Improved Random Forest Fault Diagnosis Model Based on Fault Ratio

Ziwei Ding
[1.] School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
[2.] State and Provincial Joint Engineering Lab of Advanced Network, Monitoring and Control
Xi'an, 710021, China
E-mail: 146377997@qq.com

Shunyuan Huang
[1.] School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
[2.] State and Provincial Joint Engineering Lab of Advanced Network, Monitoring and Control
Xi'an, 710021, China
E-mail: sylvianoemie@sina.com

***Abstract*—With the rapid development of information technology, the informatization, integration and complexity of more and more large equipment are increasing day by day, so it is very important to carry out fault diagnosis for such complex equipment. In the traditional way, expert system technology is usually used for fault diagnosis of complex equipment. However, with the increasing of equipment data information, traditional methods cannot solve the fault diagnosis requirements in the case of a large amount of data. Therefore, data-driven fault diagnosis method can solve this problem, The carrier of data-driven fault diagnosis is a large amount of engineering data, and its focus is to explore new methods of fault diagnosis from a large amount of historical data. In this paper, the classical random forest algorithm is selected as the basic model, and aiming at the imbalance of complex equipment data, the improved random forest voting mechanism based on the fault ratio is proposed to optimize the model, which makes the final model diagnosis accuracy more than 95%, and has good application value.**

Keywords-Complex Equipment; Fault Diagnosis; Random Forest; Unbalanced Data

I. INTRODUCTION

Along with the rapid development of information technology era, large equipment is more and more in different industries tend to electronic and complication, integration, and summarizes it is different in the field of large equipment increasingly tend to be intelligent, this development trend will largely increase the probability of equipment failure and the difficulty of the late breakdown maintenance, The traditional mode of "periodic maintenance" and "post-repair", such as manual scheduled maintenance and fault reprocessing, is no longer applicable to the current large and complex equipment. At the same time, a series of chain reaction caused by equipment failure will also cause serious safety accidents and bring huge economic losses. After long-term experience and practice, in order to ensure efficient, safe and reliable operation of equipment, reasonable use and in-depth study of fault diagnosis technology is particularly important [1].

The research on fault diagnosis was first carried out by NASA in the late 1960s. Since the research started, this technology has crossed many disciplines in other fields, and then derived many new fault diagnosis methods, attracting the attention of a large number of European developed countries. This technology has been applied to aviation, navigation, large-scale industrial projects, chemical industry and military fields in many countries. Westinghouse electric Company has been committed to the research of artificial intelligence expert system of power station since 1980s, and has achieved good results. Boeing, a civil aviation giant, has also developed IMA systems that combine artificial intelligence technology with fault diagnosis. Our country's Xiong Fanlun applied expert system in the field of agriculture, to achieve a more reasonable and convenient agricultural production [2].

Fault diagnosis refers to the technology of identifying and classifying the device status by collecting the current and historical status

information of the device. The purpose of this technology is to ensure the smooth and normal running of the system devices and avoid unnecessary emergencies. However, traditional fault diagnosis requires a high level of technical personnel in operation, and is not suitable for deeper diagnosis scenarios [3]. Therefore, with the continuous development of artificial intelligence and its derivatives, fault diagnosis technology has gradually realized the transformation to intelligent fault diagnosis. The core of intelligent fault diagnosis is to create an entity that can diagnose faults on devices as an "expert" and provide the same diagnosis results as traditional expert detection. At the same time, with the continuous development of machine learning, its performance in the field of fault diagnosis is becoming more and more excellent. Relevant research data of scholars show that the application of this technology to large and complex equipment can quickly identify faults, significantly improve the durability and reliability of equipment, and have universality and research ability [4]. In view of this equipment data imbalance, this paper choose machine learning in the classical model of random forest model based on random forest model in the history of unbalanced data sets, generally due to other machine learning model, and further introduce the basic principle of random forest algorithm, and then to the imbalance of a large number of complex equipment failure data based on the fault than improved random forest model, Experimental results show that the improved model has higher accuracy than the single random forest model.

II. BASIC PRINCIPLE OF RANDOM FOREST ALGORITHM

A. Random forest algorithm concept

Random Forest is a classification prediction model based on ensemble learning proposed by Leo Briman, academician of American Academy of Sciences in 2001. The smallest unit of the model is the decision tree. Intuitively speaking, every decision tree is a simple classifier. For an input sample, N decision trees will have N classification results. Random forest algorithm is a collection of all the classification results, the proportion of all results is the final decision results [5].

