Hindawi
Security and Communication Networks
Volume 2023, Article ID 9818912, 1 page
https://doi.org/10.1155/2023/9818912

Retraction

Retracted: Optimization and Application of Random Forest Algorithm for Applied Mathematics Specialty

Security and Communication Networks

Received 27 June 2023; Accepted 27 June 2023; Published 28 June 2023

Copyright © 2023 Security and Communication Networks. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] W. Li, "Optimization and Application of Random Forest Algorithm for Applied Mathematics Specialty," *Security and Communication Networks*, vol. 2022, Article ID 1131994, 9 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 1131994, 9 pages
https://doi.org/10.1155/2022/1131994

WILEY | Hindawi

RETRACTED

Research Article

Optimization and Application of Random Forest Algorithm for Applied Mathematics Specialty

Wei Li

Department of Marx School (Basic Teaching Department), Chongqing City Vocational College, Chongqing 402160, Yongchuan, China

Correspondence should be addressed to Wei Li; liwei2022215@163.com

Received 3 March 2022; Revised 19 April 2022; Accepted 27 April 2022; Published 21 May 2022

Academic Editor: Chin-Ling Chen

Copyright © 2022 Wei Li. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For unbalanced data classification, RF (Random forest) algorithm will cause problems such as poor classification performance and a large DT scale. With the advent of the era of big data, RF algorithms should have the ability to process large-scale data. Aiming at the problem that RF cannot handle unbalanced data well, this paper improves the feature selection method built in RF and proposes a new feature selection algorithm. On the basis of feature importance ranking, randomness is introduced to ensure the strength of each tree and reduce the correlation between trees. In the extended transform data set, the sensitivity of the RF model has exceeded 0.8, and that of other models has increased to about 0.65. The prediction accuracy of the centralized RF model for the company's credit rating reached 100%, while the CART model misjudged companies C6 and C7, while the Logit model misjudged companies C3, C5, and C8. Experiments prove the extrapolation of the RF model and its excellent prediction ability. In the practical application of applied mathematics specialty, the RF optimization algorithm proposed in this study can well handle continuous variables and improve the classification accuracy of RF. This paper holds that the advantages of the RF algorithm in data processing and model performance will make it more widely used in the field of enterprise credit risk evaluation.

1. Introduction

Supervised learning in machine learning algorithms is nothing more than solving classification problems and regression problems, among which there are many algorithms to solve classification problems, such as NB (Naive Bayesian), SVM (Support Vector Machine), DT (Decision Tree), and so on [1–3]. Obviously, these are all single classifiers, which are prone to overfitting problems, and there will be bottlenecks when improving their performance, so the ensemble learning algorithm came into being. Bayesian and DT are more representative of single classifier technology. These algorithms have promoted the development of classification technology to a certain extent, and all aspects of research and application have been comprehensively carried out. However, due to its own limitations, the performance improvement of a single classifier has reached an insurmountable bottleneck, so people began to put forward the idea of a multiclassifier combination [3]. Multiclassifier combination uses multiple base classifiers for classification and integrates all classification results to form a final result. RF (Random forest) is a multiclassifier combination produced under this background.

As a major direction in data mining, classification technology is a supervised machine learning method. It trains the training set to get the learner model and then tests the test set with this model to get the classification result. Rather et al. proposed a method to predict the activity of cannabinoid receptor agonists using RF technology [4]. BZBA and others introduced the Bagging method and systematically expounded the RF algorithm, and the RF algorithm officially became an important part of the data mining classification algorithm [5]. Vassallo et al. used the RF algorithm to study the land coverage area and found that the RF algorithm can train faster than other combination algorithms [6]. Cao et al. also applied RF to time series to detect the change points of time series [7]. Although the performance of the RF algorithm has been gradually improved, and the scenes used are more and more extensive, there are still some defects in some aspects, such as feature

selection and processing of unbalanced data sets. At home, researchers focus on the application of the RF algorithm but there is not much research on the optimization and improvement of the RF algorithm.

Because the RF algorithm has the problem of a low prediction rate of subcategories when dealing with unbalanced data sets, researchers optimized the data preprocessing process of the RF algorithm. They calculate the attribute weights, calculate the correlation between attributes according to the Chi-square test, and calculate by analyzing the correlation between each attribute and the target attribute. Finally, they sort the weights of each attribute, and in the process of feature selection, they tend to the attribute ranked first, which enhances the DT intensity of each tree and reduces the correlation coefficient between trees, thus improving the classification accuracy of the RF algorithm [8]. In this paper, on the basis of full access to relevant information at home and abroad, aiming at the problems of RF in theory and application, the optimization of RF algorithm and application for applied mathematics major is mainly carried out.

The main innovations of this paper are as follows:

(1) In the aspect of RF self-optimization, the influencing factors of classification performance of RF algorithm are analyzed in detail. Aiming at the phenomenon of different RF performances caused by different node splitting algorithms during RF generation, an RF optimization algorithm based on linear transformation is proposed.

(2) After the above optimization, this paper actively explores the application of the optimized RF algorithm in enterprise credit risk evaluation and constructs a six-level risk evaluation model based on RF. By comparing the models, it proves that RF has excellent stability, extrapolation, and predictive ability.

The paper is divided into five chapters.

