Research and application of fault diagnostics method for new energy power plant equipment based on big data mining

Y Zhang1,3, S M Sun1, P Yu1, Y J Wang1 and S Li2

1State Grid Shandong Electric Power Research Institute, Jinan 250000, Shandong Province, China
2State Grid Shandong Electric Power Company, Jinan 250000, Shandong Province, China
E-mail: wanzhangyong@163.com

Abstract. In view of the diversity, complexity and magnanimity of information data in the operation status of new energy power plant, the data mining technology is applied to fault diagnostics. On this basis, the overall process of fault diagnostics methods including clustering analysis, fault rule mining, fault modeling and other mathematical statistical theories is designed. The logical relationship between massive data is redefined at a deeper level. And the most effective fault characterization is extracted. The correctness and validity of the method are proved by comparing the curve of the fault characteristic parameter prediction results with the field measurement results. This method has high data processing efficiency, effectively improves the reliability and accuracy of fault diagnostics and early warning. And it provides a strong guarantee for the safe operation of the new energy grid-connected.

1. Introduction
The stable operation of power plant equipment is the important foundation for ensuring the safe operation of power grid. With the progress of modern industrial technology, equipment is developing towards diversification [1]. The new energy power plant includes network equipment, converter, transformer subplant main equipment, voltage transformer, current transformer, switch, arc suppression coil, dynamic reactive compensation device, box transformer, photovoltaic cell, multi winding transformer, inverter, solar battery board and so on. The possibility of equipment fault is also increasing due to the increasing type and number of equipment. The introduction of fault diagnostics technology [2] to daily equipment supervision is an effective way to ensure the safe operation of equipment and improve the efficiency of operation and maintenance. The development of modern new energy power plant equipment is increasingly efficient, complex and networked. And the structural relationship of all parts is more and more complex and close. The consequent fault is also diverse, hidden and complex. The analysis ability of traditional fault diagnostics method cannot meet the needs of equipment diagnostics in power plant. Intelligent system has become an important development direction and inevitable trend in the field of fault diagnostics research.

With the increasing number of devices and the increasing complexity of structure, the data of state parameters increase exponentially, the difficulty of equipment operation and maintenance is increasing [3]. The throughput of monitoring information of the daily status of new energy power plant equipment is huge. Facing the challenge of massive and complex information, the data mining technology came into being under the development of computer science and technology, which
provides new technical support for the intelligent equipment fault diagnostics [4]. The key to the accuracy and effectiveness of data mining technology lies in the selection of its algorithms. We can extract hidden patterns from massive, uncertain and noisy mass information only by using the correct algorithm [5,6]. However, because of the continuous change of data storage mode and the rapid intensification of the complexity of data relations, the demand for data mining technology is getting higher and higher. The selection of mining algorithms and the formulation of the mining process still need to be deeply studied. Data mining is a long way to go in the development of the fault diagnostics domain.

In order to ensure the accuracy and reliability of information processing of multi-source and massive state of power plant equipment, based on the current research status of data mining, the theories and tools of mathematical statistics and pattern recognition are introduced into the domain of fault diagnostics. A fault diagnostics method for new energy power plant equipment based on big data mining is proposed. It focuses on mining the correlation between the factors under the uncertainty model and deleting the uncertainty of complex redundant data information. Combined with the practical engineering application, a reasonable diagnostic process is developed. And it realizes function for precise positioning, fault diagnostics and fault warning. The application of the diagnostic method makes the equipment operation and maintenance change from the traditional plan maintenance to the state maintenance, ensures the stable operation of the new energy power plant equipment, and improves the safety reliability of the grid-connected. The economic losses caused by power failure in grid-connected accidents are reduced. It is conducive to the implementation of large-scale energy security grid-connected new energy, which is of great significance to the development of new energy sources in China.

2. Theoretical basis of equipment fault diagnostics based on big data mining

2.1. Clustering algorithm for data mining

The equipment fault diagnostics method based on big data mining [7-9] was mining common fault modes of equipment through clustering analysis of abnormal data parameters. The clustering method used in this paper was widely used in K-means clustering.

