A Classifier of Railway Power Supply Equipment Concern Importance Based on Ensemble Learning

Lei Zhang¹, Tianbing Wang¹, Shengcan Yu², Tao Yu²,*

¹China Railway Siyuan Survey and Design Group Co., Ltd., Wuhan 430063, China
²School of Electric Power Engineering, South China University of Technology, Guangzhou 510640, China

*Corresponding author: taoyul@scut.edu.cn

Abstract. It is of great significance to evaluate the condition of equipment and identify the degree of important concern of equipment for formulating reasonable condition maintenance strategy and ensuring the safe and stable operation of railway power grid. In view of the characteristics that the overload operation state and maintenance interval of equipment are closely related to the equipment state level, the evaluation feature vector composed of five indexes, including overload time of heavy overload equipment, whether the central power supply equipment is in the city center, unrepaired time of equipment, whether it is the power supply equipment and the voltage level of equipment is maintained, is constructed in this subject. In addition, based on the Stacking ensemble learning framework, a hybrid ensemble learning framework algorithm is proposed, which combines the single Bagging and Boosting framework algorithms, and realizes the mapping between feature vectors and importance levels. The data of actual railway power grid equipment are substituted into the model for analysis, and the model is compared with the random forest model based on single ensemble learning algorithm. The simulation results show that compared with other ensemble learning algorithms, the accuracy of equipment state evaluation is significantly improved by using the combination model of stacking ensemble learning method.

1. Introduction
Using big data technology and artificial intelligence method to monitor equipment operation condition of high-speed railway traction supply network online. The potential risks are identified, and develop overhaul plan for the reasonable equipment state [1]. Complete the efficient maintenance of the power supply equipment, and improve the work efficiency of the power supply equipment, reduce equipment maintenance costs. All of them are of great significance to achieve the purpose of efficient management of railway power grid equipment. Machine learning is the mapping relationship between the learning features and the importance degree of equipment, which completes the state assessment of equipment [2] [3].

This paper mainly analyzes five characteristics of the indicators, including the overload time of heavy overload equipment, whether it is the central power supply equipment, the unrepaired time of the equipment, whether it is the power supply equipment and the voltage level of the equipment. Then we adopt Three Sigma Guidelines and cases elimination method to clean the data, and selects One-Hot code
method to complete the discrete data coding processing. Max-min algorithm was set as the continuous data normalization method. Finally, based on the training analysis of ensemble learning method, and considering the problem that the features of adjacent importance levels are difficult to distinguish, a hybrid ensemble learning framework algorithm is proposed based on Stacking ensemble learning framework, which combines the single bagging and boosting framework algorithm, which realized high precision identification of important concern level of equipment.

2. Data preprocessing method

2.1. Abnormal data processing based on Three Sigma Guidelines

In the normal distribution, the probability of the numerical distribution in \((\mu - \sigma, \mu + \sigma)\) is 0.6826, the probability of the numerical distribution in \((\mu - 2\sigma, \mu + 2\sigma)\) is 0.9544, and the probability of the numerical distribution in \((\mu - 3\sigma, \mu + 3\sigma)\) is 0.9974. It can be considered that numerical distribution is almost all concentrated in the interval of \((\mu - 3\sigma, \mu + 3\sigma)\), and the possibility beyond this range is only less than 0.3 percent. Therefore, the error of equals is usually taken as the limit error. If the absolute value of the residual error of a measurement value \(v_i > 3\sigma\) in a set of measurement data, then the measured value is an outlier, and the corresponding sample is an abnormal sample, which should be eliminated.

2.2. Data normalization method and coding method

2.2.1. Maximum and minimum normalization of continuous variables.

We use the maximum and minimum normalization method to process continuous data. Under the premise of not changing the data distribution, the data is transformed into continuous variables between 0 and 1, so as to accelerate the convergence of machine learning.

\[
x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}
\]

Where, \(x_{\text{min}}, x_{\text{max}}\) represent the maximum and minimum values of the original data respectively.

2.2.2. One-hot coding of discrete variables.

One-hot coding method, also known as one-bit effective coding, mainly uses the N-bit status register to encode N states, each state consists of independent register bits, and the only one is valid at any time. One-hot encoding is the representation of a class variable as a binary vector. This first requires mapping the category value to an integer value. Then, each integer value is represented as a binary vector, which is zero except for the index of the integer, which is marked as 1.

3. Traditional ensemble learning algorithm

At present, with the rapid development of artificial intelligence and machine learning technology, a variety of intelligent algorithms such as KNN [4], SVM [5], and neural network [6] emerge in an endless stream, providing a new solution for equipment state evaluation. However, if a separate method is used for load forecasting, due to the large hypothesis space of classification problems such as state recognition, there may be multiple assumptions to achieve the same performance on the training set. If a single model is used, the generalization performance may be poor due to randomness. Therefore, seeking the way to use combination classifier is an inevitable choice to further improve the accuracy of model classification.

