Research on big data risk assessment of major transformer defects and faults fusing power grid, equipment and environment based on SVM

Lijuan Guo¹, Haijun Yan¹, Wensheng Gao², Yun Chen² and Yongqi Hao³*,

¹Guangxi Electric Power Research Institute, Nanning, China;
²School of Electrical Engineering, Tsinghua University, Beijing, China;
³School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China.

*Corresponding author e-mail: haoyongqi001@163.com

Abstract. With the development of power big data, considering the wider power system data, the appropriate large data analysis method can be used to mine the potential law and value of power big data. On the basis of considering all kinds of monitoring data and defects and fault records of main transformer, the paper integrates the power grid, equipment as well as environment data and uses SVM as the main algorithm to evaluate the risk of the main transformer. It gets and compares the evaluation results under different modes, and proves that the risk assessment algorithms and schemes have certain effectiveness. This paper provides a new idea for data fusion of smart grid, and provides a reference for further big data evaluation of power grid equipment.

1. Introduction
In the development of urban power grid, equipment safety warning plays a very important role. And equipment defects and risk assessment need to integrate a variety of factors, so the comprehensive assessment model is in short supply. With the development of smart grid, large and complex equipment monitoring systems have been on the line. At the same time, the species and quantity of data produced by increment of all kinds of equipment test and inspection plan both are increasing rapidly. With the increment of power system equipment related data, the disadvantages of traditional data analysis method will be highlighted. It leads to low data utilization, and makes power system staff more and more skeptical about existing equipment analysis and evaluation model. In the foreground of big data development, it is very important to integrate the data from different dimensions to mine potential information. Equipment failure risk science determine the needs of different departments in the information processing stage of data problems, such as lack of cross and miscellaneous. Aiming at this kind of problem, the traditional data modeling analysis is difficult to keep pace with the change of data type and the fallibility of partial data. The better solution is to quickly excavate the latent law of data analysis by machine learning algorithm, and form an elastic evaluation scheme. In this paper, based on the monitoring, defect, test and other data of the whole network main transformer equipment in Guangxi Province, the data are preprocessed and fused firstly. Then, this paper designs risk evaluation scheme of the defect and fault based on SVM for main transformer. Finally, the algorithm is implemented and the experimental results are analyzed.
2. Algorithm principle and data fusion

In the rapid development of power big data, the data type of power transmission equipment design is very diverse [1]. In a typical data center that does not include real-time data integration of power grid equipment environment data, there are more than 1000 database table files, the total number of data columns is more than 8000 columns, and the number of records is more than 1 billion. They have included the acquisition plans and installation data such as construction planning, ordering information [2], technical requirements, accounting supervision, information infrastructure, information engineering, factory test information of all kinds of equipment, but also contains a large amount of unstructured data such as contract information, blueprint and structure design etc. The total amount of data is near 100TB level. The log data generated in the operation and maintenance of power transmission and transformation equipment include data such as test, repair, defect, overhaul, technical transformation, movement and return, and the total amount of data is more than 100 thousand lines per year [3]. The data of Online monitoring, such as the running state of the primary or secondary equipment, is larger. In the province's quasi real-time monitoring data of transmission and transformation equipment (average 15 minutes refresh), there are more than 20 million measuring points. More than 2 billion records are generated daily, and more than 10G of data is produced daily after compressed storage. The high frequency real-time monitoring data can produce hundreds of TB level data every year [4]-[15]. Some unstructured monitoring data, such as smart substations, online video surveillance and other data, can produce nearly PB of data each year. The operating environment of power transmission and transformation equipment also affects its performance and life curve. The data include meteorological data, regional weather forecast, meteorological six factors, weather warning, typhoon, lightning situation, ice cover, GIS topographic and topographic data, pollution monitoring, remote sensing data and so on [16]-[20]. The accumulated data has been over 50TB, with an increase of about 100G order. Finally, SVM is one of the classifiers in large data machine learning algorithms, and is very suitable for nonlinear classification problems.

![Figure 1. Principle of SVM machine learning algorithm](image)

Therefore, this paper uses SVM as a machine learning algorithm to further evaluate and analyze the equipment risk status based on data fusion analysis.