The basic idea of K-means clustering algorithm is as follows: (1) From the n objects in the dataset, K centroids were randomly selected, representing the initial cluster center points of the K class clusters, respectively. (2) The Euclidean distance between each data object and each cluster center was calculated, the data object was divided into the nearest cluster. (3) For each cluster that has completed in the step (2), the internal mean was reprocessed, the new mean of the object in the cluster was calculated, the new mean was used as the center of the cluster. (4) By comparing the calculation of the cluster center, if the convergence condition was met, the algorithm would terminate, otherwise it would go back to step (2) and continue to iterate and update until convergence [10-14]. The clustering criterion function uses the error square sum criterion function, which is defined as:

$$G_k = \sum_{j=1}^{K} \sum_{i=1}^{n}|x_i^{(j)} - m_j|^2$$

(1)

For each data object i, the function of its class is calculated as follow:

$$c_i = \arg \min_{j} |x_i - m_j|^2$$

(2)

The value of $m_j$ is the center of clustering.

The center point calculation function for the cluster j center is as follow:

$$m_{new\_j} = \frac{\sum_{i=1}^{n} W_{ij} x_i}{\sum_{i=1}^{n} W_{ij}}$$

(3)
Where $x_i$ is the data object, $w_j$ is the criterion function to calculate whether the data object $x$ is in the cluster $j$. In the cluster range, $w_j=1$; not in, $w_j=0$.

2.2. A regular algorithm between information parameters

The new energy plant has many kinds of equipment. And the information parameters collected are huge and redundant. As far as the research status was concerned, the relationship between equipment parameters cannot be deeply analyzed. And there is a lack of understanding of the systematic nature for equipment parameters. This leads to the reduction of the accuracy of fault diagnostics. Mining association rules was mainly to display all possible related links between items and items in the database in the form of rules, which was an algorithm for helping people make management decisions. In the existing association rules of mining algorithm, the Apriori algorithm was the most influential and widely used.

The Apriori algorithm adopts the iterative algorithm based on layer by layer search. The basic idea of the algorithm was as follows: Firstly, find out the frequent 1 item sets $L_1$ from the transaction database $D$; Then the candidate frequent 2 items sets $C_2$ is generated according to the $L_1$ search, the item with less than the minimum support degree in $C_2$ was eliminated, the frequent 2 items sets $L_2$ was obtained; Then, the candidate frequent 3 items sets $C_3$ was generated by $L_2$ iterative search, so that the maximum frequent k items sets $L_k$ can be found [15]. When all the frequent items sets in the database were found, they can be generated according to these frequent items sets, that was, the support degree of the association rules was greater than the minimum support degree, the strong association rules of the association rule's confidence greater than the minimum confidence level can be satisfied.

- The algorithm with minimum support degree is as follows:

The transaction number in the known transaction database $D$ contains item set $A$ with CourrtA, then the A's item set support degree is defined as:

$$Support(A) = P(A) = \frac{CountA}{CountD}$$

For a given association rule $A \Rightarrow B$, it was known that the transaction number of the item set $A$ and the item set $B$ in the transaction database $D$ is $Count(A \cap B)$, then the probability of the item set $A$ and $B$ at the same time in the transaction, that is, the confidence of its association rules is as follow:

$$Support(A \Rightarrow B) = P(A \cap B) = \frac{Count(A \cap B)}{CountB}$$

Where $Support(A \Rightarrow B) \in [0,1]$

The minimum support degree is defined as $\min Sup$, where $\min Sup \in (0,1]$

- The minimum confidence algorithm is as follow:

For a given association rule $A \Rightarrow B$, the probability of the item set $A$ and $B$ at the same time in the transaction, that is, the confidence of its association rules is as follow:

$$Confidence(A \Rightarrow B) = P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{Count(A \cap B)}{CountB}$$

Where $Confidence(A \Rightarrow B) \in [0,1]$

The minimum confidence degree is defined as $\min Conf$, where $\min Conf \in (0,1]$

2.3. Mining construction of data model

A large amount of data will be generated during the operation of the equipment, resulting in large amount of computation for raw features and sparse distribution of samples. The traditional physical model cannot find out the abnormal situation quickly and make accurate diagnostics and evaluation [16]. Therefore, a data-based modeling method is proposed. On the basis of processing and analyzing data sets and obtaining the most effective feature parameters, the data model construction process is as
follows:
3 wavelets multi-resolution analysis of data based on Db10 wavelet basis was performed, 3-layer wavelet coefficients of low frequency band was obtained. Each fault data set has to compute a 3-dimensional feature vector, as shown in formula (7):

\[ f = [f_1, f_2, f_3], \quad f_i = \sum_{j=1}^{i} w_j^i \]  

(7)

Here, \( i \) is the number of wavelet decomposition layers, \( w \) is the decomposition coefficient of each layer wavelet.

