PAPER • **OPEN ACCESS**

Method of Analyzing Transformer DC Magnetic Bias Based on Big Data Clustering and Relevance Analysis

To cite this article: Pei Yuan *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **310** 032027

Method of Analyzing Transformer DC Magnetic Bias Based on Big Data Clustering and Relevance Analysis

Pei Yuan[1]*, Tao Wan[1], Xiaowen Wu[1], Huisheng Ye[1] and Wenqi Mao[2]

[1]State Grid Hunan Electric Power Company Limted Research Institute, Changsha, Hunan, 410007, China

[2] State Grid Hunan Electric Power Company Limted, Changsha, Hunan, 410004, China

*Corresponding author's e-mail: ypsnow87@sina.com

Abstract: In this paper, the big data analysis method is applied to the analysis of the factors affecting the DC magnetic bias of the transformer. The improved k-means clustering algorithm is used to identify the different magnetic reasons. The weighted relevance analysis is used to evaluate the influence of different factors on the magnetic current. Finally, the correctness of the method is verified by an example analysis, thus providing an effective basis for "one-button sequence control" DC magnetic bias treatment.

1. Introduction

DC magnetic bias hazard refers to a series of adverse consequences after the DC current invades the AC system and causes the transformer's iron core to saturate. With the rapid development of China's UHV DC transmission and urban rail transit in the past decade, transformer DC magentic has become a hot issue of electromagnetic compatibility, and related forecasting and treatment work is urgently needed. In addition, this current will also invade underground pipelines and buried communication lines, affecting the normal operation of pipeline systems and communication systems, and even causing serious safety production accidents, which will have a very bad impact on people's lives and social stability [1-2].

At present, the research on DC magnetic bias of transformers is mainly based on actual measurement, which is lack of vision. It can only judge the risk of DC magnetic damage by analyzing the DC current value at transformer neutral point. Therefore, it can only adopt measurements passively and a systematic and complete analysis method has not been formed. In the face of problems, only the treatment of 'cure where it hurts' is adopted. The cost of treatment is high, a lot of resources are wasted and the effect is not good.

This paper will creatively introduce the big data thinking into the analysis of DC magnetic bias, and combine the power industry parameters with the non-power industry parameters. Two typical algorithms of big data (data clustering and correlation analysis) are used to realize the quantitative analysis and evaluation of the multiple influencing factors of DC magnetic.

2. Data clustering

2.1. Improved k-means clustering algorithm

In this paper, the improved k-means algorithm based on the attributes of data set can be used to replace the initial center of the k-means algorithm initialization process. The proposed algorithm is more consistent with the data distribution and guarantee the correctness of the results [3-5]. The process of improving the k-means algorithm is as follows:

First, an improved k-means algorithm based on data set attributes creates a high-density data set by adding high density objects. The higher the density of sample points, the denser the data objects around the sample object data objects.

$$R = \alpha \frac{\max\left(\mathrm{dis}\left(p_i, p_j\right)\right)}{k} \tag{1}$$

$$\mathrm{Den}\left(p_i\right) = \sum_{j=1}^{n} u\left(R - \mathrm{dis}\left(p_i, p_j\right)\right), \quad u(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases} \tag{2}$$

In this formula, k is the number of clusters categories, which is an adjustable parameter, generally equals to 1. Dis (pi, pj) is the distance between point pi and point pj (generally Euclidean distance).

