Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/90/20241758

Fault diagnosis of transformer based on XGBoost

Lin Wang

School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun, China

2301854550@qq.com

Abstract. With the continuous development of technology, the fault diagnosis of transformer is also continuously advancing. Traditional fault diagnosis methods such as the three-ratio method and the characteristic gas method have the disadvantages of missing coding and limited identification capabilities. Data-driven fault diagnosis methods can integrate various information for data analysis to accurately diagnose the type of fault. This paper proposes a fault diagnosis method for transformers based on the XGBoost model. First, the input data is selected for the ratio feature engineering, and the obtained ratios are also used as input data, and then the input data is normalized for data preprocessing. Secondly, the XGBoost model is introduced, which minimizes the prediction errors in each round of gradient boosting, thereby improving the accuracy and recall rate. Finally, through case analysis, comparisons are made with other methods in various performance indicators, and the accuracy is improved. The results show that the XGBoost is more accurate in the fault diagnosis of transformers, proving that the research has certain value and significance.

Keywords: Transformer, Fault Diagnosis, XGBoost, Machine Learning.

1. Introduction

Transformers are indispensable components in power systems and are crucial in ensuring a stable power supply and improving power quality. Transformers can be divided into dry type and oil-immersed type according to the cooling medium. About 85% of transformers are oil-immersed. When operating, oil-immersed transformers may experience low-temperature overheating, medium-temperature overheating, high-temperature overheating, partial discharge, low-energy discharge, and high-energy discharge. These faults will cause changes in the oil chromatographic information of the transformer[1].Therefore, fault diagnosis can be performed through oil chromatographic analysis.

The characteristic gas method and the three-ratio method are traditional methods of fault diagnosis, but these methods have problems such as missing codes, fuzzy fault types, and diagnostic preferences. Therefore, it is particularly important to study data-driven methods of transformer fault diagnosis. Reference [1] proposed a BP neural network to diagnose transformer faults, but it has the disadvantage of slow convergence speed; Reference [2] proposed a clustering algorithm for transformer oil chromatography analysis, but it has the disadvantage of being sensitive to initial conditions; Reference [3] conducted a diagnostic analysis of transformer fault types based on an expert system, but when the knowledge base is insufficient, accurate diagnosis cannot be ensured.

DOI: 10.54254/2755-2721/90/20241758

This paper uses the XGBoost algorithm to diagnose transformer faults. First, the data is normalized to eliminate the impact of the dimension. The data is ratio-featured to improve the predictable performance of the model. Secondly, XGBoost is an ensemble learner that can combine multiple weak learners to improve the ability to generalize. In each round of iteration, the errors of prediction are minimized and the accuracy of prediction is continuously improved[2]. The case analysis verifies the accuracy and effectiveness of fault diagnosis of transformer based on XGBoost.

2. Fault diagnosis of transformer based on XGBoost

2.1. Data preprocessing and feature engineering

Characteristic gases are released, including H_2, CH_4, C_2H_2, C_2H_4 and C_2H_6. The traditional three-ratio method encodes according to the ratio range of C_2H_2/C_2H_4, CH_4/H_2, C_2H_4/C_2H_6, and determines the fault type through the encoding rules[3]. Existing data-driven methods usually diagnose directly based on gas concentration, and most of them ignore the gas ratio information. The concentration distribution of the five gases released when the transformer fails may be quite different. If they are directly input, the output results will be inaccurate. In order to improve the accuracy of the sample, these characteristic gases are usually input after the ratio is calculated. The concentrations of H_2, CH_4, C_2H_2, C_2H_4 and C_2H_6 are processed, and the values of C_2H_2/C_2H_4, CH_4/H_2 and C_2H_4/C_2H_6 are calculated and input as features. The ratio feature helps the model better capture the interaction among variables and improve the predictable accuracy of the model[4].

