Comparative Study on Defects and Faults Detection of Main Transformer Based on Logistic Regression and Naive Bayes Algorithm

Can Hu¹, Chenmeng Zhang²*, Zongxi Zhang², Shijun Xie²

¹State Grid Sichuan Electric Power Company, Chengdu, China
²State Grid Sichuan Electric Power Research Institute, Chengdu, Sichuan, China

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

Abstract. The proper work of main transformers plays an important role in the stability of the power grid. The development of power data allows us to evaluate the instant status of the main transformer. If we use appropriate data classification algorithm, the laws and values hidden behind the mass power data can be discovered. This paper took basic power data, main transformers’ status data and oil chromatography data into consideration, and used Logistic regression and Naive Bayes algorithm to evaluate 110kV main transformers’ defects and faults from regression analysis and probability statistics separately. The results of these two algorithms were obtained and compared. Finally, a more accurate algorithm of evaluating main transformer fault was selected, which provided a new reference for the application of power big data.

1. Introduction

There are a large number of transformers around the world, and many of them have been in use for many years, which causes many defects and faults. Accurate diagnosis of whether there is a fault inside the transformer is of great significance to make sure the entire power grid in order. With the development of the smart power grids, the monitoring for equipment has been improved, and the tests and inspections for various types of equipment have been increased. As a result, the quality of power data, including the main transformers’ data, has been greatly improved. As the increase in the number and types of power equipment data, traditional multiple linear regression models cannot well describe the relationship between variables, such as the cubic graph method, electrical collaborative research method, four ratio method and David triangle method. Especially, the improved three-ratio method proposed by Cao Dunkui [1] and other scholars has become the industry standard for diagnosing transformer faults in China. However, the rule-based diagnosis method requires high integrity and accuracy of the monitoring data, and it is difficult to process the boundary data, and the lack of data is not allowed. Therefore, the data utilization rate is low, and the accuracy of the model evaluation results is not high as well, which leads to a waste of data resources. To address this problem, Wu Zhongli [2] and other scholars used rough sets and SVM to evaluate faults, but the model efficiency could not meet the needs of processing mass data. In our paper, we use some other machine learning algorithms to see whether they can better fit the trend of electric power data, and better mine the internal
relationship between electric power data and main transformer faults, and form a flexible evaluation mechanism.

Based on the data which contains all the information for 110kV main transformers in the whole power grid of Guangxi Province, this paper firstly carried out variable screening and data preprocessing, and then fitted two defaults assessment models based on Logistic regression and Naive Bayes algorithm separately. The accuracy of the two models were evaluated and compared.

2. Background

Data processing technology is becoming more and more maturing. At the same time, a large amount of data will be generated when the power system is running. It is suitable for processing big data of electricity with big data method, so that big data of electricity can provide the most value. The types and amount of data about power transformation equipment are diverse [3]. The log data generated during the running and maintenance of power transformation equipment includes tests, overhauls, defects, technical changes, movements, and return shipments, which is more than 100,000 lines each year [4]. However, the amount of online monitoring data such as the working status of the primary and secondary equipment is even larger. The real-time monitoring data of one province’s power transmission and transformation equipment (refresh every 15-minute in average) has more than 20 million measurement points, generating more than 100 million records every day, which can generate about 10G of data after compression and storage. Meanwhile, high-frequency real-time monitoring data can generate hundreds of terabytes of data each year [5], and some unstructured monitoring data such as smart substation online video monitoring data can generate PBs of data every year. The environmental condition while working of power transmission and transformation equipment also affects its performance and life span. Those data covers meteorological data like typhoon, thunder and lightning, ice coating, GIS terrain data, pollution monitoring, remote sensing data, the accumulated data, which has generated 50TB and about 100G of data is added annually.

Logistic regression is a regression model in which the dependent variable ranges from 0 to 1. It is suitable for the estimation and prediction of probability, so that we can used it to estimate the happening probability of defaults in main transformers. Logistic regression classifies by fitting a curve of predicted probability, while Naive Bayes evaluates classification by considering probability distribution and realizing conditional probability maximization.

Therefore, this paper intends to compare two classification algorithms to select a better model to evaluate and analyze the defaults of main transformers based on merging data from several data sources.