Then, according to the dissimilarity matrix, a Huffman tree is constructed from a high-density group, and a cluster center selection is performed through a Huffman tree.To express the dissimilarity of the data, an improved k-means algorithm based on the data set attributes defines the data set D = (X, A), where X = {x1, x2, ..., xm}, representing the m data objects contained in the data set, A = {a1, a2, ..., am}, A corresponds to the data dimension of each x, and the dissimilarity between them is Dis (i, j):

$$\mathrm{Dis}(i, j) = \sqrt{\sum_{k=1}^{n}\left(x_{ik} - x_{jk}\right)^2} \tag{3}$$

Because in a relatively large data set, the data value of each dimension fluctuates greatly, so some changes are made to the data dissimilarity to ensure the preservation of the original information. The data dissimilarity for each dimension is:

$$ad_{ij}^{k} = \frac{\left|\left|x_{ik} - x_k\right| - \left|x_{jk} - x_k\right|\right|}{x_{\max k} - x_{\min k}} \tag{4}$$

Where x_{ik} is the value of x_i on dimension k, x_{jk} is the value of x_j on dimension k, x_k is the average of the entire data set in the kth dimension. x_{maxk} and x_{mink} are the maximum and minimum values in the kth dimension of the entire data set, respectively. According to the definition of dissimilarity, the improvement is equivalent to the constraint and normalization of data, which can better preserve the information of the original data. So, the dissimilarity of the two data objects $x_i = (x_{i1}, x_{i2}, ..., x_{in})$ and $x_j = (x_{j1}, x_{j2}, ..., x_{jn})$ is redefined as:

$$D_{ij} = \frac{\sum_{k=1}^{n} ad_{ij}^{k}}{n} \tag{5}$$

To generate a dissimilarity matrix, and the matrix formed by the dissimilarity between all the objects in the data set is called the dissimilarity matrix, which is marked as P_u.

$$P_u = \begin{vmatrix} D_{12} & D_{13} & \cdots & D_{1(n-1)} & D_{1n} \\ D_{23} & D_{24} & \cdots & D_{2n} & \\ \cdots & \cdots & & & \\ D_{(n-2)(n-1)} & D_{(n-2)n} & & & \\ D_{(n-1)n} & & & & \end{vmatrix} \tag{6}$$

The Huffman tree is constructed as shown in Figure 1.

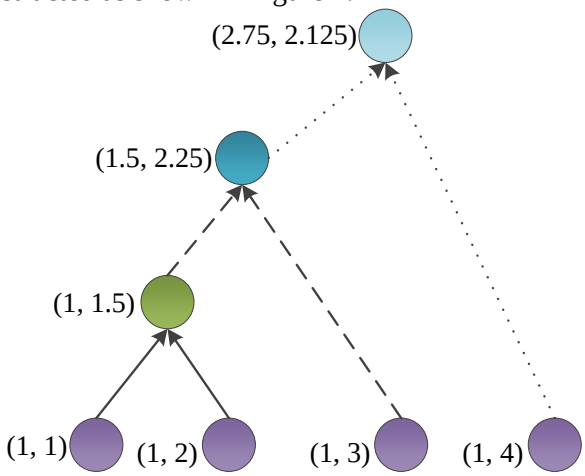

Figure 1. building process of Huffman tree

After the Huffman tree is generated, the k–1 inverse delete object can obtain the k object [6], which is the initial cluster center. When processing the density data set, the average density data set is selected to determine whether the data object is based on a high-density object, and the calculation formula of the average density D_{av} and the high density set U_1 is shown as below:

$$D_{av} = \beta \frac{1}{n} \sum_{i=1}^{n} \mathrm{Den}\left(P_i\right) \tag{7}$$

$$U_1 = \left\{ p \mid \mathrm{Den}\left(p_i\right) \geq D_{av}, p_i \in U \right\} \tag{8}$$

Among them, the adjustment factor (usually 1). When processing data sets, the density is less than the time to construct the Huffman tree. You can change to adjust the size of the high-density set to reduce the time-consuming construction of the Huffman tree. However, if the density of the high-density data set is too high, the quality of the initial cluster center will be degraded, so these two aspects need to be integrated to select the appropriate coefficient value.

2.2. Clustering application example analysis

This paper selects the neutral point current of 1# transformer grounding of 220kV substation in Changsha from 2016 to 2018 for cluster analysis and they are divided into 3 categories, using the improved k-means algorithm based on data set attributes for cluster analysis. The results are shown in Figure 2.

