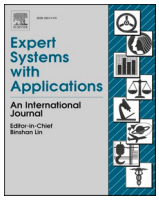
Expert Systems With Applications 237 (2024) 121549

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



An improved random forest based on the classification accuracy and correlation measurement of decision trees

Zhigang Sun [a,b,d,e], Guotao Wang [a,b,d,e,*], Pengfei Li [b,d,e], Hui Wang [c], Min Zhang [a], Xiaowen Liang [a]

[a] School of Electrical and Electronic Engineering, Heilongjiang University, Harbin 150080, China
[b] School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150001, China
[c] School of Hydraulic Science and Engineering, Yangzhou University, Yangzhou 225009, China
[d] Key Laboratory of Electrical and Electronic Reliability Technology in Heilongjiang Province, Harbin 150001, China
[e] MOE Key Laboratory of Reliability and Quality Consistency of Electronic Components, Harbin 150001, China

ARTICLE INFO

Keywords:
Classification accuracy
Correlation measurement
Dot product
Random forest
CART

ABSTRACT

Random forest is one of the most widely used machine learning algorithms. Decision trees used to construct the random forest may have low classification accuracies or high correlations, which affects the comprehensive performance of the random forest. Aiming at these problems, the authors proposed an improved random forest based on the classification accuracy and correlation measurement of decision trees in this paper. Its main idea includes two parts, one is retaining the classification and regression trees (CARTs) with better classification effects, the other is reducing the correlations between the CARTs. Specifically, in the classification effect evaluation part, each CART was applied to make predictions on three reserved data sets, then the average classification accuracies were achieved, respectively. Thus, all the CARTs were sorted in descending order according to their achieved average classification accuracies. In the correlation measurement part, the improved dot product method was proposed to calculate the cosine similarity, i.e., the correlation, between CARTs in the feature space. By using the achieved average classification accuracy as reference, the grid search method was used to find the inner product threshold. On this basis, the CARTs with low average classification accuracy among CART pairs whose inner product values are higher than the inner product threshold were marked as deletable. The achieved average classification accuracies and correlations of CARTs were comprehensively considered, those with high correlation and weak classification effect were deleted, and those with better quality were retained to construct the random forest. Multiple experiments show that, the proposed improved random forest achieved higher average classification accuracy than the five random forests used for comparison, and the lead was stable. The G-means and out-of-bag data (OBD) score obtained by the proposed improved random forest were also higher than the five random forests, and the lead was more obvious. In addition, the test results of three non-parametric tests show that, there were significant diversities between the proposed improved random forest and the other five random forests. This effectively proves the superiority and practicability of the proposed improved random forest.

1. Introduction

In 1995, Ho (1995) first proposed the concept of random decision forest. Specifically, he proposed to construct a classifier based on decision trees. The number of decision trees contained in this classifier can be expanded infinitely, and, these decision trees are combined in a complementary or weighted manner to construct a new classifier, known as the random decision forest mentioned above. It solves the problem that a single decision tree is prone to overfitting. In 1996, Breiman (1996) proposed a bagging algorithm. This algorithm extracts samples from the original sample set through bootstrap sampling, and constructs multiple sample subsets. A strong classifier is constructed by training weak classifiers on each sample subset and seeking ways to combine multiple weak classifiers. In 1998, Ho (1998) further proposed a random subspace algorithm. This algorithm extracts samples from the original feature space through bootstrap sampling, and constructs

* Corresponding author at: School of Electrical and Electronic Engineering, Heilongjiang University, Harbin 150080, China.
E-mail addresses: 22b906048@stu.hit.edu.cn (Z. Sun), wanggt@hlju.edu.cn (G. Wang), 20b906020@stu.hit.edu.cn (P. Li), 2211813@hlju.edu.cn (M. Zhang).

https://doi.org/10.1016/j.eswa.2023.121549
Received 26 August 2022; Received in revised form 24 August 2023; Accepted 9 September 2023
Available online 12 September 2023

multiple feature subspaces. By using multiple feature subspaces to train multiple classifiers, respectively, the set of these classifiers has better generalization ability. In 2001, Breiman (2001) integrated the bagging algorithm, the random subspace algorithm and the classification and regression tree (CART), thus proposed the random forest. This unoptimized random forest is also called traditional random forest. Since then, the traditional random forest has been widely used in various fields, and has achieved good results in solving conventional classification or regression problems. However, when faced with the individualized problems in different fields, such as the small size of the original data set or feature space, many and miscellaneous classification categories, the performance of the trained traditional random forest is greatly reduced, so it is difficult to achieve satisfactory results. Therefore, according to the actual needs of the field, experts have studied random forest from different perspectives and put forward many effective improvement schemes. These improved random forest effectively solved the problem of insufficient capacity of traditional random forest in different individualized problems, and also provided a reference for this study.

As mentioned above, the random forest is an ensemble learning method that combines multiple decision trees into a whole through a specific combination. Therefore, improvements to the random forest have focused on the specific combination methods and decision trees. In the traditional random forest, we often use the majority voting method to combine multiple decision trees, i.e., the so-called voting mechanism, which has been improved by many scholars. For example, Gall and Lempitsky (2009) proposed a Hough random forest. The algorithm used the Hough voting mechanism to replace the simple majority voting mechanism. Tripoliti et al. (2013) proposed modifications to the voting mechanism based on feature selection, clustering, nearest neighbors and optimization techniques. Li et al. (2018) proposed a new voting mechanism based on Potential Nearest Neighbor to replace the traditional majority voting mechanism to avoid the information loss caused by out-of-bag samples. Kim et al. (2020) proposed to use a wave voting scheme instead of the majority voting scheme to improve the random forest, further improved the achieved classification accuracy. Sun et al. (2021) calculated the similarity of decision trees based on the tree edit distance, and further performed cluster reduction based on the maximum and minimum distance algorithm, and then introduced the classification accuracy of decision trees to construct the weight matrix to achieve weighted voting in the voting stage. It is worth noting that, in recent years, there has been less and less research focusing solely on improving voting mechanisms, and more often it has been used as part of improvement research on decision trees. This is because, in terms of the construction process of random forest, the voting mechanism is at the end of the whole construction process, and its improvement space is limited. More and more attentions are paid to the front end, the construction of the decision tree.

As mentioned above, the base classifier of traditional random forest is CART in standard configuration, which performs generally in solving individualized problems, and many scholars have studied it from multiple perspectives. From the perspective of node splitting criteria of decision trees, Ma et al. (2019) improved the Grassbcrgcr entropy, and used the improved Grassberger entropy to calculate the information gain. The random forest was trained by selecting the optimal split parameters of the split node. From the perspective of the construction process of decision trees, Ishwaran et al. (2008) proposed the survival tree idea, and applied it to the generation process of the random forest, then proposed an improved random survival forest (RSF). The advantage of this algorithm is that, for any data set, a corresponding decision tree was created. In this way, when performing voting, the classification result of the random forest can be represented by the average classification results of multiple decision trees. Wang and Wang (2020) proposed a Post-Selection Boosted Random Forest (PBRF). The algorithm combined the traditional random forest with the Lasso method. It can dynamically obtain the decision trees according to different input samples, and output the prediction results without giving the number of decision trees for the final prediction in advance. On this basis, Farhadi et al. (2023) proposed a Reducing and Aggregating Random Forest Trees by Elastic Net (RARTEN). The algorithm improved the random forest by selecting the most appropriate penalized regression methods, and it is tried to improve the PBRF using Elastic Net regression. From the perspective of the feature subspace of decision trees, Amaratunga et al. (2008) introduced a new feature weight method to replace the simple random sampling method for the sampling of the feature subspace. Kulkarni and Sinha (2013) proposed a new method called disjoint partitioning. This method used disjoint partitions of the training data sets to train decision trees, thus helped create diversity among decision trees. Moreover, it also used different attribute subsets on each node of the decision tree to increase diversity. Liu and Zhao (2017) proposed a variable importance-weighted Random Forest, which instead of sampling features with equal probability at each node to construct decision trees, sampled features according to their variable importance scores, and then selected the best split from the randomly selected features. Ghosh and Cabrera (2021) developed a new method to improve the traditional random forest by reducing the contribution of decision trees whose nodes are populated with less informative features. The proposed method selected qualified feature subsets at each node by weighted random sampling, instead of simple random sampling in traditional random forests. Aiming at the problem that the random forest adopted the simple random sampling feature selection method when generating feature subspaces that cannot distinguish redundant features, resulting in low classification accuracy and large calculation loss, Wang et al. (2022) used Spark to conduct optimization research. This improved random forest extracted features according to the calculated feature importance, and formed the feature subspace. When generating the random forest, it selected the decision trees according to the similarity and classification accuracy of different decisions.

In summary, currently, there is few research focused on improving the voting mechanism and there is little room for improvement. More attention has been paid to improving the base classifier decision tree, including its own performance improvement, combination or decision method improvement, etc. The research conducted from the perspective of node splitting criteria or construction process of decision trees is more focused on constructing decision trees with excellent classification effect or broad representativeness. In this way, the random forest constructed by these decision trees can have a strong generalization ability. The research conducted from the perspective of the feature subspace of decision trees is more focused on improving the diversity between decision trees. In this way, the independence between these decision trees is greatly enhanced, therefore, the reference value of the decision results they provide is enhanced. This effectively solves the problem of construction redundancy and decision redundancy. The random forest constructed from these decision trees can have a strong classification effect and low calculation loss. To sum up, the existing research focuses on the improvement of decision trees, and is committed to improving the classification effect of decision trees and the diversity between decision trees. Unfortunately, the existing research ignores the research on the evaluation mechanism or method of the classification effect of decision trees, and generally uses their corresponding out-of-bag data (OBD) to test them, respectively. This can lead to unstable evaluation results and inconsistent standards. Meanwhile, the existing research focuses on the diversities between decision trees, but few studies quantify the diversity. They used different methods to construct diverse feature subspaces, thereby training decision trees with significant diversities. But they did not calculate and display the specific values of the diversities between these decision trees, and the improvement effect was not direct and persuasive. In addition, most existing researches only focus on improving the classification effects of decision trees, or only on improving the diversities between decision trees, and rarely combine the two for comprehensive consideration.

Therefore, in this study, the authors proposed an improved random forest based on the classification accuracy and correlation measurement

of decision trees, which considered the improvement of classification effects of decision trees and the diversities between decision trees at the same time, and committed to constructing a random forest with good generalization ability and classification effect. First, the authors improved the evaluation mechanism for the classification effect of decision trees. By using three reserved data sets to evaluate the classification effect of the decision tree, the calculation method of the classification accuracy achieved by a single decision tree is improved, and the classification effect of each decision tree in the random forest is objectively evaluated, providing a reference for subsequent screening of high-quality base classifiers. Second, the authors proposed a method to quantify the diversity between decision trees. By using the improved dot product method to calculate the cosine similarity between decision trees in the feature space, called the correlation between decision trees, thus quantifying the diversity between decision trees. By searching for the correlation threshold, reference can be provided for the subsequent screening of two decision trees with high correlation. Finally, the authors combined the classification accuracy with the correlation measurement to screen out those decision trees that have weak classification effect and high correlation with other high-quality decision trees from multiple decision trees to construct the random forest. In this way, the retained decision trees have superior classification effect and low correlation with each other, i.e., high diversity. This effectively ensures the reliability of a single decision tree and the superiority of the comprehensive performance of the random forest.

The main contributions of this paper are as follows.

(1) Concentrate the research focus on the improvement of decision trees, and considering the improvement of the classification effect of the decision tree and the diversity between decision trees at the same time, thus propose an improved random forest based on the classification accuracy and correlation measurement of decision trees.
(2) Improve the evaluation mechanism for the classification effect of decision trees. Use three reserved data sets to evaluate the classification effect of decision trees, which effectively avoids the instability and inconsistent standard problem caused by the performance evaluation based on single data set with different components.
(3) Propose a method for quantifying the diversity between decision trees. Propose an improved dot product method to calculate the cosine similarity between decision trees in the feature space, and provide specific calculation values to visually display the diversity between decision trees, providing a reliable reference for decision tree screening.
(4) Multiple test results verify the superiority of the proposed improved random forest. It is compared with five other improved random forests in twenty representative data sets in many aspects, and its excellent generalization ability and classification effect has been highlighted.

The rest of this paper is organized as follows. In Section 2, we briefly introduced the construction principle, advantages and disadvantages of the traditional random forest. In Section 3, we described in detail the working principle of the improved random forest based on the classification accuracy and correlation measurement of decision trees. In Section 4, we comprehensively verified the comprehensive performance of the improved random forest, and gave the specific comparison and analysis results. In Section 5, we gave a discussion of the improved random forest, including the experimental result analysis and future research content. In Section 6, we summarized the research content of the full text.

2. Related work

The Bagging algorithm is based on the bootstrap sampling (Ditzler et al., 2018). Assuming that the original data set contains $\boldsymbol{M}$ samples, each time, we randomly extract a sample from the original data set and put it back into the original data set. By repeating the above process $\boldsymbol{m}$ times, we obtain $\boldsymbol{m}$ samples, and use them to create a new data set. There may be cases where some samples were never drawn. Relevant research and calculation results show that, about 36.8% of the samples in the original data set will not be extracted (Martinez-Munoz & Suarez, 2010).

Decision tree is a process of setting rules to classify data. Different types of decision trees are applicable to processing different types of data. The ID3, C4.5, and CART are currently commonly used algorithms for generating decision trees (Suknovic et al., 2012). The ID3 algorithm uses information gain as the selection index of split attributions, and it selects the attribution with the largest information gain after splitting as the next division standard (Sheng & Sun, 2019; Mienye et al., 2019; Wang et al., 2015). The ID3 algorithm is easy to understand, but cannot be used to deal with continuous values (Ding et al., 2015). Then the scholars further proposed the C4.5 algorithm. The C4.5 algorithm uses the information gain rate as the selection index of split attributions, which can be used in the processing of discrete and continuous attributions at the same time, and is insensitive to the absence of attributions (Han et al., 2019; Putri & Waspada, 2018). The CART uses the Gini coefficient as the selection index of split attributions, which is mainly used in the processing of the binary classification problem (Smayra et al., 2019).