Generally speaking, each decision tree can be as a focus on a particular aspect of the referee, if only to listen to a referee rhetoric so there must be some deviation, but there are a number of referee each referee different perspective to deal with problems, eventually all vote the way the referee to determine the results, although there will be individual differences, But the overall prediction variance must be decreasing [6]. Compared with traditional classification algorithms, random forest algorithm has fast classification speed, large capacity of data processing, strong ability of error balance and difficult to over-fit. In addition, it is worth mentioning that the robustness of the random forest model depends on the number of decision trees. The more decision trees there are, the more precision of the model will increase. The random forest model is shown in the figure below.

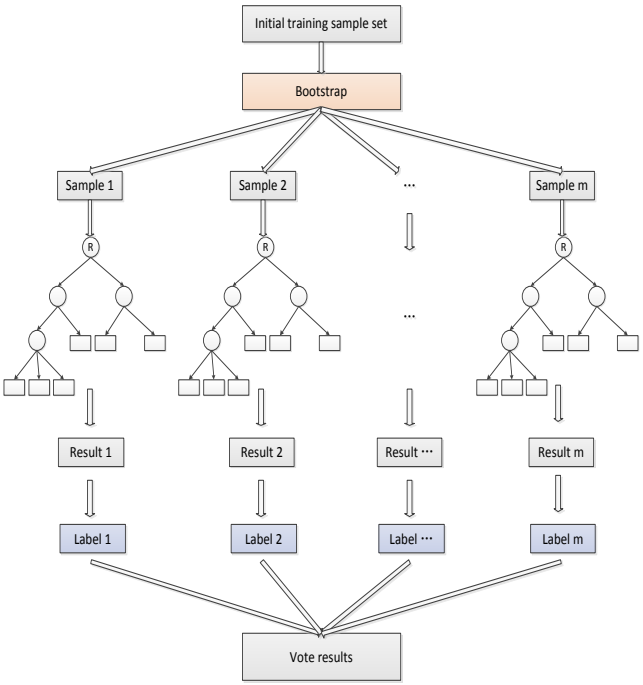

Figure 1. Random forest model

B. Random forest model construction process

The construction of random forest can be divided into four steps as shown in the following figure:

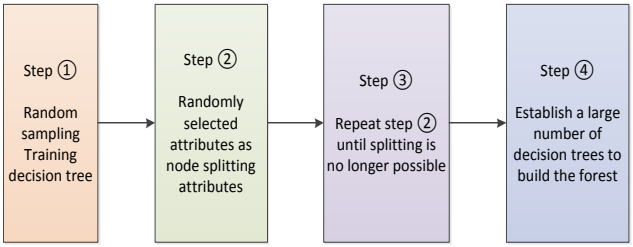

Figure 2. Random forest model construction process

1) Assume that there are currently N samples and select N samples by random extraction with put back. Random extraction with put back means that a sample is randomly selected from the data set every time. After the selection, the sample is put back into the data set and then returned to select new samples. The selected samples are used to train a decision tree;

2) When each sample has M attributes and each node in the decision tree needs to be split, M attributes are randomly selected from M attributes, and the condition to be met is M < Then, other indexes such as information gain and GINI value are used as splitting strategy to select the splitting attribute of this node;

3) A decision tree is established based on Step (2). Every node in the process of forming the decision tree should be split according to step (2). If the attribute selected by the node next happens to be the attribute used by its parent node when splitting, there is no need to split again. Repeat until it can no longer split;

4) Repeat all the above steps, and the completed decision trees constitute a complete random forest.

C. Advantages and disadvantages of random forest model

1) RF model is based on decision tree, so the model can also realize classification and regression functions [7]. However, the model is mostly applied to classification, and its advantages are as follows:

a) For features, the importance of each feature can be judged by tree structure;

b) In the case of feature missing or outliers, the model can also show good performance and still maintain accuracy;

c) The mutual influence of different features can be judged;

2) Compared with the advantages, the disadvantages are slightly insignificant. The model has the following disadvantages:

a) In dealing with regression problems, the performance of the model is not as good as that of dealing with classification problems;

b) When the external noise in the training sample set is relatively large, the over-fitting problem is more likely to occur;

c) In the case of imbalanced samples, it is impossible to accurately measure the contribution ratio of certain features to RF model.

III. IMPROVED RANDOM FOREST ALGORITHM DIAGNOSIS MODEL BASED ON FAULT RATIO

The imbalance of complex equipment failure data, the author of this paper chose random forests can better deal with unbalanced data model, but in practical engineering applications, the complex equipment failure probability is very low, so will lead to failure data and under normal circumstances the ratio of the equipment condition data samples is very small, the normal data sample size is far greater than the fault data sample size, We also refer to such data as highly unbalanced data sets [8].