The first chapter introduces the research background and outlines the main tasks of this paper. The second chapter introduces the research status of the RF algorithm. The third chapter introduces the optimization and application of the RF algorithm. The fourth chapter compares the performance of this model through experiments. The fifth chapter is the full-text summary.

2. Related Work

2.1. Optimization Method of RF Algorithm. Xu et al. proposed the random survival forest algorithm and introduced the concept of survival trees in RF [9]. In the process of the bootstrap resampling method, the algorithm has to generate a corresponding analysis tree for each training subset formed by sampling. Hao et al. put forward the quasiadaptive classification RF algorithm. First, it was found that the Adaboost algorithm has great advantages in adaptive self-help sampling weight and adaptive voting weight setting [10]. Fornaser et al. mixed the C4.5DT algorithm, and Fornaser et al. mixed the C4.5DTalgorithm and CART (Classification and Regression Tree) algorithm into one algorithm and used the mixed algorithm to generate the RF algorithm, which improved the accuracy of RF [11]. Paul et al. proposed an improved RF algorithm based on DT clustering to extract DTs with low classification accuracy and high similarity. Experiments show that this algorithm is higher than the traditional RF algorithm in integration accuracy and classification efficiency [12]. Li introduced the theory of quantile regression, applied quantile regression to the process of DT generation and decision-making, and proposed the quantile regression forest algorithm. He mathematically proved the consistency of quantile regression to the forest [13]. Chen et al. developed a new cost-sensitive RF algorithm from the perspective of applying a cost-sensitive learning algorithm to solve the classification problem of unbalanced data sets [14]. Dou et al. proposed a new RF feature selection algorithm by analyzing the relationship between the intensity and correlation coefficient of each tree in RF [15]. The main idea of this algorithm is that by analyzing the upper bound of RF generalization error, it is found that increasing the intensity of DT in the forest can reduce the generalization error of RF. Santra et al. proposed an improved RF classifier, which is classified by the minimum number of trees and limited the number of DTs in RF according to the importance of features. Experiments on different data sets show that the classification error is significantly reduced [16].

2.2. Application of RF Algorithm. RF algorithm is widely used in many fields because of its good comprehensive performance. Asadi et al. applied the RF algorithm to environmental protection, used it to predict urban smog, and finally analyzed and expounded the control measures of smog [17]. Prasad et al. established the fund rating model by using RF algorithm and thought that the information ratio was the most important index of fund evaluation, followed by a determinable coefficient. The research proved that the stability and accuracy of the model reached an excellent level [18]. Kwan et al. introduced the nonparametric RF method into the field of fund excess return direction prediction in China, which proved that the RF method was superior to random walk and support vector machine algorithms, and also proved the predictability of the domestic financial market to a certain extent [19]. Wu et al. used the RF algorithm to conduct on-site monitoring and penetration prediction of plasma arc welding [20]. Mei et al. studied the identification of commuters based on RF of smart card data and compared it with the discriminant analysis method [21].

3. Methodology

3.1. Overview of RF Algorithm. On the basis of constructing bagging integration with DT-based learners in RF, the selection of random attributes is further introduced into the training process of DT. RF algorithm is simple, easy to implement, low in computational cost, and shows strong performance in many practical tasks. Therefore, the RF

RETRACTED

method is an extension of the traditional DT method, which combines multiple DTs to improve prediction accuracy.

DT is a typical single classifier. To use it for classification, we need to build a DT model based on training data and then use this model to classify unknown sample data. The pruning process is to prune some subtrees or leaf nodes in the DT model, and the main purpose is to avoid overfitting by simplifying the DT model. Firstly, according to the selected feature evaluation criteria, child nodes are recursively generated from the root node from top to bottom until the leaf node is reached.

In the process of DT node splitting, ID3 algorithm takes the information gain of features as the feature evaluation standard, the feature with the largest information gain as the test attribute, and the calculation of information gain is based on information entropy. Let X be a discrete random variable with finite values, and its probability distribution is as follows:

$$P(X = x_i) = p_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

Then the entropy of the random variable X is defined as follows:

$$H(X) = -\sum_{i=1}^{n} p_i \log\ p_i. \tag{2}$$

The generalization error of DT in all forests converges to the following expression:

$$\lim_{n \longrightarrow \infty} PE^* = P_{xy}\left(P_\Theta (k(X, \Theta) = Y) - \max_{j \neq Y} P_\Theta (k(X, \Theta) = J) < 0 \right), \tag{3}$$

where n is the number of trees in the forest.

The basic idea of RF classification: Firstly, k samples are extracted from the original training set by bootstrap sampling, and the sample capacity of each sample is the same as that of the original training set; Then, k DT models are established for k samples, and k classification results are obtained. Finally, according to the k classification results, vote on each record to determine its final classification. The schematic diagram is shown in Figure 1.

In the process of generating the RF algorithm, bagging sampling technology is mainly used to generate training subsets from the original training set. The size of each training subset is about two-thirds of that of the original training set, and each sampling is random and put back to sampling, which makes the samples in the training subset have certain duplication, and the purpose of this is to prevent DT in the forest from generating local optimal solutions.

