IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 22, 2021, accepted August 8, 2021, date of publication August 18, 2021, date of current version August 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3105796

A Modified Random Forest Based on Kappa Measure and Binary Artificial Bee Colony Algorithm

CHEN ZHANG[1,2,3], XIAOFENG WANG[1], SHENGBING CHEN[1], HONG LI[1], XIAOXUAN WU[1], AND XIN ZHANG[1]

[1]School of Artificial Intelligence and Big Data, Hefei University, Hefei 230009, China
[2]Guochuang Software Company Ltd., Hefei 230009, China
[3]School of Computer Science and Technology, University of Science and Technology of China, Hefei 230009, China

Corresponding author: Chen Zhang (zhangchen0304@163.com)

This work was supported in part by the Major Science and Technology Projects of Anhui Province under Grant 201903a05020011, in part by the National Natural Science Foundation of China under Grant 61806068 and Grant 61672204, in part by Anhui Provincial University Outstanding Talent Cultivation Project under Grant gxgnfx2020117, and in part by the Universities Natural Science Research Project of Anhui Provincial under Grant KJ2019A0836.

ABSTRACT Random forest (RF) is an ensemble classifier method, all decision trees participate in voting, some low-quality decision trees will reduce the accuracy of random forest. To improve the accuracy of random forest, decision trees with larger degree of diversity and higher classification accuracy are selected for voting. In this paper, the RF based on Kappa measure and the improved binary artificial bee colony algorithm (IBABC) are proposed. Firstly, Kappa measure is used for pre-pruning, and the decision trees with larger degree of diversity are selected from the forest. Then, the crossover operator and leaping operator are applied in ABC, and the improved binary ABC is used for secondary pruning, and the decision trees with better performance are selected for voting. The proposed method (Kappa+IBABC) are tested on a quantity of UCI datasets. Computational results demonstrate that Kappa+IBABC improves the performance on most datasets with fewer decision trees. The Wilcoxon signed-rank test is used to verify the significant difference between the Kappa+IBABC method and other pruning methods. In addition, Chinese haze pollution is becoming more and more serious. This proposed method is used to predict haze weather and has achieved good results.

INDEX TERMS Random forest, Kappa measure, artificial bee colony algorithm, haze prediction.

I. INTRODUCTION

RF is an ensemble classifier method proposed by Breiman [1]. Random forest comprises tree classifiers, and meta-classifiers are decision trees constructed by classification and regression trees (CART). A majority vote (for classification) or an average of the outputs of all the single trees (for regression) is used to obtain the output of RF [2]. Compared with a single CART, the majority vote of several CARTs is less susceptible to outliers, which mitigates the volatility due to small data and improves the robustness. Random forest has been applied to many fields, such as the remote sensing [3], crime linkage [4], target detection [5], lymph node segmentation [6], speech emotion recognition [7], hemagglutinin sequence data [8], driver's stress level classification [9], estimation of daily PM2.5 concentrations [10]–[12], CO2 emissions [13].

Random forest can better tolerate noise, deal with continuous attributes and discrete attributes at the same time. Furthermore, RF avoids overfitting and is suited to deal with the missing value and abnormal value. Although random forest has many advantages, it still has some disadvantages, such as poor classification in unbalanced data and inability to control the specific operation inside the model. It can only be overcome by parameter adjustment and other methods. Ishwaran *et al.* [14] proposed a RF model based on survival trees, and proved that the classification ability of random survival forests (RSF) for high-dimensional samples was greater

The associate editor coordinating the review of this manuscript and approving it for publication was Valentina E. Balas.

than that of ordinary random forest. Park *et al.* [15] used quantile regression forest (QRF) to predict the extremal precipitation. Xu *et al.* [16] proposed the weighted random forest (WRF) for classifying text data. Wang *et al.* [17] proposed RF based on particle swarm optimization algorithm (PSO), PSO is applied to optimize the weight of every tree in RF. In the previous modification of the random forest, all decision trees participate in voting. As we all know, the contribution of the different trees in forest is different, and some of them may amplify the wrong prediction, which will reduce the prediction performance of the forest. Therefore, sub-forests obtained by pruning the relatively harmful trees have better performance than complete forests. How to choose the high-quality trees from the forest is the key of our study.

The diversity and average precision of decision trees in random forest are two important indexes to improve the performance of random forest. However, increasing the diversity among the classifiers will inevitably reduce the classification accuracy, and increasing the accuracy of the classifiers will also reduce the diversity. There is a balance between the diversity and the precision, which will make the generalization performance of the model better. Zhou [18] found that using some of the classifiers is better than using all the classifiers. That is, by pruning the base classifier in the ensemble system, the generalization ability will improve. Margineantu and Dietterich [19] proposed Kappa pruning, according to the order of the Kappa measure of the base classifier from small to large, the base classifiers with smaller Kappa measure of the integrated part were selected. With the help of Kappa pruning, good integration results were obtained. In addition, swarm intelligence optimization algorithm can also pruning and select some high quality decision trees to participate in integration. ABC [20] is an algorithm inspired by bee colony behaviour. Compared with particle swarm optimization, genetic algorithms and other similar evolutionary algorithms and swarm intelligence optimization techniques, ABC has good local and global searching ability and few control parameters. ABC is very effective when dealing with large and complex search spaces.

Many continuous variants of ABC have been widely used to deal with continuous optimization problems [22]–[26]. In addition, for discrete optimization problems, discrete optimization algorithms need to be used to solve. For discrete swarm intelligence algorithms, there are mainly four different approaches: genetic operators [27], [28], transfer functions [29], similarity-based approach [30], and logic gate operators [31], [32]. ABC algorithm is applied for knapsack problem, allocation (assignment) problems, facility location problems, feature/attribute space reduction, portfolio selection and so on. Some discrete ABC algorithms have been proposed [33].

For workforce scheduling problem in call centres, neighbourhood structures and abandoning mode are introduced in ABC, an enhanced ABC is tailored [34]. Cauchy OBL strategy is applied to initialize the population, and generate candidate positions, which improves the search method of ABC [35]. The single population ABC is extended to multi-hive cooperative coevolutionary model by divide-and-conquer decomposing strategy and multi-objective handling strategies, which fastens the convergence speed and improves the population diversity [36]. Orthogonal Latin squares method and reinforcement learning are introduced in ABC, which are applied to solve the RNP instances [37]. In addition, various improved ABC have also been applied to the distributed heterogeneous no-wait flowshop scheduling [38], job-shop scheduling problem [39], traveling salesman problem [40], multi-objective resource allocation problem [41]. Therefore, we use the improved ABC to select the high quality trees from the forest for voting.