A core part of the random forest model is the integrated voting process. In the basic RF model, the mode voting mechanism of "minority follows majority" is usually used to determine the final result. However, considering the extremely unbalanced data studied in this paper, the degree of imbalance will seriously affect the result of the mode voting. Random forest integrates multiple decision trees to achieve integrated learning. Although integrated learning can reduce the comprehensive error to a certain extent, the imbalance of the research object in this paper is too high, so it is particularly important to take corresponding measures. Therefore, this paper uses fault ratio to improve the voting decision [9].

A. Improved voting rules based on fault ratio

For complex equipment system, the number of fault classes is small, that is, the number of

positive classes is small. The number of normal classes is larger, that is, there are more negative classes. In this paper, the original mode voting mechanism is abandoned and the voting decision rules in RF model are improved by combining the ratio between the sample size of faulty data and the sample size of normal data, namely the fault ratio. The improved rule is described as follows: Check the category labels output by each decision tree in the forest. If the ratio of positive category labels to negative category labels is greater than the fault ratio, the final result is classified as positive category (small sample category). In this way, the original mechanism can be optimized, and the weight of positive and negative label decision trees in the forest is no longer fixed, which can better overcome the imbalance of positive and negative samples [10].

B. Improved random forest algorithm based on fault ratio

The implementation process of the improved random forest algorithm is shown in the following table:

Improved random forest algorithm based on fault ratio
Step1: The training samples were randomly put back from the data set, and were extracted for n times in total to obtain n independent training sets with repeated elements. Step2: The n decision trees are trained on different training sets. Step3: The sample category labels corresponding to N decision trees were analyzed, and the final voting induction was carried out by combining the improved voting decision method based on fault ratio.

IV. EXPERIMENT AND RESULT ANALYSIS

A. Introduction to complex equipment fault diagnosis data set

In this paper, the fault data set of a complex equipment is used, but considering the confidentiality of the relevant data of a complex equipment, the relevant feature attributes of the data set are expressed in the form of coding, and the specific situation of the data set is as follows:

1) Large amount of sample data. The data set provided is composed of training set and test set. The sample size of training set is 60000 (1000 positive classes and 59,000 negative classes). The sample size of the test set is 16000 (375 positive categories and 15625 negative categories), so the data set is rich in information and also in line with reality. Although complex equipment may have failures in real life, the occurrence of failures must be rare, so the proportion of normal and abnormal data in the original data set is 59:1, which is reasonable.

2) Rich sample attributes. Can be in the data set, each sample data are of the fire control system contains 171 features, it also means that each of the sample data is composed of 171 properties, research on the fault diagnosis of complicated equipment concerned about most is the word "complex", through the analysis of the data set can be obtained after data set features more, experimental complexity increase, It has reference value for the future study of fault diagnosis of complex equipment.

3) In essence, the fault diagnosis set is a typical binary data set with positive and negative imbalance.

B. Data preprocessing

1) Data missing value processing

Due to the high integration of complex equipment, some state data in the data set are missing. Therefore, the missing ratio of each feature is firstly calculated in feature engineering operation, and the results are shown in the form of bar graph as shown in the figure below.

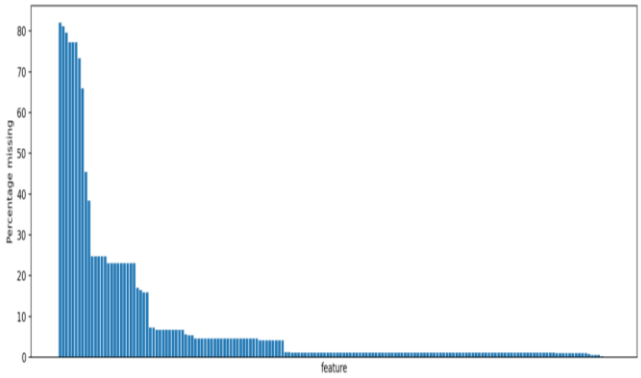

Figure 3. .Characteristic exact ratio

According to the figure above, there are a large number of missing attributes in the data set. In the operation of this paper, the attribute features with missing values greater than 75% were deleted, that is, six attribute columns were deleted, namely br_000, BP_000, BP_000, BO_000, AB_000 and CR_000.