Using Bagging to build RF has two meanings. On the one hand, it can improve the classification accuracy of the RF algorithm. Because the samples are put back, almost 37% of the samples are not in the training subset, which can prevent abnormal data and noise data from appearing in the training subset to a certain extent, and can get higher performance

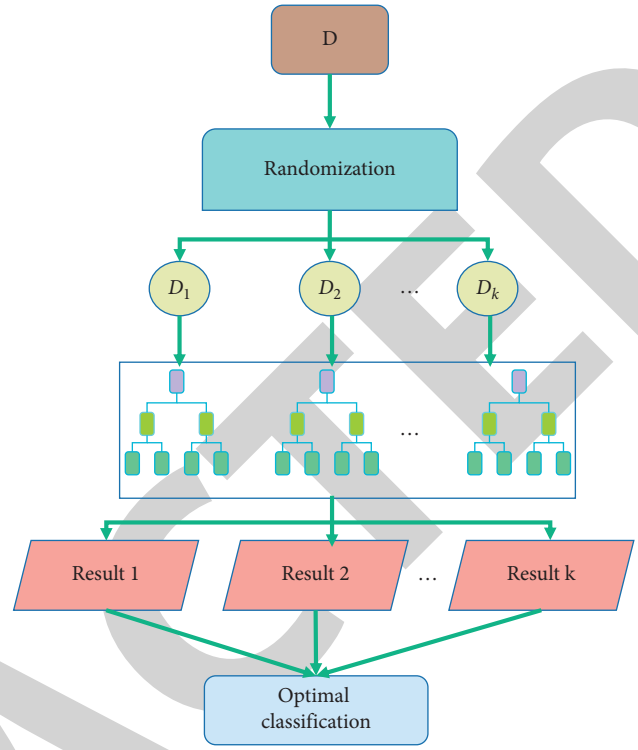

Figure 1: Schematic diagram of RF classification.

DT compared with the original data set. A random feature vector can reduce the correlation between trees in forest, thus reducing overfitting and improving the classification accuracy of forest, which introduces another random factor for RF.

In RF, if there are continuous variables, it is common practice to divide the values of these continuous variables into different intervals, that is, "discretization." However, due to the great relationship between the algorithm complexity after discretization and the reduction rate of the data set, it takes a lot of time to analyze and calculate the node splitting standard, which greatly affects the execution speed of the algorithm. Therefore, the discretization of continuous variables is content that needs to be optimized in the RF algorithm.

3.2. Optimization Method of Algorithm Based on. Generally speaking, high-dimensional data is relative to traditional low-dimensional data, and the number of attributes can often reach hundreds of thousands or even higher. Mining algorithm for high-dimensional data has always been a hot research topic, and the classification of high-dimensional data is a difficult pattern recognition problem. Traditional classification algorithms have problems such as low classification accuracy, easy overfitting, and long-running time when processing high-dimensional data. In this chapter, aiming at the problems of low classification accuracy and large generalization error of RF algorithm in high-dimensional data classification, an intelligent algorithm-based RF feature selection and parameter optimization for high-dimensional data is proposed.

RETRACTED

Feature selection can be described as a process in which an optimal feature subset $T' = \{t_{1'}, t_{2'}, \ldots t_{s'}\}$ is selected from the feature set $T = \{t_1, t_2, \ldots t_s\}$ and T' can contain most of the information of the original sample, among which $s' < s$, the purpose of feature selection is to make the classification or regression model constructed by feature subset T' achieve similar or even better prediction accuracy than before feature selection.

Therefore, according to the above two methods, we can study how to improve the treatment method of the class imbalance problem. The first scheme is to study the data distribution using the balanced RF method, which can be easily integrated into RF. Another approach is to apply the generation-sensitive algorithm to RF. This method can be accomplished using a weighted RF algorithm.

Smote (Synthetic Minority Over-sampling Technique) algorithm improves the random upsampling method. Because random upsampling is a random replication of negative samples, and many of the new data sets generated are duplicated, it is difficult to effectively solve the data imbalance problem.

SMOTE algorithm first looks for k nearest negative samples around each negative sample and then constructs a new negative sample between this sample and k adjacent samples. The process of interpolation synthesis is shown in the following formula:

$$P_{ij} = x_i + \text{rand}(0,1) \times \left(y_{ij} - x_i\right), \tag{4}$$

$x_i (i = 1, 2, \ldots, n)$ is negative samples, and n represents the number of negative samples; $y_{ij} (j = 1, 2, \ldots, m)$ is the m nearest neighbor samples adjacent to x_i; P_{ij} represents a new sample synthesized by sample x_i; $\text{rand}(0, 1)$ represents any random number between (0, 1).

However, SMOTE algorithm has two problems: First, when choosing the nearest neighbor, there is certain blindness about how much k value to take. How many nearest neighbor samples to take need to be solved by users themselves? Sometimes, we have to repeatedly test according to specific data sets, and we have to explore what kind of k value makes the algorithm optimal.

In this paper, a new discretization algorithm of continuous variables based on x^2 correction is designed, and its calculation process is as follows:

Calculate the number k of decision attributes in two adjacent intervals of a certain attribute value, and calculate the theoretical times E_{ij}, and the formula is as follows:

$$E_{ij} = R_i \times \frac{C_j}{N}. \tag{5}$$

In which: $R_i = \sum_{j=1}^{k} A_{ij}$ is the number of samples in the i interval; $C_j = \sum_{i=1}^{k} A_{ij}$ is the number of class j samples; N is the total number of samples in two adjacent intervals.

