Comprehensive analysis and evaluation of big data for main transformer equipment based on PCA and Apriority

Lijuan Guo¹, Haijun Yan¹, Yongqi Hao²⁺ and Yun Chen³

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

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

Abstract. With the power supply level of urban power grid toward high reliability development, it is necessary to adopt appropriate methods for comprehensive evaluation of existing equipment. Considering the wide and multi-dimensional power system data, the method of large data mining is used to explore the potential law and value of power system equipment. Based on the monitoring data of main transformer and the records of defects and faults, this paper integrates the data of power grid equipment environment. Apriori is used as an association identification algorithm to extract the frequent correlation factors of the main transformer, and the potential dependence of the big data is analyzed by the support and confidence. Then, the integrated data is analyzed by PCA, and the integrated quantitative scoring model is constructed. It is proved to be effective by using the test set to validate the evaluation algorithm and scheme. This paper provides a new idea for data fusion of smart grid, and provides a reference for further evaluation of big data of power grid equipment.

1. Introduction
With the development of urban power grid, the automated level of power grid equipment is also improved. As a result, the data generated by the power grid, both kinds and quantities, is increasing rapidly. Because of the universality and multi-dimension of power grid data, the disadvantages of traditional methods for data modeling and analysis will be highlighted. Due to the low utilization of data, the existing equipment quantitative evaluation system and model have been suspected by power system staff. In order to solve the above problems, in the case of the continuous expansion of data types and the unreliability of data, the machine learning algorithm is used to avoid the drawbacks of traditional data modeling and analysis. Then, the big data mining technology is used to quickly extract the potential law of data, so as to form an elastic comprehensive evaluation scheme. The data in this paper are from the monitoring, defect, test and other data of the whole network main transformer equipment in Guangxi Province. Firstly, the existing data is pretreated and fused, and then the evaluation scheme based on PCA and Apriority is designed for the main transformer. Finally, the algorithm is compiled and implemented, and the experimental results are analyzed.
2. Data fusion
Big data has been widely used in commercial, financial and other fields, the development of intelligent power grid also provides a basis for big data applications[1]-[7]. According to data sources, smart grid data can be divided into two broad categories: one is the internal data of the grid, and the other is external data. The internal data from the electricity information collection system (collection system information, CIS), marketing system, wide area monitoring system (wide area measurement system, WAMS), production management system, distribution management system (production management system, PMS), the energy management system (energy management system, EMS), detection and monitoring systems and equipment customer service system, financial management system data. The external data comes from electric vehicle charging management system, weather information system, geographic information system (GIS), public service department, Internet and so on. These data are scattered in different places, managed by different units/departments, and have the characteristics of decentralized placement and distributed management [8]-[10].

Therefore, on the basis of data fusion analysis, this paper uses Apriority algorithm as the main algorithm of equipment association analysis, and uses PCA to analyze and evaluate each attribute data. Finally, a test set is extracted to analyze the effect of this evaluation scheme.

3. Raise and analysis of the evaluation scheme
It is difficult to analyze and evaluate the main transformer, such as large amount of information, complex types, and the association analysis of multidimensional data. At the same time, due to the lack of effective integrated information management and data mining analysis methods, all kinds of data cannot be fully utilized. The cost of operation and maintenance of transmission and transformation equipment is high during the whole life cycle. Thus, to make a comprehensive evaluation of power transmission and transformation equipment, it is necessary to fuse the data first, and the logic is shown below. In the data preprocessing and fusion stage, according to the equipment data sheet, combined with accounting information, construct the data associated with a main transformer only ID, and associated monitoring, inspection and important defects and fault data.

In the data preprocessing and fusion stage, according to the equipment data sheet, the big data is combined with accounting information. The paper constructs the data associated with the main transformer only ID, to achieve data association of monitoring, inspection and defects as well as faults. Considering the related factors of environment and power grid data, the power grid data and environment data are extracted from each kind of data, which is related to the equipment ID, so as to fuse all the data and realize the pretreatment. In Figure 2, the level of pollution zone reflects the environmental factors...
corresponding to the equipment. Cumulative number of failures and the number of months running normally reflect the information storage of the power grid. The equipment itself includes the commissioning time, voltage level and type and so on. As can be seen from the fusion data, the device related factors are multi-class and cross-system.

