Proceedings of CONF-APMM 2025 Symposium: Controlling Robotic Manipulator Using PWM Signals with Microcontrollers
DOI: 10.54254/2753-8818/2025.AD26526

Fault Diagnosis of Dissolved Gas Analysis for Transformer Based on XGBoost

Chuqi Yang

School of Management and Engineering, Nanjing University, Nanjing, China
231870111@smail.nju.edu.cn

Abstract. Aiming at the fault diagnosis problem of oil-immersed transformers, this paper proposes a transformer oil dissolved gas analysis (DGA) fault diagnosis method based on the XGBoost algorithm. Traditional diagnostic methods have defects such as vague boundary values and poor interpretability. However, the XGBoost algorithm can effectively capture the nonlinear relationship between DGA data and fault types by iteratively optimizing the additive model. In the study, 1260 sets of DGA data were preprocessed, including mean normalization to eliminate the influence of dimensions, construction of gas ratio features to enhance sensitivity, and conversion of fault types into numerical labels. The model was optimized by setting hyperparameters such as learning_rate and max_ depth, and using 5-fold cross-validation and early stopping mechanism. Experimental results show that the accuracy of the XGBoost model on the test set reaches 93.6%, which is significantly higher than that of LSTM (82.3%) and PSO-LSTM (85.7%), and the RMSE and MAE indicators are better. The research shows that this method can accurately diagnose transformer faults and provide effective technical support for the safe operation of power systems.

Keywords: Transformer fault diagnosis, DGA, XGBoost algorithm, Gradient boosting decision tree, Hyperparameter setting

1. Introduction

Transformers, as essential equipment in power systems, play a crucial role in power generation, transmission, and substation operations. They are key devices for the efficient transmission and distribution of electrical energy. Transformers are primarily classified into oil-immersed and dry-type categories, with the former dominating the market in China, accounting for nearly 90% of the market share. As the service life of transformers increases, issues such as overheating and discharge faults may occur [1]. Accurately diagnosing the fault type is the prerequisite for targeted fault treatment, which is of significant importance for ensuring the safety of electricity supply. These faults often lead to changes in the dissolved gas-in-oil (DGA) content, making DGA data analysis a viable method for fault diagnosis.

Traditional fault diagnosis methods include the three-ratio method and data-driven approaches. The former involves collecting DGA data from transformers and calculating three key gas concentration ratios, which are then used to classify fault types via a coding scheme [2]. The latter

© 2026 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/).

DOI: 10.54254/2753-8818/2025.AD26526

utilizes algorithms such as machine learning or deep learning, where the collected DGA data and their corresponding fault types or states are used as training samples for model training, allowing the model to learn the mapping relationship between DGA data and faults [3].

Although traditional methods assist in determining the fault type of transformers, they all have limitations. The three-ratio method suffers from issues such as ambiguous boundary values, blind spots in coding, and large measurement errors when gas concentration is low. Data-driven approaches face challenges such as data imbalance, missing data, poor model interpretability, and overfitting. These limitations motivate the adoption of artificial intelligence (AI) methods for DGA diagnosis. Compared to traditional methods, AI-based approaches offer advantages such as faster speed and better accuracy and robustness in fault analysis.

Although AI technology has achieved certain results in the field of transformer DGA fault diagnosis, there is still room for improvement in existing research. Some methods still need to enhance diagnostic accuracy when dealing with complex fault patterns; in cases of data imbalance or noise, the stability and generalization ability of the models are insufficient. Moreover, most studies lack in-depth interpretation of diagnostic results, which makes it difficult to meet the practical requirements of fault root cause analysis in engineering. Therefore, exploring more efficient, accurate, and highly interpretable transformer DGA fault diagnosis methods is of great practical significance for ensuring the safe and stable operation of power systems. This paper, based on the XGBoost algorithm, conducts research from various aspects such as data preprocessing, model construction and optimization, and result evaluation and interpretation, aiming to provide better solutions for transformer fault diagnosis.

2. XGBoost method

2.1. Problem description

Oil-immersed transformers are critical equipment in power systems, and their operational status directly affects the stability of the power grid. Internal faults in transformers, such as partial discharge, overheating, and arc discharge, can lead to the decomposition of insulating oil, generating characteristic gases such as hydrogen (H_2), methane (CH_4), acetylene (C_2H_2), ethylene (C_2H_4), and ethane (C_2H_6). Traditional fault diagnosis methods, such as the IEC three-ratio method and the Duval triangle method, rely on expert experience and suffer from drawbacks such as strong subjectivity and poor adaptability.

In recent years, artificial intelligence (AI) technologies, such as machine learning and deep learning, have shown significant advantages in transformer fault diagnosis. This paper proposes a data-driven DGA fault diagnosis method based on the XGBoost (Extreme Gradient Boosting) algorithm to improve classification accuracy and generalization ability [4,5].

2.2. Data preprocessing

The original DGA data often suffer from issues such as noise, dimensional differences, and sample imbalance, which require the following preprocessing steps:

Mean Normalization is applied to eliminate the effect of dimensional differences. The formula is as follows:

$$x^{'} = \frac{x}{\mu}, \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

DOI: 10.54254/2753-8818/2025.AD26526

Where x represents the original gas concentration, μ is the mean of the training set, and x' is the normalized feature.