If the theoretical degree E_{ij} of a certain group is less than 5, it should be combined with its adjacent group or groups until the theoretical degree E is greater than 5 or there is only one set of data in an interval, and then the k value should be recalculated.

The value of the x^2 statistic is calculated in two cases: When $k < 2$, the formula of x^2 statistics is as follows:

$$x^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(\left|A_{ij} - E_{ij}\right| - 0.5\right)^2}{E_{ij}}, \quad j = 1; i = 1, 2. \tag{6}$$

When $k \geq 2$, the formula of x^2 statistics is as follows:

$$x^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}} \quad j = 1, 2, \ldots, k; i = 1, 2, \tag{7}$$

where k is the number of target variable categories; i number two adjacent intervals; A_{ij} is the number of class j samples in the i th interval in two adjacent intervals;

The order of merging intervals is determined by D value, and its formula is as follows:

$$D = \frac{x_a^2 - x^2}{\sqrt{2v}}. \tag{8}$$

Select the interval with the smallest D value to merge. After all the continuous attribute variables in the data set are subjected to the above steps, the reduction process of the data set is completed. That is, the discretization of the continuous variables is realized. The program flow of the continuous variable discretization algorithm based on x^2 correction is shown in Figure 2 below.

3.3. Application of Optimized RF Algorithm. Credit means that one party obtains something from the other party and promises to repay it in the future. However, in actual transactions, it is often impossible to ensure whether the debtor has sufficient repayment ability and willingness. Therefore, creditors can only evaluate the debtor's repayment ability and willingness in probability, and the uncertainty in this decision-making process will lead to credit risk.

The characteristics of credit risk mainly reflect four aspects:

(1) Credit risk is systematic. Systemic risk is the inherent risk of the financial market, which is closely related to macroeconomic changes.

(2) The formation of credit risk is related to people's subjective will. In credit transaction activities, when one party has obvious floating losses, in order to ensure personal economic interests, strategic default is often chosen.

(3) The profit and loss caused by credit risk is usually asymmetric. In the stock market, the rise and fall of the stock price can be considered symmetrical, so the profits and losses brought by it are also symmetrical in theory.

(4) The formation of credit risk is a cumulative process. Looking at the origin of credit risk events, most of them are related to the deteriorating historical business conditions. Local credit risk is transmitted to the whole financial market, causing a chain reaction and eventually even causing the disorder in the whole financial system.

RETRACTED

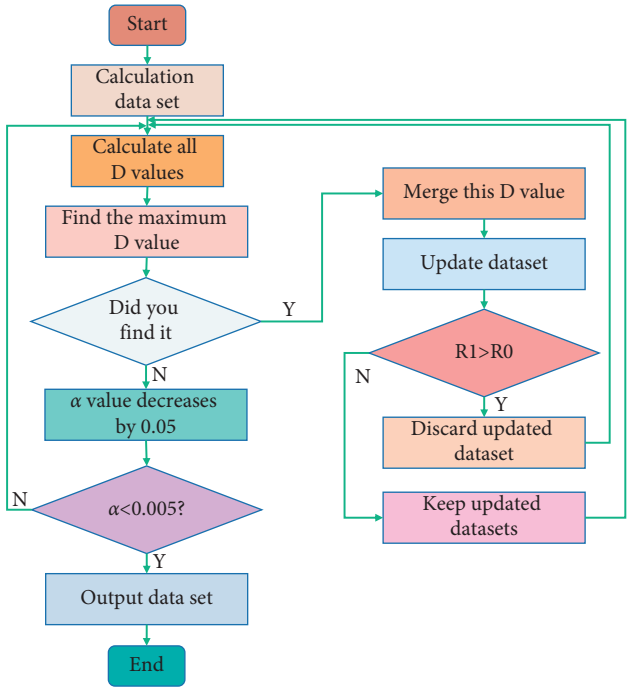

Figure 2: Algorithm program flow chart.

RF algorithm is an integrated algorithm. CART and Bagging methods are combined in the classifier, which is more adaptive than other data mining algorithms. As a nonlinear modeling tool, RF can well deal with classification problems with few rules constraints and missing data, and it is also effectively used in the credit approval of commercial banks. From the past research, the accuracy of RF algorithm is better than other algorithms such as neural network algorithm, which is also the reason for further research in this paper.

This chapter will conduct an empirical study with the manufacturing industry as the research background and corporate credit risk as to the research object. Choosing and constructing an evaluation index system is an important step of risk evaluation research. Choosing indicators to build a credit evaluation index system must be based on the purpose of evaluation, carefully analyze the things to be inspected, find out the factors that affect the evaluation objects, select some main factors from them, and build a credit evaluation index system. In this paper, 26 quantitative indicators including profitability, cash flow capacity, operational capacity, development capacity, short-term solvency, and long-term solvency and a qualitative indicator of management quality are selected.

This paper assumes that the financial statements published online by the eight companies involved in the research are consistent with those provided to credit rating companies and that the data are true and reliable. This paper introduces pseudo data. The pseudo data satisfies the following inequality conditions:

$$x(1-w) \le x' \le x(1+w). \tag{9}$$

In which: x is the original real data; w is the width of the random number; x' is false data.

