Defect analysis of the same batch of substation equipment based on big data analysis algorithm

Qiang Gao¹, Chuan Zhong¹⁺, Yong Wang², Peng Wang², Zaiming Yu² and Jiannan Zhang³

¹Electric Power Research Institute of State Grid Liaoning Electric Power, Shenyang, China
²State Grid Liaoning Electric Power Supply, Shenyang, China
³Kekai Electric Power Branch of Liaoning Electric Power Energy Development Group Co., Ltd, Shenyang, China

*Corresponding author: zc1_ldk@ln.sgcc.com.cn

Abstract. High-quality power equipment is the basis for ensuring the safe operation of the power grid and improving the reliability of power supply. In actual operation, although some electrical equipment can continue to be used, abnormal operation or hidden dangers will affect the safety of people, equipment and power grid, reliable and economic operation of power grid and equipment, equipment output or life span, and power quality. Therefore, research on multi-dimensional analysis of the same batch of substation equipment, and identification of suspected family defects and frequent defects is of great significance for improving the company's equipment management and defect management. The maturity and promotion of unstructured text data mining, graph computing technology, and semantic analysis have provided a wider dimensional space for the analysis of power equipment defect data. Aiming at the shortcomings of traditional defect data analysis, this article summarizes and analyzes the defects and hidden dangers found in the equipment, traces the distribution status of the same equipment through the physical "ID", carries out multi-dimensional analysis of the same batch of equipment, and investigates the hidden dangers of equipment family defects. A big data analysis algorithm to build automatic identification models of suspected familial defects and frequent defects in main substation equipment such as transformers, disconnectors, circuit breakers, and use graph computing technology to quickly integrate analysis capabilities in multi-source heterogeneous data fusion with hidden relationship discovery capabilities, identify equipment defects with potential impact relationships. From historical data, discover which manufacturers or purchased batches of equipment are more likely to have the same defect, and which defect types have a higher frequency, improve the recognition model of family defects and frequent defects, and improve the accuracy and comprehensiveness of defect recognition.

1. Introduction

1.1. Research background and significance
The reliability of power equipment directly affects the safe operation of the power system. Some equipment has family defects due to factors such as design, material, and production, resulting in a significantly higher failure rate after being put into operation. There are also defects due to climate, geographical environment, and personnel operations. The problem caused frequent defects. The diagnosis of familial defects and frequent defects of traditional equipment is mainly determined by experts subjectively, without quantitative standards, and the response is not timely. With the development of the company's informatization level, a large amount of substation equipment files and defect record data have been accumulated in the informatization system, and related manufacturers, models, batches and other equipment management information and equipment failures and other operating information have been linked. The identification of family defects and major defects based on data mining provides basic data support.

With the rapid development of big data technology, major breakthroughs have been made in unstructured text data mining, graph computing technology, and semantic analysis, making it possible to automate text and graph data such as defect records and equipment structures that could not be processed by information systems. Processing analysis becomes possible. This paper summarizes and analyzes the defects and hidden dangers found in the equipment, traces the distribution status of the same equipment through the physical "ID", carries out multi-dimensional analysis of the same batch of equipment, investigates the hidden dangers of equipment family defects, combs and analyzes the relevant data of each business link, and improves equipment procurement quality. And on the basis of completing the management of substation equipment ledger data and defect classification and classification, it is compatible with different data sources and types and other elements, comprehensively applying a variety of big data analysis algorithms, and constructing transformers, disconnectors, circuit breakers and other main substation equipment. The automatic recognition model of familial defects and frequent defects, and the rapid integrated analysis ability and hidden relationship discovery ability of graph computing technology in the fusion of multi-source heterogeneous data, identify equipment defects with potential influence relations, and mine historical data Which manufacturers or purchased batches of equipment are more likely to have the same defects, and which types of defects occur more frequently, improve the recognition model of family defects and frequent defects, and improve the accuracy and comprehensiveness of defect recognition.

