Fault Diagnosis of Transformer Dissolved Gas Based on Random Forest

Geer Jing

School of Electrical and Electronic Engineering, North China Electric Power University, Beijing, China
qingfengsunshine@163.com

Abstract: To improve the accuracy of transformer fault diagnosis, this paper proposes a transformer fault diagnosis model based on the Random Forest (RF) algorithm. First, the dissolved gas analysis (DGA) method is used to preprocess the concentration data of characteristic gases dissolved in transformer oil. The Random Forest model is then constructed using Bootstrap sampling and random feature selection, achieving high-precision classification of transformer fault types through the integration of multiple decision trees. To validate the effectiveness of the model, actual transformer fault operation data are selected for testing. Additionally, the Random Forest model is compared with five other models: GNB, LDA, KNN, SVM, and GBDT. The results show that the Random Forest model significantly outperforms the other models in terms of both training and testing accuracy, while also demonstrating higher efficiency in training time. The comprehensive performance of the Random Forest model is superior to the comparison models. This study demonstrates that the Random Forest-based transformer fault diagnosis method can effectively handle complex dissolved gas data, accurately identify multiple fault types, and exhibit high generalization ability and robustness, making it suitable for practical transformer fault diagnosis in power systems.

Keywords: Random Forest, transformer, fault diagnosis, DGA

1. Introduction

As a critical component of modern power systems, transformers play a vital role in voltage level conversion. By 2023, oil-immersed transformers account for 89% of the market, dominating the industry. Therefore, understanding their real-time operating status and accurately diagnosing fault types are crucial for the stable operation of power grids [1].

When faults occur in oil-immersed transformers, gases such as H_2, CH_4, and C_2H_6 are generated, causing changes in the concentration of dissolved gases in the oil. This makes DGA the primary method for detecting the operating status of oil-immersed transformers. Traditional methods for diagnosing transformer faults, such as the three-ratio method [2], analyze the ratios of characteristic gas concentrations (C_2H_2/C_2H_4, CH_4/H_2, and C_2H_4/C_2H_6) to determine fault types. However, existing data-driven methods rely solely on the concentrations of characteristic gases (CH_4, C_2H_6, C_2H_4, C_2H_2, and H_2) for judgment, ignoring the ratio information between gases. In practice, different fault types alter both the concentration and the proportional relationships of these gases, making fault diagnosis based solely on gas concentrations inaccurate. In recent years, with the advancement of artificial

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

intelligence, data-driven diagnostic techniques have been continuously improved. Literature [3] proposes a fault diagnosis model based on Support Vector Machine (SVM) and Genetic Algorithm (GA), which uses SVM for fault classification and optimizes SVM hyperparameters using GA, achieving high accuracy on small datasets. However, SVM training is time-consuming, and GA has high computational complexity. Literature [4] introduces a clustering algorithm for analyzing DGA data to identify transformer fault types. While the algorithm is simple and computationally efficient, it is sensitive to initial conditions.

In summary, this paper proposes a transformer fault diagnosis model based on the Random Forest algorithm. This model can extract effective features from complex DGA data and achieve high-precision fault classification. As an ensemble learning algorithm, Random Forest offers advantages such as fast training speed, strong resistance to overfitting, and high interpretability, making it suitable for analyzing high-dimensional data classification problems.

First, the collected dissolved gas concentration data are preprocessed, including normalization, handling missing data, and calculating ratio relationships. Then, the Random Forest model is constructed, and fault type classification is achieved through the integration of multiple decision trees. Finally, the model is compared with other models using specific data, demonstrating its superiority in testing accuracy and training time, and proving its effectiveness in fault diagnosis.

2. Transformer Fault Diagnosis Method Based on Random Forest

2.1. Problem Modeling and Data Processing

The gases dissolved in transformer oil mainly include methane, ethane, ethylene, acetylene, and hydrogen. Based on the three-ratio method, this paper calculates the ratios of CH_4/H_2, C_2H_4/C_2H_6, and C_2H_2/C_2H_4, and uses these ratios along with the gas concentrations as input data. To eliminate the influence of different gas concentration scales, the gas concentration data are normalized using the min-max normalization method, as shown in Equation (1):

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

where $X_i^{'}$ is the data in the input sample, X_{min} and X_{max} are the minimum and maximum values of the corresponding gas concentration, and $X_i^{'}$ is the normalized data.