3. Raise and analysis of the evaluation scheme (data cleaning, SVM)

The information of power transmission and transformation equipment is huge and the type is complex. Cross system uniqueness and association difficulty is one of the bottlenecks in the analysis and evaluation of the main transformer and other equipment. Lack of unified information management and effective means of data mining and analysis makes it impossible to make effective use of all kinds of data. It also directly results in the problems of high maintenance cost and difficult to control during the
whole life cycle of transmission and transformation equipment. Thus, it is necessary to fuse the data for the defects and faults of power transmission equipment, and the logic is as follows:

In the data preprocessing and fusion stage, firstly, according to the equipment data and the account information, the paper constructs the primary data associated with the unique ID as the primary key. Thus, it can be related to the data of monitoring, inspection and defects. On the basis of the above, the power grid data and environmental data are expanded according to the idea of facing the equipment. In each class of data, the unique ID column of the device is inserted as an association, thereby extending the grid data and performing the preprocessing. Finally, the environment data is connected to the corresponding device data sets according to the region and time, so as to form an association expansion between the device and the environment device. The preprocessed data, according to the day time slice, the data fusion for the "time equipment only ID Inherent attribute equipment monitoring data +environmental data” was recorded as X. The corresponding "whether there is a defect or fault" is denoted as y, as an association data set for the structure. Finally, the data are divided into training sets and the last 100 test sets. For high frequency monitoring data, such as the number of cross-border records, the number of low voltage alarms associated with the main transformer and other data takes the latest data as the standard, such as H2, C2H2, CO, CO2 and so on in the oil. Finally, combining the inherent data, such as device manufacturers, voltage levels, and run time, the data fusion for SVM is completed. At this point, the problem can be reduced to:

Defining a straight line

\[ y(x) = w^T x + b \] (1)

The distance from any point to the straight line is:

\[ \frac{1}{\|w\|}(w^T x + b) \] (2)

For all training sets, that is, the main transformer state + defect, or the fault state, are denoted as \((x_i, y_i)\). The solution process is briefly considered as follows:
arg max_{w,b} \left\{ \frac{1}{m} \min_{n} \left[ y_i (w^T x_i + b) \right] \right\} \quad (3)

On the basis of the above, first of all, this paper builds a big data computing and analysis platform based on Spark. And the machine learning is analyzed. Finally, the trained model is tested and analyzed.

4. Analysis of machine learning results
As described above, SVM is trained by using a training set, and the results are tested to obtain the following experimental results.

| Table 1. Assessment results of defects and failure for transmission and transformation equipment based on SVM |
|------------------------------------------------------------|
| SVM R-O | SVM T-O | SVM R-S | SVM T-S |
| iter | 1585 | 305 | 1519 | 292 |
| nu | 0.7784457249 | 0.7891332471 | 0.7479370952 | 0.7554980595 |
| obj | -350.6843968 | -610.0 | -342.6970074 | -584.0 |
| rho | -0.348415953 | -1.0 | -0.390821545 | -1.0 |
| nSV | 772 | 610 | 772 | 584 |
| nBSV | 298 | 610 | 284 | 584 |
| Total nSV | 772 | 610 | 772 | 584 |
| Accuracy | 72.0%(72/100) | 71.0%(71/100) | 59.0%(59/100) | 59.0%(59/100) |

According to the actual data, the C-SVC type SVM is selected as the main type. The R represents the radial basis kernel, and T represents the hyperbolic tangent kernel. The O indicates the temporal ordering, and S represents the random sort, such as SVM R-O, which represents the C-SVC+ radial kernel + time sorted data set. In the test results, ITER stands for the number of iterations that indicate the extent of the correction to the model, and nu is only the one-class SVM parameter. Obj is the minimum value obtained for the two programming solution converted into SVM file, and Rho is the constant term of the decision function. NSV is the total number of support vectors, nBSVC is the number of support vector on the boundary, and Total is the total number of nSV vectors. Finally, Accuracy represents the correct rate of testing the model by using the test set after the training. The higher the accuracy, the closer the model is to the actual situation, the more it can be used as a reference for the state assessment of the equipment.

