ARCHIVES OF ELECTRICAL ENGINEERING VOL. 69(4), pp. 793–800 (2020)

DOI 10.24425/aee.2020.134630

Research on operation fault diagnosis algorithm of power grid equipment based on power big data

JIANGUO QIAN[1], BINGQUAN ZHU[1], YING LI[1], ZHENGCHAI SHI[2]

[1] *Equipment Monitoring Department, State Grid Zhejiang Electric Power Company China*

[2] *Wenzhou Power Supply Company, State Grid Zhejiang Electric Power Company China*

e-mail: gj1owz@126.com

(Received: 17.04.2020, revised: 22.06.2020)

Abstract: Power big data contains a lot of information related to equipment fault. The analysis and processing of power big data can realize fault diagnosis. This study mainly analyzed the application of association rules in power big data processing. Firstly, the association rules and the Apriori algorithm were introduced. Then, aiming at the shortage of the Apriori algorithm, an IM-Apriori algorithm was designed, and a simulation experiment was carried out. The results showed that the IM-Apriori algorithm had a significant advantage over the Apriori algorithm in the running time. When the number of transactions was 100 000, the running of the IM-Apriori algorithm was 38.42% faster than that of the Apriori algorithm. The IM-Apriori algorithm was little affected by the value of $\text{support}_{\min}$. Compared with the Extreme Learning Machine (ELM), the IM-Apriori algorithm had better accuracy. The experimental results show the effectiveness of the IM-Apriori algorithm in fault diagnosis, and it can be further promoted and applied in power grid equipment.

Key words: association rules, big data, data mining, fault diagnosis, grid equipment

1. Introduction

With the development of technology, the power industry has also gained rapid development, for example, the power grid scale is larger, the equipment structure is more complex, and the operation mode is more and more diverse, which makes the fault diagnosis of power grid equipment in operation more and more difficult [1]. The traditional fault diagnosis method is to carry out regular maintenance of power grid equipment. No matter whether the equipment is in fault or not,

the maintenance of the equipment will be carried out through the established process and means after power failure. This way will not only cause a huge waste of human and material resources, but also cannot find the fault in time and effectively. With the development of power data, which is more and more quantitative and complex, the fault diagnosis method based on power big data has been widely concerned by researchers [2]. Aggarwal *et al*. [3] combined empirical mode decomposition (EMD) with the probabilistic neural network (PNN), studied the fault classification of transmission lines, used MATLAB/SIMULINK to carry out experiments in 500 kV, 300 km transmission lines, and verified the excellent learning speed and classification accuracy of this method. Lazim *et al*. [4] designed the fault diagnosis method of a transmission line through an artificial neural network (ANN) by using two factors, bus voltage and line fault current, and then simulated it in MATLAB 6.0. Kari *et al*. [5] designed a method based on an adaptive neural fuzzy inference system (ANFIS) and the Dempster-Shafer theory (DST) on the basis of dissolved gas in oil (DGA), compared it with an ordinary ANFIS through experiments, and verified the effectiveness of the method. Sahri *et al*. [6] designed a genetic algorithm (GA)-support vector machine (SVM) algorithm for transformer fault classification, then carried out experiments using data sets from the real world, and found that the GA-SVM method eliminated redundant features. In this study, the big data of the power system were processed using association rules, the Apriori algorithm was analyzed and improved to obtain the IM-Apriori algorithm, and the simulation experiment was carried out to verify the reliability of the algorithm. This study makes some contributions to improving the operation efficiency of power system equipment.

2. Fault diagnosis method based on Apriori algorithm

2.1. Association rule mining

The operation state of power grid equipment can be affected by many factors, such as DGA [7], partial discharge (PD) [8], etc., which are the basis of fault diagnosis. In order to get useful information from power big data, this study used the method of association rules to analyze and process these data.

Association rule [9] refers to the rule that has some association relationship in data. In fault diagnosis, different power data represent different fault information. Through association rules, the potential relationship between data information and fault can be mined, so that the fault type of equipment can be determined as soon as possible through the analysis of data information. The related concepts are as follows.