The main contributions of this paper are as follows:

(1) A novel random forest based on Kappa measure and the improved binary artificial bee colony algorithm (IBABC) is proposed, where Kappa measure is regarded as a pre-pruning way and IBABC is used for secondary pruning. The proposed method can significantly improve the generalization performance.

(2) Kappa measure is used for pre-pruning, the CARTs with poor comprehensive performance are eliminated, the number of CARTs is reduced, and the computational complexity of integrated pruning is reduced.

(3) The crossover operator and leaping operator are applied in ABC, and the improved binary ABC is used for secondary pruning, and the decision trees with better performance are selected for voting.

(4) The computational results on several UCI data sets present that the Kappa+IBABC method is superior to other random forest methods. The proposed algorithm uses fewer base classifiers and obtains better integration performance.

The reminder of the paper is as follows: Section 2 describes the related work. Section 3 presents random forest based on Kappa and IBABC. Section 4 shows the computational results on UCI datasets and haze prediction. Section 5 provides the conclusion.

II. RELATED WORK

A. RANDOM FOREST

RF [1] is an ensemble method which involves construction of several CART via bootstrap sampling. The growth of a single CART in random forest is as follows:

(1) The training set of each tree is generated by bootstrap sampling: n is the number of samples of the original training set, and randomly select n samples from the original training set using a bootstrap sampling method.

(2) The internal nodes of each tree are selected from a subset of randomly selected candidate features: let the number of features in the original dataset be M, and a positive integer $M_{try} \ll M$ is predetermined; at each internal node, M_{try} features are randomly selected from all M features as candidate features, from these M_{try} features, we choose the features that can best separate the data set.

(3) Every tree is allowed to grow without pruning.

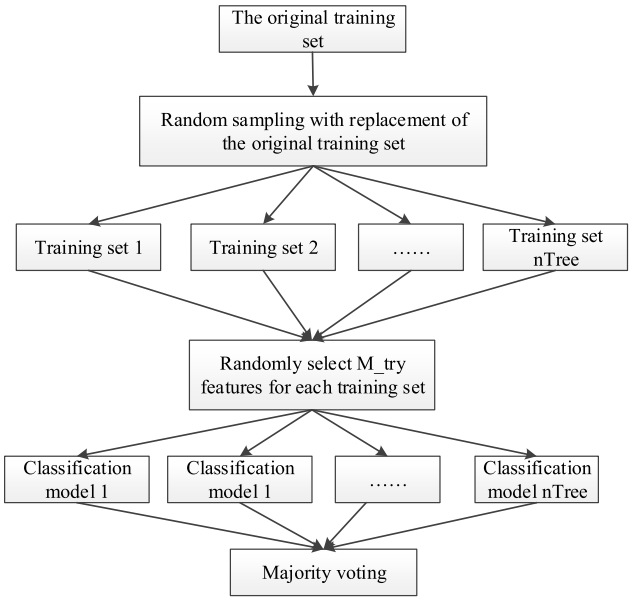

FIGURE 1. The random forest.

The steps of random forest are shown as follows.

For $i = 1$: *nTree*

Using the bootstrap method, each tree is given a training set with the size of n.

Randomly select M_{try} features at nodes, compare and select the best features

Recursively generate each decision tree without pruning.

End

The classification is determined by majority voting.

The process is shown in Figure 1.

It is worth noting that random forest employs all the CATRs to constitute an ensemble outputs. As we all know, the contribution of the different trees in the forest to the algorithm is different, and some harmful trees may amplify the wrong prediction, which will reduce the prediction performance of the forest. Therefore, sub-forests obtained by pruning some relatively harmful trees have better performance than complete forests. Zhou *et al.* [18] proposed a selective ensemble learning method, which deems that ''many can be better than the all''. This method eliminates the base classifiers with poor classification performance, and selects some base classifiers with high accuracy and great diversity to integrate, which can get better generalization performance.

B. ARTIFICIAL BEE COLONY ALGORITHM

Karaboga proposed a novel intelligent optimization algorithm called ABC [20], and it imitated the intelligent behavior of bees. Employed bee, onlooker bee and scout bee make up the entire bee colony. The steps of ABC are shown in Figure 2 [20]. In each search process, the employed bees and onlooker bees explore food sources, the scout bees observe whether they fall into local optimum, and if so, randomly searches other possible food sources. Each food source indicates a possible solution, and the amount of nectar represents the fitness of the solution.

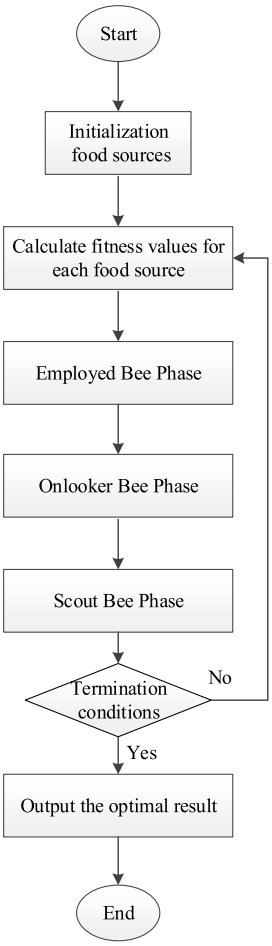

FIGURE 2. The flowchart of ABC algorithm.

The specific description of ABC algorithm is as follows:

(1) Initialize food sources x_{ik} by Eq. (1)

$$x_{ik} = x_k^{min} + rand(0, 1)(x_k^{max} - x_k^{min}) \tag{1}$$

where $i = 1, 2, \ldots, N$ (N is the number of food sources), $k = 1, 2, \ldots, E$(E is the number of parameters). x_k^{max} and x_k^{min} are upper and lower bounds of kth parameter, respectively.

(2) In employed bee phase, a new food source v_{ik} is generated by Eq. (2).

$$v_{ik} = x_{ik} + rand(-1, 1)(x_{ik} - x_{jk}) \tag{2}$$

where j is the randomly selected food source.

(3) In the onlooker bee phase, the onlooker bees wait in the hive and look for the new food sources by sharing information with the employed bee. Calculate the selection probability of each solution according to Eq. (3),

$$P_i = fit_i / \sum_{i=1}^{N} fit_i \tag{3}$$

where fit_i is the fitness value of ith food source.

Algorithm1 ABC

Inputs:

Dataset, *limit* and maximum number of iterations T

Outputs:

The optimal solution.