For missing data during collection, this paper adopts two approaches: for small amounts of missing data, they are directly deleted; for larger amounts of missing data, interpolation methods such as linear interpolation or spline interpolation are used. For data that cannot be estimated through interpolation, zero-filling is applied. The input data in this paper are the processed gas concentrations and ratio information, and the output is the transformer fault type, which is a multi-classification problem. Therefore, the output variable is set as Y, and the transformer fault types are divided into six categories and encoded, as shown in Table 1.

Table 1: Fault Type Encoding.

Fault Type	Encoding
Normal	0
Low-Temperature Overheating	1
Medium-Temperature Overheating	2
High-Temperature Overheating	3
Partial Discharge	4
Low-Energy Discharge	5
Arc Discharge	6

2.2. Random Forest Diagnosis Method

This paper adopts the Random Forest method for transformer fault diagnosis. Random Forest is an ensemble learning algorithm that uses multiple decision trees to form a combined classifier[5]. A decision tree is a tree-structured classification and regression method that builds decision rules by continuously splitting data features. The core idea of Random Forest is to construct a strong learner by integrating multiple weak learners (i.e., decision trees). Each decision tree is trained on different subsets of data and features, and these trees are independent of each other. During prediction, Random Forest aggregates the output results of all decision trees, using majority voting (for classification tasks) or averaging (for regression tasks) to obtain the final prediction result.

In transformer fault diagnosis, the workflow of Random Forest is as follows: First, Bootstrap sampling is used to generate multiple subsets from the original transformer fault dataset, and each subset is used to train a decision tree. During the training of each decision tree, for each node, a subset of features is randomly selected, and the optimal split point is determined using the Gini impurity. The Gini impurity is defined as shown in Equation (2), where p_i is the proportion of the i-th class sample in dataset D.

$$G(D) = 1 - \sum_{i=1}^{k} p_i^2 \tag{2}$$

When different features are used to train a tree, the Gini impurity varies. This paper adopts the principle of minimizing Gini impurity to determine the split point.

By continuously splitting nodes until certain stopping conditions are met (e.g., the number of samples in a node falls below a threshold or the tree depth reaches a limit), a decision tree is constructed. This process is repeated to build multiple decision trees, forming a Random Forest.

When new transformer data need to be diagnosed, the data are input into each decision tree in the Random Forest, and each tree provides a prediction result. The final fault type is determined by majority voting. For example, if 60 out of 100 decision trees predict "Low-Temperature Overheating," 20 predict "Medium-Temperature Overheating," and 20 predict other types, the final diagnosis result is "Low-Temperature Overheating."

Random Forest has many advantages. It is highly adaptable to large datasets, computationally efficient, and capable of handling high-dimensional data without the need for feature selection. Additionally, by integrating multiple decision trees, Random Forest effectively reduces the risk of overfitting, improves generalization ability, and exhibits strong robustness to noisy data and outliers. In transformer fault diagnosis, Random Forest can fully utilize historical and real-time monitoring data to accurately identify different fault types, providing strong support for the safe and stable operation of transformers.

3. Case Study

Historical data of dissolved gas concentrations in actual transformers are selected, including the six fault types mentioned above and normal state data, totaling 309 sets. Among these, 183 sets are used as the training set, and 126 sets are used as the test set. The processed data are input into the model. Additionally, five learning models—Gaussian Naive Bayes (GNB), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Gradient Boosting Decision Tree (GBDT)—are compared with the proposed model. The results are shown in Table 2 and Figure 1.