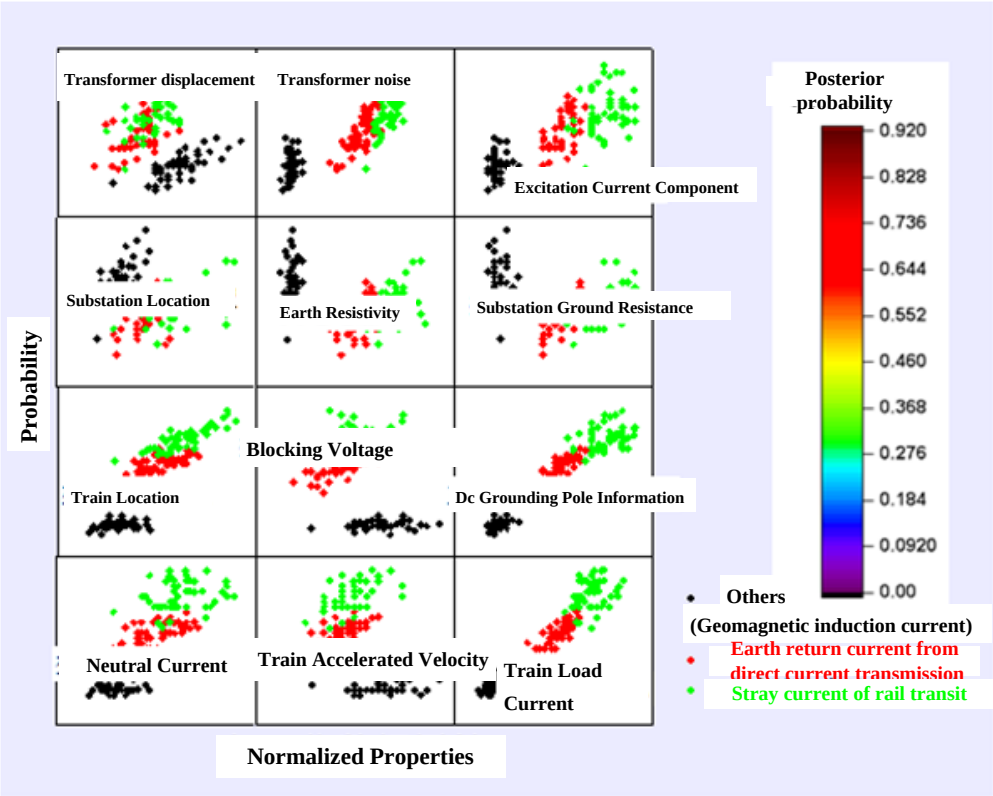

Figure 2. The data clustering results of dc magnetic bias

It can be seen from Fig. 2 that the improved k-means algorithm successfully identifies three types of events: the DC current of the neutral point of the transformer caused by the stray current of the subway is a high probability event (about 80%). The probability of return current of DC transmission is reduced to about 1/3, and the geomagnetic induced current is a small probability event, and the probability of its occurrence is almost zero. This provides a very effective basis for the DC magnetic risk management of power transformers in Changsha. The governance objectives should consider the DC magnetic damage caused by rail transit and DC transmission. It can be seen from the overall results that the clustering center selection method proposed in this paper can effectively improve the clustering effect, ensure the reliability of the results, and avoid falling into the local optimal solution. The initial clustering center obtained by the algorithm is close to the optimal clustering situation, which can reduce the number of iterations and verify the effectiveness of the algorithm.

3. Relevance analysis

Relevance analysis is an advanced process for extracting unknown, credible, novel, effective, and potentially application-oriented information or patterns from multi-system mystery big data. These useful correlation rules can be used to make better decisions for DC magnetic. From the perspective of DC magnetic risk management, prioritize risky factors and ignore low risk factors. However, the factors with high risk do not appear frequently. So by using the classical Apriori algorithm, the risks that decision makers consider important will be ignored[7].