Transaction database is written as D, subset transaction as T, and then there is $D = \{T_1, T_2, \cdots, T_N\}$, where N refers to the number of T. The total number is written as $|D|$, and the subset transaction was written as $T_n = \{\lambda_1, \lambda_2, \cdots, \lambda_M\}$, where λ stands for the item and M stands for the number of items. The set of items in D is represented by Ψ. If there is the item set $|A| = k$, then A is a k-item set. The frequency of the item set is written as $f(A)$, and its support degree, i.e. the proportion of the number of transactions containing A to the total number of transactions, is written as support(A):

$$\text{support}(A) = \frac{f(A)}{|D|} \times 100\%. \tag{1}$$

$\text{support}_{\min}$ is taken as the minimum support degree specified by a user. When $\text{support}(A) > \text{support}_{\min}$, A is a set of frequent items, which is called a frequency set for short; otherwise, it is a non-frequency set. It is assumed that there are item sets P and Q in D.

If $P \subseteq Q$, then:

1) $\text{support}(P) \geq \text{support}(Q)$;

2) P is a non-frequency set, then Q is also a non-frequency set;

3) Q is a frequency set, then P is also a frequency set.

If there are item sets P and Q and $P \cap Q \neq \varphi$, then formulas similar to $P \Rightarrow Q$ can be called as the associate rule, where P is the former term and Q is the latter term. The support degree of the item set $P \cup Q$ is the support degree of $P \Rightarrow Q$:

$$\text{support}(P \Rightarrow Q) = \text{support}(P \cup Q). \tag{2}$$

Confidence refers to the proportion that D not only includes P but also includes Q, which is written as $\text{confidence}(P \Rightarrow Q)$, and its formula is:

$$\text{confidence}(P \Rightarrow Q) = \frac{\text{support}(P \cup Q)}{\text{support}(P)} \times 100\%. \tag{3}$$

$\text{confidence}_{\min}$ is taken as the minimum confidence specified by a user.

When

$$\text{support}(P \Rightarrow Q) \geq \text{support}_{\min}$$

and

$$\text{confidence}(P \Rightarrow Q) \geq \text{confidence}_{\min},$$

then $P \Rightarrow Q$ is the strong rule; otherwise, it is the weak rule.

Therefore, the mining of association rules mainly includes:

1) finding out all the frequency sets in D according to $\text{support}_{\min}$;

2) finding out the strong rule according to the frequency set and $\text{confidence}_{\min}$.

2.2. Apriori algorithm

Apriori is a kind of association rules [10]. Its principle is as follows. Firstly the candidate item set is generated, and it is pruned to obtain the frequency set. Starting from a 1-item set, the frequency 1-item set L_1 is found. Then a candidate 2-item set is generated according to L_1, which is written as C_2, and L_2 is obtained after pruning. The rest can be done in the same manner until the frequency set L_k is obtained.

2.3. Fault diagnosis algorithm based on improved Apriori algorithm

In order to make the Apriori algorithm have a better application in the power big data, this study combined the bit string logic operation to improve the Apriori algorithm and obtain the IM-Apriori algorithm. Its principle is to use the concept of "bit" and represent transaction in D by a bit string. The steps of the algorithm are as follows:

1. Threshold $\text{support}_{\min}$ and $\text{confidence}_{\min}$ are specified.

2. The whole database is scanned. If an item appears in the transaction, "1" is written; if it does not appear, "0" is written. The number of "1s" in each item is counted and written as the support count of the item, and the candidate item whose support count is larger than the threshold is regarded as the item in L_1.
3. The item sequence H is generated according to L_1. Every item is coded to obtain a coding bit string. If an item appears in the generated item set, the corresponding item sequence position is written as "1", otherwise it is written as "0".
4. The logic "OR" operation is performed on the item bit string in L_{k-1}, and the number of "1s" is calculated. If it is k, then it is added to the candidate item set C_k.
5. The logic "OR" operation is performed in L_{k-1} to obtain the item bit string of C_k. The number of "1s" is the support count. The candidate item whose support count is larger than the threshold is taken as the item of L_k. The operations are repeated until the number of single items in L_k is smaller than $k+1$.

Transformer fault is taken as an example. Five DGA content, H_2, CH_4, C_2H_6, C_2H_4 and C_2H_5, is taken as fault indexes and numbered as P1-5. Then five fault types and one normal state shown in Table 1 are selected. One thousand known data information is searched. The calculation is carried out according to the IM-Apriori algorithm designed in this study.