Let DT have M leaf nodes, $R_m (m = 1, 2, \ldots, M)$ is the decision area under the m-th leaf node, and C_m (constant) is the decision value, which indicates the proportion of target classification under the decision area.

According to the process of DT discrimination, the new sample will eventually fall into one of the decision areas after top-down judgment, and the probability that the sample belongs to the target class is expressed by C_m. Thereby having the following leaf node discrimination function:

$$fc(X) = \sum_{m=1}^{M} C_m I(X, R_m). \tag{10}$$

In which, X is the input vector and $I(\cdot)$ is an illustrative function, which means that when the discrimination of X falls into the region R_m, the value is 1; otherwise, it is 0.

Choosing the feature-based credit risk evaluation model means choosing the best subset of indicators from the candidate evaluation index system as the evaluation index system of the model. In the traditional sense, feature selection refers to the self-correlation analysis between features to get the evaluation index system by removing the features with high linear correlation. Because the Wrapper algorithm has advantages over specific algorithms, this paper uses the Wrapper method for feature selection, and the specific process is shown in Figure 3:

The specific steps of selecting the optimal feature subset by RF are as follows:

(1) Randomly select a certain percentage of data from the sample data as training data.

(2) Use the package VarSeIRF in R language to calculate the importance of each evaluation index in training data and sort it.

(3) Carry out many experiments, compare the OOB error rate of each experiment, evaluate the influence of the number of indexes on the model performance, and take the index set with the lowest OOB error rate as the index set for feature selection.

When inputting model data, it is usually necessary to normalize the data. The advantage of normalization is that it can eliminate the dimensional influence of different data. When the data is not normalized, the larger evaluation index may weaken the influence of the smaller evaluation index on the model. The most commonly used normalization method is the maximum-minimum method. Generally, we classify the index as between [−1, 1] and [0, 1].

Because some issuers in private placement bonds did not disclose financial information, the corresponding indicators were missing, so these samples were excluded, and the sample distribution is shown in Table 1.

After SMOTE expansion of the data set, consider grouping the continuous variables to reduce the dimension. Among the 28 variables selected, there are only two values of x^2 (enterprise nature), so there is no need to group them again. For the other 27 continuous variables, most of the values range from positive infinity to negative infinity.

4. Experiment and Results

In order to study the parallel RF based on the MapReduce model, this experiment selects Ionosphere, Crowdsourced

RETRACTED

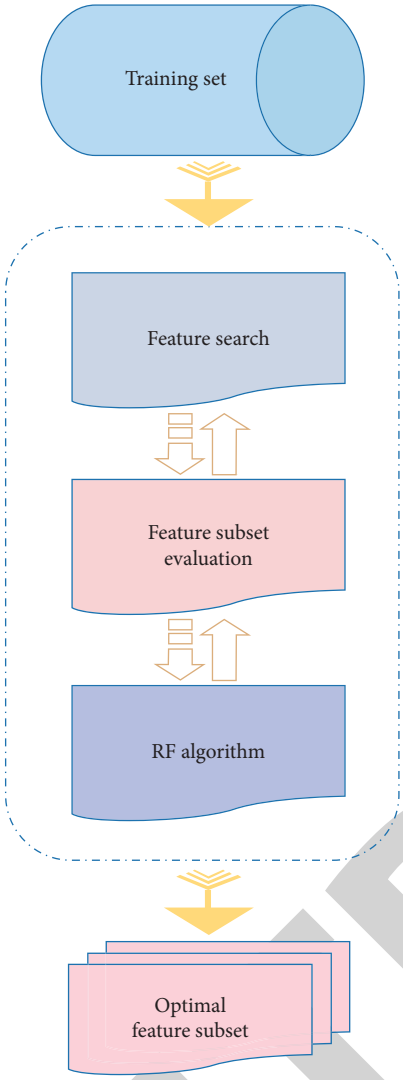

Figure 3: Feature selection process.

Table 1: Original sample distribution.

Credit event	Number of positive samples	Negative sample number	Total
Guarantor compensation	0	2	2
Guarantor's rating upgrade	147	0	147
Main body rating upgrade	10	0	10
Break a contract	0	22	22
Subject rating maintenance	206	0	206
Downgrading of subject	0	77	77

Mapping, KDDTest, and Covertype datasets in the UCI machine learning database, and the number and size of the datasets increase in turn. The data set description is shown in Table 2.

In the RF model, the number of DTs is uniformly set to 100, and the running time and acceleration of stand-alone RF and parallelized RF on Ionosphere, Crowdsourced Mapping, KDDTest, and Covertype data sets are as shown in Figures 4 and 5. "node" represents the number of data nodes.

Table 2: Data set description.