The mixed Gauss model was used to describe the complex distribution of multi-dimensional feature data, generating mathematical model of fault mode. Where, the Gauss mixture model expression is shown in formula (8).

\[ p(z) = \sum_{j=1}^{M} w_j p_j(z) = \sum_{j=1}^{M} w_j N(z; \mu_j; \Sigma_j) \]  

(8)

Where \( z \) is the characteristic data vector, \( M \) is the mixed number of the model, \( w_j \) is the weight coefficient of the mixed model:

\[ N(z; \mu; \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (z - \mu)^T \Sigma^{-1} (z - \mu)} \]  

(9)

Where \( \mu \) is the central point of the density function, \( \Sigma \) is the covariance matrix of this density function.

3. Process analysis of fault diagnostics method of new energy plant equipment

3.1. Overall design of fault diagnostics method

The research and design of fault diagnostics method of new energy power plant equipment had the modules of data preprocessing, data clustering analysis, data model fault diagnostics, fault data prediction and so on. The specific flow chart of the diagnostic method is shown in figure 1.

For all kinds of new energy plant equipment, the equipment status information was extracted according to a certain period, the equipment characteristic information parameters were obtained after pretreatment; Then, the feature data was saved to the equipment running history database, marking and classifying the fault types of the abnormal state data and constructing the data model; Next, match the real-time monitoring data with the fault running data model. If the match was unsuccessful and no actual fault occurs, the end analysis was performed. If the match was unsuccessful and there was a fault, the new fault mode was added to the fault model library after expert’s diagnostics. If the match was successful, the fault type was determined and the fault was located to the specific fault element; And the probability of occurrence of the dominant fault [17] in the whole fault library was calculated. Finally, the key factor of the fault state was used as the input value of the trained BP neural network model, the hidden fault value of the fault element was obtained and its range was determined. It can judge the development level of hidden faults and predict the development trend of hidden faults. It can also give the result of hidden fault early warning and give an alarm in time.
3.2. Data processing flow of fault diagnostics

In view of the above new energy plant equipment fault diagnostics method, taking three-phase inverter fault diagnostics [18,19] as an example, the data mining process was described in detail. First, the abnormal state data of the inverter was collected, including malfunction and defect data. The circuit connection of the three-phase inverter is shown in figure 2. This paper focuses on the open circuit and short circuit fault of the most common switching power transistor. The characteristic information parameters select the voltage and current signal. The signals were used as the known quantity to input the fault models. The output impedance of the model was compared with the actual measured output impedance to get the residual error ($\gamma$). The fault information contained in the residual signal was used for the fault decision.

![Figure 1. Overall block diagram of fault diagnostics method.](image)

![Figure 2. Photovoltaic grid-connected inverter.](image)

Based on the data collected from several inverters, the k-means algorithm was used to cluster the original data. In order to adapt to the mining algorithm, the residuals are discretized. The residual value was recorded as "0" within the threshold range and "1" outside the threshold range, which can be understood as the possibility of fault caused by the residual. The relationship between the processed
fault set and the residual set was shown in table 1. $\gamma_1, \ldots, \gamma_7$ is the acquired residual. $R_{on1}, \ldots, R_{on6}, R_{off1}, \ldots, R_{off6}$ is the fault parameters.