Aiming at this problem, this paper proposes an improved Apriori algorithm based on weighted association rules. Firstly, the matrix is introduced, which greatly reduces the number of scans of the database. Secondly, the definition of the weights given in this paper, that is, the sets that with big weight but infrequently appear, also consider the sets with small weights but frequent occurrences; Finally, the k-term support expectation method is introduced to pruning, which solves the problem that the weighted frequent sets of the algorithm do not have the monotonicity of frequent itemsets in

common association rules [8-9]. The flow chart of the improved Apriori algorithm based on weighted association rules is shown in Fig. 3.

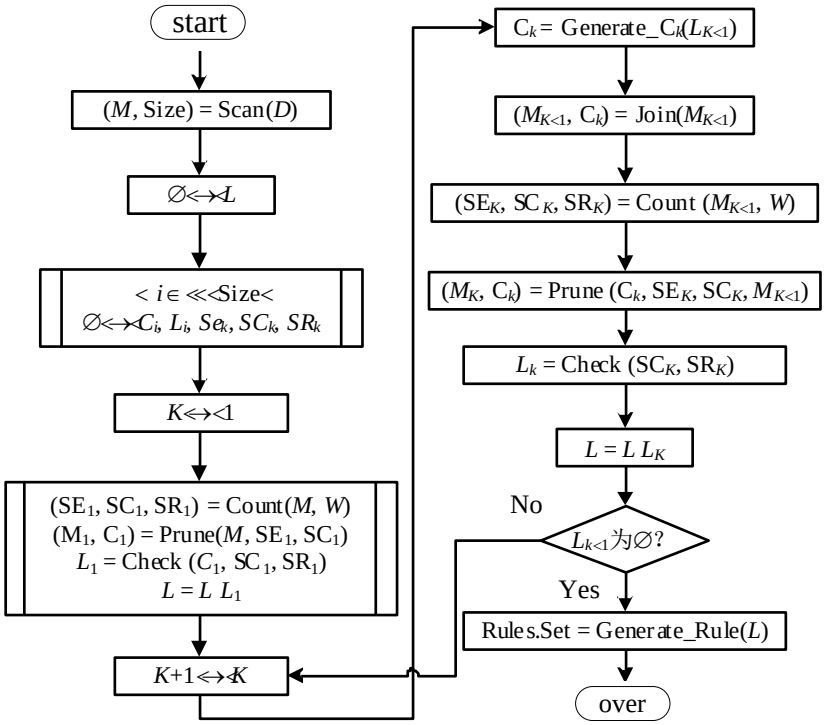

Figure 3. Flow chart of improved Apriori algorithm based on weighted relevance analysis

Improving the setting of Apriori algorithm weights will directly affect the correlation analysis results, according to the actual situation: set of given items I = {i_1, i_2, ..., i_m}, is a collection of m different items. A transaction database D is given, where each transaction T is a collection of items in I. Each transaction has an identifier called a TID. The total number of transactions in the transaction database D is n, indicating the importance of different items, and each item i is assigned a weight w, where0 ≤ w ≤ 1, j = {1, 2, ..., n}, ie W(ij) = wj.

For a certain set X, define its weight W(X) as:

$$W(X)=\frac{1}{2}\left(\frac{\sum_{i_j}\in X_j^W}{k}+\max i_j\in X\left\{w_j\right\}\right) \tag{9}$$

In the formula: k represents the number of items contained in X, 0 < k ≤ m. W(X) takes into account both the projects with significant weights and the projects that are frequently appearing and have small weights.