Before the data is input into the model, it is necessary to normalize the data. Data normalization is a commonly used data preprocessing technique, which aims to convert the data into a unified scale and eliminate the influence of different dimensions. The normalization method used in this paper is Min-Max normalization. Normalization formula:

$$X_i^{'} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where, X_i is the original value of the sample data, X_{min} is the minimum value of the sample data, X_{max} is the maximum value of the sample data, and $X_i^{'}$ is the value after normalization.

If the original data is used directly for analysis, the higher-valued indicators will have a greater impact on the results, thereby reducing the impact of the smaller-valued indicators. Normalization can convert the characteristic values of the sample to the same dimension, so that the values of each indicator are at the same quantitative level for comprehensive evaluation and analysis.

2.2. Fault diagnosis of transformer based on XGBoost

Ensemble learning is an efficient machine learning technology that improves the overall prediction performance by combining the predictions of multiple models. Compared with a single learner, an ensemble learner has significantly improved capabilities of generalization and accuracy[5]. Bagging and Boosting are two typical integrated learners. The XGBoost (eXtreme Gradient Boosting) model used in this article is an efficient and accurate application algorithm based on Boosting ideas. Its base learner is the CART (Classification and Regression Tree) decision tree. XGBoost adds a regularization term to the loss function to help to prevent model to overfit. Additionally, it can automatically handle missing values in your data by assigning a direction to the path of the missing value. First, the learner is initialized. XGBoost usually starts with constant prediction. Secondly, the existing model residuals are fitted, and in each iteration of gradient boosting, a new weak learner is found to correct the prediction error of the existing model. Secondly, the CART tree generated by fitting is added to the existing model with a certain weight to minimize the loss function of the overall model. By continuously iterating and adding a new CART tree at each step, the performance of the model will continue to improve. If the number of iterations is reached, the learner will be superimposed[6] .The diagnosis method based on the XGBoost model can accurately detect faults and maintain good performance in the face of imperfect data, without causing large prediction errors due to minor anomalies. The XGBoost prediction model is as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i) \;\; f_k \in F \tag{2}$$

$$F = \{f_{(x)} = \omega_{q(x)}\} \; q: R^m \to T, \omega \in R^T \tag{3}$$

3. Case analysis

3.1. Data processing

There are 309 data samples, of which 183 are train data and 126 are test data. The input of the data is the normalized data of the H_2, CH_4, C_2H_2, C_2H_4 and C_2H_6 gas content generated in the transformer and the ratio characteristics C_2H_2/C_2H_4, CH_4/H_2 and C_2H_4/C_2H_6. After calculation and diagnosis by the model, the fault type corresponding to each group of data can be diagnosed and the fault type is encoded. The result is output as the encoding value. The transformer fault type encoding is shown in Table 1.

Table 1. Transformer fault type coding.

Transformer fault category	Status Code
normal	0
Low temperature overheating	1
Medium temperature overheating	2
High temperature overheating	3
Partial Discharge	4
Low energy	5
High Energy Discharge	6

3.2. Training set and test set accuracy

By comparing the prediction accuracy and test time of different model methods for the same train set and test set, the prediction ability and generalization ability of different models can be shown. This paper selects eight learning models, including linear discriminant analysis (LDA), logistic regression (LR), Gaussian naive Bayes (GNB), K-nearest neighbor (KNN), decision tree (DTC), random forest (RF), gradient boosting tree (GBDT), and lightweight gradient boosting machine (lgb), to conduct comparative experiments with the model proposed in this paper. The comparison results are shown in Figure 1 and Table 2.

As shown in Figure 1,the training accuracy of the XGBoost model reaches 100%. Compared with other models, XGBoost has the best prediction performance in the test set. As shown in Table 2 ,the test time of the XGBoost model is very close to that of other models and the time is short. From the perspective of training time, the time required for XGBoost is in the middle range.