In addition to the original gas concentrations, gas ratio features (such as CH_4/H_2, C_2H_2/C_2H_4) are introduced to enhance the model's sensitivity to fault patterns.

Fault types (such as normal, partial discharge, overheating) need to be converted into numerical labels (e.g., 0, 1, 2) to adapt to the classification model.

2.3. Principle of the XGBoost method

XGBoost is an improved algorithm of Gradient Boosting Decision Tree (GBDT). Its core idea is to iteratively optimize an additive model, gradually improving prediction performance by minimizing the loss function.

2.3.1. Objective fuction

The objective function of XGBoost Method consists of the loss function and regularization:

$$L(\emptyset) = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

In which:

• $l\left(y_i, \widehat{y_i}\right)$ is the loss function(such as cross-entropy),which measures the difference between the predicted value $\widehat{y_i}$ and the true value y_i .

• $\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\lambda \| \omega \|^2$ is the regulation term, T is the number of leaf nodes, ω is the leaf weight, and γ and λ are hyperparameters.

2.3.2. Gradient boosting process

At the t-th iteration, the model's predicted value is:

$$\widehat{y_i}^{(t)} = \widehat{y_i}^{(t-1)} + f_t\left(x_i\right) \tag{3}$$

Where f_t is the newly trained tree. The objective function is approximated using a second-order Taylor expansion:

$$L^{(t)} \approx \sum_{i=1}^{n} \left[l\left(y_i, \widehat{y_i}^{(t-1)}\right) + g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right) \right] + \Omega\left(f_t\right) \tag{4}$$

In which

$$g_i = \frac{\partial}{\partial_{\widehat{y_i}^{(t-1)}}} l\left(y_i, \widehat{y_i}^{(t-1)}\right) \tag{5}$$

$$h_i = \frac{\partial^2}{\partial_{\widehat{y_i}^{(t-1)^2}}} \tag{6}$$

are the first-order and the second-order gradients, respectively.

DOI: 10.54254/2753-8818/2025.AD26526

2.3.3. Optimal leaf weight calculation

For each leaf node j, the optimal weight ω_j^* is:

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{7}$$

Where I_j belongs to the sample set of j.

2.4. Model training and optimization

2.4.1. Hyperparameter settings and training process

In the training and optimization of the model, hyperparameter settings should focus on learning_rate (which controls the weight reduction of each tree to prevent overfitting), max_depth (the maximum depth that affects the complexity of a single tree), subsample (the sample sampling ratio that enhances generalization ability), and lambda (the L2 regularization coefficient). The training process includes splitting the dataset into training and testing sets in an 80:20 ratio, selecting the optimal hyperparameters through 5-fold cross-validation, and using early stopping (where training is terminated if the performance on the validation set does not improve within 10 rounds) [6].

2.4.2. Classification decision

The final prediction result is the weighted output of all the trees:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i) \tag{8}$$

Converted into a probability distribution through the softmax function:

$$P(y = cx) = \frac{e^{\hat{y}_c}}{\sum_{j=1}^{C} e^{\hat{y}_i}} \tag{9}$$

Where C is the number of fault categories.

3. Example analysis

3.1. Dataset overview and preprocessing

The transformer DGA dataset used in this example contains 1,260 sample data points, covering 7 typical fault types, including normal operation, partial discharge, overheating, and others. The original data suffers from issues such as inconsistent dimensions and feature redundancy. To improve the model's training performance, mean normalization is first applied to dimensionless the data and eliminate the dimensional differences between various gas concentration features. Next, gas ratio features are constructed to explore the underlying relationships between gas concentrations. Finally, the dataset is split into a training set (1,008 samples) and a testing set (252 samples) in an 80:20 ratio. The training set is randomly shuffled to ensure the randomness of the data distribution and prevent any bias in model training.

DOI: 10.54254/2753-8818/2025.AD26526

3.2. Network parameter settings

This study primarily compares three methods: XGBoost, LSTM, and PSO-LSTM. The XGBoost model is configured with a learning_rate of 0.1, a maximum tree depth (max_depth) of 6, a sample sampling ratio (subsample) of 0.8, and an L2 regularization coefficient (lambda) of 0.1. The LSTM model is built with a 2-layer network structure, each layer containing 64 neurons, and the learning rate is set to 0.001 [7]. The PSO-LSTM method uses the Particle Swarm Optimization (PSO) algorithm to optimize parameters such as the learning rate and the number of hidden layer neurons of LSTM, in order to improve model performance [8].

3.3. Evaluation metrics and result comparison

Accuracy, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are used as model evaluation metrics. The experimental results show that the XGBoost model achieves an accuracy of 93.6%, with an RMSE of 0.12 and an MAE of 0.08 on the test set. The LSTM model has an accuracy of only 82.3%, with an RMSE of 0.25 and an MAE of 0.18. The PSO-LSTM model, optimized by PSO, shows an accuracy improvement to 85.7%, with an RMSE reduced to 0.22 and an MAE of 0.15 [9,10]. Compared to the other two methods, XGBoost performs the best in all metrics, fully demonstrating its effectiveness in transformer DGA fault diagnosis.