Then, for the data with the missing percentage less than 75%, the missing value is completed. The main methods of missing value completion are median completion and mean completion. Here, the two methods are used to complete the data respectively, and the final method is measured by comparing the standard deviation of the data set after the two completions.

The standard deviation of each attribute after filling is shown in the following table:

TABLE I. DATA VARIANCE COMPARISON

Attribute name	Standard deviation after median supplement	The standard deviation of the mean
aa_000	1.454301e+05	1.454301e+05
ac_000	7.767625e+08	7.724678e+08
ad_000	3.504525e+07	3.504515e+07
ae_000	1.581479e+02	1.581420e+02
af_000	2.053871e+02	2.053753e+02
...	...	
ee_007	1.718666e+06	1.718366e+06
ee_008	4.472145e+05	4.469894e+05
ee_009	4.721249e+04	4.720424e+04
ef_000	4.268570e+00	4.268529e+00
eg_000	8.628043e+00	8.627929e+00

As can be seen from Table 1, for the attribute features with missing values, the standard deviation of the data after using the mean value is less than that after using the median value. Therefore, this paper adopts the mean value supplement method for the missing data.

2) Unbalanced data processing

Random forest has a good performance on unbalanced data sets, but the corresponding unbalanced processing of data sets is also very important.

Exist unbalanced data sets is to complex equipment fault diagnosis based on machine learning research a well-known problem in the process, there is no doubt that the small sample data in the process of research has played a more important role, and attract the attention of the researchers, in real life applications such small sample is a part of the researchers are more interested in.

The processing of unbalanced data sets can be divided into data level and algorithm level in general direction, as shown in the figure below:

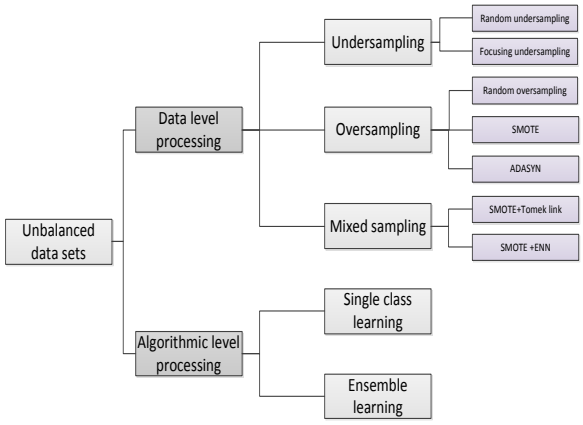

Figure 4. .Unbalanced data processing method

In this paper, ensemble learning is used at the algorithm level. Therefore, after comparing different sampling methods, we finally choose the improved method of smote borderline-smote1 as the solution for unbalanced data set processing. This method further divides a small number of samples into "safety", "danger" and "attention", and randomly selects a small number of samples of k-nearest neighbor in the "danger" attribute.

C. Introduction to complex equipment fault diagnosis data set

There are many super parameters in the random forest model. In this paper, the simple grid search method combined with 50% cross verification is used. On the premise of determining the node splitting index, the optimal parameters in the verification set are selected by changing the tree and the largest characteristic tree in the forest, and drawn into a broken line diagram, as shown below.

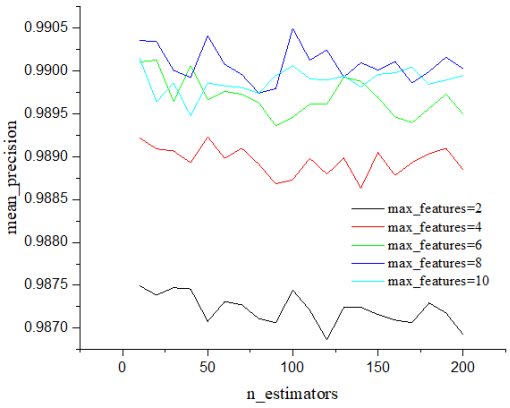

Figure 5. .Unbalanced data processing method

It can be intuitively observed from the figure that when the maximum feature number is selected differently, the corresponding average accuracy rate is also different. The average accuracy rate of the model here can be basically maintained above 0.99. To sum up, when the maximum feature number is 8, the number of decision trees in the forest is 100, and the splitting index is "Gini", the model achieves the best effect on the verification set.

D. Experimental results and comparison

After model fitting, the confusion matrix obtained on the test set is shown in the table below. According to the information in the table, the model correctly divides the normal data into 15611 and the fault data into 361. However, at the same time, 14 normal data are mistakenly divided into fault data and 14 fault data are mistakenly divided into normal data.

