Prediction Method of Enterprise Return on Net Assets Based on Improved Random Forest Algorithm

Yongsong Cai¹,a, Qi Yin²,b, Qian Su³,c, Xinyu Huang⁴,d, Yin Zhang⁵,e, Ting Liu⁶,f

¹UK Zhongshan business association, Zhongshan, China
²Harbin Institute of Technology, Harbin, China
³Central South University of Forestry and Technology, Changsha, China
⁴Shanghai Lixin University of Accounting and Finance, Shanghai, China
⁵Inner Mongolia University of Technology, Huhhot, China
⁶Sichuan University, Chengdu, China

a*28136347@qq.com, b1171510122@stu.hit.edu.cn, c1061411301@qq.com,
d2963534249@qq.com, e2504084916@qq.com, f*1186153724@qq.com

Abstract—In order to accurately predict the change trend of enterprise assets, this paper proposes a prediction method of enterprise return on net assets based on improved random forest algorithm. Firstly, this paper analyzes the research background and significance of the enterprise net income rate prediction, and on this basis, it considers to use the random forest algorithm to achieve the prediction. In view of the shortcomings of the algorithm, the paper introduces the simulated annealing algorithm to improve the random forest algorithm and optimizes the prediction performance of the algorithm in the direction of feature selection, parameter optimization and weight setting, finally establishes the prediction model of enterprise return on net assets. In this paper, the proposed algorithm is compared with other algorithms and results show that the method proposed in this paper has a good prediction effect on the enterprise return on net assets, which is helpful for the capital market to evaluate the asset value and profitability of enterprises.

1. INTRODUCTION

With the rapid development of China's economy, the influence of China's stock market in the international market is growing. At the same time, the number of investors in China's stock market is growing day by day. The stock market has become an important part of China's economic development, and also an important way for the national and investment institutions to invest. The equilibrium of capital asset pricing is the research foundation of modern finance. In recent decades, with the increasing diversification of investment portfolio, the improvement of market efficiency and information transparency, investors have paid more and more attention to the expected return rate of stock assets. For enterprises, how to let more investors invest in the company, that is, how to obtain the maximum profit has become the most important issue for entrepreneurs. A more intuitive way to judge the profitability of an enterprise is through the net rate of return, which is an important consideration of the capital market for enterprise valuation, so it is an important content of enterprise investment value fundamental analysis. Enterprise net return rate is an important index to judge the value of enterprise
stock investment. Therefore, the prediction of enterprise net return rate has become a hot spot in modern finance research [1-2].

2. TRADITIONAL RANDOM FOREST ALGORITHM

Random forest algorithm is a combination of Bagging algorithm, CART algorithm and other algorithms. It mainly uses the decision tree as the basic unit to carry out multi-style combination, and then obtains the accurate prediction of the samples of unknown categories by voting method. In summary, the random forest algorithm includes two random processes: the first is to generate the training set of each decision tree by using Bagging algorithm; the second is to select feature subset randomly and determine the best prediction feature by information gain and other indicators [3-4].

The Bagging algorithm and CART algorithm in the random forest algorithm are described in detail, and then the algorithm can be improved on this basis, so as to improve the prediction accuracy of the algorithm.

2.1 Bagging Algorithm

The main idea of Bagging algorithm is that any prediction calculation process does not depend on any previous prediction calculation process. Each prediction calculation process randomly extracts part of the data set from the original data set for independent operation, and effectively aggregates the calculation results of each classifier, so as to improve the prediction accuracy of the algorithm, and avoid over fitting and generalization Stable learning algorithm.

The calculation process of Bagging algorithm is as follows:

(1) A new training data subset is formed by random and repeatable n-times sampling from data set D;
(2) Repeat the above process m times to get m independent training data subsets;
(3) The prediction and regression algorithms are run on m training data subsets to get m results;
(4) For m results, weighted calculation, arithmetic average and majority voting method are used for final calculation to get the overall output results.

In the process of random forest algorithm implementation, Bagging method can effectively improve the effectiveness of different decision tree data sets, enhance the generalization ability of data set selection, and improve the prediction performance of the algorithm from the side; at the same time, it can effectively reduce the over fitting of training subsets of different decision trees.

2.2 CART Algorithm

The sub model of stochastic forest algorithm is CART model, so it is necessary to understand CART model before describing random forest algorithm in detail. CART algorithm mainly selects the attribute with Gini index as the splitting attribute when the decision tree node splits, and finally generates a binary decision tree. CART algorithm uses binary recursive mode to construct binary tree. Each partition will divide the whole sample set into two subsets, thus generating two branch subtrees. Gini index is mainly used to represent the impure data set. In the process of CART algorithm implementation, we need to select the attribute with the minimum Gini index as the splitting attribute. CART algorithm recursively divides the data set to be classified by dichotomy, until each leaf node of the prediction tree is the same kind of data set. When the eigenvalues of the data set are continuous values, the result is a regression tree.