Ensemble learning is a process of combining multiple single classifiers to make judgments on things, resulting in better classification effect compared to a single classifier (Schapire, 1990). Therefore, we are committed to designing base classifiers with strong generalization ability and large diversity, which are the key to construct well-performing ensemble classifiers, i.e., strong classifiers, and are also our original intention in constructing ensemble classifiers. Strong generalization ability refers to the ability of the base classifier to handle different types of data and achieve satisfactory results, which depends on its classification effect. Large diversity refers to the low correlation between the base classifiers, which results in greater independence of the classification results and does not cause decision redundancy. As mentioned above, random forest is an algorithm based on the ensemble learning idea. It is designed on the basis of random decision forest, combining the Bagging algorithm and CART. It uses the Bagging algorithm to create multiple sample sets in the sampling stage, and trains multiple CARTs, and then constructs a random forest from their effective combination.

For sample T, formula for calculating the Gini coefficient is as follow (Jiang et al., 2020).

$$Gini(T) = 1 - \sum_{i=1}^{c} p_i^2 \tag{1}$$

Where p_i represents the proportion of samples of category i to the total samples, c represents the number of categories included in the sample.

Assuming that after feature A was split, sample T is divided into K parts. For sample T_j, formula for calculating the Gini coefficient is as follow.

$$Gini(T,A) = \sum_{j=1}^{k} \frac{|T_j|}{|T|} Gini(T_j) \tag{2}$$

Then the mathematical expression of random forest is as follow (Svetnik et al., 2003).

$$\{h(x,\beta_i), i = 1,2,3,\ldots\} \tag{3}$$

Where $h(x,\beta_i)$ is the base classifier that constructs the random forest, i.e., the above-mentioned CART. It should be noted that, the base classifier here is an unpruned CART. x is the origin data set, which is a multi-dimensional vector set. β_i is a vector set randomly selected from x using the Bagging algorithm, both of which are independent identical

distribution. In fact, β_i determines the classification effect of the corresponding CART.

When the random forest is used to solve the classification problem, multiple internal CARTs make predictions on the same testing set in parallel, and give multiple classification results. On this basis, the classification result of each CART is counted, and the majority voting mechanism is used to determine the category that received the highest number of votes, which is the final classification result. The construction process is as follows.

Step 1: Use the Bagging algorithm to extract n samples from the origin data set N to create a training set. Usually, n is much smaller than N. Repeat the above process k times to form k training sets.

Step 2: On the basis of obtaining k training sets, construct corresponding k CARTs. Specifically, each node in each CART is constructed by selecting m features from the origin feature set M. Usually, m is much smaller than M. In this way, each training set has a corresponding feature subset. It should be noted that, when constructing each CART, we select the feature with the smallest Gini coefficient to split the nodes. Other nodes are constructed with the same splitting rule until all the training data of the node belong to the same class or have reached the maximum depth of the tree.

Step 3: After the above two steps, we obtain multiple training sets and feature subsets, and train multiple CARTs, and thus construct the random forest. It can be used to make predictions on the predicted data. It should be noted that, the multiple CARTs used to construct the random forest generally do not perform pruning operations.

Step 4: When making decisions on the predicted data, multiple CARTs in the random forest make decisions at the same time, and then use the majority voting mechanism to process these decision results to produce the final decision result. Fig. 1 shows the construction process of the random forest.

From the above construction process, we can see that, the random forest inevitably has the following shortcomings. First, from the perspective of construction principle and decision process, the more CARTs, the more reliable the decision results of the random forest. However, when the number of CARTs is large, the training space and training time of the random forest will increase, resulting in a large calculation loss. Second, due to the fact that the training set used to train CARTs is constructed from the original data set using the Bagging algorithm, it cannot be guaranteed that all the training data in the training set have good characteristics, which means that the quality of the training set is not uniform. This will inevitably result in the ordinary classification effect of CARTs trained on some training sets with ordinary quality, which will lead to the decline of the classification effect of the constructed random forest. Finally, since the Bagging algorithm adopts the bootstrap sampling, it is inevitable that it will result in too many overlapping training data between certain training sets. In this way, the correlation between the CARTs trained on them is greater, i.e., the diversity is smaller. Therefore, the random forest constructed by them will have a serious problem of decision redundancy, resulting in high computing loss and low generalization ability. And this clearly goes against our original intention of constructing an ensemble classifier. The above shortcomings are the focus of this research, and they are also emphasized and improved in this paper. On this basis, an improved random forest based on the classification accuracy and correlation measurement of decision trees is proposed.

3. Method

As mentioned above, in the construction process of the random forest, both the training sets and feature subsets were randomly selected. Such randomness may lead to poor classification effect of some CARTs in the random forest, and they contributed less to the classification effect of the random forest. Therefore, in the improved random forest, for each generated CART, we used three reserved data sets to evaluate its classification effect. Further, we took the average classification accuracy achieved by the CART as reference, and sort all the CARTs in descending order to highlight CARTs with excellent classification effect as the priority selection object.

Similarly, the above-mentioned randomness can lead to excessive overlapping of training data between certain training sets, and there may be situations where some CARTs are more similar, i.e., the correlation between them is high. Deleting CARTs with high correlation can effectively improve the generalization ability of the random forest and reduce the calculation loss. However, the correlation between CARTs is not as low as possible. On the one hand, if the correlation between

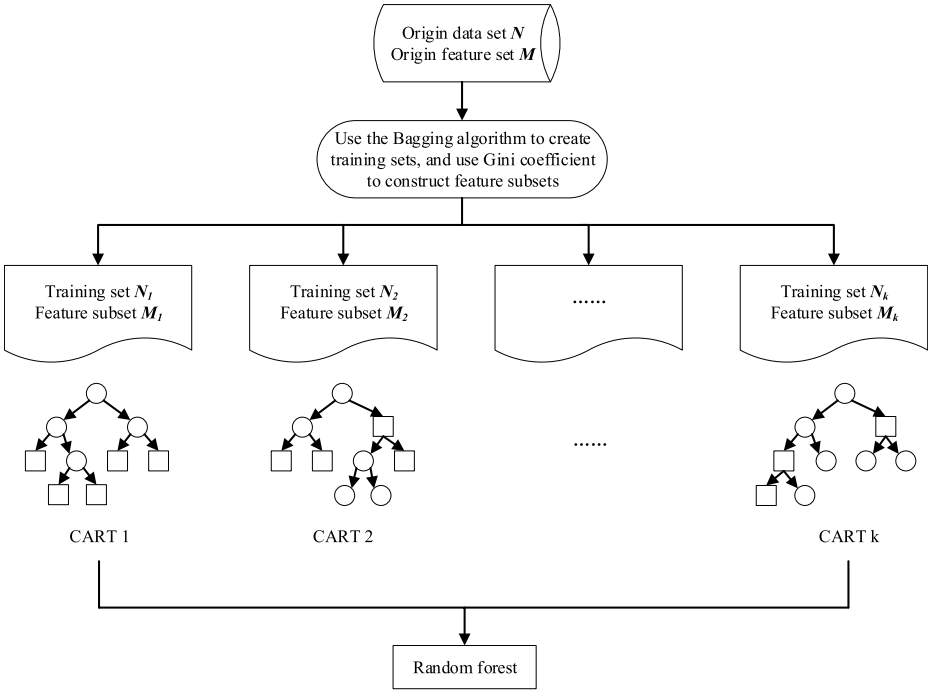


Fig. 1. Construction process of the random forest.

CARTs is required to be too low, it may lead to insufficient number of CARTs to construct the random forest, and reduce its classification effect. On the other hand, CARTs with low correlation means that there is less overlapping training data between the training sets created by the Bagging algorithm, which may result in the created data set not covering all the samples in the original data set. In this way, these CARTs are not globally representative, which will reduce the generalization ability of the constructed random forest. Aiming at these problems, in the improved random forest, we used the improved dot product method to measure the correlations of CARTs. We took the average classification accuracy achieved by the random forest as reference to find the correlation threshold, i.e., the inner product threshold. If the inner product value between a CART pair is greater than the inner product threshold, they are judged to be highly correlated with each other, and we deleted the one with a lower average classification accuracy among the CART pair.

In addition, in the proposed improved random forest, for CARTs used to construct the random forest, first, we train a batch of CARTs whose number is more than the preset number. Then, based on the above construct process, we comprehensively consider the classification accuracy and correlation of these CARTs, and delete those CARTs with higher correlation and lower classification accuracy until the remaining number of CARTs reaches the preset number. In this way, the classification effect of the retained CART is relatively high and the diversity between them is relatively large, which meets the requirements of designing ensemble classifiers, and thus can achieve the purpose of improving the comprehensive performance of the random forest.

3.1. Basic research work

3.1.1. Achieved average classification accuracy

Since the classification effect of the CART is sometimes not ideal, many scholars proposed different methods to improve it. The most common method is to use OBD for weighting. However, since the training set used to train the CART was created by using the Bagging algorithm, the multiple created data sets were different, thus their corresponding OBDs were also different. In this case, it is not appropriate to use different training sets to train CARTs and measure their classification effects. To avoid this problem, in the improved random forest, we modified the evaluation method for the classification effect of the CART. Specifically, we used three reserved data sets to evaluate the classification effect of each CART, respectively, and calculated the average classification accuracy achieved on three data sets, and then ranked multiple CARTs in descending order based on this. It can be seen that using three data sets instead of one data set or one weighted OBD to evaluate the CART can obtain more comprehensive evaluation results. Based on this, the comprehensive evaluation results are more stable and reliable, effectively avoiding the shortcomings of the existing evaluation method. The specific evaluation processes of the classification effect of the CART are as follows.

Step 1: Select three data subsets with equal numbers from the original data set as the testing sets to evaluate the classification effect of each CART, i.e., the above-mentioned three reserved data sets. It should be noted that, the three data subsets used to evaluate each CART are different, and the data in these data subsets are all labeled.

Step 2: Determine the number N of CARTs to be constructed and the number of features in each feature subset. Use the Bagging algorithm to extract samples from the corresponding remaining origin data set, and create their respective training sets, and thus train multiple CARTs on this. It should be noted that, because the three data subsets used to evaluate each CART in Step 1 are different, the corresponding remaining original data sets are also different.

Step 3: Apply each CART to make predictions on the corresponding three data subsets, and express the achieved classification accuracy as a_i^j. Among them, $i = 1, 2, \ldots, N$, represents the *ith* CART, and $j = 1, 2, 3$, represents the *jth* group of data subsets.

Step 4: Calculate the average classification accuracy of the *ith* CART, the formula is: $\overline{a_i} = \frac{a_i^1 + a_i^2 + a_i^3}{3}$.

Step 5: Arrange all the CARTs in descending order according to their achieved average classification accuracies.

3.1.2. Correlation measurement

The correlation measurement method adopts one or a set of strategies to compare the similarity between two quantities (Deng et al., 2016). At present, there are two main correlation measurement methods, one is the vector space method, and the other is the semantic dictionary method (Liu et al., 2010). Studies have shown that, vector space models have great advantages in dealing with individualized classification. This is because computers have a complete and sophisticated set of vector processing algorithms. After the data is converted into a vector, the computer can efficiently complete the vector processing, i.e., the required data processing. Therefore, the vector angle can be used to calculate the correlation between two quantities (Chen et al., 2016). In general, the smaller the vector angle between two quantities, the higher the correlation between them. Conversely, the larger the vector angle between the two quantities, the lower the correlation between them. In the vector space model, the correlation calculation methods mainly include dot product method, cosine method, Manhattan distance method, Euclidean distance method, etc. (Li et al., 2020; Jiang et al., 2017; Khoshkenar & Mahlooji, 2013; Chetlur et al., 2020; Merigo & Casanovas, 2011). Among them, Euclidean distance method and Manhattan distance method are used to measure the spatial distance between two quantities, and the calculation result is a specific distance value, which is not the method required for this study. Moreover, these two methods are influenced by the data scale and usually achieve better results after performing normalization processing. In machine learning, the data scale in the data set is non-standard. The cosine method is used to measure the cosine similarity between two quantities, and the calculation result is 1 or -1. The cosine similarity of two vectors with identical directions is 1, while the similarity of two vectors with opposite directions is -1. Although the cosine method indirectly measures the angle between two quantities, it does not provide a specific angle value and only provides a judgment of directional consistency. The dot product method can be used to directly measure the angle between two quantities. It infers the angle between two vectors by calculating their specific values. Therefore, in this study, we decided to use the dot product method to evaluate the correlation between CARTs. In the linear algebra field, the calculation formula of the conventional dot product method is as follows:

$$\theta = acos\left(\frac{a \bullet b}{||a|| ||b||}\right) \tag{4}$$

Where a and b represent two vectors, respectively. $||a||$ and $||b||$ represent the lengths of the two vectors, respectively, which are the sum of the squares of their respective components in a and b.

Referring to formula (4), we provide the calculation formula for the dot product method used to measure the correlation between two CARTs in this study, as follows:

$$Sim\left(D_i, D_j\right) = acos\left(W_i \bullet W_j\right) \tag{5}$$

Where D_i and D_j represents two CARTs, respectively, W_i represents the feature subset corresponding to CART i, W_j represents the feature subset corresponding to CART j. The detailed calculation formula for $W_i \bullet W_j$ is as follows:

$$W_i \bullet W_j = \frac{\sum_{m=1}^{p}\sum_{n=1}^{q} I(W_{im} = W_{jn})(W_{im} \bullet W_{jn})}{\sum_{m=1}^{p}\sum_{n=1}^{q}(W_{im} \bullet W_{jn})} \bullet 180^{\circ} \tag{6}$$

Where W_{im} represents the feature m in feature subset W_i, W_{jn} represents the feature n in feature subset W_j. I is the indicator function, only when $W_{im} = W_{jn}$, $I\left(W_{im} = W_{jn}\right) = 1$, otherwise $I\left(W_{im} = W_{jn}\right) = 0$.