The radial basis kernel and the hyperbolic tangent kernel have 72% and 71% accuracy in the sorted data sets. It can be seen that the power grid equipment environment fusion data can reflect the health status of the equipment to a certain extent, and it has a strong correlation.

5. Summary
The research on big data risk assessment of power transmission equipment defects and faults based on SVM shows that transformer manufacturers, running time, oil immersion data and other equipment attributes have a strong correlation with their failure rate. Moreover, the power grid operation data and environment data can provide reference for the prediction of equipment running state directly. This research provides a new idea for the grid equipment environment data fusion of the smart grid, and makes a new exploration for the equipment status assessment. It provides a reference for more efficient models in the future and a more accurate assessment of device status.

References
[1] Zheng Xiang, Zhao Fengzhan, Yang Renggang, et al 2017 An evaluation system to reactive power operation in lv distribution network based on big data[J]. Power System Technology, 1: 038.
[2] Xun Ying, Gao He, Li Kejun, et al. 2016 Condition assessment model of distribution transformer based on multi-period information fusion[J]. High Voltage Engineering, 42(7): 2054-2062.

[3] Deng Song, Yue Dong, Zhu Lipeng, et al. 2016 Framework of intelligent and efficient analysis and mining technology for power big data[J]. Journal of Electronic Measurement and Instrumentation, 30(11): 1679-1686.

[4] Wang Wei, Liu Yin, Yu Zhanpeng, et al. 2016 System design of the big data center architecture in electric power big data environment[J]. Electric Power Information and Communication Technology, 14(1): 1-6.

[5] Zheng Like, Li Shuangbing, Wang Xiaodong, et al. 2016 Insulation condition assessment for power transformer based on normal cloud model with optimal variable weights[J]. High Voltage Apparatus, 52(2): 85-92.

[6] Zhou Zhicheng, Yang Zhichao, Yang Chengshun, et al. 2016 Fault diagnosis of power transformer based on modified radag-svm[J]. Journal of Southeast University(Natural Science Edition), 46(5): 964-971.

[7] Ji Xuan, He Dongsheng, Sun Changxu. 2016 The application of svm in transformer protection is analyzed[J]. Heilongjiang Science and Technology Information, (17): 143.

[8] Wang Yiping. 2016 Transformer fault diagnosis based on least squares support vector machines[J]. Electrotechnics Electric, (6): 24-27.

[9] Zhang Yongqiang, Kou Lingfeng, Sheng Wanxing, et al. 2016 Big data analytical method for operating state assessment of distribution transformer[J]. Power System Technology, 40(3): 768-773.

[10] Xue Yusheng, Lai Yening. 2016 Integration of macro energy thinking and big data thinking (I) big data and power big data[J]. Automation of Electric Power Systems, 40(1):1-8.

[11] Wang Dewen, Zhou Qing. 2016 Distributed on-line analytical processing method for power equipment condition monitoring big data[J]. Proceedings of the CSEE, (19): 5111-5121.

[12] Ding Dengwei, Gao Wensheng, Liu Weidong. 2014 Investigation of improvement of sensitivity verification of partial discharge detection using uhf method in gas[J]. High Voltage Apparatus, 8: 003.

[13] Bai Cuifen, Gao Wensheng, Jin Lei, et al. 2013 Study on comprehensive stress life model of transformer[J]. Advanced Engineering Sciences, 2.

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

[15] Zhou R, Gao W, Zhang B, et al. 2014 Prediction of tropical cyclones’ characteristic factors on hainan island using data mining technology[J]. Advances in Meteorology, 2014.

[16] Gao W, Ding D, Liu W, et al. 2014 Investigation of the evaluation of the pd severity and verification of the sensitivity of partial-discharge detection using the uhf method in gas[J]. IEEE Transactions on Power Delivery, 29(1): 38-47.

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

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

[19] Gao Wensheng, Zhang Bowen, Zhou Ruixu, et al. 2015 Nowcasting of the thunderstorm trend based on data collected by lightning location system[J]. Power System Technology, 39(2): 523-529.

[20] Hu J, He J, Long W, et al. 2010 Temperature dependences of leakage currents of zno varistors doped with rare - earth oxides[J]. Journal of the American Ceramic Society, 93(8): 2155-2157.