# Application of Parallel Random Forest in Doubt Prediction of Audit Big Data

Jing Xiao, Yi Du*

(Mianyang Teachers' College, College of Economics & Management, Sichuan, Mianyang, 621016)

smartceci@163.com

**Abstract.** In order to make a scientific audit plan under the background of auditing big data, an audit doubt prediction algorithm based on parallel improved random forest algorithm is proposed. The goal of this algorithm is to improve the efficiency of the algorithm while ensuring the prediction accuracy, so as to meet the needs of big data processing. Firstly, a Hadoop-based big data management scheme for power enterprise audit is established. This scheme can integrate and store heterogeneous audit data from various power grid subsystems. On this basis, a parallel random forest algorithm based on three-layer MapReduce is implemented to predict the probability of audit doubts. Random forest is an integrated learning algorithm, which predicts by constructing multiple decision trees and combining them. In the parallel improvement, a three-layer MapReduce architecture is adopted, which divides the data into multiple sub-data sets and constructs multiple decision trees by parallel computing. Finally, the results of these decision trees are combined to get the final probability prediction result of audit doubt. The experimental results show that the proposed parallel algorithm meets the requirements of big data processing in running speed. At the same time, this algorithm is superior to the commonly used decision tree algorithm in prediction accuracy. This means that the algorithm can not only process a large number of audit data efficiently, but also provide accurate prediction results of audit doubts. Such an algorithm will provide scientific data support for the formulation of audit plans and meet the needs of big data processing.

**Keywords:** Random forest algorithm; Prediction of audit doubt; Hadoop

## 1 Introduction

With the arrival of the "internet plus" era, the scale of data is increasing, the types are becoming more and more diversified, and the requirements for data processing speed are also increasing. Big data has the characteristics of 5V, such as low data value density and data authenticity, which means that big data contains a lot of valuable information, but at the same time, there are challenges that it is difficult to process and verify the data. In electric power enterprises, with the improvement of information deployment and the online operation of business systems and information systems,


Y. Jiao et al. (eds.), *Proceedings of the 3rd International Conference on Internet Finance and Digital Economy (ICIFDE 2023)*, Atlantis Highlights in Economics, Business and Management 1,
https://doi.org/10.2991/978-94-6463-270-5_42

more electronic data are generated than before. In particular, the audit data shows a trend of massive growth, and the storage scale has increased from the previous GB level to TB or even PB level, forming a huge audit database. These audit big data contain various types of data, including structured, semi-structured and unstructured data. In the face of these massive and rapidly growing audit big data, how to store, manage and analyze these data reliably and efficiently, so as to give full play to the advantages of big data in finding data evidence, has become an important research topic at present [1].

In order to solve this problem, based on the existing big data related technologies, the researchers put forward a management scheme of auditing big data for power enterprises, that is, using Hadoop cluster to build a data storage platform to integrate the data generated and stored by various power enterprises in different subsystems. This platform can help power enterprises to store big data reliably and manage and analyze these data effectively. On this basis, in order to support the formulation of the audit plan, the researcher proposed a method to predict the probability of audit suspects in audit big data based on parallel improved random forest. This method takes the audit doubt probability as the target variable, and analyzes and predicts the audit big data by improving the random forest algorithm in parallel. Through experimental verification, the researchers used large data sets of different scales to verify the algorithm, and the results confirmed the effectiveness of the algorithm. To sum up, building a data storage platform through Hadoop cluster and applying the method based on parallel improved random forest to predict the probability of audit doubt can effectively manage and analyze the audit big data of power enterprises and provide data support for the formulation of audit plans. The implementation of this method can make full use of the advantages of big data in finding data evidence and promote the efficiency and accuracy of the audit work of power enterprises [2].