| $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\gamma_4$ | $\gamma_5$ | $\gamma_6$ | $\gamma_7$ |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| $R_{on1}$ | 1         | 1         | 0         | 0         | 1         | 0         |
| $R_{on2}$ | 0         | 0         | 0         | 0         | 0         | 1         |
| $R_{on3}$ | 1         | 0         | 1         | 0         | 0         | 1         |
| $R_{on4}$ | 0         | 0         | 0         | 1         | 0         | 0         |
| $R_{on5}$ | 1         | 0         | 0         | 1         | 0         | 1         |
| $R_{on6}$ | 0         | 0         | 0         | 1         | 1         | 0         |
| $R_{off1}$ | 1         | 1         | 0         | 0         | 1         | 0         |
| $R_{off2}$ | 0         | 0         | 0         | 1         | 0         | 0         |
| $R_{off3}$ | 1         | 0         | 1         | 0         | 0         | 1         |
| $R_{off4}$ | 0         | 1         | 0         | 1         | 0         | 0         |
| $R_{off5}$ | 1         | 0         | 0         | 1         | 0         | 1         |
| $R_{off6}$ | 0         | 0         | 1         | 0         | 0         | 1         |

From the above table, we can see that the dimension of fault data is high and contains more redundant information. SIE algorithm was used to reduce the dimensionality of high-dimensional data. And the adaptive maximum likelihood estimation method was used to estimate the dimension of dimensionality reduction. The results of the reduction are shown in following table 2. Data dimensionality was reduced and a lot of useless information was removed, which provides a prerequisite for mining subsequent fault rules.

| $\gamma_1$ | $\gamma_2$ | $\gamma_5$ | $\gamma_7$ |
|-----------|-----------|-----------|-----------|
| $R_{on1}$ | 1         | 1         | 0         |
| $R_{on2}$ | 0         | 0         | 0         | 1         |
| $R_{on3}$ | 1         | 0         | 0         | 0         |
| $R_{on4}$ | 0         | 0         | 1         | 0         |
| $R_{on5}$ | 1         | 0         | 0         | 1         |
| $R_{on6}$ | 0         | 0         | 1         | 0         |
| $R_{off1}$ | 1         | 1         | 1         | 0         |
| $R_{off2}$ | 0         | 0         | 0         | 1         |
| $R_{off3}$ | 1         | 0         | 0         | 0         |
| $R_{off4}$ | 0         | 1         | 1         | 0         |
| $R_{off5}$ | 1         | 0         | 0         | 1         |
| $R_{off6}$ | 0         | 0         | 0         | 0         |

In order to improve the comprehensiveness and accuracy of fault diagnostics, the Apriori algorithm was used to mine fault rules after reducing the fault data information. In order to dig out valuable and meaningful strong association rules, we need to set confidence and support values reasonably. Because the equipment has many state parameters, the threshold of support should not be too large. The 0.1 was set as minimum support. At the same time, in order to ensure the high reliability of the obtained association rules, 0.85 was set as the minimum confidence. The minimum fault characteristics obtained after the above processing are shown in table 3 for fault decision making.
Table 3. Minimum fault characterization after integration.

|       | $\gamma_1$ | $\gamma_2$ | $\gamma_5$ | $\gamma_7$ |
|-------|-------------|-------------|-------------|-------------|
| $R_{on1}$ | 1           | 1           | 1           | 0           |
| $R_{on2}$ |             |             |             | 1           |
| $R_{on3}$ | 1           |             | 0           |             |
| $R_{on4}$ |             |             |             | 1           |
| $R_{on5}$ | 1           |             |             | 1           |
| $R_{on6}$ |             |             |             | 1           |
| $R_{off1}$ |             |             | 1           | 0           |
| $R_{off2}$ | 0           |             | 0           |             |
| $R_{off3}$ |             |             | 1           | 0           |
| $R_{off4}$ |             |             | 1           | 1           |
| $R_{off5}$ | 0           |             |             | 1           |
| $R_{off6}$ |             |             |             | 0           |

4. Application examples

The new energy plant equipment fault diagnostics method was tested on photovoltaic power plant in June 2017. The key characteristic information parameters of three-phase inverter-voltage and current signals were predicted. Meanwhile, the effective data of the 9–18h of the inverter under the clear weather condition was obtained by using the rate of 5000 Hz, each 6min was measured in a group, each group measures 40s. The prediction results and the analysis as well as the comparison curves with the field measurement results are shown in figures 3 and 4.