Procedure:

1: Generate the initial solutions by Eq. (1).
2: Calculate fitness of the solutions.
3: **While** $t \leq T$ **do**
4: //**Employed Bee Phase**
5: **for** i=1: N **do**
6: Search new solution (t) by **Algorithm2**
7: **end for**
8: //**Onlooker Bee Phase**
9: Calculate probabilities P_i by Eq. (3).
10: **for** i=1: N **do**
11: Select a candidate x_i by probability P_i.
12: Search new solution (t) by **Algorithm2**
13: **end for**
14: //**Scout Bee Phase**
if max Trial≥*limit*,
15: Randomly generate a new solution by Eq. (1).
16: $Trial_{max}$=0
17: **end if**
18: **end while**

FIGURE 3. The ABC algorithm.

The quality of the food source is measured by Eq. (4)

$$fit_i = \begin{cases} \dfrac{1}{1+f_i} & \text{if } f_i \geq 0 \\ 1 + abs\,(f_i) & otherwise \end{cases} \tag{4}$$

where f_i is the objective function of the food source x_i.

(4) In scout bee phase, if a certain food source has not been replaced by a better food source after a continuous "*limit*" cycle, the employed bee will be transformed into a scout bee, which randomly generates a new food source by Eq. (1).

The pseudo codes of the ABC algorithm is shown as Algorithm 1 (Figure 3) and Algorithm 2 (Figure 4) [21].

The steps for searching new solution (t) in ABC are shown in Figure 4 [21].

Choosing the better decision trees from the forest can be considered as a discrete optimization problem. The process of bees searching for honey is an optimization process. The concepts mapping between the behavior of bees and an optimization problem is shown in Table 1. Therefore, ABC algorithm is applied to deal with the optimization problems.

III. RANDOM FOREST BASEN ON KAPPA PRUNING AND THE IMPROVED BINARY ARTIFICIAL BEE COLONY ALGORITHM

Random forest integrates all decision trees to get the final result. Some low-quality decision trees will reduce the accuracy of random forest. To improve the accuracy of random forest, a novel RF based on Kappa pruning and IBABC is proposed. In this method, the random forest is pre-pruned by the kappa measure, the CARTs with poor comprehensive performance are eliminated, and the complexity is significantly reduced. Then, an improved binary ABC is proposed by improving the movement mode and search process of bee colony, and crossover operator and leaping operator is introduced. Final, the retaining CARTs are further pruned using IBABC, which can achieve the optimal results.

A. KAPPA PRUNING

Kappa pruning method selects the subset of the classifiers [19]. The diversity is measured by κ statistic. Set two classifier h_1 and classifier h_2, a data set containing m examples and L categories, structure a table where cell C_{ij} contains

Algorithm2 Search new solution (t)

1: **repeat**
2: Generate the candidate solution y by Eq.(1).
3: Calculate the fitness value of y.
4: **if** the current solution x is not better than the fitness of y **then**
5: $x = y$.
6: **end if**
7: **if** solution y is improved **then**
8: Reset *limit* to 0.
9: **else**
10: *limit* ←*limit*+ 1.
11: **end if**
12: **until** repeat t times

FIGURE 4. The steps of search new solution (t) in ABC algorithm.

TABLE 1. The concepts mapping.

The behavior of bees	Optimisation problem
Food source	The possible solution
The location of the food source	The location of the solution
Amount of nectar from food sources	The fitness value
The process of searching food sources	Problem solving process
Maximum nectar quantity nectar source	The optimal solution

the number of examples x for which $h_1(x) = i$ and $h_2(x) = j$.

$$p_0 = \frac{\sum_{i=1}^{L} C_{ii}}{m} \tag{5}$$

$$p_e = \sum_{i=1}^{L} \left(\sum_{j=1}^{L} \left(\frac{C_{ij}}{m}\right) \cdot \sum_{j=1}^{L} \left(\frac{C_{ji}}{m}\right)\right) \tag{6}$$

p_0 is the probability that the two classifies agree(this is just the sum of the diagonal elements divided by m), p_e is the probability that the two classifiers agree by chance, given the observed counts in the table.

Then, the κ statistic is defined as Eq. (7)

$$k(h_1, h_2) = \frac{p_0 - p_e}{1 - p_e} \tag{7}$$

where $\kappa = 0$ means classifiers whose agreement equals that expected by chance. $\kappa = 1$ means classifiers that agree on every example in dataset.

The steps for Kappa pruning are shown as Algorithm 3 (Figure 5).

B. THE IMPROVED BINARY ARTIFICIAL BEE COLONY ALGORITHM

In essence, the classifier selection is a combinatorial optimization problem. In order to make traditional ABC suitable for solving discrete problems, discretization is carried out on ABC. For binarization of the swarm intelligence algorithm, there are mainly four approaches: genetic operators, transfer functions, similarity-based approach and logic gate operators. Compared with the other three methods, genetic operators are simple and easy to operate. Therefore, genetic operation is introduced into ABC algorithm in this paper. The search behavior of ABC is improved, the crossover and leaping behaviors are introduced to increase the diversity of the population and avoid falling into the local optimum, which can improve the search efficiency.

1) INITIALIZATION PHASE

The solution (food source) is expressed by a binary string.

Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{iE})$ be the ith solution, which can be generated by Eq. (8).

$$x_{ik} = \begin{cases} 1, & rand \geq 0.5 \\ 0, & rand < 0.5 \end{cases} \tag{8}$$

where E is the number of CARTs, *rand* is a random number between 0 and 1.

If $x_{ik} = 1$, it indicates that the corresponding CART is selected; If $x_{ik} = 0$, it indicates that the corresponding CART is excluded;

For example,

$$x = (1, 0, 0, 0, 0, 1, 1, 1, 0),$$

x means the first, sixth, seventh and eighth CARTs are selected.

Algorithm3: Kappa pruning

Inputs:

D_v: dataset;

H = $\{h_1, h_2, \cdots, h_L\}$: Set of classifiers to be selected, |H| = L

KM: Kappa measure

Outputs:

S = $\{h_1^*, h_2^*, \cdots, h_M^*\}$: Set of classifiers is selected

M: The size of the subset of classifiers

Procedure:

1: S ← {h}, H ← H\S, where h is the classifier with the smallest error rate

2: **for** i = 1: M **do**

3: Calculate the average diversity between classifier h_j and all classifiers h_k in S :

4: $divh_{j,S} \leftarrow \frac{1}{|S|}\sum_{h_k \in S} KM(h_j, h_k)$

5: Select the classifier h with the smallest diversity $divh_{j,S}$

6: S ← S ∪ {h}, H ← H\{h}

7: **end**

FIGURE 5. The steps for kappa pruning.

2) THE FITNESS FUNCTION

The fitness function F is shown in Eq. (9)

$$F = \frac{1}{A} \tag{9}$$

where A is the classification accuracy, $A = \frac{1}{m}\sum_{j=1}^{m} L(f(x_j), y_j)$, $L\left(f\left(x_j\right), y_j\right) = \begin{cases} 1, f\left(x_j\right) = y_j \\ 0, f\left(x_j\right) \neq y_j \end{cases}$, m is the number of samples, $f\left(x_j\right) = \text{argmax}(N_i)$ (N_i is the number of CARTs that predict x_j as class i) is the actual output of the integration to the sample x_j, and y_j is the real class of the sample x_j.