3. Variable settings and data preprocessing

The complex and mass main transformer information is stored in different databases, and the recording time nodes of different attributes are different, which makes cross-system association difficult. Therefore, in order to evaluate the defects and faults of power transmission and transformation equipment, it is necessary to correlate and fuse the data first, and then model the data appropriately. After merging the data from all resources, we must use unified information management system and effective data mining and analysis methods to make the data used effectively. Figure 1 shows the whole process described above:
In data preprocessing, since each main transformer has a unique ID number, we first combined the data with the main transformer ID as the primary key. We recorded the faults of the main transformers into our data set according to the history inspection information of the main transformers. Secondly, we added environmental information and the basic information of the power grid into the data set according to the corresponding region, ID number and time. Due to the high frequency of data records, we got a lot of superfluous data in the same time period. As a result, for the data from the same time period, we only keep one and delete other duplicate records. Since attributes recorded in different systems cannot correspond exactly, each record may be incomplete, so there are missing values in the data set. To solve this issue, we use the average value of each attribute to fill in missing values. After data preprocessing, we set variables like shown in Table 1.

### Table 1 variable Settings

| Symbol | Meaning                                      | Attribute          |
|--------|----------------------------------------------|--------------------|
| 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  |
| Freq   | Vibration frequency of tubing (Hz)           | numeric/regressor  |
| Stat   | 110kV main transformer condition            | categorical/depende|

We randomly select 80% of the data in the data set as the training set to estimate model parameters, and the remaining 20% of the data as the test set to illustrate the accuracy of the model.
4. Model theoretical analysis

The difference between logistic regression and linear regression is that the dependent variable in Logistic is a categorical variable rather than a quantitative variable, which means that it has a better use scenario in analyzing influencing factors. In logistic regression, we assume that each record is independent and classify transformers by a model ranging from 0 to 1, estimated by MLE. The general form of the model is as follows.

$$
\Pr(y = 1|x = x_i) = \frac{\exp(\beta^T x_i)}{1 + \exp(\beta^T x_i)}
$$

Figure 2 Schematic diagram of Logistic regression function

First, we estimate $\beta^T$ through MLE and substitute the obtained $\beta^T$ into the model. Then, we estimate the probability of the occurrence of the event according to the given regressor $x_i$.

Different from the logistic regression model which get the classification by substituting the value of the independent variable into the estimated model, the Naive Bayes algorithm is solved by the Bayes formula, and we calculate the conditional probability value through the distribution function each time.

Naive Bayes algorithm is one of the most widely used classification algorithms. The Naive-Bayes method is based on the Bayes algorithm which, given a target value, assumes that the attributes are independent of each other. The proportion of each regressor will not be particularly heavy or light. Although this simplified method reduces the classification effect of the Bayesian classification algorithm to some extent, it greatly simplifies the use of the Bayesian method in practical application scenarios. The core of the Naive Bayes algorithm is the Bayes formula in probability statistics, and the cornerstone of the Bayes formula is conditional probability. The Bayes formula is as follows.

$$
P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)}, \text{ Where } \{B_i\} \text{ is a full partition of the space}
$$

The $B_i$ that maximizes the conditional probability $P(B_i|A)$ is the category to which we estimate. Among this function, $A$ is the value of the independent variable we know, $B_i$ is the occurrence of the $i–$th category. For a fixed $A$, the numerator of the conditional probability is constant, so the classification is equivalent to the following formula.

$$
\hat{y} = \arg \max_{B_i \in B} P(B_i) \prod_{i=1}^{d} P(x_i|B_i)
$$

Among them, we assume that the regressors in the $i–$th category follows a multivariate normal distribution, and we obtain the multivariate normal parameters for each category through the training data set. Given $x_i$, we can obtain $P(x_i|B_i)$.

Analysis of model results
As mentioned above, we applied stepwise logistic regression and Naive Bayes model to the training data set, and obtained the following experimental results.