Serial number	Dataset name	Number of samples	Number of attributes	Data set size
1	Ionosphere	350	33	0.0771
2	Crowdsourced mapping	10553	29	1.12 M
3	KDDTest	125677	40	3.36 M
4	Covertype	590161	55	70.82 M

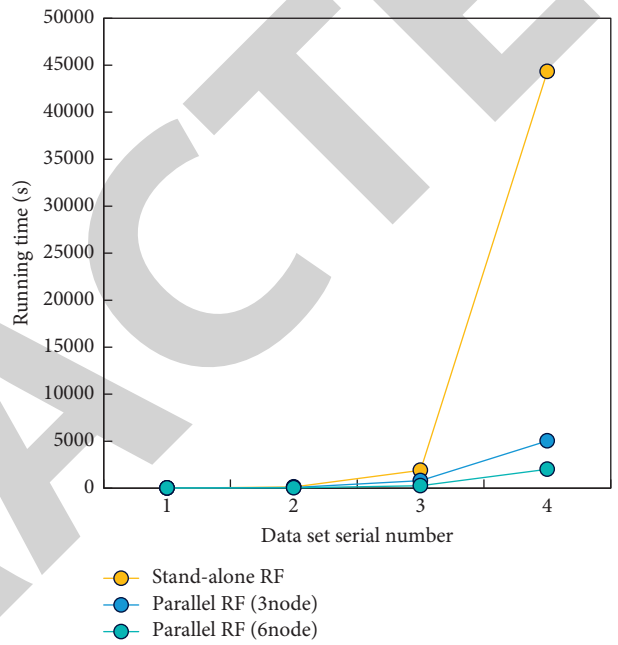

Figure 4: Run time comparison.

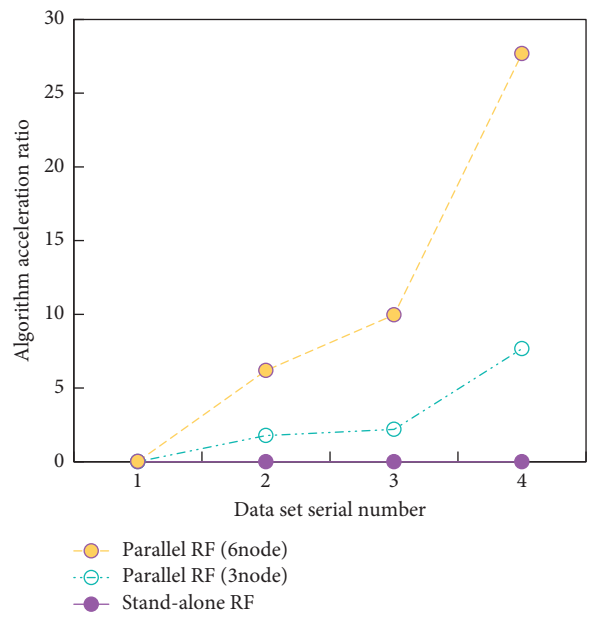

Figure 5: Comparison of algorithm speedup ratio.

RETRACTED

It can be seen that the speedup ratio of the parallel RF algorithm based on the MapReduce model in a large-scale data environment has increased on each data set, and the speedup ratio is larger with the larger data set, and the speedup ratio is larger with more nodes.

Comparing the running time of parallel RF with that of stand-alone RF, we can see that with the increase of data set size, the running time of stand-alone RF will increase rapidly, while that of parallel RF will increase very slowly, and with the increase of node number, the running time of RF algorithm will drop more significantly. This shows that the parallel RF algorithm can better adapt to large-scale and massive data, and the efficiency of the algorithm will be higher.

When RF algorithm selects features, the Gini index is used as the measurement standard, and the random forest contains many DTs. In the process of establishing DTs, it is usually necessary to calculate the corresponding information bribe of each feature, that is, the reduction of impurity. The information gain is to calculate the information bribe of the class label on the whole data set and then calculate the difference according to the information bribe of each feature.

As the feature is selected, the impurity will be greatly reduced, so the importance of other associated features will be greatly reduced, and it is difficult for other associated features to be selected again. Thus, the initially selected features are very important, while the other associated features are of low importance. This phenomenon reduces the importance of other features, but in fact, the importance of these features is similar.

When selecting features, on the one hand, select important features according to the Gini index; on the other hand, select features in a certain proportion in two-interval features, respectively, and do partial random, so as to balance the strength and correlation of features.

Here, we use the ratio of 3 : 7 to segment the data, randomly extract 90% of the features in the top 30% of importance, and randomly extract 60% of the features in the bottom 70% of importance so as to ensure that the extracted features account for about 70% of the total features and build a DT, as shown in Figure 6.

The time for RF modeling and prediction is relatively fast, and the time cost for evaluating features by chi-square test is also very small, so this method can ensure the running efficiency of the model.

The transformation of data sets has little influence on the importance of variables. Next, we will further explore the change in model performance before and after the expansion transformation. Set the original data to 7 : 3 according to the distribution ratio of the training set and test set and establish three models of RF, DT, and logit, respectively. After 100 simulations, the comparison results of the average accuracy of each model are shown in Table 3.

From the analysis in Table 3, it can be seen that the extended transformation of the data set is beneficial in improving the accuracy of the model. The reason is that the extended transformation of data balances the structure of sample data and makes the performance of the classifier more stable. At the same time, continuous variables are discretized to avoid the influence of extreme values to a certain extent, which makes the model more adaptable. Therefore, the extended transformation optimization measures for data sets improve the overall accuracy of each model.