From formula (6), it can be seen that, the denominator part calculates the sum of values between all features in feature subset W_i and feature subset W_j, while the numerator part calculates the sum of values between the same features in feature subset W_i and feature subset W_j. From this, it can be seen that, the key to measuring the correlation between two CARTs lies in the molecular part of formula (6). It is worth noting that, in this study, we used the numerical values between the feature subsets corresponding to CARTs, which are actually features, to measure the correlation between the two. In the study conducted by Cervantes et al. (2017), they detailed various relationships between decision trees, not just the correlations used in this study.

In this way, for the multiple constructed CARTs, we can use the dot product method to calculate the correlation between them. If we set an inner product threshold, we can compare the inner product value of each CART pair with the inner product threshold. From this, we can obtain CART pairs with high correlation and label them as pending CART pairs. For CART pairs with lower correlation, those marked as safe CART pairs will retain the two CARTs they contain.

3.2. Improved random forest

On the basis of Section 3.1, in the improved random forest proposed in this paper, we combined the classification accuracy with the correlation measurement of decision trees to comprehensively evaluate the CARTs used to construct the random forest, with a view to improving its comprehensive performance. As the first step to construct the random forest, it is necessary to determine the number of its internal base classifiers, i.e., CARTs. At present, there are many methods to determine the optimal number of base classifiers, we can also set them based on empirical values. Considering that the purpose of this study is to select CARTs with better classification effect and low correlation from multiple CARTs, therefore, if the optimal number of base classifiers is considered as the preset number for constructing CARTs, we should set a value slightly larger than the preset number in advance. In this way, the number of CARTs corresponding to the difference between a slightly larger value and the preset number is the maximum range for us to screen out CARTs. Thus, after determining the optimal number of base classifiers, when constructing the random forest, we constructed a certain percentage of CARTs based on the preset number. The numerical setting of this percentage can be adjusted according to different application scenarios. In previous studies, after a series of experimental analysis, we suggested that the value of this percentage can be set to 0.1 or 0.15. On this basis, according to the basic research work described in Section 3.1, we comprehensively considered the classification accuracy and correlation measurement of decision trees, and deleted those CARTs with high correlation and low classification accuracy until the number of remaining CARTs met the preset number.

The specific implementation steps of the improved random forest are as follows.

Step 1: Select three data subsets with equal numbers from the original data set as testing sets to evaluate the classification effect of each CART. Other descriptions are consistent with Step 1 in Section 3.1.1.

Step 2: Determine the number N of CARTs to be constructed and the number of features in each feature subset. Use the Bagging algorithm to perform$N + m^*N$times sampling from the corresponding remaining original data set to create$N + m^*N$training sets. On this basis, train$N + m^*N$CARTs. Other descriptions are consistent with Step 2 in Section 3.1.1.

Step 3: Apply each CART to make predictions on the corresponding three data subsets, and express the achieved classification accuracy as a_i^j. Among them, $i = 1, 2, \ldots, N$, represents the *ith* CART, and $j = 1, 2, 3$, represents the *jth* group of data subsets. Other descriptions are consistent with Step 3 in Section 3.1.1.

Step 4: Calculate the average classification accuracy of the *ith* CART, the formula is: $\overline{a_i} = \frac{a_i^1 + a_i^2 + a_i^3}{3}$.

Step 5: Arrange all the CARTs in descending order according to their achieved average classification accuracies.

Step 6: Use the improved dot product method to calculate the inner product values between the CARTs and save them, and use the grid search method to find the optimal inner product threshold t. Then, do not process those CART pairs whose inner product value is less than the inner product threshold. For those CART pairs whose inner product value is greater than the inner product threshold, mark the one with low average classification accuracy in the pair as deletable.

Step 7: Arrange all the CARTs marked as deletable in ascending order based on the achieved average classification accuracy, and delete them in sequence until the number of remaining CARTs is N. That is, the number of remaining CARTs reaches the preset number. It should be noted that, if the number of the remaining unmarked CARTs is greater than N after all the CARTs marked as deletable are deleted, we arrange the remaining CARTs in ascending order based on the achieved average classification accuracy, and continue to delete them in sequence until the number of remaining CARTs is N.

Step 8: Use the remaining N CARTs to construct the random forest. After processing by the majority voting mechanism, the majority preference of the CART voting results is the decision result of the random forest. The construction process of improved random forest proposed in this paper is shown in Fig. 2.

4. Experiment and analysis

4.1. Data set

As mentioned above, the base classifier of the random forests is CARTs. CARTs are quality classifiers that have been widely validated to solve binary classification problems. In machine learning, a multi-classification problem can be transformed into a combination of multiple binary classification problems.Since the random forest is a strong classifier consisting of multiple CARTs, therefore, the random forest can be used to efficiently solve not only binary classification problems, but also multi-classification problems. Therefore, in this study, the authors used representative data sets containing two categories and multiple categories to verify the proposed improved random forest. Among them, data sets containing two categories come from two sources, private data sets and public data sets.

One of the authors' affiliations is the Institute of Reliability in Electrical Apparatus and Electronics at Harbin Institute of Technology, and the research interest is loose particle detection research for sealed electronic equipment. During the manufacturing process of the sealed electronic equipment, some particles may be encapsulated inside the equipment due to process constraints or operational errors. These particles are called loose particles (Sun et al., 2021). When the sealed electronic equipment containing loose particles is in a motion state by external excitation, the signal generated by the collision or sliding of loose particles is called loose particle signal, and the signal generated by the vibration of internal components is called component signal (Liang et al., 2020). These two signals are similar in many aspects, and it is difficult to distinguish them. One of the authors' studies is the identification of loose particle siganls and component signals. In previous studies, the authors extracted sound features from both signals to create the data set, and then trained the classifier to try to identify the loose particle signals and the component signals. It can be seen that, the data set created from the loose particle signals and component signals also contains two categories. Meanwhile, in previous studies, the classifier trained by the authors was based on the traditional random forest. Therefore, we can use the data set created from the loose particle signals and component signals to verify the proposed improved random forest. If its classification effect is verified to be advantageous, it can be used to solve the loose particle signal and the component signal identification

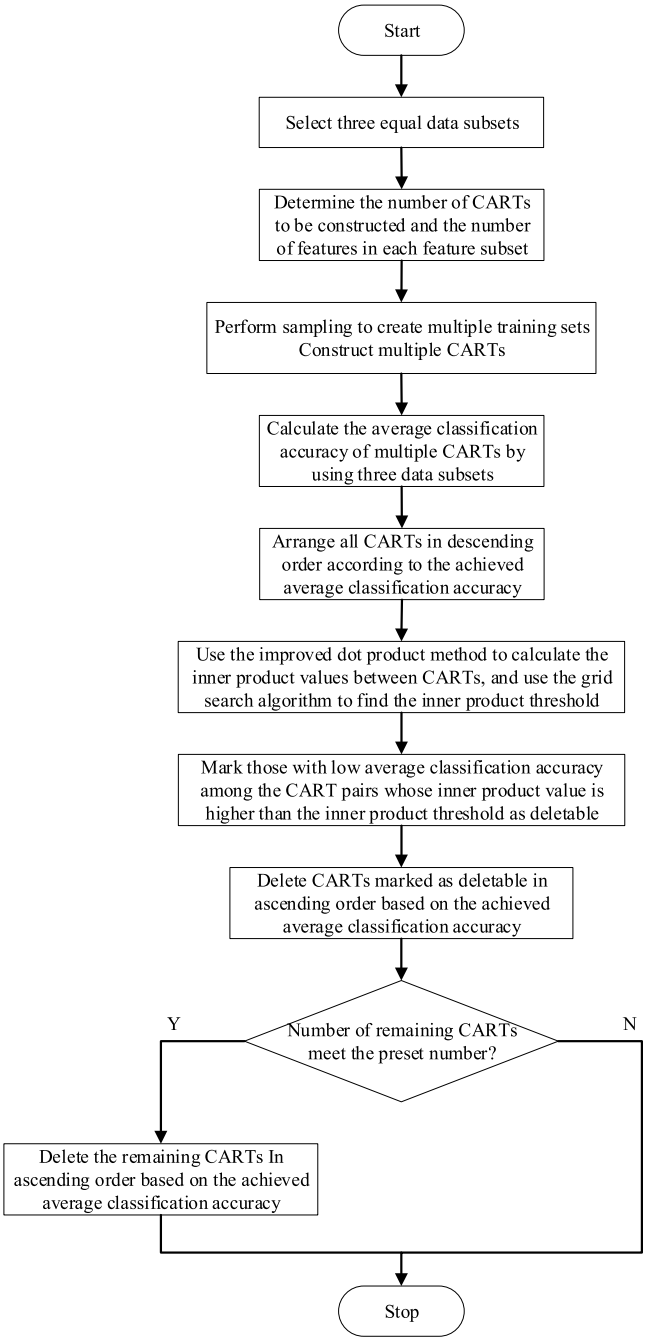


Fig. 2. Construction process of the improved random forest.

problem, which has important practical value. It should be noted that, in previous studies, a total of six data sets were created by the authors because of the differences in the extracted sound features. In this study, they are referred to as private data set 1, private data set 2, private data set 3, private data set 4, private data set 5 and private data set 6. The public data sets were selected from Haberman's Survival Data Set (abbr. Haberman), Diabetes Data Set (abbr. Diabetes), Blood Data Set (abbr. Blood), and Abalone Data Set (abbr. Abalone). Their specific descriptions are shown in Table 1.

Based on the effective identification of the loose particle signals and component signals, we can further analyse the properties of the loose particle from the loose particle signals, including its material and location. For example, there are differences in the acoustic characteristics of the loose particle signals generated by loose particles of different materials, and there are also differences in the acoustic characteristics of

Table 1
Ten data sets containing two categories.

Data set	Number of data	Number of features	Number of categories
Private data set 1	14,375	14	2
Private data set 2	11,563	14	2
Private data set 3	16,684	14	2
Private data set 4	16,392	14	2
Private data set 5	13,813	14	2
Private data set 6	26,157	14	2
Haberman	306	3	2
Diabetes	768	8	2
Blood	748	4	2
Abalone	731	7	2

the loose particle signals generated by loose particles at different locations. These differences in acoustic characteristics can also be reflected by extracting sound features, thus we can create the material data set and the localization data set, respectively. It can be seen that, the data sets created from the loose particle signals generated by different materials or at different locations contain multiple categories. Similarity, on the one hand, they can be used to verify the proposed improved random forest. On the other hand, the improved random forest is of great relevance to solve the loose particle material identification problem and the loose particle localization problem. It should be noted that, in previous studies, a total of three localization data sets were created by the authors because of the differences in the extracted sound features. In this study, they are referred to as localization data set 1, localization data set 2 and localization data set 3. Meanwhile, the authors created a material data set. The public data sets were selected from Wine Quality Data Set (abbr. Wine), Diabetes Health Indicators Data Set (abbr. Diabetes_HI), Microsoft Malware Sample Data Set (abbr. Microsoft_MS), UrbanSound8K Data Set (abbr. UrbanSound8K), Wine Quality N Data Set (abbr. Wine_QN), and Machine Predictive Maintenance Classification Data Set (abbr. Machine_PMC). Their detailed descriptions are shown in Table 2.

To verify the comprehensive performance of the proposed improved random forest, in addition to the traditional random forest (abbr. Tra_RF), the authors also chose the parameter-optimized random forest (abbr. PO_RF), and the improved random forests proposed by Wang and Wang (2020), Farhadi et al. (2023) and Wang et al. (2022) in the Introduction, called PBRF, RARTEN and Spark_RF, respectively. It should be noted that, both Wang and Wang (2020) and Farhadi et al. (2023) focused on improving the classification effect of decision trees, while Wang et al. (2022) focused on improving the diversity between decision trees. It can be found that, their research focus overlaps with that of this study, providing a basis for comparison. In addition, these three improved random forests have achieved satisfactory results on some of the public data sets used in this paper, and have a relatively complete comparative basis. Moreover, they are the latest improved

Table 2
Ten data sets containing multiple categories.

Data set	Number of data	Number of features	Number of categories
Localization data set 1	102,267	50	8
Localization data set 2	81,768	11	8
Localization data set 3	76,784	11	16
Material data set	1,039,776	14	6
Wine	1143	11	6
Diabetes_HI	253,680	21	3
Microsoft_MS	1642	257	9
UrbanSound8K	8732	6	10
Wine_QN	6497	12	7
Machine_PMC	10,000	6	3

random forests proposed in recent two years, which are widely representative and progressiveness, and are suitable for comparison with the improved random forest proposed in this paper.

It should be noted that, the PO_RF refers to the optimization of the internal parameters of *RandomForestClassifier()* on the Sklearn platform using the grid search method (Li et al., 2012), and thus obtain the optimal parameter combination. In this study, the authors focused on optimizing the *n_estimators*, *max_depth* and *max_features* of *RandomForestClassifier()*. Where *n_estimators* denotes the number of decision trees, *max_depth* denotes the maximum depth of the decision tree, and *max_features* denotes the maximum number of features of the decision tree. Taking the localization data set 1 as an example, the authors combined with the grid search method to obtain the optimal parameter combination of the PO_RF, as shown in Table 3. The optimal parameter combinations for other data sets can be derived according to the same steps. Considering the research focus of this paper and the length of this paper, the optimal parameter combinations of random forests applicable to each data set are not given here.

It is worth noting that, in order to ensure the stability and reliability of the experimental results, we performed 10-fold cross-validation on each data set, and took the average of the ten results as the final experimental result. Moreover, for the six private data sets, we regarded the data representing the loose particle signal as the positive class data and the data representing the component signal as the negative class data.

