Research on main transformer defect detection methods based on Conditional Inference Tree and AdaBoost Algorithm

Chenmeng Zhang1,*, Can Hu2, Zongxi Zhang1 and Shuping Cao1
1State Grid Sichuan Electric Power Research Institute, Chengdu, Sichuan, China
2State Grid Sichuan Electric Power Company, Chengdu, China

*Corresponding author e-mail: 916188925@qq.com

Abstract. With the development of data science, there are more and more ways to dig for patterns hidden in the data. If we can apply the advanced model of data science to the power data, we can mine the potential value of power data. In this paper, the monitoring and inspection data of 110kV main transformers and related basic environmental data are fused to establish conditional inference tree model and AdaBoost algorithm to evaluate whether the 110kV main transformers have defects and faults. The accuracy of the two algorithms is compared. Finally, we select the AdaBoost algorithm with higher accuracy for building the 110kV main transformer fault evaluating model. This paper provides a reference for the automatic detection of power grid faults, and provides new ideas for the application of power grid big data.

1.Introduction
The development of power grid automation has made many equipment monitoring systems launched one after another, and the tests for power equipment have increased as well. At the same time, the type and quantity of power grid data have been continuously enriched. With the continuous enrichment of power equipment data, the relationship between variables has become more and more complicated, which makes the accuracy of traditional data modeling for equipment fault evaluation getting lower and lower, leading to inefficient use of data. To deal with problems mentioned above, we need to consider algorithms that perform better in high-dimensional data modeling. The conditional inference tree and the AdaBoost algorithm exactly have a good performance on high-dimensional data classification. Therefore, we consider that these two algorithms into consideration to mine the potential value of grid data efficiently and form an evaluation model for the main transformer fault. Based on the power equipment test data and related environmental data of all 110kV main transformers in Guangxi province, this paper firstly carried out data fusion and data preprocessing on the obtained data sets, and then established two defect and fault risk evaluation models based on conditional inference tree and AdaBoost for the main transformer Scheme, and finally implemented the algorithm and analyzed the experimental results.

2.Background
At present, there are lots of methods for transformer fault diagnosis in which the key gas method and IEC three-ratio method are commonly used in practice [1]. Although these methods have effect to some extent, there are problems such as low accuracy and low speed. In response to this problem, in
recent years, scholars have proposed some machine learning algorithms to upgrade those existing diagnostic methods. However, there are still some shortcomings, such as that ordinary decision tree algorithm is difficult to make full use of data, causing insufficient information and that fuzzy clustering method is not ideal for large-scale sample classification [2]. Similarly, SVM is also difficult to apply to large-scale samples [3]. Therefore, a method based on the traditional decision tree is proposed where the significance of the relationship between the dependent and independent variables is used to classify. Meanwhile, we try to use the AdaBoost algorithm to continuously update the model.

When running, the insulating oil from main transformers will decompose a lot of gases suffering from electricity and being heated, such as H2, CH4, C2H6, C2H4, C2H2, etc. [4]. Through spectral analysis of the gases decomposed, we can obtain the data related to the running condition of the main transformers. What’s more, the recording methods of power grid basic data and related environmental data have also made great progress, which has enriched the data that can be used to evaluate transformer faults.

This paper intends to evaluate and analyze power equipment status based on fused power-environmental-parameters data and compares two classification algorithms.

3. Variable settings and data preprocessing

Since the information about the main transformers is stored in many different databases, and the recording time point of different datasets is different, so the fusion of attributes is very difficult. Therefore, how to perform data fusion is the key to data preprocessing. After fusing various data sources, we need to use effective data modeling methods to make the data effectively used.

In a word, to evaluate the defects and faults of power transmission and transformation equipment, it is necessary to correlate and fuse the data first, and then to model the data appropriately. The logic is shown in the figure below.

Figure.1. Data processing and modeling

In data preprocessing, since each main transformer has a unique ID number, we first established the fused working condition data with the main transformers’ ID as the primary key. After that, we recorded the faults of the main transformers into our data set according to the history inspection information of the main transformers. Next, we added the basic information and environmental
information of the power grid into our data set to form a fused power-environmental-parameters data set. Due to the high frequency of power grid data records, there was a large amount of duplicate data to deal with. For the data from the same time period, we kept one left and deleted other duplicate records. After we finished the data cleaning, we need to deal with the problem with missing values. Due to the fact that the data from different systems cannot be completely matched at the time, some attributes in some records are missing. To address this problem, we use the average of each attribute to fill in the missing values. After data preprocessing, the variable settings are shown in Table 1.