2.3 Flow of Random Forest Algorithm

Combined with the above algorithms, we will detail the specific work flow of random forest algorithm. The specific work flow of random forest algorithm is as follows:

(1) Generating the training subset. Using Bagging algorithm, the original training set is randomly sampled and the training subset is obtained. At the same time, the basic attributes are selected and analyzed.
(2) For each training subset, cart algorithm is used to generate an uncut binary recursive decision tree. CART tree needs to continue to split nodes until a specific termination condition is reached. Finally, the corresponding decision tree is generated for each training subset.

(3) The decision trees generated in the previous step are combined to form a random forest, and the corresponding prediction results are obtained based on the test set samples.

(4) For the prediction results of decision tree, the majority voting method and weighted voting method are used to determine the final prediction results for the test data, and the prediction error needs to be calculated.

Out of bag error (OOB) is defined as: in the process of generating each tree, we get training subsets based on Bagging algorithm for classifier training. According to probability statistics, only 63% of the data is repeatedly extracted, while other data is never extracted. This part of data that has never been extracted is called OOB. In the actual implementation of random forest algorithm, OOB can be used to check the quality of each decision tree, and the corresponding calculation error is OOB error. OOB Error calculation is a kind of unbiased estimation, which can replace the cross validation method of data sets, and the actual experience shows that the calculation value is usually quite accurate, so it is usually used as the evaluation index of random forest algorithm to evaluate the prediction performance of random forest algorithm. The smaller the OOB error value, the better the classification performance of the random forest algorithm. OOB error usually uses the average OOB error of the decision trees as the OOB error estimate of the random forest algorithm.

3. IMPROVEMENT OF ALGORITHM
Compared with other prediction algorithms, the original random forest algorithm has achieved better results, but it still has several problems: first, the default parameters are not the optimal selection of multiple parameters, and the traditional grid search method is too time-consuming to make optimal parameter decision, especially the weighted weight is difficult to be applied in practice. Secondly, the prediction error of the algorithm is relatively large, which needs to be further reduced. Thirdly, the number of feature selection in the algorithm is usually preset and used as the objective function of the algorithm, which may be far from the accurate value. Therefore, this paper proposes an improved random forest algorithm to correct the above problems. In the subsequent improvement process, the simulated annealing algorithm is integrated into the random forest algorithm, and the algorithm is improved from the aspects of feature selection, parameter optimization and weight setting.

3.1 Simulated Annealing Algorithm
The main goal of simulated annealing algorithm is to solve the problem of local minimum in the optimization process and overcome the dependence of initial value. Simulated annealing algorithm imitates the physical process of annealing. Starting from a certain initial temperature, with the continuous decline of temperature, combined with the characteristics of probability jump, the global optimal solution is randomly found in the solution space. The reason why simulated annealing algorithm can be applied in optimization problems is mainly due to the high similarity between physical annealing process and combinatorial optimization problem [5-6].

The basic flow of simulated annealing algorithm is shown as follows, where k is the iteration step, \( d(x) \) is the temperature cooling update formula, and \( E(x) \) is the energy function. From the flow of the algorithm, we can see that the algorithm includes two-layer cycle, the inner loop represents the state random search under the same temperature, and the outer loop is the change when the temperature drops and the stop condition of the algorithm.

Step 1: Randomly generating an initial solution \( x_0 \), let \( x_{best} = x_0 \), and calculating the objective function value \( E(x_0) \), and setting \( k = 0 \), initial temperature \( t_0 = T_{max} \).

Step 2: If the internal circulation stop condition is reached within this temperature, jump to step 3, otherwise, the current \( x_{next} \) is selected from the neighborhood \( N(x_{best}) \) of optimal solution \( x_{best} \), and
the correlation objective function value $E(x_{best})$ is calculated. From this, we can calculate the increment value $\Delta E = E(x_{new}) - E(x_{best})$. Updating $x_{best}$ as follows:

$$x_{best} = \begin{cases} x_{new}, & \Delta E < 0 \\ x_{new}, & \Delta E > 0 \text{ and } e^{-\Delta E} > \text{random}(0,1) \\ x_{best}, & \text{else} \end{cases}$$

(1)

Step 3: Setting $t_{k+1} = d(t_k), k = k + 1$. If the final stop condition is met, terminate the calculation and output optimal solution $x_{best}$, otherwise return to step 2 to continue execution.