DOI: 10.54254/2755-2721/90/20241758

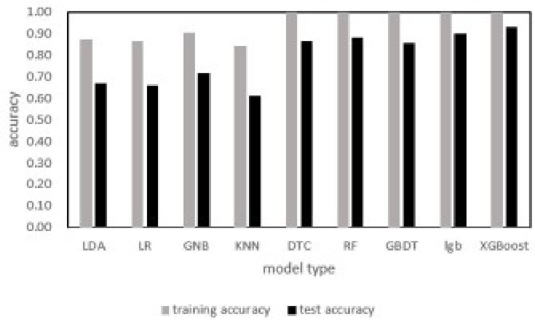

Figure 1. Comparison of accuracy of different models.

Table 2. Comparison of different models training time and test time.

Model	Training time(s)	Testing time(s)
LDA	0.018	0.000
LR	0.030	0.001
GNB	0.001	0.001
KNN	0.005	0.004
DTC	0.008	0.001
RF	0.113	0.008
GBDT	1.622	0.004
lgb	0.327	0.006
XGBoost	0.131	0.002

Precision, recall, and Macro F 1 Score are three important indicators for evaluating multi-classification models. Precision focuses on the accuracy of the prediction model, and recall focuses on the proportion of positive class samples captured by the model. Macro F 1 Score is an important indicator for measuring multi-classification models. It comprehensively considers precision and recall and balances the two. The comparison results of the three indicators of different models are shown in Figure 3. According to Figure 3, since Macro F 1 Score can evaluate the overall performance, and the higher the score, the better the overall performance, it can be observed that XGBoost has the best overall performance.

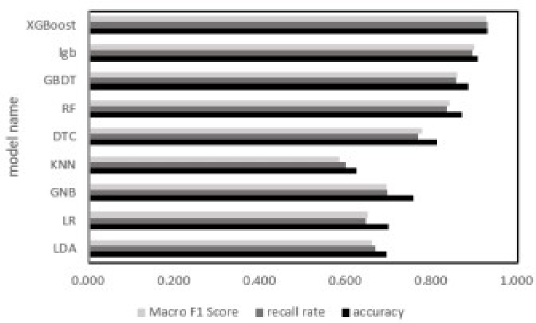

Figure 3. Comparison of evaluation classification model indicators.

4. Conclusion

In addressing the issue of fault diagnosis of transformer, the data-driven approach XGBoost is proposed. The total of 309 data entries are divided into training and testing sets, followed by data preprocessing and ratio feature engineering. The processed data is then input into the model. Comparative experiments clearly demonstrate the superior comprehensive performance of XGBoost, leading to the selection of the XGBoost model for transformer fault diagnosis. The model exhibits excellent performance on both the training and testing sets, enhancing the accuracy and efficiency of transformer fault detection, with the accuracy of prediction reaching 92.9%.

This research showcases the potential application of XGBoost in the field of transformer fault diagnosis, providing favorable technical support for the safe operation and maintenance of power systems. It holds significance in both theoretical research and practical engineering applications.

References

[1] Zhou, W, Z., Ma, Y., & Deng, J, H. (2023). Transformer DGA fault diagnosis based on FedAVG-SA-BP. *Transformer*, 60(09), 31-35.

[2] Li, E, W., Wang, N., & Song, B. (2018). Transformer oil chromatographic analysis based on improved fuzzy clustering algorithm. *Transactions of China Electrotechnical Society*, 33(19), 4594-4602.

[3] Gao, S., Liu, Z, G., & Xu J, F. (2013). Research on traction transformer fault diagnosis based on model diagnosis and expert system. *Journal of the China Railway Society*, 35(07), 42-49.

[4] Lin, B, M. (2023). Analysis of typical transformer fault cases based on the three-ratio method. *Electrical Technology*, 24(10), 63-67.

[5] Jing, L, T., Zhang, Y., & Zhang, B., Transformer fault identification method based on improved LightGBM hybrid ensemble model. *High Voltage Technology*, 1-13.

[6] Duan, X, Y. (2024) Research on the Improvement of LIBS Steel Classification Accuracy by Feature Selection. *Changchun University of Technology*.