Table 1. Fault types

Number	Fault type
D1	Normal
D2	Medium and low temperature overheating
D3	High temperature overheating
D4	Low energy discharge
D5	High-energy discharge
D6	Winding fault

Taking winding fault as an example, there are 152 known data of this type. Among the indexes related to this fault, H_2, C_2H_6 and C_2H_4 are abnormal for 126, 9 and 8 times, respectively, while the other indexes are normal. According to the IM-Apriori algorithm, there is transaction database

$$D_j = \{\text{the } j\text{-th fault type exceeds standard}\},$$

the item set

$$P_{j,k} = \{\text{the } k\text{-th index in the } j\text{-th fault type is abnormal}\}$$

and

$$Q_j = \{\text{the } j\text{-th fault appears}\}.$$

For $P_{j,k} \Rightarrow Q_j$, its correlation degree is:

$$\text{support}\left(P_{j,k} \Rightarrow Q_j\right) = \frac{f\left(P_{j,k} \cup Q_j\right)}{\left|Q_j\right|} \times 100\%,$$

in the winding fault

$$|D_6| = |Q_6| = 152,$$
$$f(P_{6,1} \cup Q_6) = 129,$$
$$f(P_{6,2} \cup Q_6) = 0,$$
$$f(P_{6,3} \cup Q_6) = 9,$$
$$f(P_{6,4} \cup Q_6) = 8$$

and

$$f(P_{6,5} \cup Q_6) = 0.$$

Through calculation, there is:

$$\text{support}(P_{6,1} \Rightarrow Q_6) = \frac{f(P_{6,1} \cup Q_6)}{|Q_6|} \times 100\% = \frac{129}{152} \times 100\% = 84.87\%, \tag{4}$$

$$\text{support}(P_{6,3} \Rightarrow Q_6) = \frac{f(P_{6,3} \cup Q_6)}{|Q_6|} \times 100\% = \frac{9}{152} \times 100\% = 5.92\%, \tag{5}$$

$$\text{support}(P_{6,4} \Rightarrow Q_6) = \frac{f(P_{6,4} \cup Q_6)}{|Q_6|} \times 100\% = \frac{8}{152} \times 100\% = 5.26\%. \tag{6}$$

It is found that the value of support $(P_{6,1} \Rightarrow Q_6)$ is the largest, indicating that there is a strong correlation between the index $P_{6,1}$ and fault type Q_6. When the index has abnormality, it can be determined that the fault type of the transformer is Q_6. The other corresponding relationships can be obtained by calculating in a similar way.

3. Simulation results

The simulation experiment was carried out in the Eclipse environment. The algorithm was deployed in the Storm platform. The CPU of the computer was 2.1 GHz, and the memory was 8 GB. Taking the transformer fault in the power equipment as an example, the fault type is shown in Table 1. Then the collected 1 000 data were randomly simulated to form 100 000 data. The value of $\text{support}_{\min}$ was set as 0.1. The performance of Apriori and IM-Apriori algorithms under different number of transactions was compared, and the results are shown in Fig. 1.

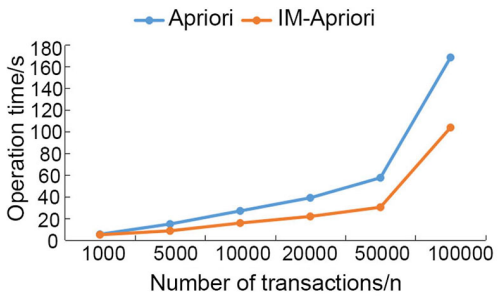

Fig. 1. The running time under different number of transactions

It was seen from Fig. 1 that the algorithm difference was small when the number of transactions was small; when the number of transactions was 1 000, the running time of the IM-Apripri algorithm was 7.84% less than that of the Apriori algorithm, and the gap was small; with the increase of the number of transactions, the gap between the two algorithms gradually increased; when the number of transactions was 10 000, the gap was 11.2 s, and when the number of transactions was 100 000, the gap was 64.7 s, i.e. the running time of the IM-Apripri algorithm was 38.42% less than that of the Apriori algorithm, which showed that the IM-Apripri algorithm designed in this study had excellent performance in big data and could significantly reduce the running time and improve the efficiency of fault diagnosis.

Taking 10 000 transactions as an example, the algorithms under different $support_{min}$ were compared, and the results are shown in Fig. 2.