Table 1. Evaluation metrics and result comparison

Method	Accurray	RMSE	MAE
XGBoost	93.6%	0.12	0.08
LSTM	82.3%	0.25	0.18
PSO-LSTM	85.7%	0.22	0.15

3.4. Result summary

Based on the comprehensive experimental results, the XGBoost model demonstrates significant advantages in transformer DGA fault diagnosis tasks. Compared to LSTM and its optimized model (PSO-LSTM), XGBoost not only achieves an accuracy improvement of 6.5%-11.3%, but also significantly reduces both RMSE and MAE, reflecting stronger prediction accuracy and stability. This result can be attributed to the gradient boosting mechanism and regularization design of XGBoost, which enables it to better capture the nonlinear relationship between DGA data and fault types, while effectively preventing overfitting. Moreover, XGBoost has superior training and testing efficiency compared to deep learning models, making it more suitable for real-time fault diagnosis scenarios in engineering practice. Therefore, the XGBoost method provides a more reliable and efficient solution for transformer DGA fault diagnosis.

4. Conclusion

This paper focuses on the oil-immersed transformer DGA fault diagnosis problem and proposes a solution based on the XGBoost algorithm. The system completes tasks such as data preprocessing, model construction, training optimization, and performance verification. The main conclusions are as follows:

In terms of method design, addressing the drawbacks of traditional fault diagnosis methods, such as strong subjectivity and poor adaptability, and leveraging the advantages of the XGBoost

DOI: 10.54254/2753-8818/2025.AD26526

algorithm in classification tasks, a fault diagnosis model based on DGA data as input is constructed. Mean normalization is applied to eliminate dimensionality effects, gas ratio features are introduced to enhance fault pattern sensitivity, and fault types are converted into numerical labels, effectively improving data quality and model adaptability.

In model optimization, hyperparameters such as learning_rate and max_depth are reasonably set, and 5-fold cross-validation along with an early stopping mechanism are employed to achieve a balance between model complexity and generalization ability. The example analysis shows that the proposed XGBoost model performs excellently on a dataset containing 1,260 samples, with a test set accuracy of 93.6%, significantly outperforming LSTM (82.3%) and PSO-LSTM (85.7%), and with better RMSE and MAE metrics. This verifies its efficiency and stability in fault diagnosis tasks.

The study confirms that the XGBoost algorithm can effectively capture the nonlinear relationships between DGA data and fault types, providing an accurate and reliable technical solution for transformer fault diagnosis. Future research can further expand the dataset size, deepen feature engineering by incorporating fault mechanisms, and explore model interpretability methods to better meet practical engineering needs.

References

[1] Wang Hao. Common Faults and Handling Methods of Oil-immersed Transformers [J]. Shandong Industrial Technology, 2018, (22): 184.
[2] Lin Beimin. Case Analysis of Typical Transformer Faults Based on the Three-ratio Method [J]. Electrical Technology, 2023, 24(10): 63-67.
[3] Hu Daofu, Wen Shanshan, He Yiming. Transformer Fault Diagnosis Based on BP Neural Network and Its Application [J]. Journal of Electric Power Science and Technology, 2008, (02): 72-75.
[4] Hong Yitian, Chen Weihua, Chen Tongzhu, et al. Operation Fault Diagnosis Method for Large Power Transformers Based on XGBOOST Algorithm [J]. Electrical Engineering, 2024, (11): 56-59.
[5] Ruan Yi, Zhang Haotian, Sun Jian, et al. A NRBO-XGBoost Transformer Fault Diagnosis Method Based on DGA [J]. Journal of Chaohu University, 2024, 26(06): 87-93+128.
[6] Jia Haoyang, Qian Yu. Transformer Fault Diagnosis Based on Bayesian Optimized XGBoost Algorithm [J]. Journal of Yellow River Conservancy Technical Institute, 2023, 35(02): 37-43.
[7] Li Yiming, Liang Zhiqing, Xu Mian. Malicious DGA Domain Name Detection Based on LSTM and Attention Mechanism [J]. Network Security Technology and Application, 2024, (12): 32-34.
[8] Chen Zesheng, Zhou Min, Feng Lichun, et al. Malicious Domain Name Recognition of DGA Based on XGBoost and Particle Swarm Optimization Algorithm [J]. Journal on Communications, 2024, 45(S2): 27-32
[9] Wang Mingyang, Ma Xuejun, Ge Lijuan, et al. Research on Transformer Fault Diagnosis Based on KOA-CNN-LSTM [J]. Journal of Inner Mongolia Agricultural University (Natural Science Edition), 2025, 46(04): 65-73
[10] Jia Rubin, Zhang Yajun, Tian Feng, et al. Application of Convolutional Neural Network in Fault Diagnosis of Oil-immersed Transformers [J]. Electrical Engineering, 2024, (10): 89-93.