4. Ensemble learning algorithm based on Stacking framework

Based on the above analysis, in order to balance the bias and variance in statistical learning algorithms, and taking into account the difference in sensitivity of Bagging ensemble learning algorithm and Boosting ensemble learning algorithm to the bias and variance, this paper effectively combines Bagging
and Boosting ensemble learning methods based on the Stacking framework[8], and give full play to the advantages of the two models to get a more stable model with stronger generalization ability.

4.1. Stacking ensemble learning framework
Stacking [11] ensemble learning framework first divides the original data set into several sub-data sets, which are input into each base learner of the first-layer prediction model, and each base learner outputs its own prediction results. Then, the output of the first layer is used as the input of the second layer to train the meta-learner of the second layer prediction model, and then the final prediction result is output by the model located in the second layer. Stacking learning framework generalizes the output results of multiple models to improve the overall prediction accuracy, as shown in Fig 1.

The specific training method of Stacking ensemble learning is: for the data set $S = \{(y_n, x_n), n = 1, \cdots, N\}$, where $x_n$ is the feature vector in the sample, $y_n$ is the corresponding label value in the Nth sample, $p$ is the number of features contained, each feature vector is $(x_1, x_2, \cdots, x_p)$. Divide the data randomly into K subsets of approximately equal size $S_1, S_2, \cdots, S_K$. Where $S_{-K} = S - S_K$, $S_K$ and $S_{-K}$ are respectively defined as the K-fold test set and training set in K-fold cross validation. For the first layer prediction algorithm contains K basic learners, the training set $S_{-K}$ is trained with the Kth algorithm to obtain the basic model $L_k, k = 1,2, \cdots, K$.

For each sample $x_n$ in the k-fold test set $S_K$ in k-fold cross validation, the prediction of the base learner $L_k$ is expressed as $z_{kn}$. After the cross-validation process is completed, the output data of K base learners are constructed into a new data sample, that is, $S_{\text{new}} = \{(y_n, z_{1n}, z_{2n}, \cdots, z_{kn}), n = 1, \cdots, N\}$.

The newly generated data set is the layer 2 input data in Stacking learning. The meta-learner $L_{\text{new}}$ is obtained by using the Layer 2 prediction algorithm to summarize these data. Stacking configuration makes the training results of the first layer can be fully used in the induction process of the second layer algorithm, the second layer algorithm can find and correct the prediction error in the first layer learning algorithm, in order to improve the accuracy of the model.

4.2. Stacking learning framework cascaded RF classifier
Considering that in the actual classification problem, there are some similar feature vectors of adjacent categories, and the classification is difficult. This paper proposes to series RF classifier after Stacking framework, and build the cascaded RF classifier Stacking integrated learning framework. A single dichotomous RF model was trained to classify samples of adjacent categories, and its learning framework was obtained, as shown in Fig 1.
5. Case analysis
In order to verify the effectiveness of the hybrid integration algorithm proposed in this paper, 1,000 sample data of actual railway power grid equipment operation are selected for example simulation analysis. The statistical results of the data are shown in Table 1. Among them, there is a small amount of data with importance level 5, and the foreseeability of classification results to distinguish 5 is poor. 900 pieces of data are randomly selected from the sample data as the training set, while the remaining 100 pieces of data are used as the test set. The importance of equipment is classified by using Random Forest, GBDT, XGBoosting and LightGBM [11-12] single ensemble learning models and the mixed ensemble model based on Stacking framework, and the results are compared and analyzed.

| Level of importance | The number of data |
|---------------------|--------------------|
| 1                   | 125                |
| 2                   | 304                |
| 3                   | 328                |
| 4                   | 199                |
| 5                   | 44                 |

A variety of trained models were used to classify the test set data, and the error classification results were compared with the correct results to obtain the distribution diagram of the error classification categories, as shown in Figure 2. In addition, the error situation of the distribution of the misclassification categories can be counted, and the results shown in Table 2 can be obtained.
According to Table 4, the results obtained by the traditional ensemble learning algorithm and the hybrid ensemble learning algorithm based on Stacking framework are analyzed respectively in this paper.

1) Classification results based on traditional ensemble learning algorithm: Its error classification distribution diagram is shown in Fig.2 (a)-(d), in which there is no cross-level error classification such as 1-3 and 1-4 in the classification results of all algorithms. For 1-2 classification, all algorithms have high accuracy, Random Forest and XGBoosting can achieve complete accuracy classification. For 4-5 classification, all traditional ensemble learning algorithms cannot accurately classify. On the one hand, this may be caused by too little sample data of 5; on the other hand, it also indicates that the corresponding device characteristic parameters of 4 and 5 are very similar. In particular, the accuracy of all the traditional ensemble learning algorithms cannot reach more than 90%.