For a weighted association rule of the form X⇒Y, define its support as:

$$W_{\sup}=W(X)\times\operatorname{support}(X\bigcup Y) \tag{10}$$

The improved Apriori algorithm defines the credibility of the weighted association rules as:

$$W_{\text{conf}}=\frac{\operatorname{support}(X\bigcup Y)}{\operatorname{support}(X)} \tag{11}$$

The improved Apriori algorithm satisfies the minimum weighted support (Wmin_sup) and the minimum weighted confidence (Wmin_conf) as a weighted association rule, and defines its actual weighted support number SR(X) as:

$$SR(X)=\frac{W_{\min\_sup}\times n}{W(X)} \tag{12}$$

Improved Apriori algorithm defines k-term support expectation: For any k-term set X, if X is a weighted frequent set, its weighted support number SC(X) should satisfy the following formula:

$$SC(X)\geq\frac{W_{\min\_sup}\times n}{W(X)} \tag{13}$$

Let I be the set of all items, Y is a q-term (q<k), and in the remaining item set (I–Y), the first (k–q) term with the highest weight is {wr1, wr2,..., wr (k–q)}. Then the possible values for the maximum weight of any k-item set containing item set Y are:

$$W(Y,k)=\frac{1}{2}\left(\frac{\sum_{i_j\in Y}W_j+\sum_{j=1}^{k-q}w_{r_j}}{k}+\max_{i_j\in X}\{W_j\}\right) \tag{14}$$

If the k-item set X containing Y is a frequent set, then there must be:

$$\mathrm{SC}(X)\geq\frac{W_{\min\_sup}\times n}{\frac{1}{2}\left(\frac{\sum_{i_j\in Y}W_j+\sum_{j=1}^{k-q}w_{r_j}}{k}+\max_{i_j\in X}\{W_j\}\right)} \tag{15}$$

$$B(Y,k)\geq\lceil\mathrm{SC}(X)\rceil \tag{16}$$

B(Y,k) is the expectation of the k-term of Y, which means that if the k-item set containing Y is weighted frequently, the support count of the k-item must be not less than B (Y, K)). That is, there is support for expectation SE=B(Y,k), which is the basis for improving the pruning step in the algorithm.

4. Conclusion

Taking the correlation analysis of transformer vibration displacement as an example, the weighted frequent set X of the take-off dimension analysis is the transformer neutral point current, DC magnetic current and transformer load. Using the improved Apriori algorithm based on weighted association rules, the correlation analysis result with 95% confidence threshold is obtained. The schematic diagram is shown in Fig. 4. It can be seen that the transformer vibration displacement is positively correlated with the transformer neutral point current, DC magnetic current and transformer load.

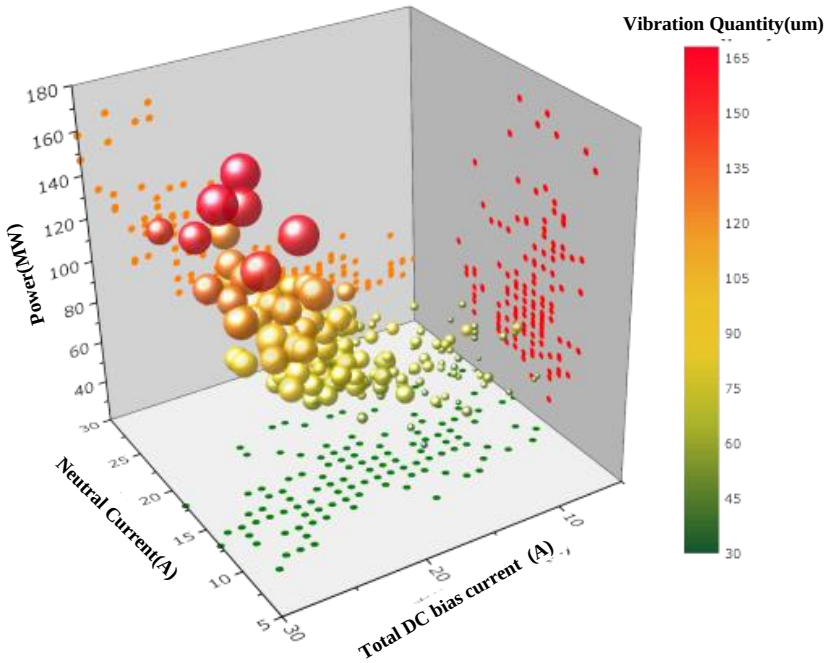

Figure 4. Relevance analysis results of vibration displacement of transformer

The correlation analysis results of DC magnetic current and excitation current peak, excitation current harmonic, vibration displacement and noise are shown in Figure 5. The confidence threshold is 95%.