### Table 1 Variable settings

| Symbol | Meaning | Attribute  |
|--------|---------|-----------|
| Con    | 110kV main transformer condition | categorical/dependent |
| V      | Oil breakdown voltage (kV) | numeric/regressor |
| FP     | Oil closing flash point (°C) | numeric/regressor |
| MW     | Oily moisture content (mg/L) | numeric/regressor |
| TEMP   | Sampling oil temperature (°C) | numeric/regressor |
| H2     | Hydrogen by oil chromatography | numeric/regressor |
| C2H2   | Acetylene by oil chromatography | numeric/regressor |
| C2H4   | Ethylene by oil chromatogram | numeric/regressor |
| CH4    | Methane by oil chromatographic | numeric/regressor |
| C2H6   | Ethane by oil chromatography | numeric/regressor |
| CO     | CO by oil chromatography | numeric/regressor |
| CO2    | Carbon dioxide by oil chromatography | numeric/regressor |
| ACID   | Oil acid ester (mgKOH/g) | numeric/regressor |
| PH     | Ph of oil | numeric/regressor |

We randomly select 80% of the data in the data set as the training data set to estimate model parameters, and the remaining 20% of the data as the test data set to check the accuracy of the model.

4. Model theory analysis

Conditional inference tree is an improved model of traditional decision tree. The selection of conditional inference tree variables and segmentation is based on the significance test. The steps to generate a conditional inference tree are as follows.

1. Calculate the p-value for the relationship between the dependent variable and each regressor.
2. Select the variable with the smallest p value.
3. Try all possible binary splits on the selected variable and choose the split that minimizes the loss function Gini (p).

$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

Each division corresponds to a loss function value, and we choose the discrimination condition that minimizes the Gini coefficient as the current node.

Among the expression of Gini, $K$ represents the total number of categories existing. In category $k$, $p_k$ represents the ratio of the number of records belonging to category $k$ which satisfy the current discriminant condition and the number of records belonging to category $k$ in all data that meets the forward discrimination conditions.

4. Divide the data set into two groups and repeat the above steps for each subgroup.
5. Repeat until all splits are not significant or the smallest node has been reached.

$$F_M(x; P) = \sum_{m=1}^{n} \beta_m h(x; a_m)$$

We randomly select 80% of the data in the data set as the training data set to estimate model parameters, and the remaining 20% of the data as the test data set to check the accuracy of the model.
Among them, \( h(x; \alpha_m) \) is a weak classifier, \( \alpha_m \) is the optimal parameter learned by the weak classifier, \( \beta_m \) is the proportion of the weak classifier in the overall strong classifier, \( P \) represents all combination. These weak classifiers are linearly combined to form a strong classifier.

AdaBoost put emphasis on the samples that were misclassified in the previous model, which means that it reduces the weights of the samples that were correctly classified and increases the weights of the samples that were misclassified in the next model. Then, models are formed based on some basic machine learning algorithms. For the weak classifier we have learned, we increase the weight of the weak classifier with a small classification error rate, and reduce the weight of the weak classifier with a large classification error rate. In other words, the generation process of AdaBoost model is a process of verifying, improving and integrating based on existing machine learning algorithms.

5. Model result analysis

As mentioned above, we apply the conditional inference tree and random forest model to the training data set, and obtain the following experimental results.

![Figure 2. Conditional inference tree model results](image1)

![Figure 3. The relationship between the AdaBoost error and the number of weak classifiers](image2)

From the conditional inference tree model, we can see that the temperature of the sampled oil, the content of methane and carbon dioxide in the sampled oil chromatogram have a strong correlation with the condition of main transformers.
As can be seen from the error of the AdaBoost model, when we set the number of weak classifiers to 4, the model's error reaches the lowest point. Therefore, in AdaBoost model establishing, we set the parameter of the number of weak classifiers to 4.

Next, we apply the trained model to the test data set to illustrate the accuracy of the two models in evaluating the state of the main transformers.

Table 2 Test results of the conditional inference tree model on the test data set

| Actual | Predicted | Accuracy |
|--------|-----------|----------|
| No     | 92        | 0        | 1        |
| Yes    | 15        | 0        | 0        |
|        |           |          | 0.860    |

Table 3 Test results of AdaBoost model on the test data set

| Actual | Predicted | Accuracy |
|--------|-----------|----------|
| No     | 82        | 10       | 0.891    |
| Yes    | 8         | 7        | 0.467    |
|        |           |          | 0.832    |

As can be seen from the above table, the comprehensive accuracy of the conditional inference tree algorithm is 0.860, and that of AdaBoost model is slightly lower, reaching 0.832. However, compared to predicting the normal states of the main transformers, it’s more meaningful to evaluate the faulty main transformers, so we are more concerned about the accuracy of the fault evaluation. The AdaBoost algorithm can achieve an accuracy of 0.467 in predicting faults. In the contrary, the conditional inference tree model does not play any role in faults evaluation. Therefore, we believe that the AdaBoost algorithm performs much better than the conditional inference tree algorithm in main transformer fault evaluation, which means that AdaBoost algorithm is more suitable for the prediction of 110kV main transformer defects and faults.

6. Conclusion
This paper did a research on 110kV main transformer defects and fault evaluation based on conditional inference trees and AdaBoost algorithm and we can learn from the research that the model based on AdaBoost shows higher accuracy and that the best number of weak classifiers when predicting main transformer defects and faults is 4.