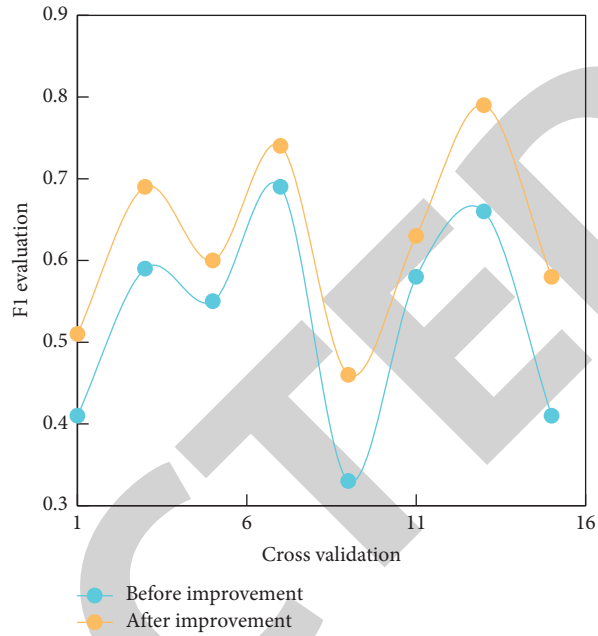

Figure 6: Comparison of results before and after feature improvement.

ROC curve takes FPR (false positive rate, the classifier mistakenly judges the actual normal as the number of default accounts for the actual total normal of the sample) as the abscissa axis and TPR (sensitivity, the classifier predicts the default accounts for the total number of actual defaults, that is, the coverage rate of default discrimination in the sample) as the ordinate axis. The ROC curve of each model is shown in Figure 7.

In the extended transformed data set, the sensitivity of the RF model has exceeded 0.8, and the sensitivity of other models has increased to about 0.65. This shows that under a certain risk tolerance, the accuracy of the model has also been greatly improved, which is particularly important in determining the appropriate risk preference.

In order to confirm the robustness and extrapolation of the RF model more reliably, this paper makes several experiments on the forecast set (the forecast set refers to the eight companies selected for forecasting in this paper) by using the RF model, CART model, and Logit regression model respectively, and the experimental results are shown in Figure 8:

From Figure 8, it can be seen that the prediction accuracy of the RF model for the company's credit rating in the prediction set reached 100%, while the CART model misjudged company C6 and C7, while the Logit model misjudged companies C3, C5, and C8. The experiment further proved the extrapolation of the RF model and the excellent prediction ability of the RF model.

RETRACTED

Table 3: Comparison of average accuracy of models under different data sets (%).

Contrast model	Training set			Test set		
	Positive sample accuracy	Negative class sample accuracy	Overall accuracy	Positive sample accuracy	Negative class sample accuracy	Overall accuracy
RF	95.36	90.01	93.36	91.96	86.54	88.17
CART	90.06	89.77	89.06	88.91	80.24	81.14
Logit	80.13	80.09	77.86	77.96	75.23	79.64
Optimize RF	97.63	95.52	96.17	92.33	91.69	94.98

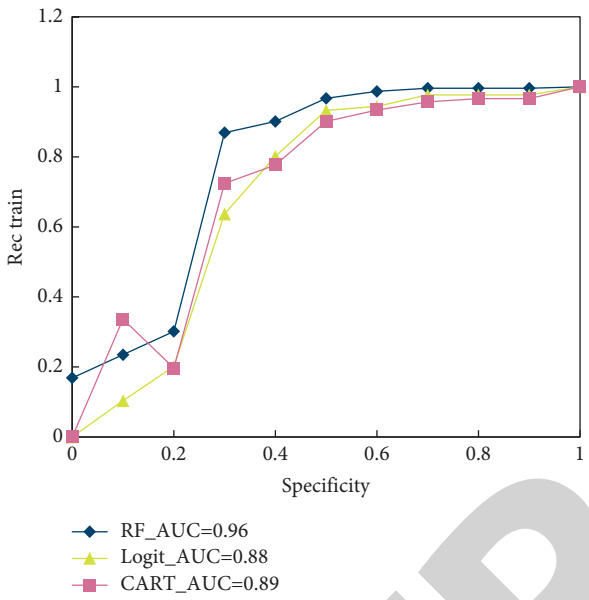

Figure 7: ROC curve based on extended transformation data set.

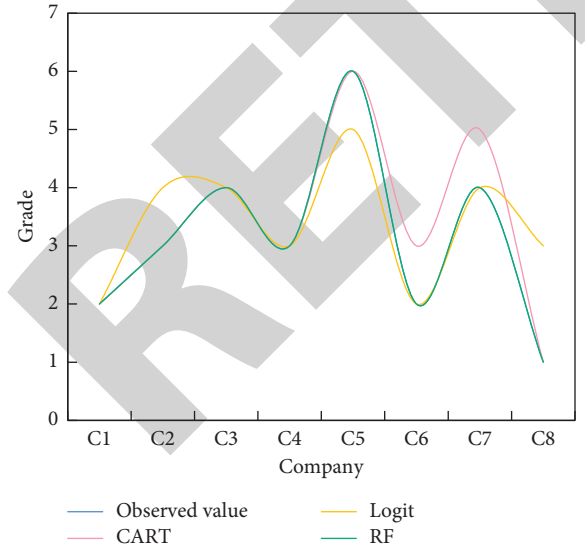

Figure 8: Comparative analysis of prediction ability of models.