4.2. Inner product threshold

As mentioned above, in the improved random forest proposed in this paper, we used the improved dot product method to measure the correlations between CARTs. For a certain data set, we took the average classification accuracy achieved by the random forest as reference, and combined the grid search method to find the inner product threshold. Taking private data set 1 as an example, we first trained multiple CARTs based on traditional random forest, and used the improved dot product method to calculate the inner product values between CARTs. According to experience, we set the inner product threshold from 5 to 29. Thus, when the inner product threshold is set as the value within the value range in turn, we deleted the CARTs with low average classification accuracy among CART pairs whose inner product values are higher than the inner product threshold, and used the remaining CARTs to construct a new random forest, and got its average classification accuracy. In this way, by using the grid search method, we set the search range from 5 to 29, searched and compared twenty-five new random forests, and got twenty-five average classification accuracies. At this time, which new random forest achieved the highest average classification accuracy means that its corresponding inner product threshold is optimal, and can be set as the final inner product threshold applicable to this data set. Fig. 3 shows the relationship between the inner product threshold and the average classification accuracy achieved by the new random forest.

It can be seen from the figure that, when the inner product threshold is 21, the average classification accuracy achieved by the new random forest is relatively high. At this time, when the inner product threshold increases, the average classification accuracy achieved by the new random forest decrease slightly or remained unchanged. Therefore, the inner product threshold applicable to private data set 1 can be set to 21. Thus, deleting those CARTs with low average classification accuracy among CART pairs whose inner product values are higher than 21 can effectively improve the classification effect of the constructed new random forest on this data set.

Table 3
Optimal parameter combination of the PO_RF applicable to localization data set 1.

Parameter	Optimal value
n_estimators	165
max_depth	21
max_features	16

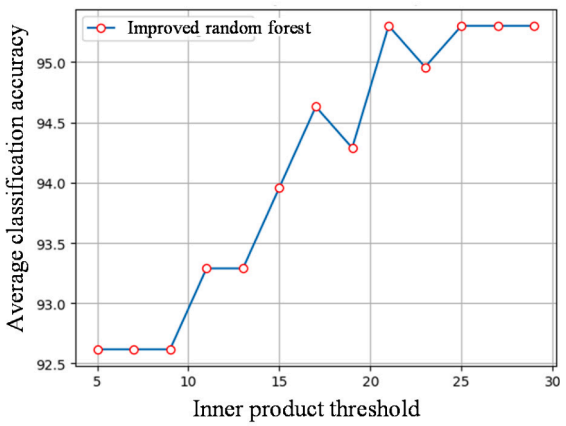


Fig. 3. Relationship between the inner product threshold and the average classification accuracy achieved by the new random forest.

The above case is about private data set 1, and the determination of the inner product thresholds for the other data sets follows the same principle. When scholars who read this paper apply the improved random forest proposed in this paper to other data sets, they also need to go through the above steps.

4.3. Evaluation index

In order to evaluate the performance of the improved random forest, it is necessary to have an evaluation index to measure the classification effect of the classifier. The authors mainly selected classification accuracy, G-means and OBD score as evaluation indexes.

Suppose the data set is $D = \{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$, where y_i is the true label of the data x_i, and $f(x_i)$ is the predicted label of the data x_i given by the classifier f. Classification accuracy can be expressed as the ratio of the number of correctly predicted data to the total number of data (Demsar & Schuurmans, 2006), its calculation formula is as follow.

$$acc(f:D) = \frac{1}{m}\sum_{i=1}^{m} I(f(x_i) = y_i) \tag{7}$$

Where I is the indicator function, when $f(x_i) = y_i$, $I(f(x_i) = y_i) = 1$.

From the definition of classification accuracy, we can conclude that the higher the classification accuracy achieved by a certain classifier, the better the classification effect.

Table 4 gives the confusion matrix for binary classification (Zeng, 2020).

On this basis, the calculation formula of G-means is as follow (Kulkarni et al., 2021).

$$G-means = \sqrt{\frac{TP}{TP+TN}\times\frac{TN}{FP+TN}} \tag{8}$$

From the calculation formula we can see that, only when the classification accuracy of the two categories of data is relatively high, G-means will be relatively high. Therefore, instead of using classification accuracy to give feedback on the overall classification results of the data

Table 4
Confusion matrix for binary classification.

True result	Prediction result	
	Positive	Negative
Positive	TP	FN
Negative	FP	TF

set, G-means can be used to evaluate the classification results of data with different categories. In general, the higher the G-means obtained by a classifier, the better its classification effect.

OBD refers to the fact that in the process of constructing the CART, bootstrap sampling results in repeated sampling of some data in the original data set, and the remaining data is not extracted (Stekhoven & Buehlmann, 2012). The collection of these unextracted data is called OBD. Therefore, the OBD score is defined as the average classification accuracy obtained by a classifier trained on the sampled data set on the corresponding OBD. From the definition of OBD score, it can be seen that the higher the OBD score obtained by a classifier, the better its classification effect.

In addition to this, the authors also chose statistical tests as an auxiliary performance evaluation index to assess the variability between the random forests. Considering that none of the twenty data sets in Table 1 and Table 2 conformed to a standard normal distribution, the authors chose the non-parametric test. Specifically, the authors chose the Wilcoxon Signed Rank Test (Crichton, 2000), the Friedman Test (Pereira et al., 2015) and the Nemenyi Test (Liu & Chen, 2012). Given the focus and the length of this paper, the specific test steps of the Wilcoxon Signed Rank Test, the Friedman Test and the Nemenyi Test are not described here, but the null hypothesis H_0 and the alternative hypothesis H_1 for carrying out the above three tests are defined below, respectively.

In Wilcoxon Signed Rank Test, For a data set, suppose we use n random forests to make predictions on it, and achieve n classification accuracies. Among them, the random forest that achieves the highest classification accuracy, i.e., the best classification effect, is M_{opt}, and the remaining $n-1$ random forests are $M_i(i=1,2,\cdots,n-1)$. Then the null hypothesis H_0 is that, M_{opt} and M_i have the same classification effect on the data set. The alternative hypothesis H_1 is that, there is a difference in classification effect between M_{opt} and M_i on the data set. Thus, we need to calculate the statistic $T_i(i=1,2,\cdots,n-1)$ as the significance level between the classification effect of M_{opt} and M_i on the data set. Typically, we set the threshold for the significance level to 0.05. Thus, if the calculated $T_i > 0.05$, then we accepted the null hypothesis and rejected the alternative hypothesis. Conversely, we accepted the alternative hypothesis.

In Friedman Test, for n random forests and N data sets, we obtain the classification accuracy achieved by each random forest on each data set. Sort random forests based on classification accuracies, and obtain the sequence numbers $1,2,\ldots,n$. It should be noted that, if multiple random forests achieve the same classification accuracy, they are evenly divided into the sum of two sequence numbers. Assuming the average sequence number of the $i-th$ random forest is r_i, thus r_i follows a normal distribution.

$$\tau_{\chi^2} = \frac{12N}{n(n-1)} \times \left(\sum_{i=1}^{n} r_i^2 - \frac{n(n+1)^2}{4}\right) \tag{9}$$

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(n-1)-\tau_{\chi^2}} \tag{10}$$

τ_F follows the F distribution with $n-1$ and $(n-1)(N-1)$ freedom degrees. At this point, the null hypothesis H_0 is that, there is no difference in the classification effects of these n random forests on N data sets. The alternative hypothesis H_1 is that, there are differences in the classification effects of these n random forests on N data sets. Based on the specific values of n and N, the threshold for the significance level can be obtained by looking up the table. Therefore, if the calculated τ_F is less than the threshold, we accepted the null hypothesis and rejected the alternative hypothesis. Conversely, we accepted the alternative hypothesis. On the basis of using the Friedman Test to determine the differences in classification effects among n random forests, it is necessary to continue using the Nemenyi Test to further distinguish each random forest.

In Nemenyi Test, we use formula (11) to calculate the critical threshold CD for the average sequence number difference of random forests mentioned above:

$$CD = q_\alpha \times \sqrt{\frac{n(n+1)}{6N}} \tag{11}$$

Based on the specific value of n, q_α can be obtained by looking up the table. Similarly, we express the random forest with the best classification effect as M_{opt}, and express the remaining $n-1$ random forests as $M_i(i=1,2,\cdots,n-1)$. At this point, the null hypothesis H_0 is that, M_{opt} and M_i have the same classification effect on the data set. The alternative hypothesis H_1 is that, there is a difference in classification effect between M_{opt} and M_i on the data set. Thus, we need to calculate the statistic $T_i(i=1,2,\cdots,n-1)$ as the significance level between the classification effect of M_{opt} and M_i on the data set. The significance level at this point is the difference between the average sequence numbers of other random forests and the random forest with the best classification effect, and the threshold for the significance level is CD. Therefore, if the calculated $T_i < CD$, then we accepted the null hypothesis and rejected the alternative hypothesis. Conversely, we accepted the alternative hypothesis.

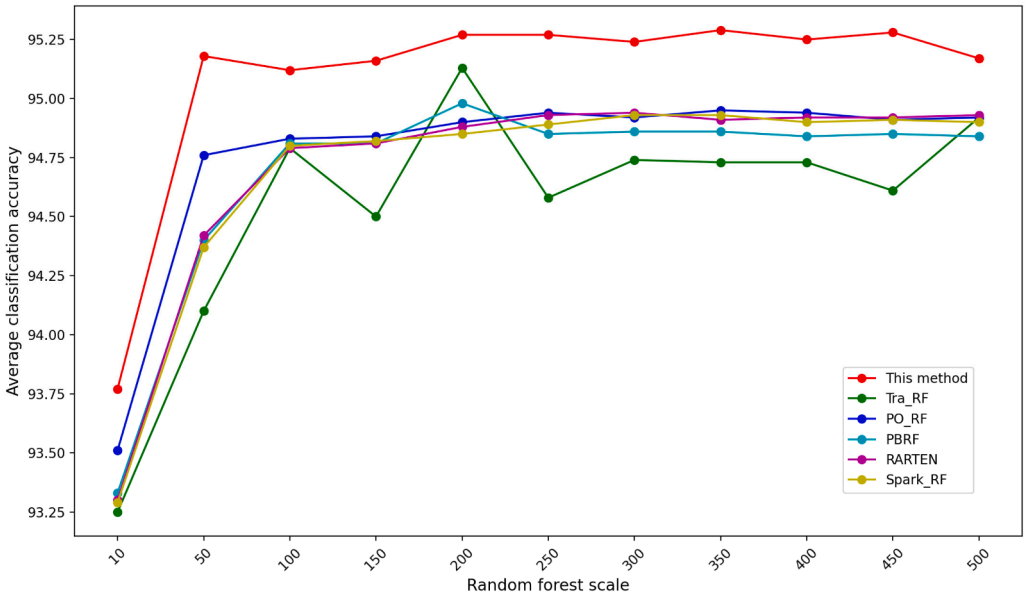
4.4. Experimental results

In order to verify the effectiveness of the improved random forest proposed in this paper, referring to the scheme designed by Canete-Sifuentes et al. (2019), the authors carried out experiments on the above-mentioned twenty data sets, and compared the comprehensive performance of the improved random forest and other five random forests from the classification accuracy, G-means, OBD score and three non-parametric tests, respectively, to highlight the advantages of the improved random forest proposed in this paper.

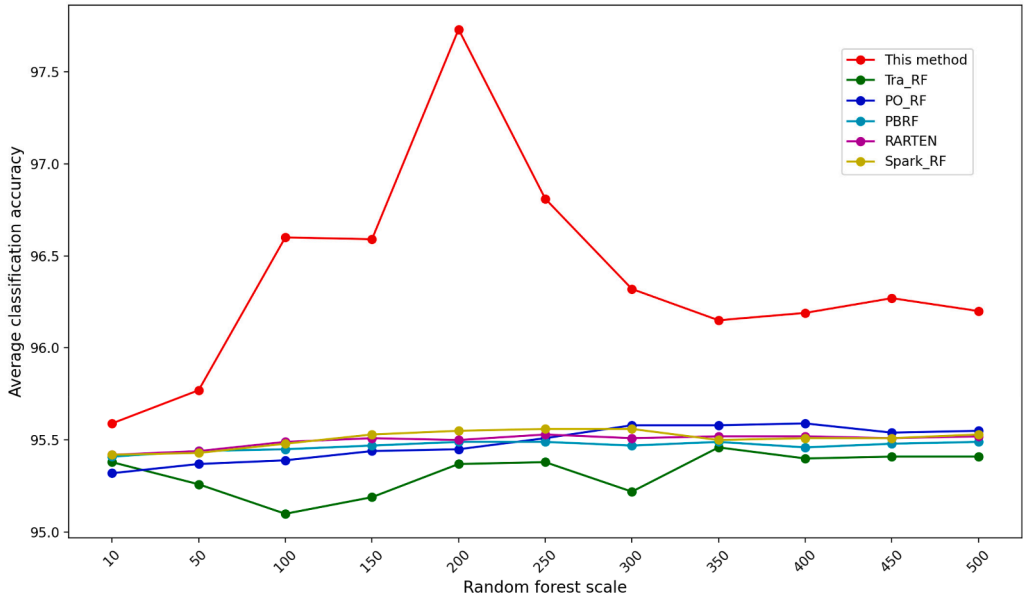
4.4.1. Comparison of classification accuracy

Following the determination principle of the inner product threshold applicable to a certain data set in Section 4.2, we obtained inner product thresholds applicable to the twenty data sets in Table 1 and Table 2. On this basis, in this section, we compared the average classification accuracy achieved by the proposed improved random forest with other five random forests on various data sets (with different random forest scales). It is worth noting that, the random forest scale here is actually the preset number of CARTs mentioned above. According to the amount of data contained in each of the twenty data sets, we set the value range of the random forest scale from 10 to 500, and set the value interval to 50. It should be noted that, except that the first value interval was set to 40 (according to 10 to 50), the rest of the value intervals were all set to 50. We applied the grid search method, and set the search range and search step as the value range and value interval, respectively, to achieve the average classification accuracy corresponding to each search. In this way, we obtained the average classification accuracy achieved by the proposed improved random forest and other five random forests on each data set (with different random forest scales). We selected six representative data sets from Table 1 and Table 2, and draw the comparison chart of the average classification accuracy achieved by six random forests, as shown in Fig. 4 and Fig. 5. Where the horizontal coordinate is the random forest scale, i.e., the preset number of CARTs, and the vertical coordinate is the achieved average classification accuracy.