On the basis of preprocessing, firstly, the fusion data is analyzed by Apriori algorithm. Apriori algorithm is one of the most influential algorithms for mining frequent itemsets of Boolean Association rules. Its core is a recursive algorithm based on the idea of two stage frequency set. In the process of association analysis, the program needs to identify specific indicators to measure the intensity of association rules. If there is a combination of the event M and the event N in the record of the total transaction, the probability of the combination is represented by the support. Support is also a measure of the importance of association rules. The accuracy of association rules is expressed in confidence, and the intensity of association rules is also expressed. Specific definitions are as follows:

\[ \text{Sup}(M \rightarrow N) = \frac{\text{count}(M \cup N)}{\text{count}(T)} \]  

\[ \text{Con}(M \rightarrow N) = \frac{\text{count}(M \cup N)}{\text{count}(N)} \]  

Where Sup and Con indicate support and confidence respectively. Count is the number of events. The minimum support and minimum confidence need to be determined before the association analysis is carried out. When the support of the event set is greater than the minimum support, then the set of events becomes the frequent set.

After analyzing the data of each attribute, it is necessary to determine the specific quantitative evaluation model for the main transformer equipment. The fusion data contains 12 categories, so the PCA algorithm is used to achieve dimensionality reduction and determine the relevant principal components. First, the first linear combination \( F_1 \) is selected, and when the \( \text{Var}(F_1) \) is larger, it means that the more information \( F_1 \) contains. If the \( F_1 \) variance is the largest in all linear combinations, then \( F_1 \) is the first principal component. In the process of defining the other principal components, you need to make sure \( \text{Cov}(F_1, F_i) = 0 \) to avoid the information already in the \( F_1 \).

On the basis of the above algorithm, a big data computing and analysis platform based on Clementine is built, and the corresponding value rules are extracted by the algorithm analysis results. Finally, the training model is tested and analyzed.

### 4. Analysis of machine learning results

As mentioned above, the Apriori algorithm is used to do the association analysis of the fused data, and get the following experimental results.

| Consequent | Antecedent | Support% | Confidence% |
|------------|------------|----------|-------------|
| C8=c       | C7=500000.0 | 5.854481407 | 86.31355932 |
| C8=c       | C7=500000.0 and C9=1.0 | 5.854481407 | 86.31355932 |
| C8=c       | C7=500000.0 and C1=0.0 | 5.581603036 | 85.73333333 |
| C8=c       | C7=500000.0 and C1=0.0 and C9=1.0 | 5.581603036 | 85.73333333 |

Table 1 shows that equipment status, equipment failure, and voltage level are closely related to the level of pollution zone. Because the environmental impact can affect the health status of equipment and evaluation results, so the above analysis results are in line with the actual situation. To better analyze the relationships among other factors, the minimum support and minimum confidence limits are reduced.
and level of pollution zone are filtered. Table 2 is the correlation result of the Apriority algorithm without environmental factors. As can be seen from the table below, vendor information is associated with unit type. Therefore, it is possible that the same device type may come from the same vendor during the equipment purchase process. The real time failure of the main transformer equipment has a potential connection with the manufacturer and equipment type. The analysis results of the relevant equipment also provide a new procurement analysis thinking for the purchasing department.

Table 2. Association rules analysis of different influencing factors of equipment without environmental factors

| Consequent                | Antecedent                                                                 | Support%     | Confidence%  |
|---------------------------|----------------------------------------------------------------------------|--------------|--------------|
| C12=SZ9-40000/110         | C6=TBEA Xinjiang transformer factory of Limited by Share Ltd               | 3.041353477 | 54.81239804  |
|                           | C6=TBEA Xinjiang transformer factory of Limited by Share Ltd and C7=110000.0| 3.041353477 | 54.81239804  |
|                           | C6=TBEA Xinjiang transformer factory of Limited by Share Ltd and C9=1.0    | 3.041353477 | 54.81239804  |
|                           | C6=TBEA Xinjiang transformer factory of Limited by Share Ltd and C7=110000.0 and C9=1.0 | 3.041353477 | 54.81239804  |