5. Conclusions

RF algorithm is an algorithm with high classification accuracy and high efficiency, and its theory and method research have been mature, and it has been applied in many fields with good results. In the aspect of the optimization of unbalanced data sets, this paper proposes a new algorithm to solve the unbalanced problem, which better solves the unbalanced problem of data sets and significantly improves the classification performance of the RF algorithm on unbalanced data sets. The experimental results show that the parallel RF algorithm has greatly improved the processing ability of large-scale data. CART and Logit are used as experimental reference models, which prove that the stability, extrapolation, and prediction ability of the RF model are far superior to the experimental reference model, and there is no overfitting phenomenon in the experimental process. It is proved that RF has better performance when it is used to build the credit risk evaluation model of listed companies in the production industry.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The author declares no competing interests.

References

[1] M. Kang, S. K. Gonugondla, S. Lim, and N. R. Shanbhag, "A 19.4-nJ/Decision, 364-K decisions/s, in-memory random forest multi-class inference accelerator," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 7, pp. 2126–2135, 2018.

[2] S. Ma, M. Chen, J. Wu, Y. Wang, B. Jia, and Y. Jiang, "Intelligent fault diagnosis of HVCB with feature space optimization-based random forest," *Sensors*, vol. 18, no. 4, p. 1221, 2018.

[3] Z. Tang, Z. Mei, W. Liu, and Y. Xia, "Identification of the key factors affecting Chinese carbon intensity and their historical trends using random forest algorithm," *Journal of Geographical Sciences*, vol. 30, no. 5, pp. 743–756, 2020.

[4] T. A. Rather, S. Kumar, and J. A. Khan, "Multi-scale habitat modelling and predicting change in the distribution of tiger and leopard using random forest algorithm," *Scientific Reports*, vol. 10, no. 1, Article ID 11473, 2020.

[5] Z. Bei, Z. Yu, N. Luo, C. Jiang, C. Xu, and S. Feng, "Configuring in-memory cluster computing using random forest," *Future Generation Computer Systems*, vol. 79, pp. 1–15, 2018.

[6] D. Vassallo, R. Krishnamurthy, T. Sherman, and H. J. S. Fernando, "Analysis of random forest modeling strategies for multi-step wind speed forecasting," *Energies*, vol. 13, no. 20, p. 5488, 2020.

RETRACTED

[7] X. Cao, R. Li, Y. Ge, B. Wu, and L. Jiao, "Densely connected deep random forest for hyperspectral imagery classification," *International Journal of Remote Sensing*, vol. 40, no. 9, pp. 3606–3622, 2019.

[8] J. Gao, D. Nuyttens, P. Lootens, Y. He, and J. G. Pieters, "Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery," *Biosystems Engineering*, vol. 170, pp. 39–50, 2018.

[9] R. Xu and F. Luo, "Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest," *Safety Science*, vol. 135, no. 24, Article ID 105125, 2021.

[10] J. Hao, C. Zhu, and X. Guo, "Wind power short-term forecasting model based on the hierarchical output power and Poisson Re-sampling random forest algorithm," *IEEE Access*, vol. 9, pp. 6478–6487, 2021.

[11] S. Pasinetti, A. Fornaser, M. Lancini, M. De Cecco, and G. Sansoni, "Assisted gait phase estimation through an embedded depth camera using modified random forest algorithm classification," *IEEE Sensors Journal*, vol. 20, no. 6, pp. 3343–3355, 2020.

[12] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved random forest for classification," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4012–4024, 2018.

[13] X. Li, "Random forest is a specific algorithm, not omnipotent for all datasets," *Journal of Applied Entomology*, vol. 50, no. 4, pp. 170–179, 2019.

[14] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.

[15] C. Dou, S. Teng, T. Zhang, B. Zhang, and K. Ma, "Layered management and hybrid control strategy based on hybrid automata and random forest for microgrid," *IET Renewable Power Generation*, vol. 13, no. 16, pp. 3113–3123, 2019.

[16] B. Santra, A. Paul, and D. P. Mukherjee, "Deterministic dropout for deep neural networks using composite random forest," *Pattern Recognition Letters*, vol. 131, pp. 205–212, 2020.

[17] S. Asadi, S. Roshan, and M. W. Kattan, "Random forest swarm optimization-based for heart diseases diagnosis," *Journal of Biomedical Informatics*, vol. 115, no. 24, Article ID 103690, 2021.

[18] R. Prasad, R. C. Deo, Y. Li, and T. Maraseni, "Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach," *Catena*, vol. 177, pp. 149–166, 2019.

[19] R. Prasad, M. Ali, P. Kwan, and H. Khan, "Designing a multistage multivariate empirical mode decomposition coupled with ant colony optimization and random forest model to forecast monthly solar radiation," *Applied Energy*, vol. 236, pp. 778–792, 2019.

[20] D. Wu, M. Hu, Y. Huang, P. Zhang, and Z. Yu, "In situ monitoring and penetration prediction of plasma arc welding based on welder intelligence-enhanced deep random forest fusion," *Journal of Manufacturing Processes*, vol. 66, no. 9, pp. 153–165, 2021.

[21] Z. Mei, W. Ding, C. Feng, and L. Shen, "Identifying commuters based on random forest of smartcard data," *IET Intelligent Transport Systems*, vol. 14, no. 4, pp. 207–212, 2020.

RETRACTED