3) EMPLOYED BEES PHASE

Employed bee search phase, a new candidate solution v_i is generated by the randomly selected food source x_k in population and the current food source x_i.

To avoid falling into local optimum, the single-point crossover operator and leaping operator are introduced to produce the new candidate solution.

The single-point crossover operator:

For example:

$$\begin{aligned} & x_i = (1, 1, 0, 1, 1, 0, 0, 1, 0) \\ \text{crossover} \quad & x_k = (1, 1, 0, 1, 0, 1, 1, 0, 1) \\ & v_i = (1, 1, 0, 1, 1, 1, 1, 0, 1) \end{aligned}$$

The cross-over point is randomly set. The bits after the cross-over point of the current solution x_i and the randomly selected solution x_k are interchanged, and the interchanged x_i is the new candidate food source v_i

Leaping operator:

Change the jth bit of v_i from 0 (1) to 1 (0). j is randomly selected.

If the current solution x_i is superior to the new candidate solution v_i, then x_i remains unchanged. Otherwise, v_i replaces x_i.

4) ONLOOKER BEES PHASE

The onlooker bees wait round hive and look for the new food sources by sharing information with the employed bee. Calculate the selection probability of each solution by Eq. (10),

$$P_i = fit_i / \sum_{i=1}^{N} fit_i \tag{10}$$

where fit_i is the fitness of jth food source.

The onlooker bee selects the largest probability solution as current solution x_i, and then randomly selected food source x_k in population. The next steps are the same as the employed bee phase.

5) SCOUT BEES PHASE

If a certain solution has not been replaced after a continuous "*limit*" cycle, the employed bee will be transformed into a scout bee. The scout bee randomly produces s a new solution by Eq. (8).

6) REPAIR THE INFEASIBLE SOLUTION

In the searching process, if the number of "1" (or "0") in the new solution is equal to E(E is the number of CARTs in random forest), we consider it to be an infeasible solution. The infeasible solution will be replaced by a new generated solution by Eq. (8).

7) ALGORITHM STEPS

The pseudo-code of random forest based on Kappa measure and IBABC is demonstrated as Algorithm4 (Figure 6).

Algorithm4 Random forest based on Kappa measure and IBABC

Inputs:

Training set, testing set.

Outputs:

The optimal sub random forest, and its accuracy.

Procedure:

1: The random forest with L CARTs is construced, $H = \{h_1, h_2, \cdots, h_L\}$

2: The random forest is pre-pruned by Kappa measure, $S=\{h_1^*, h_2^*, \cdots, h_M^*\}$

3: The random forest is secondary-pruned by IBABC

3.1: Initialization: Produce the 2N initial population by Eq.(8)

3.2: The population are divided into employed bees and onlooker bees by the fitness value

3.3: **While** $t \le T$ **do**

// Employed bee phase

3.4: **for** i=1*:* N **do**

3.5: Produce a candidate solution v_i for x_i using the crossover and leaping operator in section 3.B.3

3.6: **if** $f(v_i) \le f(x_i)$

3.7: Replace x_i by v_i

3.8: Trial (i)=0

3.9: **else**

3.10: Trial (i)= Trial (i)+1

3.11: **end**

3.12: **endfor**

3.13: Calculate selection probabilities P_i by Eq.(10)

// Onlooker bee phase

3.14: **for** i=1 : N **do**

3.15: Select a candidate x_i by probability P_i

3.16: Produce a candidate solution v_i for x_i using the method in section 3.B.4

3.17: **if** $f(v_i) \le f(x_i)$

3.18: Replace x_i by v_i

3.19: $Trial_i = 0$

3.20: **else**

3.21: $Trial_i = Trial_i + 1$

3.22: **end**

3.23: **endfor**

// **Scout bee phase**

3.24: **if** max Trial≥*limit*,

3.25: Replace v_i with a randomly generated solution by Eq.(8),

3.26: $Trial_{max}$=0

3.27: **endif**

3.28: Save the optimal solution

3.29: **endwhile**

FIGURE 6. The pseudo-code of random forest based on kappa measure and IBABC.

IV. EXPERIMENTAL RESULTS

Two sets of experiments are tested the capabilities of the proposed Kappa+IBABC algorithm. The first set of experiments is carried out on UCI data set. The second set of experiments use the Kappa+IBABC to predict the haze weather in China.

TABLE 2. Experimental data set.

Data set	Classes	Instances	Attributes
Abalone	3	4177	8
Banks	2	45211	17
Car Evaluation	4	1728	6
Letter	26	20000	16
Wine Quality	7	4898	11
Yeast	4	1484	8

The experiment is programmed by MATLAB 2016a. Experimental results are the average over the 30 runs of 5-fold CV.

A. EXPERIMENTAL RESULTS ON THE BENCHMARK DATA-SETA

The proposed method was compared with some single classifier and other improved random forest methods on several UCI data set. Table 2 shows the properties of data sets.

To show the performance of the proposed method, the classification accuracy is used as the evaluation criteria. $A = 1/\mathrm{m}\sum_{j=1}^{m} L(f(x_j), y_j)$, $\mathrm{L}\left(f\left(x_j\right), y_j\right) = \begin{cases} 1, f\left(x_j\right) = y_j \\ 0, f\left(x_j\right) \neq y_j \end{cases}$, m is the number of samples, $f\left(x_j\right) = \mathrm{argmax}(N_i)$ (N_i is the number of CARTs that predict x_j as class i) is the actual output of the integration to the sample x_j, and y_j is the real class of the sample x_j.

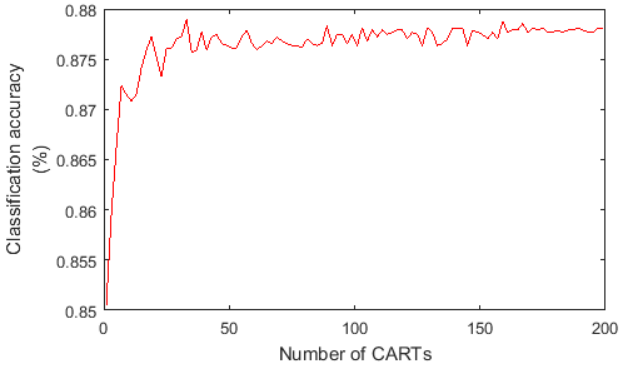

FIGURE 7. The classification accuracy of different sizes on banks.