![Figure 3. The comparison curve between voltage prediction and measured results (standard value).](image1)

![Figure 4. The comparison curve between the predicted current and the measured results.](image2)

By comparing the voltage and current signals of the measured and predicted models in the above charts, we can see that the predicted results of voltage and current data are similar to the actual measured results. The trend of the curve is almost the same, which shows that the fault diagnostics method designed in this paper has more accurate prediction ability of fault data. It is possible to detect hidden fault of new energy power plant equipment early. And it can judge its development degree and stop the spread of faults in time. The correctness and effectiveness of the fault diagnostics method are proved.

5. Conclusions

The normal operation of new energy plant equipment will directly affect the security and reliability of the new energy grid-connected. The status diagnostics and early warning evaluation are the key issues in the current new energy grid-connected. For this reason, this paper proposes a fault diagnostics method for new energy power plant based on big data mining. Through a series of processing modes
such as cluster analysis, fault rule mining and fault data modeling, the relationship between fault features and equipment is extracted, which reduces the redundancy of diagnostics information. The accurate diagnostics of equipment failure and accurate prediction of fault data are realized. Finally, the reliability of fault data prediction ability is verified through the comparison of prediction results of fault feature information and field measurement results. The design and successful application of this method is of practical value for new energy plant equipment. It can effectively improve its operation reliability, prolong the service life of the equipment. It is playing an active role in reducing economic losses and improving the safety and stability of new energy grid-connected.

References
[1] Yu B W 2015 Application of data mining technology in fault diagnostics (Daqing: Northeast Petroleum University)
[2] Zhang X, et al 2013 Research process and prospect of fault diagnostics of power grid Power Sys. Techno. 10 2745-53
[3] Zhou L L, Sun J and Wang T 2015 Research on data mining technology and methods for equipment fault diagnostics Techniques of Automation and Applications 34 18-22
[4] Yuan J 2012 Research and implementation of key technology of data mining for equipment fault diagnostics (Xi’an: Xidian University)
[5] Wu Y Y 2017 Data mining based algorithm for detecting and identifying bad data in power system (Nanjing: Nanjing University of Science and Technology)
[6] D’Antona G and Perfetto L 2015 Bad data detection and identification in power system state estimation with network parameters uncertainty Int. Conf. on Knowledge-Based Eng and Innovation (Tehran, Iran) pp 26-31
[7] Jiang B B 2011 Research on fault diagnostics method based on Data Mining (Tianjin: Institutes of Technology of Tianjin)
[8] Li R, Zhang L Y and Gu X P 2010 Distributed power system fault diagnostics based on rough set and joint rule mining algorithm Proceedings of the CSEE 30 28-34
[9] Yang C and Zhang L 2016 Application of data mining technology in power equipment fault diagnostics Electric Switchgear 2 83-6
[10] Meng J J and Liu D C. 2016 A new method for identifying bad data in power system based on Spark and cluster analysis Power Sys. Protec. Control 03 85-91
[11] Zhao J, et al 2015 Multistage phasor-aided bad data detection and identification General Meeting of the IEEE Power & Energy Soc. (Denver, CO, USA) pp 1-5
[12] Thong Y and Liu D 2016 The application of K-means clustering algorithm based on Hadoop Cloud 2016 IEEE Int. Conf. on. Computing and Big Data Analysis (Chengdu, China) pp 88-92
[13] Mohamed A and Wesam A 2013 Efficient data clustering algorithms: Improvements over K-means Int. J. Intelligent Sys.(IJISA) 5 37-49
[14] Xu X D, Li K and Zhu S R 2010 Research on the improvement of Apriori algorithm in Web usage mining Comput. Eng. Des. 31 539-41
[15] Chen W, et al 2011 Parallel spectral clustering in distributed systems IEEE Trans. Pattern Analysis and Mac. Intelligence 33 568-86
[16] Calderaro V, et al 2011 Failure identification in smart grids based on petri net modeling industrial IEEE Trans. Electro. 58 4613-23
[17] Xiong W H, et al 2013 Probabilistic assessment of transformer potential faults based on Bayesian network Electric Application S2 706-10
[18] Masrur M A, Chen Z and Murphey Y 2010 Intelligent diagnostics of open and short circuit faults in electric drive inverters for real-time applications IET Power Electro. 3 279-91
[19] Choi U M, et al 2012 Method for detecting an open-switch fault in a grid-connected NPC inverter system IEEE Trans. Power Electro. 27 2726-39