2) Classification results of mixed ensemble learning based on Stacking framework: The error classification distribution obtained by classifying only with the Stacking framework is shown in Figure 2 (e), with the number of validation sets remaining at 100. As we can see, the total number of misclassifications has dropped to single digits, to just six. Among them, there were 4 errors in 2-3 and 2 errors in 4-5, which were still difficult to distinguish. Therefore, the next step is to optimize specifically for these two types of errors.

The improved StackingRF framework of cascading 4-5 RF classifiers and 2-3 RF classifiers is used for instance analysis, and the number of verification sets is still 100. The results are shown in Fig. 2 (f). At this time, the accuracy rate reached 99%, which significantly improved the identification accuracy of 2-3 categories and 4-5 categories. Thus, the machine learning effect based on StackingRF framework has reached the expectation.
The accuracy of all algorithms in the training set and test set is shown in Table 3.

|               | RF   | GBDT | XGBoosting | LightGBM | Stacking | StackingRF |
|---------------|------|------|------------|----------|----------|------------|
| The training set | 1.00 | 1.00 | 1.00       | 1.00     | 1.00     | 1.00       |
| The test set   | 0.90 | 0.84 | 0.89       | 0.88     | 0.93     | 0.99       |

Through comparison, it can be seen that under Stacking, Bagging and Boosting three frameworks, the accuracy of the training set is 100%, and the learning effect of Stacking framework in the test set is better than that of Bagging and Boosting framework, and the learning effect of improved Stacking RF framework is the best, which is 99 percent. All the learning processes of this paper, the sample data of the training set are all 900, and the accuracy of device classification can be further improved by increasing the number of sample data in the future.

6. Conclusions

In this paper, based on the characteristics of feature vectors in equipment state assessment, the data preprocessing method was designed scientifically. Three Sigma Guidelines and the case elimination method were used to clean the data. One-hot coding was selected to complete the discrete data coding processing, and Max-min method was set as the continuous data normalization method. In addition, based on ensemble learning method of traditional framework for training analysis, and considering the problem that the features of adjacent important levels are difficult to distinguish, a hybrid ensemble framework algorithm – Stacking-RF algorithm is proposed based on Stacking ensemble learning framework, which combines the single Bagging and Boosting framework algorithms. According to the results of a practical example, it is proved that this method can realize the high precision identification of important concern level of equipment.

Acknowledgments

This work was financially supported by Key Equipment Prediction and Health Management (PHM) System for Power Supply and Transformation (2019C16) fund.

References

[1] ENDRENYI J, ANDERS G J, LEITE DA SILVA A M. Probabilistic evaluation of the effect of maintenance on reliability: an application to power systems [J]. IEEE Transactions on Power Systems, 1998, 13(2): 576-583.
[2] WELTE T M. Using state diagrams for modeling maintenance of deteriorating systems [J]. IEEE Transactions on Power Systems, 2009, 24(1): 58-66.
[3] ABEGUNAWARDANE S K, JIRUTITIJAROEN P. New state diagrams for probabilistic maintenance models [J]. IEEE Transactions on Power Systems, 2011, 26(4): 2207-2213.
[4] Cordeiro de Amorim Renato, Mirkin Boris. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering [J]. Pattern Recognition, 2012, 45(3): 1061-1075.
[5] Li Guangli, Cui Guangshun. An improved Algorithms of Fuzzy Support Vector Machines [J]. Computer Measurement & Control, 2011, 19 (4): 908-10, 914.
[6] David S, Huang A, Maddidon C J. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529 (7587) : 484-489.
[7] Breiman L. Bagging predictors [J]. Machine Learning. 1996, 24(2): 123-140.
[8] KEARNS M J, VALIANT L G. Learning boolean formulae or finite automata is as hard as factoring [R]. Cambridge, USA: Harvard University, 1988.
[9] A robust semi-supervised learning approach via mixture of label information [J]. Yun Yang, Xingchen Liu. Pattern Recognition Letters. 2015.
[10] Wolpert D H. Stacked generalization[M]. Boston: Springer,2017:6-10.
[11] Study on A Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms according to Different High Dimensional Data Cleaning[J]. Xiaojun Ma, Jinglan Sha, Dehua Wang, Yuanbo Yu, Qian Yang, Xueqi Niu. Electronic Commerce Research and Applications. 2018
[12] Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination[J]. Wenjie Zhang, Hao Quan, Dipti Srinivasan. Energy. 2018