In order to determine the number of CARTs in random forest, Figure 7 shows the trend between the different sizes of classifiers in random forest and the classification accuracy on dataset Banks. The classification accuracy increases with the growth of the number of classifiers. When the number of classifiers reaches 70, the accuracy is the highest, and then tends to be stable. For Banks the number of CARTs in the random forest is set to 70.

After determining the number of CARTs in random forest, kappa measure is used for pre-pruning. In order to determine the number of CARTs retained by Kappa measure, take Banks as an example. Figure 8 shows the trend between the number of CARTs retained by Kappa measure and the classification accuracy on Banks. It can be seen from Figure 8, with the increase of the number of reserved classifiers, the classification accuracy increases rapidly. When the number of classifiers reaches a certain number, the classification accuracy declines. The possible reason is that as the number of classifiers increases, the degree of diversity also increases, so the accuracy increases. However, when the number of classifiers reaches a certain amount, even if the number of classifiers continues to increase, it has no effect on the accuracy. From Figure 8, it is found that the performance is better when the number of reserved classifiers is 18 for Banks dataset. Finally, IBABC is used to perform secondary pruning of the retained classifier.

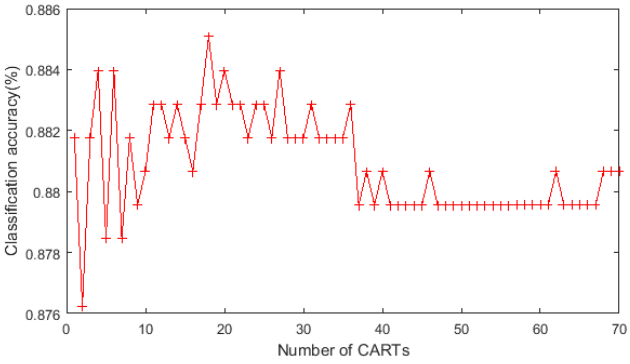

FIGURE 8. The trend between the number of CARTs retained by kappa measure and the classification accuracy on banks.

Table 3 presents the classification accuracy of different methods on UCI data sets. "Max", "Mean" and "Min" represent the maximum, average and minimum values of classification accuracy of CARTs in random forest, respectively. RF represents the traditional random forest method. Kappa represents the classification accuracy obtained by using Kappa pruning for random forest. IBABC means only using IBABC algorithm to prune the random forest. Kappa+IBABC means that the base classifier pool is firstly pre-pruned by Kappa measure, and then the secondary pruning is performed by IBABC algorithm. From Table 3 we can find that the accuracy of RF is better than that of any trees in the forest. The classification accuracy using only Kappa or IBABC is better than that of RF. The accuracy of Kappa+IBABC is optimal. The possible reason is that RF uses all the decision trees to participate in the voting, Kappa and IBABC only have one pruning. While Kappa+IBABC has two pruning. Firstly, Kappa measure is used for pre-pruning, and the decision trees with a larger degree of diversity are selected from the forest. Then, the IBABC is used for secondary pruning, and the decision trees with better performance are selected for voting.

In addition, the numbers in parentheses represent the number of CARTs participating in voting. The proposed method has the least number of decision trees and the highest accuracy.

To further verify the performance of the Kappa+IBABC method, which is compared with the conventional classifiers

TABLE 3. Classification accuracy of different methods on UCI data sets.

Data set	max	mean	min	RF	Kappa	IBABC	Kappa+IBABC
Abalone	0.7412	0.6835	0.6146	0.7780(70)	0.7802(31)	0.7989(45)	**0.8076(25)**
Banks	0.8854	0.8601	0.7945	0.8997(70)	0.8987(18)	0.9083(31)	**0.9125(15)**
Car Evaluation	0.9670	0.9510	0.9179	0.9689(80)	0.9780(38)	0.9759(41)	**0.9821(27)**
Letter	0.8361	0.7917	0.7209	0.9558(70)	0.9615(45)	0.9620(47)	**0.9652(25)**
Wine Quality	0.5851	0.4989	0.4061	0.6752(70)	0.6781(37)	0.7297(43)	**0.7310(27)**
Yeast	0.5460	0.4698	0.3778	0.6114(60)	0.6265(29)	0.6679(45)	**0.6963(21)**

Note: Bold represents the optimal accuracy of each dataset.

TABLE 4. The comparison of different algorithms on data-sets from Wang *et al.* [17].

Data set	Abalone	Banks	Car Evaluation	Letter	Wine Quality	Yeast
Kappa+IBABC	**0.8076**	**0.9125**	**0.9821**	**0.9652**	**0.7310**	**0.6963**
PSOWRF	0.8012	0.9005	0.9753	0.7567	0.6037	0.6798
WRF	0.7841	0.8993	0.9657	0.714	0.5905	0.6772
RSF	0.7627	0.9042	0.9673	0.7325	0.6072	0.6786
QRF	0.7706	0.8975	0.971	0.6872	0.5894	0.6743
RF	0.7535	0.8977	0.9602	0.6664	0.5877	0.676
DT	0.7184	0.8854	0.9313	0.5354	0.4977	0.5723
SVM	0.7768	0.9014	0.9728	0.5993	0.5568	0.6822
BP	0.6529	0.8643	0.6951	0.067	0.0893	0.4516

such as general weighted random forest (WRF), quantitive regression forest (QRF), random survival forest (RSF), traditional random forest (RF), C4.5 decision tree classifier (DT), support vector machine (SVM) and BP neural network mentioned in literature [17]. We find that the Kappa+IBABC achieves the optimal results on all data set in Table 4.

In addition, the Kappa+IBABC is compared with those methods in literature Daho and Chikh [42], which are a single CART tree, Bagging (CART trees), PERT (Perfect Random Trees), SubBag (Subspaces Bagging), RSM (Random Subspaces Method classical RF) and Sub_RF (Sub Spaces Random Forests). Table 5 presents the data-sets from Daho and Chikh [42], and Table 6 shows the results of the comparative experiment, which shows that the Kappa+IBABC achieves the optimal results on most data sets expect the Yeast data set.

In this section, explore the influence of parameters on ABC algorithm. Take Banks dataset as example, Figure. 9 shows the relationship between population size and classification accuracy on Banks. As the population size increases, the classification accuracy increases significantly. But when the population size reaches a certain level, the accuracy rate will no longer improve.

Figure 10 shows the relationship between the classification accuracy and the number of iterations on Banks. As the number of iterations increases, the classification accuracy increases significantly. But when the population size reaches a certain level (170), the accuracy rate will no longer improve.

TABLE 5. Data-sets from Daho and Chikh [42].