TABLE II. MODEL RESULTS

Forecast	Physical truth	
	Normal data	Fault data
Normal data	15611	14
Fault data	14	361

Combined with the model evaluation criteria, the evaluation results obtained according to the confusion matrix are as follows:

TABLE III. MODEL METRICS

Evaluation criteria	Result
ACC	96.94%
Precision	96.27%
Recall	96.27%
F1 score	0.9627
AUC	0.9701

In order to highlight the feasibility of the improved model based on fault ratio, the unmodified random forest model is used as the control experiment. The comparison of the experimental results is shown in the figure below.

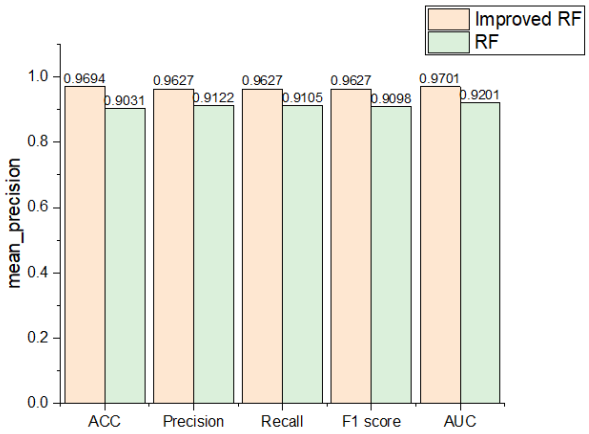

Figure 6. .Comparison of experimental results

Through the comparison of experimental results and indicators, it can be concluded that all indicators of the improved random forest model are better than the basic random forest model. The experiment shows that the random forest algorithm can better diagnose and process the unbalanced data, and the precision can reach more than 90%, but the accuracy rate of the model based on fault is better than that after improving the decision rules, and the precision can reach more than 96%. The improvement is feasible.

V. CONCLUSION

In view of the extremely unbalanced fault data of complex equipment, this paper proposes an improved random forest model based on fault ratio to realize the diagnosis task. The experimental results show that the performance of the model after using the improved method is better. The

importance of fault diagnosis in real life is self-evident. At present, there are many mainstream diagnosis methods, but it is very important to realize the analysis of "adjusting measures to local conditions" for different types of fields. Different machine learning algorithms may show different abilities in different data. This paper provides a new idea for the establishment of extremely unbalanced data model, In the follow-up research, We can study the following points: ①If the amount of data is large enough, fault diagnosis models can be built with the help of more emerging deep learning methods in recent years. ②In the creation of machine models, there are many choices for the selection of model parameters. In future research, intelligent optimization algorithms can be applied to the optimization of model hyperparameters. ③More sampling methods can be tried to solve the data imbalance problem.

REFERENCES

[1] Liu Zhantao Research on integrated method of condition monitoring and fault diagnosis of large equipment system [D] Beijing University of chemical technology, 2009.

[2] Xiong Fanlun Architecture and implementation of intelligent system technology for agricultural field [J] Pattern recognition and artificial intelligence, 2012, 25 (05): 729-736.

[3] Guo Zhi Research on fault diagnosis method of chemical machinery and equipment based on big data [J] Information recording materials, 2021,22 (09): 233-235.

[4] Zhang Peilin, Cao Jianjun, Ren Guoquan Research on condition monitoring system of large mobile complex equipment [J] Journal of Gun Launch and control, 2006 (03): 15-18.

[5] Xu Dongpo,Liu Yunqing,Wang Qian. Random forest-based human pose detection system for through-the-wall radar [J]. Journal of Physics: Conference Series, 2021, 1966(1).

[6] Sherif Ahmed Abu El-Magd, Sk Ajim Ali, Quoc Bao Pham. Spatial modeling and susceptibility zonation of landslides using random forest, naïve bayes and K-nearest neighbor in a complicated terrain [J]. Earth Science Informatics, 2021 (prepublish).

[7] Wu Weijie Research on application and optimization method of random forest algorithm [D] Jiangnan University, 2021.

[8] Dong Hongyao, Wang Yidan, Li Lihong. Overview of Random Forest Optimization Algorithms [J]. Information and Computer (Theoretical Edition), 2021, 33(17): 34-37.

[9] Sun Mingzhe, Bi Yaojia, Sun Chi. Overview of Improved Random Forest Algorithm [J]. Modern Information Technology, 2019,3(20):28-30.

[10] Wang Yisen, Xia Shutao. Overview of Random Forest Algorithm for Ensemble Learning [J]. Information and Communication Technology, 2018, 12(01): 49-55.