3.2 Improvement Idea

The improved random forest algorithm proposed in this paper is mainly used for feature selection, weight modification and parameter optimization to realize the prediction. The goal of the algorithm is to remove useless features and obtain efficient feature subset; analyze the relationship between tree size parameter $K$, super parameter $N$ and algorithm performance, and analyze the setting of different decision tree weight values, and finally find the optimal parameter combination configuration.

In the process of introducing simulated annealing algorithm to improve random forest, we mainly improve the algorithm from the aspects of feature selection, parameter optimization and weight correction. Feature selection is to select a part of feature attributes from all attribute sets of data set to form feature subset and represent the whole data set. From the historical experience, feature selection can effectively reduce the correlation between different trees in random forest, and improve the overall effect of the algorithm.

There are two important parameters in the efficiency of random forest algorithm: the size of decision tree $K$ and the size of attribute feature subset $O$. $K$ represents the number of decision trees, and the higher the value, the more classifiers. The more classifiers, the better the classification performance. The disadvantage is that the higher the time and space cost, it will also affect the interpretability of the algorithm. If the K value is small, the performance will be worse, and the accuracy of the corresponding prediction results will be lower. The size of attribute feature subset $O$ represents the size of attribute number of feature subset, and represents the number of attributes of randomly selected data scale without replacement, which is usually far less than the total number of attributes in the data set, and keeps a fixed value during the execution of random forest algorithm. The main meaning of this parameter is to increase the diversity of different decision trees, enhance the independence of different decision trees, and reduce the correlation between different trees. If the $O$ value is too high, it means that almost all the attribute features in the data are included in the attribute feature set to split the tree nodes, which will lead to the high similarity of different decision trees and affect the final prediction effect. If the $O$ value is too small, it will lead to many important features can not be adopted as a basis for decision-making, also will reduce the prediction accuracy. An appropriate parameter $o$ can directly improve the overall performance of the algorithm, which has a great impact on the performance of the algorithm.

In the stage of decision tree result statistics, it is necessary to carry out weighted statistics for decision tree results, and the weighted weight of each decision tree is $W$. The traditional random forest algorithm will deal with each classification tree with a weight of 1. It is difficult to reflect the difference between strong classification tree and weak classification tree by statistics. Therefore, the concept of weighted weight is introduced here to strengthen the classification effect of strong classifier and reduce the classification effect of weak classification tree.
3.3 Improved Algorithm Flow

In order to realize feature selection, parameter optimization and weight adjustment, binary coding, OOB error minimization and simulated annealing are used to obtain the optimal combination values of feature selection $O$, the size of decision tree $K$, the size of feature subset $N$ and decision tree weight $W$.

In order to achieve the above objectives, we set the objective function of the improved random forest algorithm as follows:

$$
\begin{align*}
\text{argmin}(\text{avg OOB error}) &= f(K', O', \{\text{Attribute}_i \mid i = 1, 2, \cdots, M\}, \{W_j \mid j = 1, 2, \cdots, K\}) \\
K', O', \{\text{Attribute}_i \mid i = 1, 2, \cdots, M\}, \{W_j \mid j = 1, 2, \cdots, K\} &= \text{random} \quad \text{(2)}
\end{align*}
$$

Optimization variables: $K$, $O$, $\{\text{Attribute}_i \mid i = 1, 2, \cdots, M\}$, $\{W_j \mid j = 1, 2, \cdots, K\}$, where $K$, $O$ are real numbers and the value range of $K$ is $[0, 500]$, $W_j$ is an integer with a value range of $[0, 15]$. The value of Attributes is 0 or 1, where 0 means that the feature is not selected, and 1 means that the feature has been selected. The optimization variable can be expressed in binary code, and is mainly composed of four segments, as shown in Table 1. The number of coding digits is fixed, a total of 22 digits.

| Optimization variables | Binary code |
|------------------------|-------------|
| $K$                    | 1 1 0 0     |
| $O$                    | 0 1 1 0     |
| $W$                    | 1 0 0 0 1 1 0 1 |
| Attribute              | 0 1 1 0 0 1 |

Based on the binary code of the above variables, in order to obtain the optimal solution of the above objective function, the improved random forest algorithm flow is as follows:

Step 1: Setting the initial temperature $t = t_{\text{max}}$, the number of iterations $\text{maxgen}$, $k=1$, randomly generate an initial solution $X_0$, and $X_{\text{best}} = X_0$.

Step 2: Combining the random forest classifier to calculate $E = \text{max}(1/f)$.