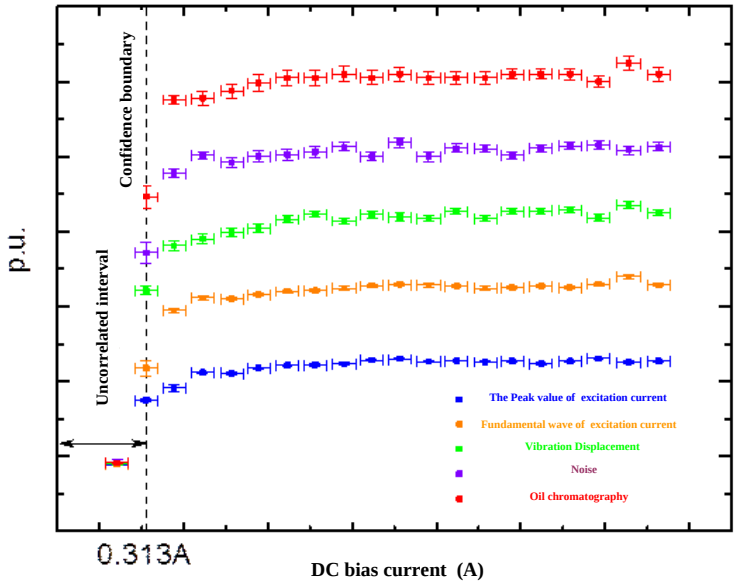

Figure 5. Relevance analysis results of dc bias current with excitation current peak, excitation current harmonic, vibration displacement and noise

The correlation analysis results of DC magnetic current and excitation current peak, excitation current harmonic, vibration displacement and noise are shown in Figure 5. It can be seen that when the DC magnetic current of the transformer exceeds 0.313A, there is a significant correlation between the data, indicating that the transformer is in a DC magnetic state, so the project method provides an effective reference for suppressing device selection and input.

$$C_T(\text{date}) = \sum_t \left| I_n(T,t) \right| \tag{17}$$

Where, I_n is the neutral point DC current of the transformer, and T is the set of all neutral grounding running transformers in Changsha, t is time (s), and date is date.

The analysis results of the correlation between the daily total amount of DC magnetic current and the running data of all transformers in Changsha are shown in Figure 6. The confidence of the correlation analysis is still 95%. The daily total DC magnetic current of the transformer is expressed by the radius of the radar chart. The specific date is expressed by the angle. Since Metro Line 1 was put into operation in the first half of 2016, only the data of 2017 was selected for analysis, and Metro Line 2 selected data from 2016 to 2018 for analysis.