Data set	Instances	Classes	Attributes
Breast	699	2	9
Pendigits	10992	10	16
Habermann	306	2	3
Liver(bupa)	345	2	6
Vehicle	846	4	18
Pima	768	2	8
Segmentation	2310	7	19
Isolet	7797	26	617
Ecoli	366	8	7

Wilcoxon signed-rank test at significance level of 5% is used to test the significance of difference between Kappa +IBABC and other methods. The p-values are reported in Table 7, which shows that the Kappa+IBABC gets better results with statistical difference in all cases.

B. THE HAZE FORECAST BASED ON KAPPA+IBABC

Haze pollution [43]–[45] is increasingly serious in China. When haze appears, it often accompanied by reduced visibility and poor air quality. Haze pollution seriously affects people's daily life and health. Therefore, it is necessary to forecast the haze and minimize its negative effects. The formation of haze weather is mainly related to

TABLE 6. The comparison of different methods on the data-sets from Daho and Chikh [42].

	One Tree	Bagging	PERT	RSM	RF	SubBag	Sub_RF	Kappa+IBABC
Breast	0.3689	0.9617	0.9697	0.9652	0.9600	0.9700	0.9597	**0.9817**
Ecoli	0.7568	0.8124	0.7445	0.8411	0.8328	0.8449	0.8501	**0.8751**
Habermann	0.6993	0.6988	0.7356	0.7397	0.7449	0.7418	0.7589	**0.8269**
Isolet	0.8018	0.8890	0.8291	0.911	0.9101	0.9192	0.9219	**0.9534**
Liver	0.5882	0.7021	0.7184	0.6884	0.7188	0.7397	0.7499	**0.9134**
Pendigits	0.7988	0.8694	0.8678	0.875	0.877	0.8713	0.8810	**0.9027**
Pima	0.6681	0.7594	0.7459	0.7548	0.7613	0.7689	0.7717	**0.8474**
Segment	0.8886	0.9695	0.9535	0.9675	0.9727	0.9746	0.974	**0.9859**
Vehicle	0.6679	0.7432	0.7276	0.7404	0.7481	0.7393	0.7472	**0.9082**
Yeast	0.5673	0.5997	0.5756	**0.752**	0.6127	0.6275	0.6356	0.6972

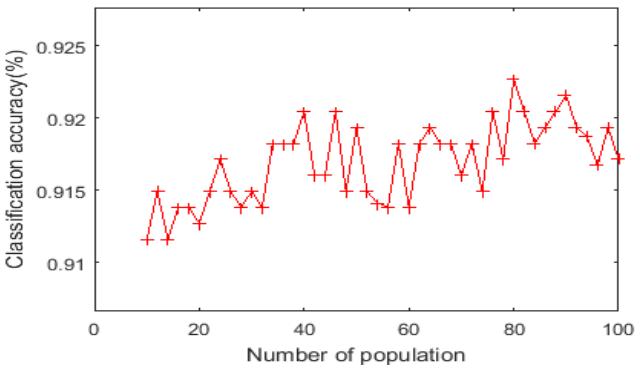

FIGURE 9. The relationship between the number of population and classification accuracy of ABC algorithm on banks.

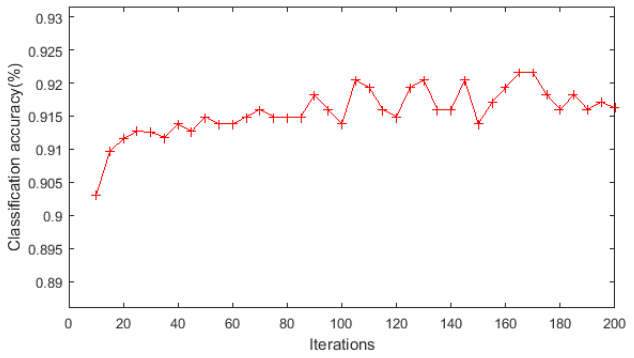

FIGURE 10. The relationship between the number of iterations and classification accuracy of ABC algorithm on banks.

meteorological conditions and air pollutants [46]–[48]. The main influencing factors of haze include PM2.5 concentration, PM10 concentration, CO concentration, NO2 concentration, SO2 concentration, O3 concentration, relative humidity, wind speed, visibility, rainfall and air temperature. In this paper, the forecast factors of previous day are used to predict whether the next day is haze weather.

The data came from the Beijing and Shenzhen meteorological bureaus from January 2018 to December 2019. Figure 11 shows the distribution figure of monthly average

TABLE 7. Results of the wilcoxon signed-rank test.

Comparison	*p*-value	Significant or not
Kappa+IBABC vs WRF	0.0003	yes
Kappa+IBABC vs RSF	0.0008	yes
Kappa+IBABC vs QSF	0.0017	yes
Kappa+IBABC vs PSOWRF	0.0007	yes
Kappa+IBABC vs Bagging	0.0026	yes
Kappa+IBABC vs RSM	0.0006	yes
Kappa+IBABC vs RF	0.0009	yes
Kappa+IBABC vs SubBag	0.001	yes
Kappa+IBABC vs Sub_RF	0.0003	yes
Kappa+IBABC vs Kappa	0.0027	yes
Kappa+IBABC vs IBABC	0.0016	yes

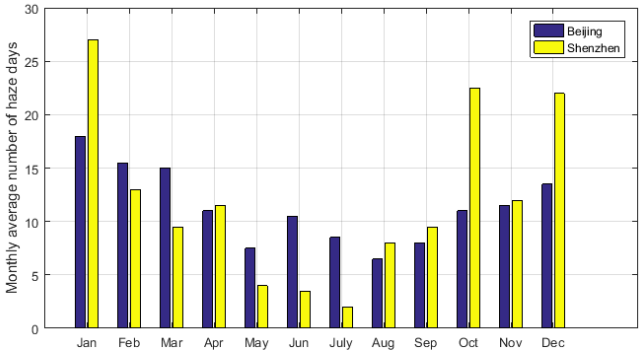

FIGURE 11. The distribution figure of monthly average of haze days in Beijing and Shenzhen.

of haze days in Beijing and Shenzhen. Beijing and Shenzhen have the most haze days in winter and the least in summer, which shows the distribution characteristics of high in winter and low in summer.

Table 8 shows the haze's forecast accuracy of Beijing and Shenzhen in various methods, Kappa+IBABC achieves the optimal results.

TABLE 8. The haze's forecast accuracy of Beijing and Shenzhen in various methods(%).