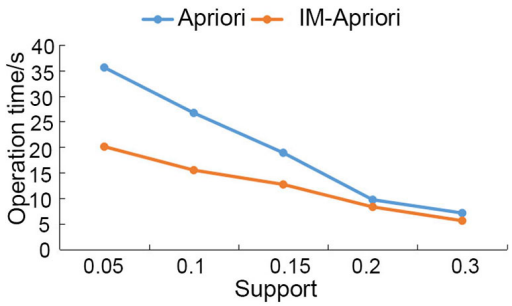

Fig. 2. The operation time under different support degrees

It was seen from Fig. 2 that the smaller the value of $support_{min}$, the longer the running time, which was because that many candidate item sets generated and the time of scanning database was long when the value of $support_{min}$ was small. Under different $support_{min}$, the running time of the IM-Apriori algorithm was shorter than that of the Apriori algorithm. Moreover the influence of the value of $support_{min}$ on the Apriori algorithm was larger than that on the IM-Apriori algorithm, indicating that the IM-Apriori algorithm reduced the number of candidate sets and improved the efficiency of the algorithm in the fault diagnosis.

To verify the accuracy of the algorithm in the fault diagnosis, data of different fault types were used, 1 000 data for each type, and the results were compared with the ELM algorithm [11], shown in Fig. 3.

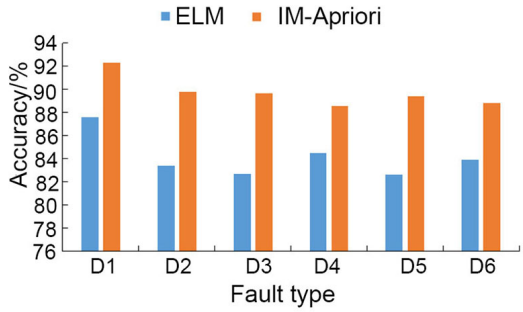

Fig. 3. Comparison of fault diagnosis accuracy

It was seen from Fig. 3 that there was a significant gap. Firstly, the accuracy of the algorithms in the diagnosis of a normal type (D1) was higher than that in the diagnosis of other faults; the accuracy of the IM-Apriori algorithm was higher than that of the ELM algorithm, for example, the accuracy of the ELM for D6 was 83.9%, while that of the IM-Apriori algorithm was 88.8%, which was 5.84% higher than that of the ELM algorithm, which showed that the IM-Apriori algorithm had better accuracy.

4. Discussion

The power grid equipment will produce a large number of data in the process of operation [12]. With the expansion of the power data scale, traditional methods have not been able to reasonably use these massive data. Therefore, big data technology is needed. Big data technology can extract valuable information from a large number of data, which has good applications in many aspects of power grid equipment, such as the diagnosis of distribution network conditions [13], power grid load prediction [14], transmission line evaluation [15], user electricity behavior analysis [16], reactive power optimization problems [17], etc. This study mainly analyzed the fault diagnosis of power grid equipment operation.

For power big data, there are many methods of fault diagnosis, such as a support vector machine [18], decision tree [19], neural network [20], Bayes [21], etc. In this study, association rules were selected for research, and the Apriori algorithm was improved to obtain better fault diagnosis performance. It was seen from the experimental results that the performance of the IM-Apriori algorithm had a significant improvement compared to the traditional Apriori algorithm. Fig. 1 shows that the IM-Apriori algorithm had an obvious advantage for power big data; under the large data volume, the IM-Apriori algorithm remained at a relatively high operation speed; for example, when the number of transactions was 100 000, the operation time of the IM-Apriori algorithm was 38.42% shorter than that of the Apriori algorithm, indicating that the efficiency of the IM-Apriori algorithm was higher when the data volume was large. It was found from Fig. 2 that the IM-Apriori algorithm was little affected by $\text{support}_{\text{min}}$, and the accuracy of the IM-Apriori algorithm was higher compared to the ELM algorithm. The experimental results showed that the IM-Apriori algorithm had a good performance in fault diagnosis.

Although some achievements have been obtained from the research of fault diagnosis of power grid equipment operation, there are some shortcomings, which need to be further improved in future work:

1) studying the application of other big data technologies;
2) searching for other methods to improve the Apriori algorithm, such as the selection of an optimization threshold;
3) carrying out experiments on more fault types.

References

[1] Xu Y., Sun Y., Wan J., Liu X., Song Z., *Industrial Big Data for Fault Diagnosis: Taxonomy, Review, and Applications*, IEEE Access, vol. 5, pp. 17368–17380 (2017).

[2] Wang H., *Fault diagnosis of analog circuit based on wavelet transform and neural network*, Archives of Electrical Engineering, vol. 69, no. 1, pp. 175–185 (2020).