From Fig. 4 and Fig. 5, it can be seen that the proposed improved random forest achieves significantly better average classification accuracy than the other five random forests on private data set 1, private data set 2, Abalone, localization data set 3, material data set and Diabetes_HI. After reaching a certain forest scale, the proposed improved random forest also achieves significantly better average classification accuracy than the other five random forests on Diabetes, Microsoft_MS, UrbanSound8K Data Set and Wine_QN. It is worth noting that, the improved



a. Private data set 1

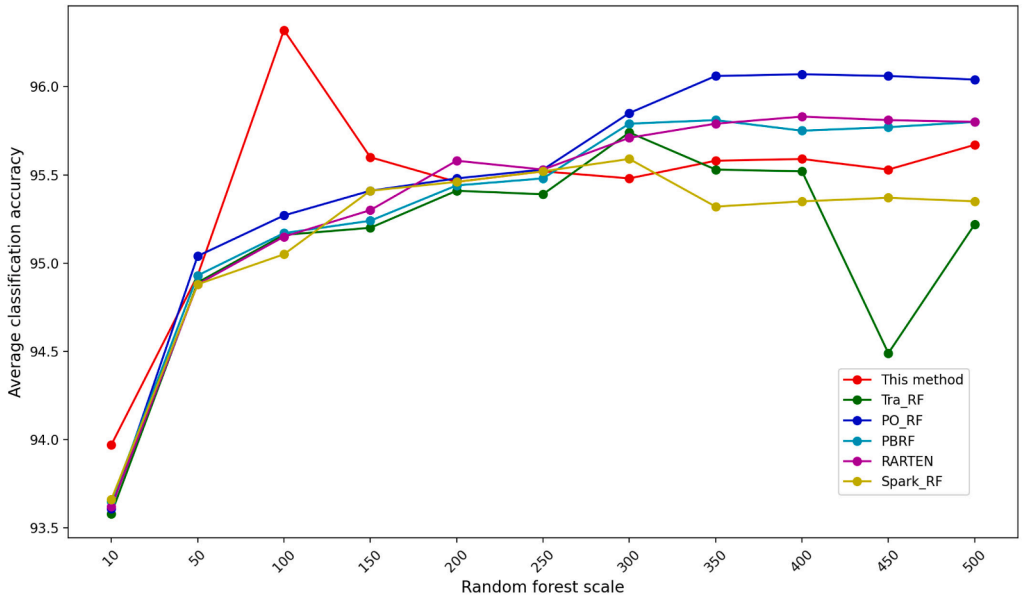


b. Private data set 2

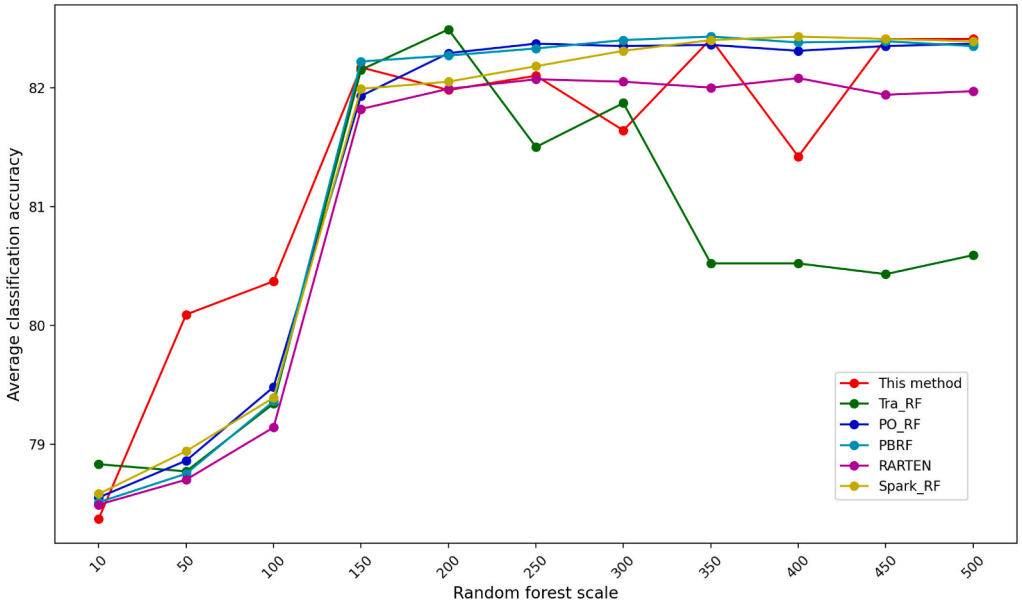
Fig. 4. Average classification accuracy achieved by six random forests on data sets containing two categories (partial).

random forest does not have a clear lead in average classification accuracy on private data set 6 and Haberman, but both outperform the Tra_RF. Specifically, on private data set 6, when the random forest scale is less than 200, the proposed improved random forest has a clear lead, and when the random forest scale is larger than 200, the average classification accuracy achieved by the proposed improved random forest is slightly lower than that of PO_RF, PBRF and RARTEN, but the difference is small. On Haberman, when the random forest scale is less than 150, the proposed improved random forest has an obvious leading edge. When the random forest scale is larger than 150, the average classification accuracy achieved by the proposed improved random forest is slightly lower than that of PO_RF, PBRF and Spark_RF, again with a small difference, but the proposed improved random forest shows an unstable and fluctuating state. Overall, the proposed improved random forest achieves a stable and significant performance advantage on six data sets containing multiple categories and four data sets containing two categories, and achieves a higher performance advantage than the Tra_RF on two data sets containing two categories. This indicates that the proposed improved random forest is more advantageous on data sets containing multiple categories.

Table 5 shows the highest average classification accuracies achieved by six random forests on the twenty data sets. From the table, it can be seen that, the proposed improved random forest achieves the highest average classification accuracy higher than the other five random forests on all data sets except Haberman and Microsoft_MS. On Microsoft_MS, the improved random forest achieves the same highest average classification accuracy as RARTEN. On Haberman, the highest average classification accuracy of 84.41% is achieved by the proposed improved random forest, which is close to the highest of 84.49%. It should be noted that, as mentioned above, the proposed improved random forest



c. Private data set 6



d. Haberman

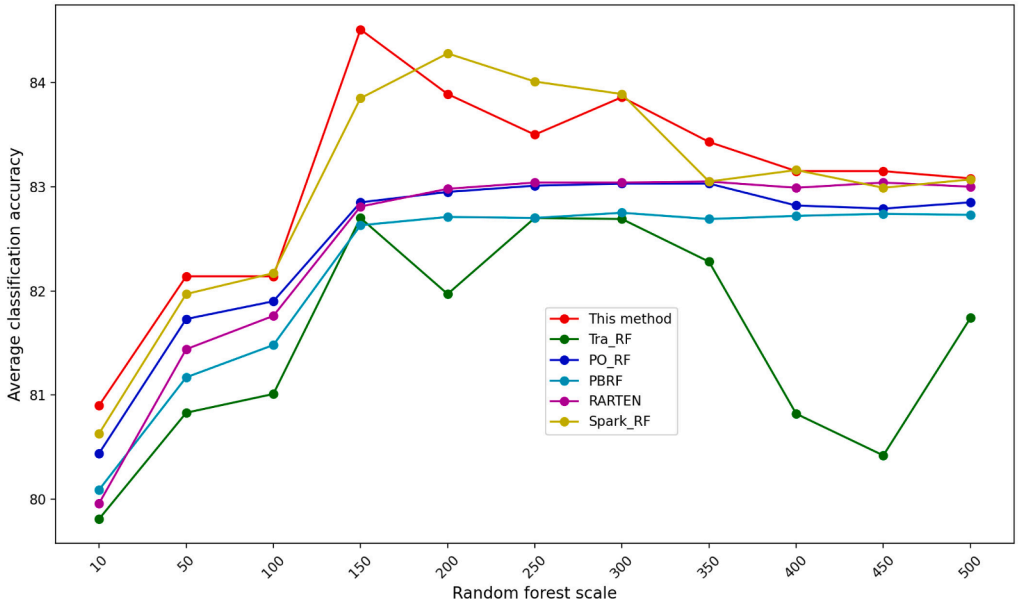
Fig. 4. (*continued*).

does not have a obvious lead in terms of the average classification accuracy achieved on Haberman. However, when compared the achieved highest classification accuracies, the other random forests have a very weak lead. Further, as mentioned above, the average classification accuracy achieved by the proposed improved random forest also does not have a obvious lead on the private data set 6. However, when comparing the achieved highest classification accuracy, the highest average classification accuracy achieved by the proposed improved random forest is higher than that of the other random forests, and the advantage is obvious. The above analysis fully illustrates the superiority of the proposed improved random forest.

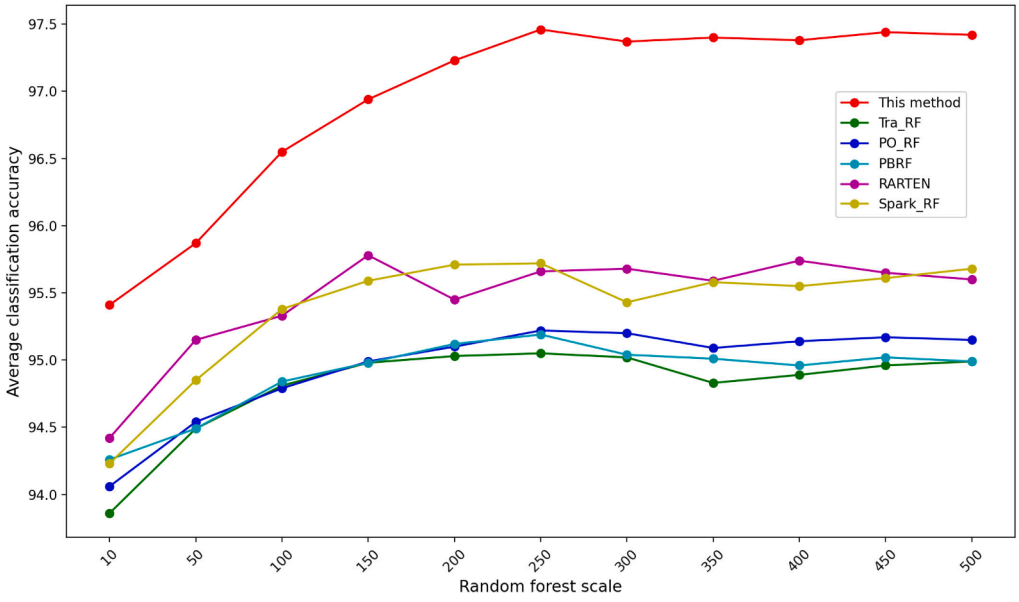
4.4.2. Comparison of G-means and OBD score

In order to further compare the performance between the proposed improved random forest and other five random forests, in this section, we calculated the G-means and OBD score obtained by six random forests on twenty data sets, respectively. The calculation results are shown in Table 6.

As can be seen from the table, the G-means and OBD scores obtained by the proposed improved random forest are higher than those of the other five random forests on the ten data sets containing multiple categories. For the ten data sets containing two categories, the OBD scores obtained by the proposed improved random forest are higher than the other five random forests. Except for the private data set 6 and Haberman, the G-means obtained by the proposed improved random forest are higher than the other five random forests on the other eight data sets containing two categories. On the private data set 6 and Haberman, the G-means obtained by the proposed improved random forest are the same as the other random forests. Overall, the G-means and OBD scores obtained by the proposed improved random forest have a obvious lead.



e. Diabetes



f. Abalone

Fig. 4. (*continued*).

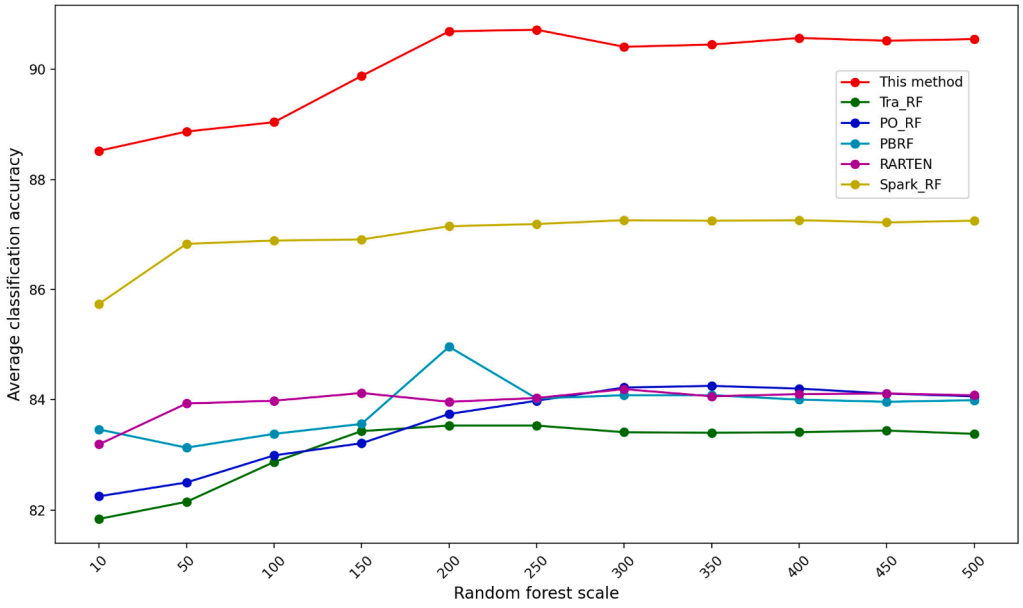
The above analysis again fully illustrates the superiority of the improved random forest proposed in this paper.