Data set	max	mean	min	RF	Kappa	IBABC	Kappa+IBABC
Beijing	0.7274	0.6236	0.5066	0.7307(70)	0.7558(35)	0.8144 (47)	0.8310 (27)
Shenzhen	0.7764	0.6921	0.6173	0.7517(70)	0.7613(36)	0.8148 (41)	0.8313 (27)

V. CONCLUSION

Some of poor CARTs in RF will reduce the prediction performance of the random forest. The diversity and average precision of decision trees in random forest are two important indexes to improve the performance of random forest. To deal with the above issue well, we propose the novel method using the combination of Kappa measure and IBABC. Comparative studies on several UCI datasets were implemented. Computational results demonstrate that the Kappa+IBABC method improves the performance on most of datasets with less number of decision trees. In addition, haze pollution in China is becoming increasingly serious. The Kappa+IBABC method is used to predict the haze and reduce the impact of haze on people's daily life.

The next research work is to study other difference measures and apply them to pre-pruning, so as to provide a base classifier with good performance and large difference for integrated pruning technology.

REFERENCES

[1] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[2] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J., Promoting Commun. Statist. Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020.

[3] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine vs. random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6308–6325, 2020.

[4] Y.-S. Li, H. Chi, X.-Y. Shao, M.-L. Qi, and B.-G. Xu, "A novel random forest approach for imbalance problem in crime linkage," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105738.

[5] Y. Dong, B. Du, and L. Zhang, "Target detection based on random forest metric learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 4, pp. 1830–1838, Apr. 2017.

[6] W. Zhao and F. Shi, "A method for lymph node segmentation with scaling features in a random forest model," *Current Proteomics*, vol. 15, no. 2, pp. 128–134, Mar. 2018.

[7] S. Yan, L. Ye, S. Han, T. Han, and Y. Li, "Speech interactive emotion recognition system based on random forest," in *Proc. IWCMC*, New York, NY, USA, 2020, pp. 585–590.

[8] Y. Yao, X. Li, B. Liao, L. Huang, P. He, F. Wang, J. Yang, H. Sun, Y. Zhao, and J. Yang, "Predicting influenza antigenicity from hemagglutintin sequence data based on a joint random forest method," *Sci. Rep.*, vol. 7, no. 1, May 2017, Art. no. 1545.

[9] H. Neska, J.-M. Poggi, R. Ghozi, S. Sevestre-Ghalila, and M. Jaïdane, "Random forest-based approach for physiological functional variable selection for driver's stress level classification," *Stat. Methods Appl.*, vol. 28, pp. 157–185, Feb. 2017.

[10] M. Stafoggia, T. Bellander, S. Bucci, M. Davoli, K. de Hoogh, F. de' Donato, C. Gariazzo, A. Lyapustin, P. Michelozzi, M. Renzi, M. Scortichini, A. Shtein, G. Viegi, I. Kloog, and J. Schwartz, "Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model," *Environ. Int.*, vol. 124, pp. 170–179, Mar. 2019.

[11] Y. Wang, Y. Du, and J. Wang, "Calibration of a low-cost $PM_{2.5}$ monitor using a random forest model," *Environ. Int.*, vol. 133, Oct. 2019, Art. no. 105161.

[12] C. Zhao, Q. Wang, J. Ban, Z. Liu, Y. Zhang, R. Ma, S. Li, and T. Li, "Estimating the daily $PM_{2.5}$ concentration in the Beijing-Tianjin-Hebei region using a random forest model with a 0.01×0.01 spatial resolution," *Environ. Int.*, vol. 134, Jan. 2020, Art. no. 105297.

[13] L. Wen and X. Yuan, "Forecasting CO_2 emissions in China commercial department, through BP neural network based on random forest and PSO," *Sci. Total Environ.*, vol. 718, May 2020, Art. no. 137194.

[14] H. Ishwaran, U. B. Kogalur, and E. H. Blackstone, "Random survival forests," *Ann. Appl. Statist.*, vol. 2, no. 3, Nov. 2008, Art. no. 841860.

[15] S. Park, J. Kwon, J. Kim, and H.-S. Oh, "Prediction of extremal precipitation by quantile regression forests: From SNU Multiscale Team," *Extremes*, vol. 21, no. 3, pp. 463–476, May 2018.

[16] B. Xu, X. Guo, Y. Ye, and J. Cheng, "An improved random forest classifier for text categorization," *J. Comput.*, vol. 7, no. 12, pp. 2913–2920, Dec. 2012.

[17] J. Wang, X. X. Cheng, and J. Z. Peng, "A weighted random forest model based on particle swarm optimization," *J. Zhengzhou Univ.*, vol. 50, no. 1, pp. 72–76, May 2018.

[18] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, no. 1, pp. 239–263, May 2002.

[19] D. D. Margineantu and T. G. Dietterich, *Pruning Adaptive Boosting*. NewYork, NY, USA: CML, 1997, pp. 211–218.

[20] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Dept. Comput. Eng., Eng. Faculty, Erciyes Univ., Kayseri, Turkey, 2005.

[21] G. Chen, P. Sun, and J. Zhang, "Repair strategy of military communication network based on discrete artificial bee colony algorithm," *IEEE Access*, vol. 8, pp. 73051–73060, 2020.

[22] H. Xing, F. Song, L. Yan, and W. Pan, "On multicast routing with network coding: A multiobjective artificial bee colony algorithm," *China Commun.*, vol. 16, no. 2, pp. 160–176, Feb. 2019.

[23] S. Panda, "Performance improvement of optical CDMA networks with stochastic artificial bee colony optimization technique," *Opt. Fiber Technol.*, vol. 42, pp. 140–150, May 2018.

[24] J.-M. Huang, R.-J. Wai, and G.-J. Yang, "Design of hybrid artificial bee colony algorithm and semi-supervised extreme learning machine for PV fault diagnoses by considering dust impact," *IEEE Trans. Power Electron.*, vol. 35, no. 7, pp. 7086–7099, Jul. 2020.

[25] H. Wang, W. Wang, S. Xiao, Z. Cui, M. Xu, and X. Zhou, "Improving artificial bee colony algorithm using a new neighborhood selection mechanism," *Inf. Sci.*, vol. 527, pp. 227–240, Jul. 2020.

[26] H. Jia, H. Miao, G. Tian, M. Zhou, Y. Feng, Z. Li, and J. Li, "Multiobjective bike repositioning in bike-sharing systems via a modified artificial bee colony algorithm," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 2, pp. 909–920, Apr. 2020.

[27] C. Ozturk, E. Hancer, and D. Karaboga, "A novel binary artificial bee colony algorithm based on genetic operators," *Inf. Sci.*, vol. 297, pp. 154–170, Mar. 2015.

[28] T. Bayraktar, F. Ersöz, and C. Kubat, "Effects of memory and genetic operators on artificial bee colony algorithm for single container loading problem," *Appl. Soft Comput.*, vol. 108, Sep. 2021, Art. no. 107462.