This study provides new ideas for the fault prediction of the power equipment with fusion grid-equipment-environment data, and a reference for more efficient models for accurate assessment of the status of the power equipment in the future.

7. Acknowledgment
This work was supported by Sichuan Electric Power Company Research Project 52199718002Q.

References
[1] Yang T, Liu P, Li J, et al. FCM combined with IEC ratio method to diagnose transformer faults[J]. High Voltage Engineering, 2007, 33 (8): 66-70.
[2] Li Y, Shu N. Transformer fault diagnosis based on fuzzy clustering and complete binary tree support vector machine [J]. Transactions of China Electrotechnical Society, 2016, 31(04): 64-70.
[3] Wang S, Liao R. Transformer fault diagnosis method based on AdaBoost optimization cloud theory[J]. High Voltage Engineering, 2015, 41(11): 3804-3811.
[4] Tao W, Gu B, Xu X, Liu L. Fault detection of power grid transformer based on adaptive RBF neural network[J]. Bulletin of Science and Technology, 2019, 35(12): 110-113.
[5] Zhu D. Research on common transformer faults and their diagnosis techniques[J]. Technology Wind, 2017(15): 171.
[6] Li E, Wang L, et al. Chromatographic analysis of transformer oil based on improved fuzzy
clustering algorithm [J]. Transactions of China Electrotechnical Society, 2018, 33(19):4594-4602.

[7] Zheng X, Zhao F, et al. Evaluation system of reactive power operation of low voltage distribution network based on big data [J]. Power System Technology, 2017, 1: 038.

[8] Sun Y, Gao H, et al. Operation status evaluation model of distribution transformer based on multi-time information fusion [J]. High Voltage Engineering, 2016, 42(7): 2054-2062.

[9] Deng S, Yue D, Zhu L, et al. Intelligent and efficient analysis and mining technology framework of power big data [J]. Journal of electronic measurement and instrumentation, 2016, 30(11): 1679-1686.

[10] Wang W, Liu M, Yu Z, et al. Design of big data center architecture in electric power big data environment [J]. Power Information and Communication Technology, 2016, 14(1): 1-6.

[11] Liu Y, Hu F, Gu L, et al. Analysis on causes of total hydrocarbon exceeding the standard in 110kV transformer [J]. Electrical Engineering, 2014(01): 81-83.

[12] Zheng L, Li S, Wang X, et al. Insulation state evaluation of power transformer based on optimal variable weight normal cloud model [J]. High Voltage Apparatus, 2016, 52(2): 85-92.

[13] Zhang Y, Kou L, Sheng W, et al. Big data analysis method of operation status evaluation of distribution transformer [J]. Power System Technology, 2016, 40(3): 768-773.

[14] Xue Y, Lai Y. The fusion of big energy thinking and big data thinking [J]. 2016.

[15] Wang D, Zhou Q. The invention relates to a distributed on-line analytical processing method for big data of power equipment state monitoring [J]. Proceedings of the CSEE, 2016 (19): 5111-5121.

[16] Ding D, Gao W, Liu W. Study on the test scheme of uhf detection sensitivity of partial discharge in GIS [J]. High Voltage Apparatus, 2014, 8: 003.

[17] Bai C, Gao W, Jin L, et al. Study on the comprehensive stress life model of transformer [J]. Journal of Sichuan University (Engineering Science), 2013, 2.

[18] Gao W, Zhao D, Ding D, et al. Investigation of frequency characteristics of typical PD and the propagation properties in GIS [J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2015, 22(3): 1654-1662.

[19] Zhou R, Gao W, Zhang B, et al. Prediction of Tropical Cyclones’ Characteristic Factors on Hainan Island Using Data Mining Technology [J]. Advances in Meteorology, 2014, 2014.

[20] Gao W, Ding D, Liu W, et al. Investigation of the Evaluation of the PD Severity and Verification of the Sensitivity of Partial-Discharge Detection Using the UHF Method in GIS [J]. IEEE Transactions on Power Delivery, 2014, 29(1): 38-47.

[21] Gao W, Ding D, Liu W, et al. Propagation attenuation properties of partial discharge in typical in-field GIS structures [J]. IEEE Transactions on Power Delivery, 2013, 28(4): 2540-2549.

[22] Liu J, He J, Hu J, et al. Statistics on the AC ageing characteristics of single grain boundaries of ZnO varistor [J]. Materials Chemistry and Physics, 2011, 125(1): 9-11.

[23] Gao W, Zhang B, Zhou R, et al. Prediction of thunderstorm cloud trend based on lightning location system monitoring data [J]. Power System Technology, 2015, 39(2): 523-529.

[24] Hu J, He J, Long W, et al. Temperature Dependences of Leakage Currents of ZnO Varistors Doped with Rare Earth Oxides [J]. Journal of the American Ceramic Society, 2010, 93(8): 2155-2157.