[3] Aggarwal A., Malik H., Sharma R., *Feature extraction using EMD and classification through Probabilistic Neural Network for fault diagnosis of transmission line*, IEEE International Conference on Power Electronics (2017).

[4] Lazim F.B., Hamzah N., Arsad P.M., *Application of ANN to power system fault analysis*, Conference on Research and Development (2016).

[5] Kari T., Gao W., Zhao D., Zhang Z., *An integrated method of ANFIS and Dempster-Shafer theory for fault diagnosis of power transformer*, IEEE Transactions on Dielectrics and Electrical Insulation, vol. 25, no. 1, pp. 360–371 (2018).

[6] Sahri Z., Yusof R., *Fault diagnosis of power transformer using optimally selected DGA features and SVM*, Asian Control Conference, IEEE (2015).

[7] Ab Ghani S.S., Muhamad N.A., *Review on Dissolved Fault Gases in Monitoring Bio-Oil Filled Transformer*, Applied Mechanics and Materials, vol. 818, pp. 69–73 (2016).

[8] Shanker T., Narasimhaiah H.N., Punekar G., *Acoustic emission partial discharge detection technique applied to fault diagnosis: Case studies of generator transformers*, Serbian Journal of Electrical Engineering, vol. 13, pp. 4–4 (2016).

[9] Altuntas S., Dereli T., Kusiak A., *Analysis of patent documents with weighted association rules*, Technological Forecasting and Social Change, vol. 92, pp. 249–262 (2015).

[10] Niu Z., Nie Y., Zhou Q., Zhu L., Wei J., *A brain-region-based meta-analysis method utilizing the Apriori algorithm*, BMC Neuroscience, vol. 17, no. 1, p. 23 (2016).

[11] Scardapane S., Comminiello D., Scarpiniti M., Uncini A., *Online Sequential Extreme Learning Machine With Kernels*, IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 9, pp. 2212–2220 (2015).

[12] Arkovi M., Stojkovi Z., *Analysis of artificial intelligence expert systems for power transformer condition monitoring and diagnostics*, Electric Power Systems Research, vol. 149, pp. 125–136 (2017).

[13] Ma Y., Yu X., Niu Y., *A parallel heuristic reduction based approach for distribution network fault diagnosis*, International Journal of Electrical Power and Energy Systems, vol. 73, pp. 548–559 (2015).

[14] Hasan H., Munawar M.R., Siregar R.H., *Neural network-based solar irradiance forecast for peak load management of grid-connected microgrid with photovoltaic distributed generation*, International Conference on Electrical Engineering and Informatics, IEEE (2017).

[15] Zhu Y., Yan J., Tang Y., Sun Y.L., He H.B., *Joint Substation-Transmission Line Vulnerability Assessment Against the Smart Grid*, IEEE transactions on Information Forensics and Security, vol. 10, no. 5, pp. 1010–1024 (2017).

[16] Li H., Zhang Z., Wang X., Zhou M., Li S., *Electricity Consumption Behaviour Analysis Based on Time Sequence Clustering*, Journal of Physics Conference Series, vol. 1168, p. 032011 (2019).

[17] Li Y.C., Yang R.Y., Zhao X.Y., *Reactive power convex optimization of active distribution network based on Improved Grey Wolf Optimizer*, Archives of Electrical Engineering, vol. 69, no. 1, pp. 117–131 (2020).

[18] Zhang Y.X., Cheng Z.F., Xu Z.P., Bai J., *Application of Optimized Parameters SVM Based on Photoacoustic Spectroscopy Method in Fault Diagnosis of Power Transformer*, Spectroscopy and Spectral Analysis, vol. 35, no. 1, p. 10 (2015).

[19] Wang L., Shang L.L., Ma M.C., Ma Z.G., *Fault Diagnosis and Trace Method of Power System Based on Big Data Platform*, Iop Conference, vol. 394, no. 4, p. 042116 (2018).

[20] Li L., Zhang X., Wang Z., *Fault Diagnosis in Solar Photovoltaic Grid-Connected Power System Based on Fault Tree and BAM Neural Network*, Transactions of China Electrotechnical Society, vol. 30, no. 2, pp. 248–254 (2015).

[21] Lakehal A., Ghemari Z., Saad S., *Transformer fault diagnosis using dissolved gas analysis technology and Bayesian networks*, International Conference on Systems and Control (2015).