## 2 Research methods

### 2.1 Hadoop-based analysis platform construction

Aiming at the massive and heterogeneous audit big data generated by different subsystems of power enterprises, how to construct a model expressed in the same specification to realize data integration is an urgent problem to be solved. Therefore, before the prediction of big data audit doubt, on the basis of cloud computing, combined with the actual needs of audit big data management and analysis, Hadoop is used to build the audit big data management platform as shown in Figure 1. The platform consists of application layer, cloud computing data processing layer and management layer. The storage system is established by HDFS, HBase and Hive, and the analysis and processing of big data is completed by MapReduce and Spark parallel computing framework [3].

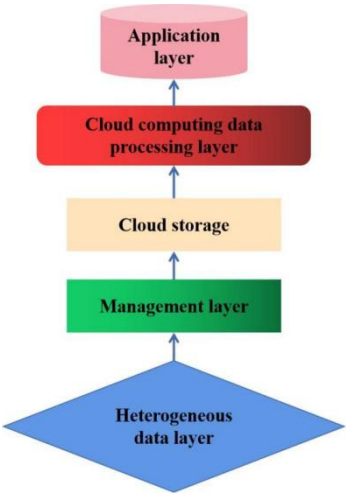


**Fig. 1.** Extraction and Integration of Audit Big Data

### 2.2 Prediction of Doubts in Parallel Random Forest Audit

The prediction of audit doubt provides scientific and effective data support for the formulation of audit plan. At present, decision tree is widely used in traditional prediction and has achieved good research results. The algorithm is composed of a series of classified regression trees with independent sample training sets TS, in which the independent training sample sets are extracted according to Bagging algorithm and have the same size as the total sample sets, while the internal nodes are composed of random attribute subsets selected by Ho theory, thus forming a tree group, and the final result is obtained by voting or averaging the results of each tree. Single-tree construction mainly branches according to attribute subsets, and then repeats the division process on subtrees until the growth stop condition is met. Gini index measures the impurity of nodes, which is suitable for the construction of classification tree in the algorithm, and its calculation formula is:

$$G(t) = 1 - \sum_j p^2(j/t) \tag{1}$$

Where: t is the node attribute, and p( j/t) represents the proportion of j-class targets at the current node. The least square deviation is suitable for the construction of regression tree in the algorithm[4-5]. After fitting the node t, the error is expressed as:

$$\mathrm{E}(t) = \frac{1}{n_t} \sum (y_i - k_t)^2 \tag{2}$$

Where nt is the number of node data instances, kt is the average value of the target value $k_t = \left(\sum y_i\right)/n_t$ of the instance data, and the least square deviation standard of node t is to maximize Equation (3).

$$E(s,t) = \frac{S_L^2}{n_{tL}} + \frac{S_R^2}{n_{tR}} \tag{3}$$

Where $S_L = \sum_{D_{iL}} y_i, S_R = \sum_{D_{iR}} y_i$ is the attribute value.

### 2.3 Parallel Random Forest Audit Doubt Prediction

The idea of ensemble learning is the basis for the parallel processing of random forest algorithm, but it is not just a simple combination of k tree models. Bagging has put back samples, which makes the data difference in the training sample set about 37%, and the construction process is independent, which ensures the parallelization of the training process data, thus improving the speed of model generation. However, the random subspace method avoids the reading and over-fitting of all attributes when testing the node attributes. Based on this, this paper proposes a parallel improved random forest algorithm based on MapReduce to predict the audit doubts of big data. In the training stage, the improved algorithm is completed by three MapReduce job classes in turn[6]. In the training stage, data dictionary generation, decision tree generation and random forest formation are completed in turn by three MapReduce job classes, and the trained model is stored in Hadoop distributed cluster. The generation of data dictionary is to describe the conditions and decision attributes of training samples, and form a description file to record the types of conditions and decision attributes and the final model attributes. The data dictionary generation process is completed by the first MapReduce, and one of its Map processes completes the reading of part of the data and the generation of the description file, which is saved in HDFS in the key/value data mode in Hadoop for subsequent MapReduce calls [7].

