Research on the application of Decision Tree and Random Forest Algorithm in the main transformer fault evaluation

Research on the application of Decision Tree and Random Forest Algorithm in the main transformer fault evaluation

Chenmeng Zhang[1], Can Hu[2], Shijun Xie[1,*], Shuping Cao[1]
[1]State Grid Sichuan Electric Power Research Institute, Chengdu, Sichuan, China
[2]State Grid Sichuan Electric Power Company, Chengdu, China

*Corresponding author e-mail: sj-xie@163.com

Abstract. The detection and monitoring methods of power equipment are relatively complex. The analysis of a large amount of data obtained from detection and monitoring and the development of data recording methods make it possible to obtain higher-dimensional power data. If we adopt appropriate data analysis method to the data sets, we can better get the potential laws and value of power data. In this paper, we took various basic monitoring data and faults records of the 110kV main transformers into consideration, and the grid-equipment-environment data was fused. Based on the fused data, we used the decision tree and random forest algorithm to evaluate the main transformers' defects and faults. The evaluation results of the two algorithms were obtained and compared, which proved the effectiveness of the fault evaluation algorithm and selected a more accurate fault evaluation algorithm. This paper provides new ideas for smart fault detection for power grid, and provides a reference for a more in-depth evaluation of power grid equipment.

1.Introduction

With the development of smart power grids, large and complex equipment monitoring systems have been launched, and various kinds of equipment tests and inspection plans have been executed. Therefore, the data is increasingly generated regardless of the type or quantity. Since the data related to power system equipment has rapidly increased, the disadvantages of traditional data modeling and analysis methods are highlighted, causing low data utilization rate and making the existing equipment analysis and evaluation systems more and more questionable. Under the condition that the data types are continuously expanded and some parts of data are unreliable, traditional data modeling and analysis is difficult to keep up. To solve above issue, machine learning algorithms are supposed to be used to quickly mine the potential rules and laws under data and form a flexible evaluation plan. In this pare, we take the monitoring, defect, and test data related to the main transformer equipment of the whole power network in Guangxi Province into account. We first performed data preprocessing and fusion on the data set we obtained, and then designed the defects and faults evaluation models based on decision tree and random forest algorithm. The algorithms were finally implemented and the experimental results were analyzed.

2.Background

In recent years, the development of the computer and Internet industries, the popularity of high-performance computing clusters and the accumulation of massive data resources have laid a solid foundation for the development of big data technology. At present, big data technology has been

widely used in transportation, medical treatment, manufacturing and many other fields, and has achieved good application results. In 2017, the State Council also added big data application to the national strategic level. Therefore, in the context of power big data, it is important to promote the comprehensive application of big data analysis technology in the field of equipment condition evaluation. At the same time, realizing the organic combination of data and models is of great significance for the intelligent management of the transformer operation and maintenance.

In the context of power big data, the types of data related to power transmission and transformation equipment are diverse [1]. A typical data center that does not contain real-time data but integrates grid-equipment-environment fusion data and other data has more than 1,000 database table files whose total number of data columns exceeds 8,000 and the number of records exceeds 1 billion. Among the purchase and installation phases, there are data generated such as construction planning, ordering technical requirements, manufacturing supervision information, infrastructure construction information, factory test information, and account information of various types of equipment [2], and there are also a large number of unstructured data such as contract information, blueprints, structural design as well. The total amount of the data mentioned above is near 100TB. The log data of power transmission and transformation equipment which is generated during the operation and maintenance includes information of tests, overhauls, defects, major repairs, technical changes, movements, and return shipments. The total amount of this kind of data recorded is more than 100,000 lines per year[3]. The amount of real-time monitoring data such as the status of primary and secondary equipment is even larger. The quasi-real-time monitoring for a province ’s power transmission and transformation equipment (refresh every 15-minute in average) has more than 20 million measurement points, generating more than 2 billion records which generate 10G of data per day after compression and storage. Meanwhile, high-frequency monitoring data can generate hundreds of terabytes of data every year [4], and some unstructured monitoring data such as smart substation online video monitoring data can generate nearly PBs of data every year.

As we all know, the environment condition also affects the performance and life span of power transmission and transformation equipment. As a result, the data including meteorological data which contains regional meteorological forecast, six meteorological factors, meteorological warning, typhoon, thunder and lightning, ice coating, etc., GIS terrain data, pollution monitoring, remote sensing data, etc., should be taken into consideration. The related data has exceeded 50TB, and about 100G of data has been adding annually.

Based on a large amount of power grid data, this paper executed data fusion, compared the two classification algorithms, and used rich power data to evaluate and analyze the status of main transformers.

3.Variable settings and data preprocessing

The bottlenecks in fault evaluation of the main transformers are the huge amount of information, complex types of data, and the difficulty to fuse the cross-system data of power transmission and transformation equipment. The lack of a unified information management system and effective data mining and analysis methods leads to the problem that various types of data cannot be effectively used, causing high operation and maintenance costs and difficulty to manage power equipment throughout its entire life cycle. Therefore, to evaluate the defects and faults of power transformers, it is necessary to correlate and fuse the data, the logic of which is shown in the figure below.