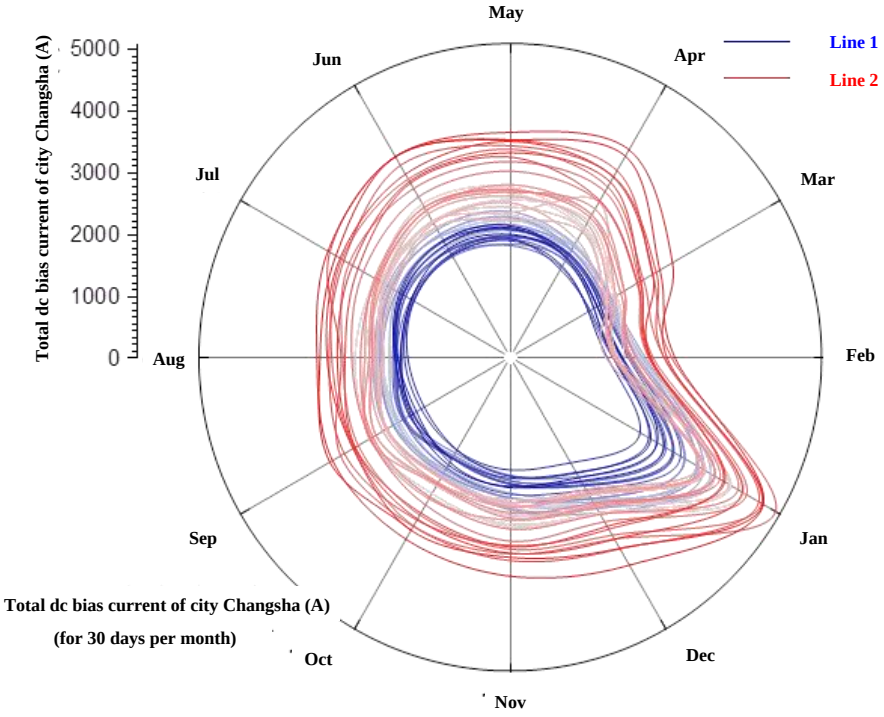

Figure 6. Relevance analysis results between the daily total dc bias current of transformer and metro operation data

The results in Figure 6 show that the risk of the DC magnetic caused by Changsha Metro Line 2 is higher than that of Line 1, and the total daily DC magnetic current is the largest in December and January. It was the smallest at the beginning of February. Combined with the original data, the weighted correlation analysis of the total daily DC magnetic current of the transformer is carried out. The results show that:

1) Changsha Metro Line 2 is closer to the AC system than Changsha Metro Line 1.

2) Due to the connection of high-speed railway station and airport, the load of Changsha Metro Line 2 is heavier and the load line current is larger.

3) Due to temperature and other reasons, people tend to take the subway in winter, and the total daily DC magnetic current of the transformer at the end of the year is larger.

4) Due to the Spring Festival holiday, the subway load is reduced and the daily DC magnetic current of the transformer is low.

From the above, the big data method is applied to provide the" one-button sequence control " operation and solution for the prevention and control of transformer DC magnetic bias in this paper.

Acknowledgments
This paper thanks for the financial support of the 2018 science and tchnology guide project of State Grid Corporation of China《Research and application of the key technology of one-button sequence control》 (Project Number: 5216A0180002).

References
[1] Jinliang He. Grounding Technology of Power System [M], Beijing: Science Press, 2006: 124-210
[2] Zhiwei Lu, Yanling Shi, Xishan Wen. Measurement And Error Analysis of Grounding Resistance in Vertical Layered Soil [J]. High Voltage Technology, 2001, 27(3):4
[3] Yi Feng. Research on Power System Bad Data Identification Algorithm Based on Cloud Computing [D]. Nanjing University Of Science And Technology, 2013
[4] Cheng Ge. Research on Power System Bad Data Identification Method Based on GSA [D]. Nanjing University of Science and Technology,2005
[5] Qi Xiao. Research on Power Load Classification Based on Optimization K-Means Algorithm [D]. Dalian University of Technology,2015
[6] Liu L, Zhai DH, Jiang XL. Current situation and development of the methods on bad-data detection and identification of power system. Power System Protection and Control 2010;38(5):143–52
[7] Buxiang Zhou, Yue Yuan, Zhiqiang Zhang, He Huang. Comprehensive Probabilistic Graph Model and Transformer State Parameter Association Rule Mining Based on Improved Apriori Algorithm [J]. Science of Hydropower And Energy, 2019,39 (3) :164-167
[8] Qian Ding. Transformer Fault Diagnosis Based on Association Rules [M]. North China Electric Power University,2009
[9] Youyuan Wang, Liwei Zhou, Xuanhong Liang, Hang Liu Et Al. Markov Prediction Model of Power Transformer Fault Based on Association Rule Analysis [J]. High Voltage Technology, 2018, 4: 1051-1058