## 3 Result analysis

In order to verify the effectiveness of the algorithm, an experimental platform composed of 40 PCs is built. One of the platforms is the main node, which allocates and schedules all resources and manages the file system, and the other is the data node, which completes the storage and prediction operations. Audit data distributed in independent power enterprise subsystems are migrated to Hadoop cluster through open source Sqoop tools. The audit data of an electric power enterprise in 2013-2017 was used for the experiment, and the audit problems were divided into 15 categories. A total of 192,020 risk statistics were finally sorted into 6,000 records after de-duplication and exception handling, and 1,300 records were sampled as a test set. The experiment is divided into two parts. Firstly, the experimental data is artificially expanded to a large data scale, and the average of 50 running results is taken. Secondly, according to whether the existence of future audit problems is related to whether similar problems are found in previous years and the frequency of their discovery, the historical data of various types of audit problems are sorted according to the time (year) dimension, and

the frequency of audit problems in the most recent year is taken as the target variable, that is, the frequency of audit problems in 2017 is taken as the target variable, and the other years are taken as the analysis fields for prediction and comparison, thus detecting the prediction accuracy of the algorithm [8].

### 3.1 Comparison experiment of algorithm running speed

The results shown in Figure 2 are the experimental results of this algorithm and the traditional random forest algorithm under different scales of auditing large data sets. It can be seen that when the data scale is small, the prediction time of the two algorithms is similar, and the traditional method has a slight time advantage, mainly because the cost of parallel data blocking and communication between nodes affects the prediction speed; However, with the increase of sample size, the time advantage of the algorithm in this paper is more and more obvious, and the time required for iterative prediction is far less than that of traditional methods.

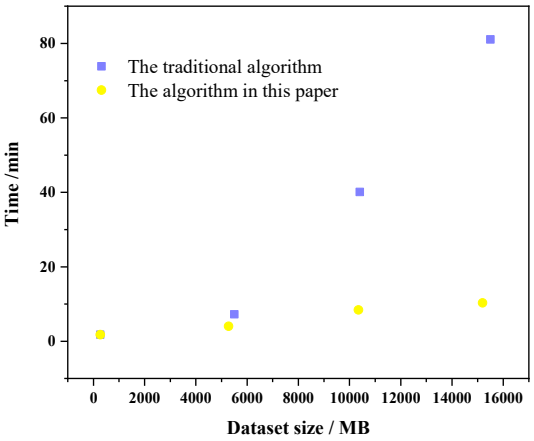


**Fig. 2.** Time required for prediction of different data sets by two algorithms

### 3.2 Comparison experiment of prediction accuracy of the algorithm

As shown in Figure 3, the parallel algorithm and the traditional decision tree algorithm in this paper, based on the statistics of audit problems in previous years, take the frequency of audit problems in 2017 as the doubtful point to predict the results, and the prediction results are still the average after many experiments. It can be seen that the algorithm in this paper is more accurate than the traditional method in predicting audit doubts, mainly because the parallel improved algorithm in this paper predicts through a number of decision trees generated by random sampling with replacement, which not

only retains the advantages of decision trees but also overcomes some of their shortcomings, showing better prediction performance [9-10].

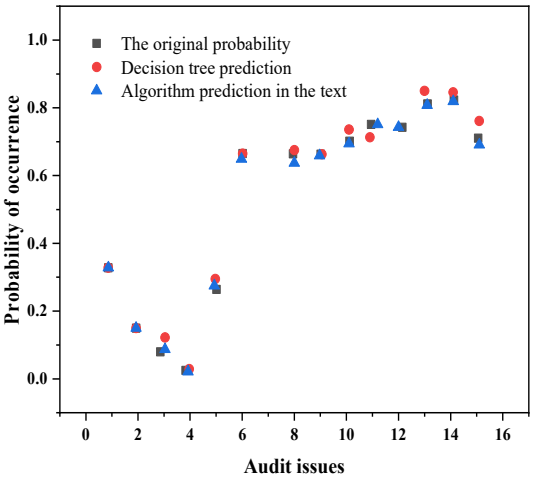


**Fig. 3.** Prediction results of audit doubts by two algorithms

The experimental results show that the average relative error of the parallel random forest algorithm is 1.43%, while the average relative recognition of the decision tree is 2.12%, which shows that the prediction accuracy of the algorithm in this paper is slightly better than that of the decision tree. Based on the above two experimental results, it can be seen that the parallel random forest algorithm in this paper has achieved good experimental results in both running speed and prediction accuracy, which shows that the algorithm in this paper is suitable for predicting the probability of audit doubts under audit big data and can provide data support for the formulation of audit plans.