[29] D. C. Tran and Z. Wu, "Binary artificial bee colony algorithm for solving 0-1 knapsack problem," *Int. J. Adv. Inf. Sci. Service Sci.*, vol. 4, no. 22, pp. 464–470, Dec. 2012.

[30] M. H. Kashan, N. Nahavandi, and A. H. Kashan, "DisABC: A new artificial bee colony algorithm for binary optimization," *Appl. Soft Comput.*, vol. 12, pp. 342–352, Jan. 2012.

[31] M. S. Kiran, "A binary artificial bee colony algorithm and its performance assessment," *Expert Syst. Appl.*, vol. 175, Aug. 2021, Art. no. 114817.

[32] M. S. Kiran and M. Gündüz, "XOR-based artificial bee colony algorithm for binary optimization," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 21, no. 10, pp. 2307–2328, Sep. 2012.
[33] B. Akay, D. Karaboga, B. Gorkemli, and E. Kaya, "A survey on the artificial bee colony algorithm variants for binary, integer and mixed integer programming problems," *Appl. Soft Comput.*, vol. 106, Jul. 2021, Art. no. 107351.
[34] Y. Xu and X. Wang, "An artificial bee colony algorithm for scheduling call centres with weekend-off fairness," *Appl. Soft Comput.*, vol. 109, Sep. 2021, Art. no. 107542.
[35] Z. Ren, L. Zhang, J. Tang, and T. Liu, "Improved artificial bee colony algorithm based on Cauchy OBL," *J. Phys., Conf. Ser.*, vol. 1920, no. 1, May 2021, Art. no. 012108.
[36] L. Ma, K. Hu, Y. Zhu, and H. Chen, "Cooperative artificial bee colony algorithm for multi-objective RFID network planning," *J. Netw. Comput. Appl.*, vol. 42, pp. 143–162, Jun. 2014.
[37] L. Ma, X. Wang, M. Huang, Z. Lin, L. Tian, and H. Chen, "Two-level master–slave RFID networks planning via hybrid multiobjective artificial bee colony optimizer," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 5, pp. 861–880, May 2019.
[38] H. Li, X. Li, and L. Gao, "A discrete artificial bee colony algorithm for the distributed heterogeneous no-wait flowshop scheduling problem," *Appl. Soft Comput.*, vol. 100, Mar. 2021, Art. no. 106946.
[39] N. Sharma, H. Sharma, and A. Sharma, "Beer froth artificial bee colony algorithm for job-shop scheduling problem," *Appl. Soft Comput.*, vol. 68, pp. 507–524, Jul. 2018.
[40] Y. Zhong, J. Lin, L. Wang, and H. Zhang, "Hybrid discrete artificial bee colony algorithm with threshold acceptance criterion for traveling salesman problem," *Inf. Sci.*, vol. 421, pp. 70–84, Dec. 2017.
[41] Z. Yilmaz Acar and F. Baçiftçi, "Solving multi-objective resource allocation problem using multi-objective binary artificial bee colony algorithm," *Arabian J. Sci. Eng.*, vol. 5, pp. 1–3, Apr. 2021.
[42] M. E. H. Daho and M. A. Chikh, "Combining bootstrapping samples, random subspaces and random forests to build classifier," *J. Med. Imag. Health Informat.*, vol. 5, no. 3, pp. 1–6, Jul. 2015.
[43] J. R. Wei, N. F. Bei, and B. Hu, "Aerosol-photolysis interaction reduces particulate matter during wintertime haze events," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 18, Apr. 2020, Art. no. 201916775.
[44] L. Chang, Z. Wu, and J. Xu, "A comparison of haze pollution variability in China using haze indices based on observations," *Sci. Total Environ.*, vol. 715, May 2020, Art. no. 136929.
[45] M. M. Haque, C. Fang, J. Schnelle-Kreis, G. Abbaszade, X. Liu, M. Bao, W. Zhang, and Y.-L. Zhang, "Regional haze formation enhanced the atmospheric pollution levels in the Yangtze river delta region, China: Implications for anthropogenic sources and secondary aerosol formation," *Sci. Total Environ.*, vol. 728, Aug. 2020, Art. no. 138013.
[46] C. Zhang, X. Wang, S. Chen, L. Zou, and C. Tang, "Coupling detrended fluctuation analysis of the relationship between $PM_{2.5}$ concentration and weather elements," *Phys. A, Stat. Mech. Appl.*, vol. 531, Oct. 2019, Art. no. 121757.
[47] C. Zhang, Z. Ni, L. Ni, and N. Tang, "Feature selection method based on multi-fractal dimension and harmony search algorithm and its application," *Int. J. Syst. Sci.*, vol. 47, no. 14, pp. 3476–3486, Oct. 2016.
[48] C. Zhang, Z. Ni, and L. Ni, "Multifractal detrended cross-correlation analysis between $PM_{2.5}$ and meteorological factors," *Phys. A, Stat. Mech. Appl.*, vol. 438, pp. 114–123, Nov. 2015.

CHEN ZHANG was born in Anhui, China. She received the M.S. degree in computational mathematics from Anhui University, in 2011, and the Ph.D. degree in information management and system from Hefei University of Technology, China, in 2016. She is currently an Associate Professor with the School of Artificial Intelligence and Big Data, Hefei University, China. Her research interests include machine learning and artificial intelligence.

XIAOFENG WANG received the B.Sc. degree in computer and science technology from Anhui University, Hefei, China, in 1999, the M.Sc. degree in pattern recognition and intelligent system from the Institute of Intelligent Machines, Graduate University of Chinese Academy of Sciences, Hefei, in 2005, and the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, in 2009. He is currently working as a Professor with Hefei University. His current research interests include image processing and image segmentation.

SHENGBING CHEN was born in Anhui, China. He received the M.S. degree from the University of Science and Technology of China, in 2005, and the Ph.D. degree from Anhui University, China, in 2010. He is currently a Professor with the School of Artificial Intelligence and Big Data, Hefei University, China. His research interests include big data mining and swarm intelligence.

HONG LI received the M.S. degree in computer science and technology from Hefei University of Technology, in 2003. She is currently a professor. Her research interests include data science, and data mining: methods and applications.

XIAOXUAN WU was born in Anhui, China. She received the M.S. degree in applied mathematics and the Ph.D. degree in management science and engineering from Hefei University of Technology, China, in 2010 and 2015, respectively. She is currently a Lecturer with the School of Artificial Intelligence and Big Data, Hefei University, China. Her research interests include machine learning and artificial intelligence.

XIN ZHANG received the bachelor's degree from Shandong University and the Ph.D. degree from the University of Chinese Academy of Sciences, China. She is currently a Lecturer with the Department of Artificial Intelligence and Big Data, Hefei University. Her research interest includes big data information.