, 2001.

[8] W. He, P. L. Qiao, Z. J. Zhou, G. Y Hu, Z. C Feng, and H. Wei, A new belief-rule-based method for fault diagnosis of wireless sensor network, *IEEE Access*, vol. 6, pp. 9404–9419, 2018.

[9] Z. J. Zhou, G. Y. Hu, B. C. Zhang, C. H. Hu, Z. G. Zhou, and P. L. Qiao, A model for hidden behavior prediction of complex systems based on belief rule base and power set, *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 48, no. 9, pp. 1649–1655, 2018.

[10] X. Yin, B. Zhang, Z. Zhou, Z. Wang, and G. Hu, A novel health estimation model for CNC machine tool servo system based on belief-rule-base, in *Prognostics and System Health Management Conference*, 2017, p. 8079212.

[11] Z. J. Zhou, Z. C. Feng, C. H. Hu, F. J. Zhao, Y. M. Zhang, and G. Y. Hu, Fault detection based on belief rule base with online updating attribute weight, in *Proc.* 32^{nd} *Youth Academic Annual Conference of Chinese Association of Automation*, Hefei, China, 2017, pp. 272–276.

[12] L. L. Chang, Z. J. Zhou, Y. W. Chen, T. J Liao, Y. Hu, and L. H. Yang, Belief rule base structure and parameter joint optimization under disjunctive assumption for nonlinear complex system modeling, *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 48, no. 9, pp. 1542–1554, 2018.

[13] Y. M. Wang, J. B. Yang, D. L. Xu, and K. S. Chin, The evidential reasoning approach for multiple attribute decision analysis using interval belief degrees, *Eur. J. Oper. Res.*, vol. 175, no. 1, pp. 35–66, 2006.

[14] F. Luna, A. J. Nebro, and E. Alba, Observations in using grid-enabled technologies for solving multi-objective optimization problems, *Parallel Comput.*, vol. 32, nos. 5&6, pp. 377–393, 2006.

[15] S. Rostami and A. Shenfield, CMA-PAES: Pareto archived evolution strategy using covariance matrix adaptation for multi-objective optimization, in *Proc.* 12^{th} *UK Workshop on Computational Intelligence*, Edinburgh, UK, 2012, pp. 1–8.

[16] Q. M. Fan, Multi-objective optimization design of vehicle transmission system based on Pareto optimal theory, in *Proc.* 2^{nd} *Int. Conf. Intelligent Computation Technology and Automation*, Changsha, China, 2009, pp. 198–201.

[17] Y. Y. Jiang, Selective ensemble learning algorithm, in *Proc. 2010 Int. Conf. Electrical and Control Engineering*, Wuhan, China, 2010, pp. 1859–1862.

[18] D. W. Corne, J. D. Knowles, and M. J. Oates, The pareto envelope-based selection algorithm for multiobjective optimization, in *Proc. 2000 Int. Conf. Parallel Problem Solving from Nature*, Paris, France, 2000.

[19] R. F. Huang, X. M. Luo, B. Ji, P. Wang, A. Yu, Z. H. Zhai, and J. J. Zhou, Multi-objective optimization of a mixed-flow pump impeller using modified NSGA-II algorithm, *Sci. China Technol. Sci.*, vol. 58, no. 12, pp. 2122–2130, 2015.

[20] X. H. Wu and Q. Xu, Optimization model of multi-objective distribution based on adaptive grid particle swarm optimization algorithm, (in Chinese), *J. Highway Transp. Res. Dev.*, vol. 27, no. 5, pp. 132–136, 2010.

[21] Z. H. Zhou, J. X. Wu, and W. Tang, Ensembling neural networks: Many could be better than all, *Artif. Intell.*, vol. 137, nos. 1&2, pp. 239–263, 2002.

[22] L. L. Chang, Z. J. Zhou, Y. You, L. H. Yang, and Z. G. Zhou, Belief rule based expert system for classification problems with new rule activation and weight calculation procedures, *Inform. Sci.*, vol. 336, pp. 75–91, 2016.

[23] Q. Q. Ye, L. H. Yang, Y. G. Fu, and X. C. Chen, Classification approach based on improved belief rule-base reasoning, (in Chinese), *J. Front. Comput. Sci. Technol.*, vol. 10, no. 5, pp. 709–721, 2016.

[24] R. Y. Rubinstein and D. P. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer, 2004.

Wanling Liu received the MSc degree from Fuzhou University in 2017. She is currently a teacher at Fuzhou University. Her research interests include multi-objective optimization, intelligent decision technology, rule-based inference, big data analysis and data mining, etc.

Weikun Wu received the MSc degree from Fuzhou University in 2017. Now he is a project engineer at Industrial Securities Ltd. His research interests include intelligent decision technology, data mining, etc.

Yingming Wang received the MSc from Huazhong University of Science and Technology, China, in 1987, and the PhD degree from Southeast University, China, in 1991. He is currently a full distinguished professor of Changjiang Scholars Program at Fuzhou University, China. He has published over 123 SCI and 28 SSCI-indexed journal papers and has been the most cited Chinese researchers since 2014. His research interests include multiple criteria decision analysis, data envelopment analysis, rule-based inference, and quality function deployment.

Yanggeng Fu received the PhD degree from Fuzhou University, China, in 2013. He is currently a full associate professor at Fuzhou University, China. His research interests include decision theory and methods, data mining and machine learning, intelligent systems, etc.

Yanqing Lin received the MSc degree from Fuzhou University in 2018. Now she is a project engineer at JingDong Company. Her research interests include decision theory and methods, intelligent systems, data mining and machine learning, etc.