4.4.3. Comparison of non-parametric tests

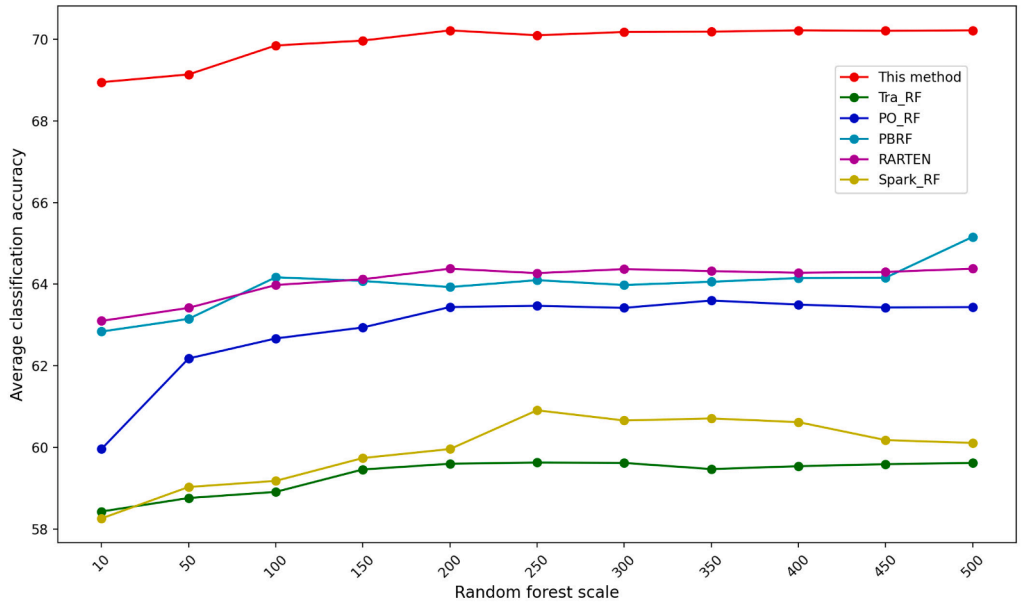
As mentioned above, in this study, the authors also chose three non-parametric tests as auxiliary performance evaluation indexes to assess the diversity between the proposed improved random forest and the other five random forests. Referring to Section 4.3, we can make the following settings. In Wilcoxon Signed Rank Test, assume that M_{opt} is the proposed improved random forest, M_1 is the Tra_RF, M_2 is the PO_RF, M_3 is the PBRF, M_4 is the RARTEN, and M_5 is the Spark_RF. In this way, the detailed description of the statistics $T_i(i = 1, 2, 3, 4, 5)$ that we need to compute is shown in Table 7.

Table 5 gives the highest average classification accuracies achieved by six random forests on the twenty data sets, with each highest average classification accuracy corresponding to a particular random forest at a given forest scale. In fact, the corresponding random forest is also optimal for the current scale. Therefore, we calculated the diversity between the proposed improved random forest and the other five random forests, as shown in Table 8.

As can be seen from the table, the calculated statistics T_1, T_2, T_3 and T_4 are all less than 0.05 on the twenty data sets. This means that they all accepted the alternative hypothesis, i.e., the classification effect of the proposed improved random forest differs from that of Tra_RF, PO_RF, PBRF and RARTEN. This indicates that the proposed improved random forest is different from the improved ideas of PO_RF, PBRF and RARTEN. As described in the Introduction, PBRF is an improvement of the Tra_RF combined with the Lasso method, and is a way of introducing an optimization algorithm, while RARTEN is an improvement of the random forest by using penalized regression methods and Elastic Net regression, which is also a way of introducing mathematical methods. PO_RF is a stepwise tuning of parameters on the Sklearn platform using the grid



a. Localization data set 3

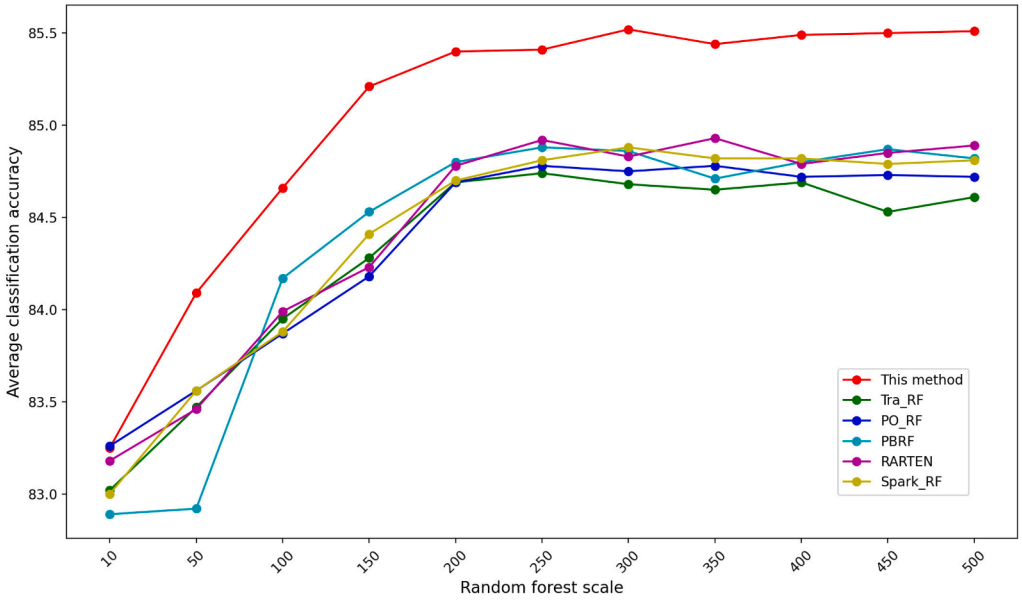


b. Material data set

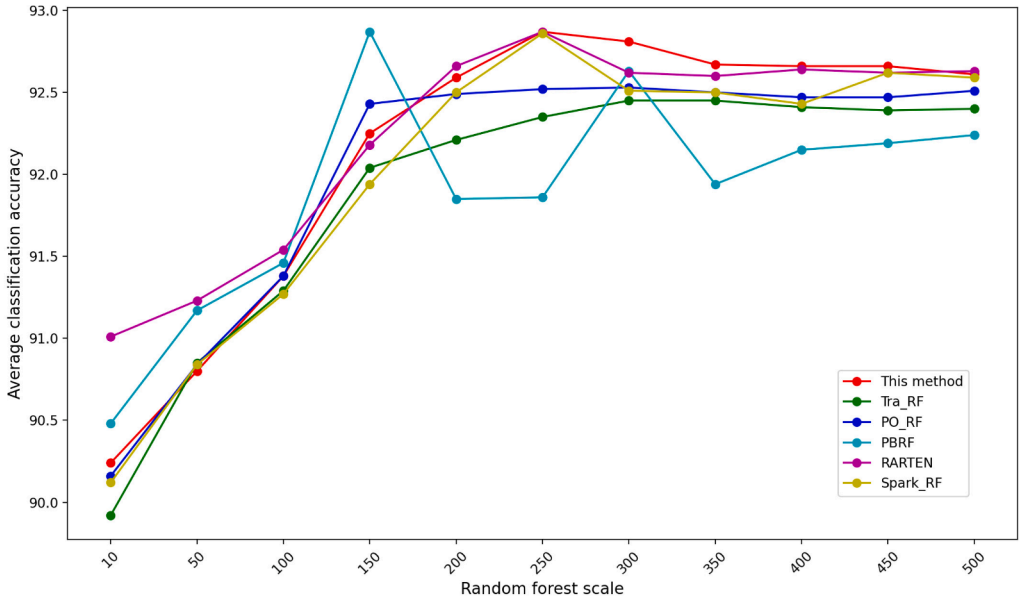
Fig. 5. Average classification accuracy achieved by six random forests on data sets containing multiple categories (partial).

search method, and does not have an interpretable set of logic or methodology, but is simply an application of the technique. In addition, the vast majority of the calculated statistics T_5 are less than 0.05, but three are equal to 0.05 and one is greater than 0.05, which is 0.051. This suggests that the classification effect of the proposed improved random forest differs from that of Spark_RF on private data set 3, private data set 6 and Wine, but not significantly, especially on Abalone, which better illustrates the small diversity between the two random forests. Referring to the descriptions in Introduction, we learn that one of the core ideas of the improved random forest proposed by Wang et al. (2022) is to select decision trees based on the similarity and classification accuracy of different decisions, and then construct a random forest. This is similar to part of the improved idea in this paper, although the methods used to calculate similarity and filter decision trees are different. This is the main reason for the small difference in classification effect between the two random forests. However, overall, the calculated statistic T_5 is less than or equal to 0.05 on the majority of the data sets, indicating that the classification effect of the proposed improved random forest is widely different from that of Spark_RF.

In Friedman Test, we also obtained the sequence numbers of six random forests on each data set based on their highest average classification accuracy achieved on twenty data sets shown in Table 5, as shown in Table 9. From this, we calculated that $\tau_{\chi^2} = 75.56$, $\tau_F = 58.74141$. In this study, the number of random forests was 6, the number of data sets was 20, then the threshold for the significance level was 2.310 through table lookup. At this point, because $\tau_F > 2.310$, thus



c. Diabetes_HI



d. Microsoft_MS

Fig. 5. (*continued*).

we accepted the alternative hypothesis, i.e., there are differences in the classification effects of six random forests. Next, we used the Nemenyi Test for further processing.

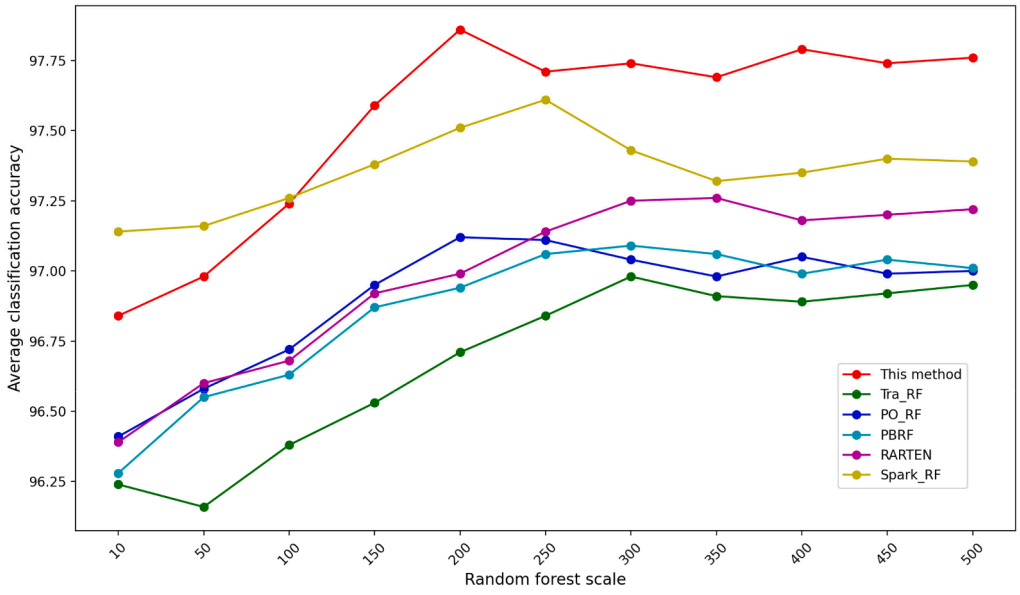
In Nemenyi Test, we also assume that M_{opt} is the proposed improved random forest, M_1 is the Tra_RF, M_2 is the PO_RF, M_3 is the PBRF, M_4 is the RARTEN, and M_5 is the Spark_RF. The detailed description of the statistics $T_i(i = 1, 2, 3, 4, 5)$ that need to compute is also shown in Table 7. In this study, the number of random forests was 6, then $q_\alpha = 2.850$ was obtained through table lookup, thus $CD = 1.686083$ was calculated. On the basis of Table 9, we calculated the average sequence numbers of six random forests on twenty data sets to be 1.175, 5.35, 3.825, 4.075, 3.075, and 3.5, respectively. In this way, we calculated the statistics $T_i(i = 1, 2, 3, 4, 5)$, as shown in Table 10.

From Table 10, it can be seen that since $T_i(i = 1, 2, 3, 4, 5) > 1.686083$, and the gap is significant, thus we accepted the alternative hypothesis. That is, the classification effects of the proposed improved random forest differs from the other five random forests. By synthesizing the test results of three non-parametric tests, it can be fully demonstrated that the proposed improved random forest has significant differences from the other five random forests, indicating its innovation.

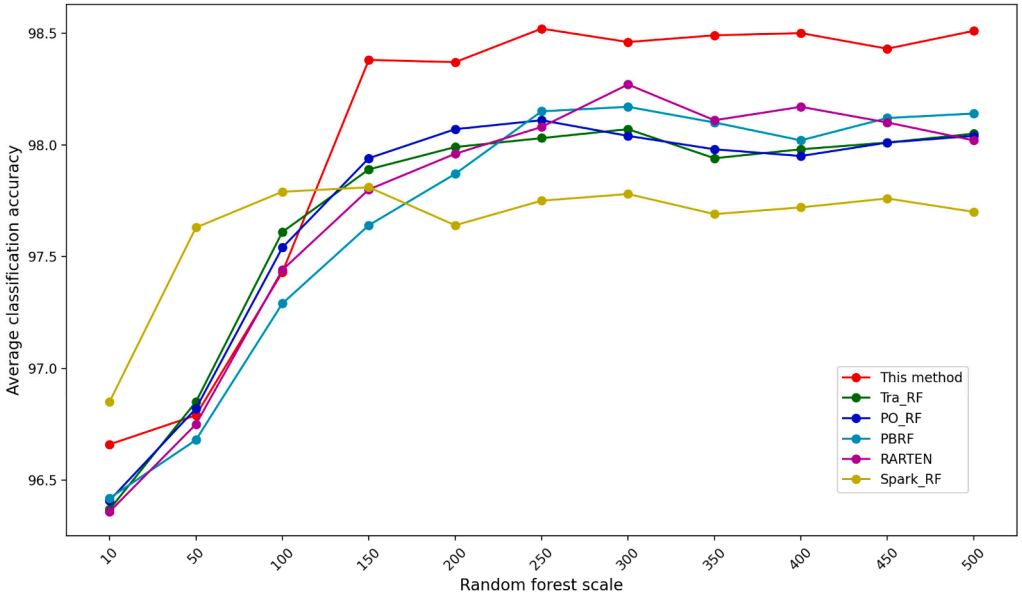
Combining Section 4.4.1, Section 4.4.2 and Section 4.4.3, from six evaluation indexes of average classification accuracy, G-means, OBD scores, Wilcoxon Signed Rank Test, Friedman Test and Nemenyi Test, compared with the other five random forests, the proposed improved random forest shows significant performance advantages, which fully demonstrates its superiority and feasibility.