Figure.1. data processing and modeling

In data pre-processing, we first fused the data from various sources by the unique ID of the main transformer to associate monitoring, inspection data with transformer defects and faults, which expanded the grid data and completes the data preprocessing. Similarly, we fused the environmental data with corresponding device data set according to the area and time, which formed an expansion between the transformer data and the environmental data. For the high-frequency monitoring data, such as the number of crossing the boundary, the number of low voltage alarms related to the main transformations and the amount of H2, C2H2, CO, CO2 in the oil chromatography should be obtained from the latest data. Finally, the inherent data, such as equipment manufacturer, voltage level and operation time, are combined to complete the data fusion. On this basis, according to the fused panel data, we set "time + unique ID of the device + inherent attributes + grid data + device monitoring data + environmental data" as regressor x, and the status of power transformer “Whether there is a defect or failure” is recorded as dependent variable y.

Next, we process the fused data further. Since each attribute recorded in different systems cannot correspond exactly in the recording time, there are missing values in our data set. We use the average of each attribute to fill in the missing values.

According to the processed data, the variable settings are shown in the table below.

Table.1. Variable settings

SYMBOL	MEANING	ATTRIBUTE
TEMP	Sampling oil temperature (℃)	numeric/regressor
H_2	Hydrogen by oil chromatography	numeric/regressor
C_2H_2	Acetylene by oil chromatography	numeric/regressor
C_2H_4	Ethylene by oil chromatogram	numeric/regressor
CH_4	Methane by oil chromatographic	numeric/regressor
C_2H_6	Ethane by oil chromatography	numeric/regressor
CO	CO by oil chromatography	numeric/regressor
CO_2	Carbon dioxide by oil chromatography	numeric/regressor
ACID	Oil acid ester (mgKOH/g)	numeric/regressor
PH	Ph of oil	numeric/regressor
V	Oil breakdown voltage (kV)	numeric/regressor
FLASH_POINT	Oil closing flash point (℃)	numeric/regressor
MACRO_WATER	Oily moisture content (mg/L)	numeric/regressor
STATUS	110kV main transformer condition	categorical /dependent

Finally, we set the last 100 records as the test data set and the remaining data as the training data set.

4.Theoretical analysis of the model

Decision tree is a kind of nonlinear supervised classification model which has a tree-like structure as a classifier. The connection points between branches represent the conditions for discrimination, and the leaf nodes at the ends of the branches represent the categories records belong to. When using the decision tree to classify, according to whether the data meets the conditions shown at the node, we select different branches to continue and repeat the above steps until we reach a leaf node.

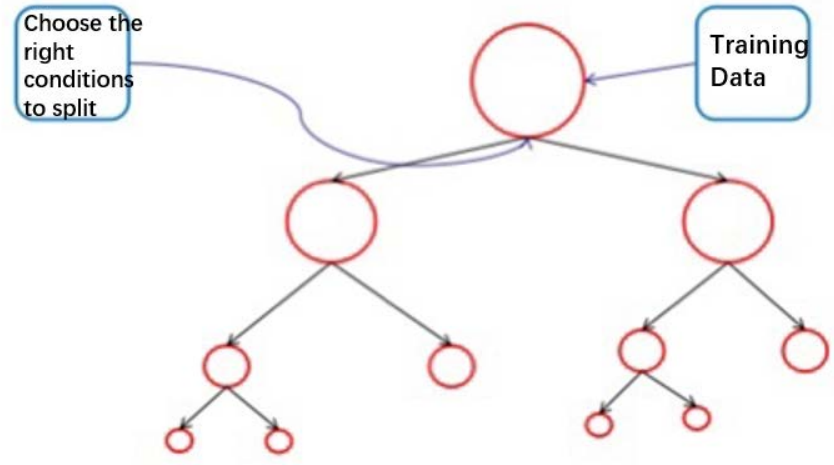

Figure.2. Decision tree generation

When generating a decision tree, we follow the principle of distinguishing the data belonging to different categories to the greatest extent. As a result, we introduce Gini coefficient as a loss function to construct a CART decision tree.

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

Each discriminant condition corresponds to a loss function value. We choose the discriminant condition that minimizes the $Gini$ coefficient as the discriminant condition of 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.

When the data set is large, we randomly select a part of the whole data set to generate a decision tree. Repeating the above process, we can generate a bunch of decision trees with different nodes and shapes. These trees combine to form a random forest. We first select n sub-samples from the sample set by Bootstrap sampling. For each sample subset, we randomly select K attributes and create a decision tree. Repeat the above two steps n times and we will build n CART decision trees which form a random forest. The voting results of n decision trees determine which type a record belongs to.

5.Analysis of model results

As mentioned above, we applied the decision tree and the random forest model to the training set, and obtained the following experimental results.