## 4 Conclusion

In this paper, a probability prediction algorithm of audit doubt based on parallel improved random forest algorithm is proposed. Based on Hadoop, the algorithm establishes a big data management scheme for power enterprise audit, and integrates and stores different types of audit data collected from various power grid subsystems. On this basis, the parallel improvement of the algorithm is realized by adopting the three-tier MapReduce architecture to improve the running efficiency of the algorithm in the big data environment. The core of this algorithm is the random forest algorithm based on parallel improvement, which can predict the probability of audit doubt. Random forest is an integrated learning algorithm, which predicts by constructing

multiple decision trees and combining them. In the parallel improvement, the data is divided into several sub-data sets by parallel computing, and each sub-data set is processed by a MapReduce task. In this way, multiple decision trees can be built in parallel, and finally they can be combined to get the final prediction result. The effectiveness of the algorithm can be verified by comparing the measured data. The experimental results will show the accuracy and efficiency of the algorithm in predicting the probability of audit doubts. This algorithm can provide scientific data support for the formulation of audit plan and meet the operational efficiency requirements in big data environment.

## References

1. Liang, J. . (2022). Problems and solutions of art professional service rural revitalization strategy based on random forest algorithm. Wireless Communications and Mobile Computing, 2022(1), 1-11.
2. Xie, H. , Dong, J. , Deng, Y. , & Dai, Y. . (2022). Prediction model of the slope angle of rocky slope stability based on random forest algorithm. Mathematical Problems in Engineering, 2022.
3. Ning, F. , Cheng, Z. , Meng, D. , & Wei, J. . (2021). A framework combining acoustic features extraction method and random forest algorithm for gas pipeline leak detection and classification. Applied Acoustics, 182, 108255-.
4. Zhang, Z. , & Cai, Z. . (2021). Permeability prediction of carbonate rocks based on digital image analysis and rock typing using random forest algorithm. Energy & Fuels(3).
5. Kim, S. , Gülay Kuja Karahan, Sharma, M. , & Pachepsky, Y. . (2021). The site-specific selection of the infiltration model based on the global dataset and random forest algorithm. Vadose Zone Journal, 20(12).
6. Inaguma, D. , Hayashi, H. , Yanagiya, R. , Koseki, A. , Iwamori, T. , & Kudo, M. , et al. (2022). Development of a machine learning-based prediction model for extremely rapid decline in estimated glomerular filtration rate in patients with chronic kidney disease: a retrospective cohort study using a large data set from a hospital in japan. BMJ open, 12(6), e058833.
7. Nuraini, A. , Leon, F. , & Usman, B. . (2021). Analysis of the effect of governance and research and development on probability of default. The International Journal of Science and Society.
8. Mcalinn, J. , & Feakins, R. . (2022). P132 an audit of quality of histopathology reporting of colorectal mucosal biopsies for the diagnosis and assessment of inflammatory bowel disease. Journal of Crohn's and Colitis.
9. Abid, A. , & Abid, F. . (2023). A methodology to estimate the optimal debt ratio when asset returns, and default probability follow stochastic processes. Journal of Industrial and Management Optimization, 19(10), 7735-7752.
10. Wu, T. G. , Chen, Y. D. , Chen, B. H. , Harada, K. H. , Lee, K. , & Deng, F. , et al. (2022). Identifying low-pm2.5 exposure commuting routes for cyclists through modeling with the random forest algorithm based on low-cost sensor measurements in three asian cities. Environmental Pollution, 294, 118597-.