5. Discussion

As mentioned above, the average classification accuracy achieved by the proposed improved random forest on private data set 6 and Haberman does not hold a obvious lead and shows an erratic fluctuation. In



e. UrbanSound8K



f. Wine_QN

Fig. 5. (*continued*).

this regard, the authors conducted research on the source data set. We found that there was a data imbalance between the private data set 6 and the Haberman. Specifically, in the private data set 6, the number of positive class data representing the loose particle signal was almost twice as many as the negative class data representing the component signal. In the Haberman data set, there were almost three times as many positive data representing "patients survived 5 years or more" as negative data representing "patients died within 5 years". Therefore, the authors made a bold inference, i.e., the smaller the imbalance of the data set, the more stable the classification effect advantage of the improved random forest. In fact, this conjecture was tested on the ten data sets containing multiple categories shown in Table 2. In the Microsoft_MS, UrbanSound8K and other data sets, there is a problem of one, two or three categories having less data compared to other categories, and the degree of imbalance has reached two or even three times. However, because these data sets contain more categories, the impact of this imbalance is minimal. In permutations, the larger the n, the larger the C_n^2. Therefore, the more categories the data set contains, the more correlation relationships there are between the different categories. The deterioration of individual relationships among them does not affect the proper functioning of the overall system relationships. However, in a data set containing only two categories, there is only one correlation relationship, thus a deterioration of this relationship can have a significant impact on the overall system relationship. Therefore, upgrading the existing improved random forest in conjunction with the imbalance degree in the data set or the number of categories contained in the data set is a worthwhile study in future.

Computational loss is another important reference for measuring the comprehensive performance of the improved random forest, and in machine learning, computational loss usually translates into the time taken by the algorithm to run. In Section 4, the authors did not compare the proposed improved random forest with the other five random forests

Table 5
Highest average classification accuracy achieved by six random forests on twenty data sets (%).

Data set	This method	Tra_RF	PO_RF	PBRF	RARTEN	Spark_RF
Private data set 1	95.29	95.13	94.95	94.98	94.94	94.93
Private data set 2	97.73	95.46	95.59	95.49	95.53	95.56
Private data set 3	97.68	97.47	97.51	97.51	97.55	97.60
Private data set 4	94.71	92.82	93.04	93.45	93.42	93.49
Private data set 5	94.66	93.62	93.68	93.61	93.87	93.73
Private data set 6	96.32	95.74	96.07	95.81	95.83	95.59
Haberman	82.41	82.49	82.37	82.43	82.08	82.43
Diabetes	84.51	82.70	83.03	82.75	83.05	84.28
Blood	82.18	81.02	82.11	82.05	82.07	81.92
Abalone	97.46	95.05	95.22	95.19	95.78	95.72
Localization data set 1	96.09	87.01	95.23	90.77	91.34	88.82
Localization data set 2	98.50	83.67	95.51	96.26	98.05	92.43
Localization data set 3	90.72	83.53	84.25	84.08	84.19	87.26
Material data set	71.22	59.63	63.60	64.17	64.38	60.91
Wine	62.03	56.42	56.82	59.17	60.40	60.85
Diabetes_HI	85.52	84.74	84.78	84.88	84.93	84.88
Microsoft_MS	92.87	92.45	92.01	92.82	92.87	92.86
UrbanSound8K	97.86	96.98	97.12	97.09	97.26	97.61
Wine_QN	98.52	98.07	98.11	98.17	98.27	97.81
Machine_PMC	89.92	88.94	89.65	89.30	89.31	89.87

Table 6
G-means and OBD scores obtained by six random forests (%).

Data set	G-means						OBD scores					
	This method	Tra_RF	PO_RF	PBRF	RARTEN	Spark_RF	This method	Tra_RF	PO_RF	PBRF	RARTEN	Spark_RF
Private data set 1	95.17	93.28	93.76	93.54	93.66	94.29	94.37	92.93	92.90	92.88	93.36	93.45
Private data set 2	96.12	95.47	95.53	95.49	95.51	95.71	94.68	93.55	93.82	93.39	94.02	93.91
Private data set 3	95.29	94.25	94.43	94.45	95.02	95.07	95.03	92.71	94.74	93.40	93.81	94.58
Private data set 4	93.96	93.71	93.82	93.94	93.94	93.95	95.57	94.16	95.47	94.73	94.64	95.02
Private data set 5	96.89	94.31	94.91	95.02	95.00	95.16	97.64	96.17	97.51	96.02	96.79	96.31
Private data set 6	93.30	92.59	92.99	92.85	93.30	93.23	96.92	92.48	95.33	93.19	95.22	95.84
Haberman	80.37	76.83	80.17	80.37	80.04	80.36	80.75	77.23	78.98	77.34	77.41	79.12
Diabetes	82.52	79.10	81.12	80.95	81.17	82.06	78.67	73.86	78.30	73.96	75.84	77.12
Blood	82.77	81.11	81.95	81.44	81.62	81.80	83.41	79.44	82.50	79.39	81.26	82.70
Abalone	97.11	96.88	97.09	96.79	96.87	96.88	97.07	94.33	96.69	94.59	94.76	94.81
Localization data set 1	94.86	86.45	90.55	87.77	88.27	88.98	96.16	86.92	88.92	88.77	90.53	91.38
Localization data set 2	97.92	83.51	92.04	87.69	87.53	88.49	98.29	83.20	91.39	88.11	90.16	93.59
Localization data set 3	90.15	83.04	85.41	87.01	87.26	86.88	90.53	85.26	88.82	86.99	89.76	88.54
Material data set	70.29	58.92	62.61	65.78	65.35	65.32	71.02	60.37	66.83	64.65	64.68	62.85
Wine	60.23	56.45	59.12	57.77	58.12	57.92	63.85	56.83	60.44	61.32	62.18	61.87
Diabetes_HI	83.18	83.97	83.14	82.81	83.07	82.90	85.46	84.11	84.52	84.46	84.52	84.21
Microsoft_MS	89.47	86.28	88.27	88.28	88.57	89.02	92.83	91.96	92.06	91.85	92.16	92.77
UrbanSound8K	95.99	93.06	94.45	93.87	93.86	94.20	97.98	96.08	96.99	96.89	96.78	97.17
Wine_QN	97.14	95.95	97.10	96.87	97.03	97.05	98.57	97.64	98.50	97.63	97.62	97.69
Machine_PMC	86.84	82.80	86.23	86.06	86.28	86.79	89.80	88.28	88.77	88.60	89.56	89.15

Table 7
Description of the statistic for the Wilcoxon Signed Rank Test.

Statistics $T_i (i = 1, 2, 3, 4, 5)$	Meaning
T_1	Differences between this method and Tra_RF
T_2	Differences between this method and PO_RF
T_3	Differences between this method and PBRF
T_4	Differences between this method and RARTEN
T_5	Differences between this method and Spark_RF

in terms of computational loss. In fact, the authors originally designed the improved random forest to solve the practical problems, includes the loose particle signal and component signal identification pronlem, the loose particle material identification problem and the loose particle localization problem. Therefore, the authors is more interested in the improvement of the classification effect. Moreover, with the rapid development of semiconductor and integrated circuit manufacturing technologies, the computational effect of existing CPUs or GPUs is powerful enough to perform fast processing of large amounts of data. To fully validate the superiority of the proposed improved random forest, the authors selected three representative data sets from Table 1 and Table 2, which have a large amount of data or multiple categories, and used them to test the six random forests. The computational losses of the six random forests are given in Table 11. It should be noted that, the above tests were all conducted on the same computer, which used an i7-12700H processor with 16G of RAM and an RTX3060 discrete graphics card.

As can be seen from the table, the Tra_RF has the obvious advantage of low computational loss, with very little computational loss on all data sets, while the computational loss of the remaining four random forests, except Spark_RF, is significantly greater than that of Tra_RF. Among them, PO_RF is a stepwise method for tuning parameters using the grid search method on the Sklearn platform, whose main computational loss is in parameter search. However, because the Sklearn platform is rich in function resources, the process has little computational loss. PBRF and RARTEN both introduce an optimization algorithm or mathematical method. They both require algorithmic or mathematical computation in the process of constructing a random forest, which can lead to large computational losses. The computational loss of the proposed improved random forest is mainly in the correlation calculation and selection of CARTs, as well as the random forest sacle. Compared to algorithmic or mathematical computations, the computational loss of these is relatively

Table 8
Statistics T_i calculated by six random forests on twenty data sets.

Data set	T_1	T_2	T_3	T_4	T_5
Private data set 1	0.043	0.046	0.041	0.041	0.049
Private data set 2	0.046	0.048	0.048	0.042	0.046
Private data set 3	0.034	0.037	0.038	0.038	0.050
Private data set 4	0.040	0.039	0.040	0.048	0.048
Private data set 5	0.037	0.041	0.038	0.047	0.048
Private data set 6	0.038	0.034	0.043	0.039	0.050
Haberman	0.041	0.044	0.047	0.041	0.046
Diabetes	0.045	0.042	0.049	0.046	0.046
Blood	0.046	0.047	0.045	0.037	0.047
Abalone	0.041	0.037	0.043	0.045	0.051
Localization data set 1	0.045	0.048	0.040	0.045	0.049
Localization data set 2	0.037	0.043	0.041	0.049	0.047
Localization data set 3	0.042	0.041	0.044	0.044	0.048
Material data set	0.038	0.046	0.037	0.041	0.047
Wine	0.047	0.042	0.043	0.040	0.050
Diabetes_HI	0.044	0.041	0.038	0.038	0.049
Microsoft_MS	0.046	0.045	0.047	0.046	0.047
UrbanSound8K	0.047	0.040	0.038	0.046	0.049
Wine_QN	0.038	0.038	0.043	0.041	0.049
Machine_PMC	0.044	0.047	0.044	0.038	0.045

Table 9
Sequence numbers of six random forests on twenty data sets.

Data set	This method	Tra_RF	PO_RF	PBRF	RARTEN	Spark_RF
Private data set 1	1	2	4	3	5	6
Private data set 2	1	6	2	5	4	3
Private data set 3	1	6	4.5	4.5	3	2
Private data set 4	1	6	5	3	4	2
Private data set 5	1	5	4	6	2	3
Private data set 6	1	5	2	4	3	6
Haberman	4	1	5	2.5	6	2.5
Diabetes	1	6	4	5	3	2
Blood	1	6	2	4	3	5
Abalone	1	6	4	5	2	3
Localization data set 1	1	6	2	4	3	5
Localization data set 2	1	6	4	3	2	5
Localization data set 3	1	6	3	5	4	2
Material data set	1	6	4	3	2	5
Wine	1	6	5	4	3	2
Diabetes_HI	1	6	5	3.5	2	3.5
Microsoft_MS	1.5	5	6	4	1.5	3
UrbanSound8K	1	6	4	5	3	2
Wine_QN	1	5	4	3	2	6
Machine_PMC	1	6	3	5	4	2

Table 10
Statistics T_i of the average sequence numbers of six random forests.

T_1	T_2	T_3	T_4	T_5
4.175	2.65	2.9	1.9	2.325

small. It is worth noting that, Spark_RF is based on the Spark platform, which has abundant computational resources, thus the computational loss on each data set is minimal and the performance advantage is very obvious. In summary, the proposed improved random forest achieves moderate computational effect with guaranteed classification effect, which can prove another advantage over PBRF and RARTEN to some extent. In fact, if we deploy the propose improved random forest on a high-performance computing platform like Spark, it can also achieve satisfactory computational effect advantage. This is what the authors need to consider when applying the proposed improved random forest to the loose particle detection field.

Table 11
Computational loss of six random forests on six data sets (ms).

Data set	This method	Tra_RF	PO_RF	PBRF	RARTEN	Spark_RF
Private dta set 6	6870	5920	5980	7170	9370	4210
Diabetes	550	270	270	550	810	190
Blood	500	250	260	510	800	190
Localization data set 1	58,560	46,380	49,560	60,180	88,260	30,170
Localization data set 3	28,130	20,270	22,880	28,320	32,540	16,850
Diabetes_HI	58,820	51,650	53,850	61,840	93,280	36,380

Following the descriptions in Introduction, it is also optional to incorporate improvements to the voting mechanism into the proposed improved random forest. For example, a weighted voting mechanism can be used to increase the influence of high-quality samples. In addition, in the improved random forest, we used the CARTs to make predictions on three reserved data sets, respectively, and obtained the selection order of the CARTs in term of the achieved average classification accuracy. In addition to the PBRF and RARTEN mentioned above, in a heuristic parallel selection ensemble algorithm based on clustering and improved simulated annealing proposed by Wu (2020), he selected the optimal classifier sequence set based on the improved simulated annealing heuristic selective ensemble algorithm. Therefore, the introduction of high-performance optimization algorithms into the improvement of random forest is also worth considering in future research. Combining the aforementioned high-performance computing platforms, on the one hand, we can improve the classification effect of random forests by introducing optimization algorithms, and on the other hand, we can take advantage of computing platforms to improve the computational effect of random forests.