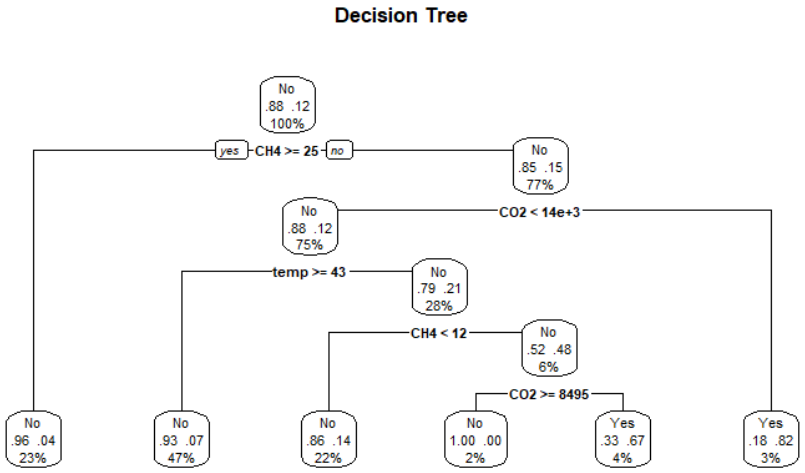

Figure 3. Decision tree model results

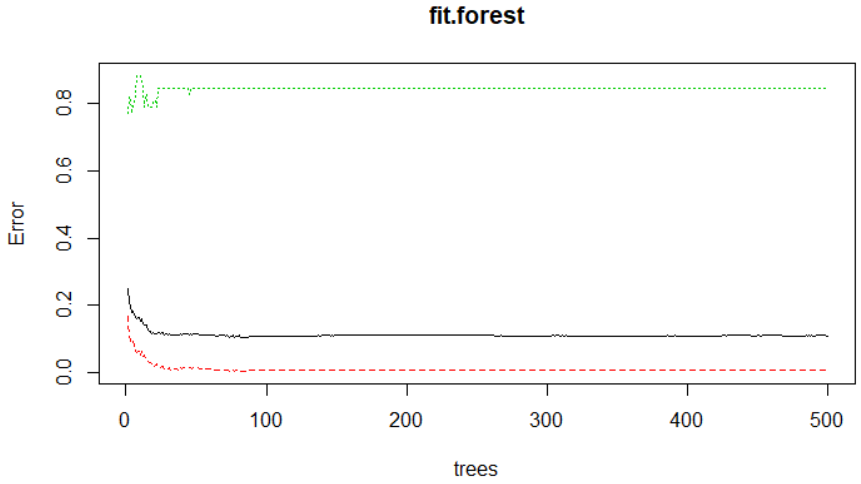

Figure 4. random forest model results

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

Table 2 test results of the decision tree model on the test data set

	PREDICTED			ACCURACY
ACTUAL		**No**	**Yes**	
	No	130	8	0.942
	Yes	14	9	0.391
				0.863

Table3 test results of random forest model on test data set

	PREDICTED			ACCURACY
ACTUAL		**No**	**Yes**	
	No	138	0	1
	Yes	19	4	0.174
				0.882

It can be seen from the result table that the comprehensive classification accuracy of the decision tree algorithm is 0.863, and that of the random forest is slightly higher, reaching 0.882. What counts most for us is the ability to predict defective main transformers, and in this respect, the decision tree algorithm performs much better than random forest. Meanwhile, considering the simplicity and the characteristics of faster speed of the decision tree algorithm, we believe that the decision tree algorithm is more suitable for the evaluation of 110kV main transformer defects and faults. From the results of the decision tree algorithm, we can see that the temperature of the sampled oil and the content of methane and carbon dioxide in the sampled oil chromatography have great significance for evaluating the defects and faults of the main transformers.

6.Conclusion

The comparative study of 110kV main transformer defects and fault evaluation based on decision tree and random forest shows that the model based on decision tree has higher accuracy, and that there is a great association between sample oil temperature, the content of methane and carbon dioxide in sample oil chromatography and transformer faults. This study also provides a new idea for the fault evaluation by the fusion data of grid-equipment-environment, and a new exploration for the evaluation of the status of the power equipment. This paper provides a reference for more efficient models and more accurate evaluation of the status of the power equipment as well.

7. Acknowledgment

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

References

[1] Li W, Song Y, Guo W, Song S, Wang J. State evaluation of power transformer based on decision tree integration algorithm[J]. Power System and Clean Energy,2017,33(10):50-55.
[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] 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.
[5] 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.
[6] Zhu Y, Liu S, Wang F. Parallel transformer fault diagnosis based on the combination of three ratios of Spark and random forest [J]. Computer Knowledge and Technology, 2017,13(27):221-224.
[7] Shan L. Fault analysis of distribution transformer and realization of random forest diagnosis[J]. Technology Wind,2016(24):120.
[8] 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.
[9] Chen J , Wu C. Improvement and application of decision tree C4.5 algorithm [J]. Software Guide. 2018(10)
[10] Wang T , Sun Z, et al. Study on fault diagnosis of power transformer based on classification decision tree algorithm[J]. Electrical Engineering,2019,20(11):16-19.
[11] 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.
[12] Xue Y, Lai Y. The fusion of big energy thinking and big data thinking[J]. 2016.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] 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.