Step 3:

① If the internal circulation stop condition is reached within this temperature, jump to ②, otherwise, the current $X_{\text{new}}$ is selected from the neighborhood $N(X_{\text{best}})$ of optimal solution $X_{\text{best}}$, and the correlation objective function value $E(X_{\text{new}})$ is calculated. From this, we can calculate the increment value $\Delta E = E(X_{\text{new}}) - E(X_{\text{best}})$. Updating $X_{\text{best}}$ as follows:

$$
X_{\text{best}} = \begin{cases} 
X_{\text{new}}, & \Delta E < 0 \\
X_{\text{best}}, & \Delta E > 0 \text{ and } e^{-\frac{\Delta E}{t_k}} > \text{random}(0,1) \\
X_{\text{best}}, & \text{else}
\end{cases}
$$

② Setting $t_{k+1} = d(t_k)$, $k = k + 1$. If the final stop condition is met, terminate the calculation and output optimal solution $X_{\text{best}}$, otherwise return to step ① to continue execution.

Step 4: If gen$>\text{maxgen}$, output the best $X^* = (K, O, \text{Attribute}, W)$, otherwise return to step 2.

4. Empirical Analysis

4.1 Data Selection and Explanation

Randomly select data samples of 800 listed companies from 2012 to 2013, of which the number of training samples is 770 and the number of test samples is 30. Select ATO, PM, LEV, GROWTH, PB, ARR, ASSET, ROEt as independent variables, ROEn (the company’s return on net assets for the next
year) as the dependent variable for analysis, where ATO represents the company’s asset turnover rate. This financial indicator is used to comprehensively evaluate the utilization efficiency of all assets of the company; PM stands for profit rate, reflecting the company’s basic debt status; LEV stands for debt-to-capital ratio, reflecting the company’s basic debt status; GROWTH represents the growth rate of the company, which is used to reflect the company PB stands for price ratio, which reflects the expected future growth rate of the company; ARR stands for income quality, which reflects the company’s unrealized main business income that year, and explains the company’s earnings quality to a certain extent; ASSET stands for asset scale, also It is called the total assets and reflects the company’s size; ROEt represents the company’s return on net assets in the year, and it directly reflects the company's profitability that year.

4.2 Experimental Process
This paper uses Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and the method proposed in this paper (IRF) to compare and analyze. The first 770 data are used as training samples, and the last 30 data are used as test samples. The data is normalized first, matrix analysis is performed, and the mean squared error (MSE) and operation time (OT) are calculated respectively, and the comparison results is shown in Table 2.

| Algorithm | MSE  | OT   |
|-----------|------|------|
| DT        | 0.062| 5.41s|
| RF        | 0.014| 8.46s|
| SVM       | 0.091| 4.86s|
| IRF       | 0.003| 2.34s|

From Table 2, it can be found that the prediction performance of random forest is improved compared with decision tree and support vector machine, and the prediction effect is better. However, due to the low efficiency of its parameter selection process, the calculation time is long. The method proposed in this paper solves this problem well, and further improves the prediction accuracy, which has good practical application effect.

5. CONCLUSION
In this paper, the traditional random forest algorithm is briefly introduced, and the core algorithm Bagging and cart algorithm and the algorithm flow are described in detail. In the subsequent improvement process, the simulated annealing algorithm is integrated into the stochastic forest algorithm, and the algorithm is improved from the aspects of feature selection, parameter optimization and weight correction, forming a prediction method of enterprise return on net assets. The results of empirical analysis show that the method proposed in this paper has high precision and high efficiency, and has great advantages compared with other algorithms.

ACKNOWLEDGMENT
This work was supported by UK Zhongshan business association

REFERENCES
[1] H. Yuzhu, L. Zengxin and H. Zaiqiang, "Financial Distress Prediction Model of Small and Medium-sized Listed Companies," 2011 International Conference on Information Management, Innovation Management and Industrial Engineering, Shenzhen, 2011, pp. 189-192.
[2] H. Lan and Y. Pan, "A Crowdsourcing Quality Prediction Model Based on Random Forests," 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, 2019, pp. 315-319.
[3] T. Lu, Y. Huang, W. Zhao and J. Zhang, "The Metering Automation System based Intrusion Detection Using Random Forest Classifier with SMOTE+ENN," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 2019, pp. 370-374.

[4] H. Yi, Q. Xiong, Q. Zou, R. Xu, K. Wang and M. Gao, "A Novel Random Forest and its Application on Classification of Air Quality," 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), Toyama, Japan, 2019, pp. 35-38.

[5] G. Wang, C. Xu and G. Liu, "The transient electromagnetic inversion based on the simplex-simulated annealing algorithm," 2018 37th Chinese Control Conference (CCC), Wuhan, 2018, pp. 4321-4324.

[6] L. Xiuqin, H. Jinfeng, X. Yongpeng, Y. Ye, Z. Yiming and C. Xiaoxin, "A wavelet de-nosing algorithm of XLPE cable partial discharge signals based on chaotic simulated annealing," 2015 IEEE 5th International Conference on Electronics Information and Emergency Communication, Beijing, 2015, pp. 333-336.