Table 2 Test results of stepwise logistic regression on the training data set

| Variable | Estimation | Z-value | P-value  |
|----------|------------|---------|----------|
| (Intercept) | -1.975 | -7.475 | 7.72E − 14*** |
| H2       | -0.0202 | -1.599 | 0.110    |
| C2H2     | 0.0307  | 1.726  | 0.084    |
| C2H4     | 0.0166  | 2.697  | 0.007**  |
| C2H6     | -0.108  | -2.270 | 0.0232*  |
| CO2      | 0.0000555 | 1.700  | 0.0891   |

Significance: 0.001'***'; 0.01'**'; 0.05'*'; 0.1'.'; 1''

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

Table 3 Test results of stepwise logistic regression on the test data set

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

0.860

Table 4 The test results of the Naive Bayes model on the test set

| Actual | Predicted | Accuracy |
|--------|-----------|----------|
| No     | No        | 91       | 1        | 0.989    |
| Yes    | Yes       | 0        | 15       | 1        |

0.991

As can be seen from the table above, the overall classification accuracy of logistic regression is 0.860, and that of Naive Bayes algorithm is much higher, reaching 0.991. However, the ability to predict a faulty main transformer is more important to us than to predict a normal main transformer and the Naive Bayes algorithm performs better in this regard, reaching 100% in accuracy. Although logistic regression model performs not as well as the Naive Bayes algorithm, it also seems to be a simple and good model to predict defaults. Considering the strong theoretical support of the Naive Bayes algorithm, we believe that Naive Bayes is a suitable algorithm for the prediction of 110kV main transformer defects and faults. At the same time, from the results of the logistic regression, it can be seen that the acetylene (C2H2), ethylene (C2H4), ethane (C2H6), and carbon dioxide (CO2) in the sampling oil chromatography have great significance for predicting the main transformer defects and faults.

5. Conclusion

A comparative study of 110kV main transformer defects and faults evaluation based on logistic regression and Naive Bayes algorithm shows that the model based on Naive Bayes has a higher accuracy rate in predicting main transformer defects and faults. It also shows that the acetylene, ethylene, ethane, and carbon dioxide in the sampling oil chromatography have a strong correlation with the 110kV main transformer defects and faults.

The analysis and quantification of the main transformer working status in our paper are universal and can be used to evaluate the status of other important equipment and the whole power grids. At the same time, it provides a new exploration for equipment status evaluation, and provides new ideas for the smart grid with the fusion data of power -equipment-environment, which has huge engineering application value.
6. Acknowledgment

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

References

[1] Wu Z, Yang J, Zhu Y, et al. Transformer fault diagnosis based on rough set theory and support vector machine[J]. Power System Protection and Control, 2010, 38(18): 80-83.

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

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

[4] Cui J, Yang B. A review of bayesian optimization methods and applications [J]. Journal of software engineering, 2018, 29(10): 3068-3090.

[5] Yang G. Research on transformer fault diagnosis method based on multiple bayesian classifiers[D]. XMU, 2017.

[6] Tian B, Gao X. Fault diagnosis of power transformer equipment based on bayes classifier[J]. Electronic Design Engineering, 2017, 25(15): 54-57+61.

[7] Chen X, Pan X, et al. The diagnosis model of partial discharge based on naive bayes[J]. Computer Applications and Software, 2016, 33(09): 51-55.

[8] Wang X, Han T. Transformer fault diagnosis based on bayesian stochastic forest optimization[J]. Electrical Measurement & Instrumentation, 2020, 1:1-8

[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 Y, 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] 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.

[12] Zhou Z, Yang Z, Yang C, et al. Power transformer fault diagnosis based on improved RADAG-SVM[J]. Journal of Southeast University: Natural Science Edition, 2016, 46(5): 964-971.

[13] Ji X, He D, Sun C. The application of SVM in transformer protection is analyzed[J]. Heilongjiang Science and Technology Information, 2016 (17): 143.

[14] Wang Y. Transformer fault diagnosis based on least squares support vector machine[J]. Electrical & Electronics, 2016 (6): 24-27.

[15] 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.

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

[17] 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.

[18] Ding W, 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.
[23] 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.

[24] Xiao L, Ya C. The relationship between ovarian endometriosis and clinical pregnancy and abortion rate based on logistic regression model[J]. Saudi Journal of Biological Sciences, 2020, 27(1).

[25] 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.

[26] 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.

[27] 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.