The frequent set of association analysis by Apriority algorithm can get more potential rules related to equipment. But according to the actual situation of the data, it is necessary to make a quantitative assessment of the main transformer equipment. Because the equipment type, ID and the manufacturer are non-numerical information, and the impact on the comprehensive evaluation is small. Thus, the data is filtered. The processed data can be used to quantify the level of pollution zone (the higher the value, the more serious the polluted area), and the following principal component analysis results can be obtained:

Table 3. Total variance explained

| Component | Extraction sums of squared loadings |
|-----------|------------------------------------|
|           | Total | Of variance% | Cumulative% |
| 1         | 2.399 | 29.991       | 29.991      |
| 2         | 1.153 | 14.409       | 44.4        |
| 3         | 1.001 | 12.51        | 56.9        |
As you can see in Table 3, the three principal components represent the information in the data well. In the principal component expression, the expression coefficients of equipment run time, voltage level and time stamp are very small, so these factors are not the dominant factor of scoring. The principal component 1 mainly considers the scores under various factors, while the main component 2 focuses on the environmental impact score. The principal component 3 can be used as the real-time risk assessment score of the equipment. By extracting the test set, table 5 shows that the first device has a better overall score, but there is also an impact on the risk of failure. The risk of the second equipment is small and the operating environment is poor. The final equipment risk assessment score is greater, indicating that the equipment has a greater operational risk.

5. Summary

Based on PCA and Apriority power transmission equipment analysis and evaluation program research shows: Equipment status, equipment failure, and pollution level have a strong correlation. At the same time, there is a certain interconnection between the manufacturer and the equipment type as well as the real-time health status of the equipment. The quantitative evaluation constructed by principal component analysis can directly provide reference for all aspects of equipment. This research provides a new way for grid equipment environment data fusion of smart grid. A new exploration is made for the assessment of equipment, which provides a reference for more efficient models in the future and a more accurate assessment of device status.

References
[1] H. Jiang, K. Wang, Y. Wang, M. Gao and Y. Zhang, "Energy big data: A survey," in IEEE Access, vol. 4, no., pp. 3844-3861, 2016.
[2] K. Wang; H. Li; Y. Feng; G. Tian, "Big Data Analytics for System Stability Evaluation Strategy in the Energy Internet," in IEEE Transactions on Industrial Informatics , vol.PP, no.99, pp.1-1
[3] J. Hu and A. V. Vasilakos, "Energy Big Data Analytics and Security: Challenges and Opportunities," in IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2423-2436, Sept. 2016.
[4] J. Choi, M. Kim and J. Yoon, "Implementation of the Big Data Management System for Demand Side Energy Management," 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, 2015, pp. 1515-1520.
[5] M. Shah, P. K. Shukla and R. Pandey, "Phase level energy aware map reduce scheduling for big data applications," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, Odisha, India, 2016, pp. 532-535.

[6] F. Xiao, S. Wang and C. Fan, "Mining Big Building Operational Data for Building Cooling Load Prediction and Energy Efficiency Improvement," 2017 IEEE International Conference on Smart Computing (SMARTCOMP), Hong Kong, China, 2017, pp. 1-3.

[7] Z. Zhou et al., "Game-Theoretical Energy Management for Energy Internet With Big Data-Based Renewable Power Forecasting," in IEEE Access, vol. 5, no. , pp. 5731-5746, 2017.

[8] K. Yang, Q. Yu, S. Leng, B. Fan and F. Wu, "Data and Energy Integrated Communication Networks for Wireless Big Data," in IEEE Access, vol. 4, no. , pp. 713-723, 2016.

[9] E. Casalicchio, L. Lundberg and S. Shirinbad, "An Energy-Aware Adaptation Model for Big Data Platforms," 2016 IEEE International Conference on Autonomic Computing (ICAC), Wurzburg, 2016, pp. 349-350.

[10] Z. Niu and B. He, "A Study of Big Data Computing Platforms: Fairness and Energy Consumption," 2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW), Berlin, 2016, pp. 207-209.