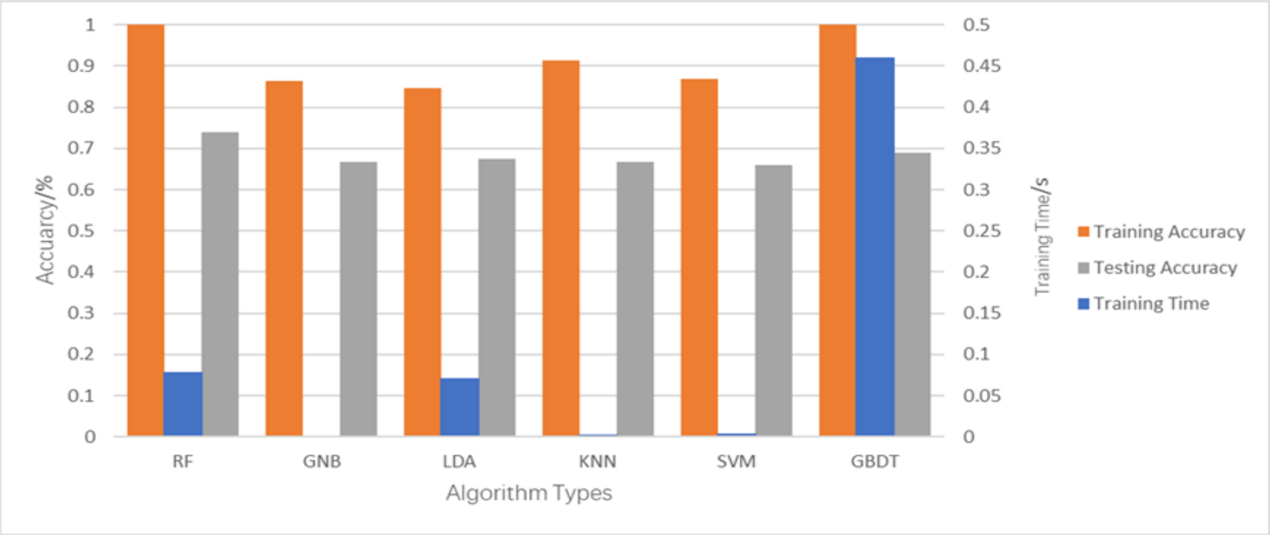

Figure 1: Comparison of Different Models with the RF Model.

Table 2: Comparison of RF with Other Models.

Model Name	Training Time (s)	Training Accuracy (%)	Testing Accuracy (%)
GNB	0.001	0.863	0.667
LDA	0.071	0.847	0.675
KNN	0.003	0.913	0.667
SVM	0.004	0.869	0.659
GBDT	0.461	1	0.69
RF	0.079	1	0.738

From the bar chart and table data, it can be observed that in terms of training accuracy, both the RF model and the GBDT model achieved a perfect fit of 100%, significantly outperforming other models (such as KNN at 91.3% and GNB at 86.3%). This indicates that the RF model is capable of effectively capturing the patterns in the training data. Secondly, in terms of testing accuracy, the RF model's 73.8% is the highest among all models, far exceeding GBDT's 69% and the approximately 66% of other models. This demonstrates that the RF model has stronger generalization capabilities, enabling it to better adapt to new data and avoid overfitting. Finally, regarding training time, the RF model's 0.079 seconds, while not as fast as GNB (0.001 seconds) and KNN (0.003 seconds), is significantly faster than GBDT's 0.461 seconds and remains within an acceptable range.

In summary, the RF model performs optimally in both training accuracy and testing accuracy. Although its training time is slightly longer, in practical applications, the model's accuracy and generalization capabilities are more critical. While the GBDT model also achieves high training accuracy, its training time is excessively long, and its testing accuracy is lower than that of the RF model. The GNB, KNN, LDA, and SVM models do not reach the level of the RF model in terms of either training or testing accuracy.

4. Conclusion

This paper proposes a transformer fault diagnosis method based on Random Forest. During the research process, to address the shortcomings of traditional methods and existing data-driven methods, the collected transformer gas concentration data are first normalized, and key gas ratios are calculated and used as input data along with gas concentrations. Then, the processed data are input into the Random Forest model for training, constructing multiple decision trees based on different subsets of training data and random feature selection. Finally, through actual test cases, the RF model is compared with other models, and various evaluation metrics such as testing accuracy and training time are analyzed. The results validate the effectiveness and accuracy of the Random Forest model in transformer fault diagnosis, achieving accurate identification of six fault types and normal states, providing reliable support for the safe and stable operation of transformers.

References

[1] Wang, X.J. (2024) Research on Fault Diagnosis of Oil-Immersed Transformers. Industrial Innovation Research, (24), 44-46.

[2] Wan, Q.H., Ye, K.D. and Zhang, S. (2024) Transformer Oil Chromatography Analysis Based on the Three-Ratio Method. Electrical Switchgear, 62(03), 65-67.

[3] Zheng, H., Li, Y., et al. (2019) A Hybrid Approach for Transformer Fault Diagnosis Using Support Vector Machine and Genetic Algorithm. IEEE Transactions on Power Delivery, 34(2), 789-797.

[4] Shao, L., Yu, J.J., Liu, H.L., et al. (2025) Transformer Fault Diagnosis Method Based on Hierarchical Clustering and Improved Support Vector Machine. Journal of Tianjin University of Technology, 1-8. Retrieved March 10, 2025, from https://kns-cnki-net.webvpn.ncepu.edu.cn/kcms/detail/12.1374.N.20241202.1110.021.html

[5] Li, W.J., Li, T.Y. and Su, J. (2025) Series Arc Fault Detection Method for Photovoltaic DC Side Based on Random Forest. Distribution & Utilization, 42(02), 108-115. DOI:10.19421/j.cnki.1006-6357.2025.02.012