6. Conclusion

In the process of constructing the traditional random forest, bootstrap sampling will make the quality of the data subsets used to train the base classifiers vary, thus the constructed decision trees may not all have better classification effects. At the same time, this sampling method will cause the overlapping of samples between some data subsets, resulting in high correlations between the constructed decision trees. Aiming at these problems, in this paper, the authors proposed an improved random forest based on the classification accuracy and correlation measurement of decision trees. Those decision trees with better classification effects and low correlations were retained to construct the high-quality random forest. Specifically, we used three reserved data sets to evaluate the classification effects of each CART, and achieved their average classification accuracies. On this basis, we ranked all the CARTs in descending order based on the achieved average classification accuracy. We used the improved dot product method to calculate the correlations between CARTs. By using the average classification accuracy achieved by the random forest as reference, combined with the grid search method, the inner product thresholds applicable to each data set was obtained. Then we marked the CARTs with low average classification accuracy among CART pairs whose inner product values are higher than the inner product threshold as deletable. Considered the average classification accuracy and correlation of each CART, we deleted those CARTs with high correlation and low average classification accuracy until the number of remaining CARTs met the preset number. Therefore, the remaining CARTs all have high classification effects and low

correlations, and did not have a negative impact on the classification effect or decision redundancy of the final constructed random forest. The experimental results on twenty data sets show that, the proposed improved random forest achieved better evaluation index results than the other random forests. This fully proves the effectiveness of the proposed improved random forest. In principle, it can be applied to data classification research in various fields, and provides new ideas for the improvement of random forests.

CRediT authorship contribution statement

Zhigang Sun: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Funding acquisition. **Guotao Wang:** Conceptualization, Validation, Formal analysis, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Pengfei Li:** Methodology, Validation, Formal analysis, Visualization, Supervision. **Hui Wang:** Methodology, Formal analysis, Visualization, Supervision. **Min Zhang:** Software, Validation, Investigation, Data curation. **Xiaowen Liang:** Conceptualization, Software, Data curation.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Guotao Wang reports financial support was provided by National Natural Science Foundation of China. Guotao Wang reports financial support was provided by Key Research and Development Program of Heilongjiang Province; Guotao Wang reports financial support was provided by Heilongjiang Science Foundation. Guotao Wang reports financial support was provided by Heilongjiang Postdoctoral Fund. Guotao Wang reports financial support was provided by Fundamental Research Funds for Universities of Heilongjiang Province. Guotao Wang reports financial support was provided by Scientific and Technological Achievement Cultivation Funds of Heilongjiang Provincial Education Department. Zhigang Sun reports financial support was provided by Graduate Innovative Research Funds of Heilongjiang University.

Data availability

Data will be made available on request.

Acknowledgments

This study was co-supported by the National Natural Science Foundation of China (Nos. 51607059); the Key Research and Development Program of Heilongjiang Province (Nos. 2022ZX03A06); the Heilongjiang Science Foundation (Nos. QC2017059 and Nos. JJ2020LH1310); the Heilongjiang Postdoctoral Fund (Nos. LBH-Z16169); the Fundamental Research Funds for Universities of Heilongjiang Province (Nos. HDRCCX-201604 and Nos. 2020-KYYWF-1006); the Scientific and Technological Achievement Cultivation Funds of Heilongjiang Provincial Education Department (Nos. TSTAU-C2018016), the Graduate Innovative Research Funds of Heilongjiang University (Nos. YJSCX2021-067HLJU). In addition, Zhigang Sun would like to express his gratitude for meeting Miss Jing Xu in the most beautiful years. Thank you for your continuous care, support, trust, and waiting. Are you willing to marry me?

References

Amaratunga, D., Cabrera, J., & Lee, Y. S. (2008). Enriched Random Forests. *Bioinformatics, 24*(18), 2010–2014.
Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32.
Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123–140.
Canete-Sifuentes, L., Monroy, R., Medina-Perez, M. A., Loyola-Gonzalez, O., & Voronisky, F. V. (2019). Classification Based on Multivariate Contrast Patterns. *IEEE Access, 7*, 55744–55762.
Cervantes, B., Monroy, R., Medina-Perez, M. A., Gonzalez-Mendoza, M., & Ramirez-Marquez, J. (2017). Some Features Speak Loud, But Together They all Speak Louder: A Study on the Correlation Between Classification Error and Feature Usage in Decision-tree Classification Ensembles. *Engineering Applications of Artificial Intelligence, 67*, 270–282.
Chen, Z., Zhang, L. P., Wang, H., Zhang, J. J., & Wang, C. H. (2016). Clone Group Mapping Method Based on Improved Vector Space Model. *Journal of Computer Applications, 36*(7), 2031–2037.
Chetlur, V. V., Dhillon, H. S., & Dettmann, C. P. (2020). Shortest Path Distance in Manhattan Poisson Line Cox Process. *Journal of Statistical Physics, 181*(6), 2109–2130.
Crichton, N. (2000). Wilcoxon Signed Rank Test. *Journal of Clinical Nursing, 9*(4).
Demsar, J., & Schuurmans, D. (2006). Statistical Comparison of Classifier Over Multiple Data Sets. *Journal of Machine Learning Research, 7*(1), 1–30.
Deng, X. Y., Hu, S. Y., Li, Z., Sui, Z. S., & Sun, D. H. (2016). Similarity Matching Algorithm of Equipment Fault Case Based on SVSM. *Radio Engineering, 46*(2), 31–35.
Ding, B., Wu, Y. P., & Li, B. Y. (2015). Parking Plot Recognition Based on C4.5 Algorithm. *Electronic Measurement Technology, 38*(8), 64–68.
Ditzler, G., LaBarck, J., Ritchie, J., Rosen, G., & Polikar, R. (2018). Extensions to Online Feature Selection Using Bagging and Boosting. *IEEE Transactions on Neural Networks and Learning Systems, 29*(9), 4504–4509.
Farhadi, Z., Bevrani, H., & Feizi-Derakhshi, M. R. (2023). *Improving Random Forest Algorithm by Selecting Appropriate Penalized Method*. Early Access: Communications in Statistics - Simulation and Computation.
Gall, J., & Lempitsky, V. (2009). Class-specific Hough Forests for Object Detection. *2009 IEEE Conference on Computer Vision and Pattern Recognition, 10835939. USA*. IEEE.: Piscataway.
Ghosh, D., & Cabrera, J. (2021). Enriched Random Forest for High Dimensional Genomic Data. *IEEE-ACM Transactions on Computational Biology and Bioinformatics, 19*(5), 2817–2828.
Han, L., Li, W., & Su, Z. (2019). An Assertive Reasoning Method for Emergency Response Management Based on Knowledge Elements C4.5 Decision Tree. *Expert Systems with Application, 122*(5), 65–74.
Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832–844.
Ho, T.K. (1995). Random Decision Forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 278-282. Canada. Piscataway: IEEE.
Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random Survival Forests. *The Annals of Applied Statistics, 2*(3), 841–860.
Jiang, T., Jia, H. J., Li, G. Q., Chen, H. H., & Jin, X. L. (2017). Cross-Correlation Coefficient-Based Coherency Identification in Bulk Power System Using Wide-Area Measurements. *Transactions of China Electrotechnical Society, 32*(1), 1–11.
Jiang, Z. L., Ji, R., & Chang, K. C. (2020). A Machine Learning Integrated Portfolio Rebalance Framework with Risk-Aversion Adjustment. *Journal of Risk and Financial Management, 13*(7), 155.
Khoshkenar, A., & Mahlooji, H. (2013). A New Test of Randomness for Lehmer Generators Based on the Manhattan Distance Between Pairs of Consecutive Random Numbers. *Communications in Statistics - Simulation and Computation, 42*(1), 202–214.
Kim, A., Myung, J., & Kim, H. (2020). Random Forest Ensemble Using a Weight-adjusted Voting Algorithm. *Journal of the Korean Data and Information Science Sociaty, 31*(2), 427–438.
Kulkarni, R., Revathy, S., & Patil, S. (2021). A Novel Approach to Maximize G-mean in Nonstationary Data with Recurrent Imbalance Shifts. *International Arab Journal of Information Technology, 18*(1), 103–113.
Kulkarni, V. Y., & Sinha, P. K. (2013). Efficient Learning of Random Forest Classifier Using Disjoint Partitioning Approach. *Lecture Notes in Engineering & Computer Science, 2205*(1), 1–5.
Li, R., Li, F., Zhang, J. Q., & Shen, F. T. (2012). Study on Boundary Search Method for DFM Mesh Generation. *China Foundry, 9*(3), 231–233.
Li, Y. Q., Yan, C., Liu, W., & Li, M. Z. (2018). A Principle Component Analysis-based Random Forest with the Potential Nearest Neighbor Method for Automobile Insurance Fraud Identification. *Applied Soft Computing, 70*, 1000–1009.
Li, Y. L., Kuang, Z. H., Li, J. Y., & Kang, K. G. (2020). Improving Random Projections With Extra Vectors to Approximate Inner Products. *IEEE Access, 8*, 78590–78607.
Liang, X. W., Jiang, A. P., Li, T., Xue, Y. Y., & Wang, G. T. (2020). LR-SMOTE—An Improved Unbalanced Dataset Oversampling Based on K-means and SVM. *Knowledge-Based Systems, 196*, Article 105845.
Liu, W. Y., Quan, X. J., Feng, M., & Qiu, B. T. (2010). A Short Text Modeling Method Combining Semantic and Statistical Information. *Information Sciences, 180*(20), 4031–4041.
Liu, Y. W., & Chen, W. H. (2012). A SAS Macro for Testing Differences among Three or More Independent Groups Using Kruskal-Wallis and Nemenyi Tests. *Journal of Huazhong University of Science and Technology - Medical Sciences, 32*(1), 130–134.
Liu, Y. Y., & Zhao, H. Y. (2017). Variable Importance-weighted Random Forests. *Quantitative Biology, 5*(4), 338–351.
Ma, J. J., Pan, Q., Liang, Y., Hu, J. W., Zhao, C. H., & Guo, Y. N. (2019). Object Detection Based on Improved Grassberger Entropy Random Forest Classifier. *Zhongguo Jiguang/Chinese Journal of Lasers, 46*(7), 0701011.
Martinez-Munoz, G., & Suarez, A. (2010). Out-of-bag Estimation of the Optimal Sample Size in Bagging. *Pattern Recognition, 43*(1), 143–152.
Merigo, J. M., & Casanovas, M. (2011). Induced Aggregation Operators in the Euclidean Distance and its Application in Financial Decision Making. *Expert Systems with Applications, 38*(6), 7603–7608.

Mienye, I. D., Wang, Z., & Sun, Y. (2019). Prediction Performance of Improved Decision Tree-based Algorithms: A Review. *Procedia Manufacturing, 35*, 698–703.
Pereira, D. G., Afonso, A., & Medeiros, F. M. (2015). Overview of Friedman's Test and Post-hoc Analysis. *Communications in Statistics-Simulation and Computation, 44*(10), 2636–2653.
Putri, R. P. S., & Waspada, I. (2018). Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika. *Khazanah Informatika Jurnal Ilmu Komputer dan Informatika, 4*(1), 1–7.
Schapire, R. E. (1990). The Strength of Weak Learnability. *Machine Learning, 5*(2), 197–227.
Sheng, W. S., & Sun, Y. W. (2019). An Improved ID3 Decision Algorithm and Its Application. *Computer & Digital Engineering, 47*(12).
Smayra, T., Charara, Z., Sleilaty, G., Boustany, G., Menassa-Moussa, L., & Halaby, G. (2019). Classification and Regression Tree (CART) Model of Sonographic Signs in Predicting Thyroid Nodules Malignancy. *European Journal of Radiology Open, 6*, 343–349.
Stekhoven, D. J., & Buehlmann, P. (2012). MissForest-non-parametric Missing Value Imputation for Mixed-type Data. *Bioinformatics, 28*(1), 112–118.
Suknovic, M., Delibasic, B., Jovanovic, M., Vukicevic, M., Becejski-Vujaklija, D., & Obradovic, Z. (2012). Reusable Components in Decision Tree Induction Algorithms. *Computational Statistics, 27*(1), 127–148.
Sun, T. B., Liu, J. H., Kan, J. M., & Sui, T. T. (2021). A Study on the Classification of Vegetation Point Cloud Based on Random Forest in the Straw Checkerboard Barriers Area. *Journal of Intelligent & Fuzzy Systems, 41*(3), 4337–4349.
Sun, Z. G., Jiang, A. P., Gao, M. M., Gao, L. Z., & Wang, G. T. (2021). Technology of Locating Loose Particles Inside Sealed Electronic Equipment Based on Parameter-Optimized Random Forest. *Measurement, 186*, Article 110164.
Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences, 43*(6), 1947–1958.
Tripoliti, E. E., Fotiadis, D. I., & Manis, G. (2013). Modifications of the Construction and Voting Mechanisms of the Random Forests Algorithm. *Data & Knowledge Engineering, 87*, 41–65.
Wang, D. M., Lu, C. H., Jiang, W. W., Xiao, M. X., & Li, B. R. (2015). Study on PSO-based Decision-tree SVM Multi-class Classification Method. *Journal of Electronic Measurement and Instrumentation, 29*(4), 611–615.
Wang, H., & Wang, G. Z. (2020). Improving Random Forest Algorithm by Lasso Method. *Journal of Statistical Computation and Simulation, 91*(2), 353–367.
Wang, S. Z., Zhang, Z. F., Geng, S. S., & Pang, C. Y. (2022). Research on Optimization of Random Forest Algorithm Based on Spark. *CMC-Computers Materials & Continua, 71*(2), 3721–3731.
Wu, M. H. (2020). Heuristic Parallel Selective Ensemble Algorithm Based on Clustering and Improved Simulated Annealing. *The Journal of Supercomputing, 76*(5), 3702–3712. https://doi.org/10.1007/s11227-018-2633-x
Zeng, G. P. (2020). On the Confusion Matrix in Credit Scoring and its Analytical Properties. *Communications in Statistics - Theory and Methods, 49*(9), 2080–2093.