Therefore, based on the company’s equipment ledger, equipment structure, defect records and other data, with text mining, semantic analysis, association analysis and graph computing techniques as means, the same batch of multi-dimensional analysis of substation equipment, suspected family defects and frequency. The research on the identification of emergent defects is of great significance for improving the company's equipment management and defect management, promoting the company's data application level, and promoting the company's innovation and development.

2. Overview of research methods

2.1. Cluster analysis
Clustering analysis (Clustering) is given a data sample set \( X = \{X_1, X_2, \cdots, X_n\} \). According to the degree of similarity between samples, the data set is divided into \( K \) clusters \( \{C_1, C_2, \cdots C_k\} \) and satisfies: \( C_i(i = 1,2,\cdots,k) \) is a subset of \( X \); \( C_1 \bigcup C_2 \bigcup \cdots \bigcup C_k = X \); \( C_i \bigcap C_j = \phi, \ i \neq j \) where clusters \( C_i = \{X_{i1}, X_{i2}, \cdots, X_{in}\} \).

According to the main idea of the algorithm, data type and clustering purpose, clustering can be divided into: partition-based method, hierarchy-based method, density-based method, grid-based method, and model-based method.

2.1.1. Method based on division. The basic idea of the division method is: for a data set containing \( N \) objects, use the strategy of minimizing the objective function to divide the data set into \( K \) clusters
(k \leq n), and this cluster should meet two conditions: 1) The cluster cannot be empty, that is each cluster contains at least one object; 2) Each object must belong to one and only one cluster. Given the number of clusters \( k \), first create an initial partition, and then iterate repeatedly to change the partition so that the changed partition satisfies the similarity of the objects in the cluster and the differences between the objects as much as possible until the result of the partition is relatively stable.

Use the partition method to divide the data set \( D = \{d_1, d_2, \ldots, d_n\} \) into \( k \) clusters. The specific steps are as follows:

1) Generate \( k \) initial cluster centers \( S = \{s_1, s_2, \ldots, s_k\} \) according to a certain principle;
2) Calculate the similarity \( D \) between each data \( d_i \) and each cluster center \( s_j \) in turn \( \text{sim}(d_i, s_j) \);
3) For the data \( d_i \), find the cluster center \( s_j, j = 1, 2, \ldots, k \) that maximizes the value of \( \text{sim}(d_i, s_j) \), and divide into the cluster centered on \( s_j \) to obtain a cluster \( C = \{C_1, C_2, \ldots, C_k\} \);
4) recalculate the new cluster center \( C \) according to the cluster;
5) If the clustering result is stable or the cluster center does not change, stop; otherwise, go to 3).

The advantage of this type of method is that the clustering speed is fast and it is suitable for large-scale model data sets; the disadvantage is that the number of clusters \( k \) is difficult to determine, and the selection of the initial clustering center has a greater impact on the clustering results. The most representative algorithms in the division method are the \( k \)-means algorithm and the \( k \)-medoids algorithm.

The biggest difference between the two is that the method of selecting cluster representative points is different. The \( k \)-means algorithm uses the mean value of the objects in the cluster to represent the cluster, while the \( k \)-medoids algorithm uses the object closest to the center of the cluster to represent the cluster.

### 2.1.2. Level-based approach

The basic idea of the hierarchy method is to decompose a given data set hierarchically. Finally, a tree with clusters as nodes is built. The root of the tree is a cluster that contains all objects, and the top of the tree is a cluster that contains only a single object. According to the process of building a tree, the hierarchical method can be divided into agglomerated hierarchical clustering method and split hierarchical clustering method. The algorithm idea is shown in Figure 1.

Agglomerative Hierarchical Clustering Method (AGNES) is also called a bottom-up method. First, each object is regarded as a cluster, and then adjacent clusters are continuously merged according to a certain standard, until all clusters are merged into a cluster that may satisfy a certain termination condition.

Divisive Hierarchical Clustering Method (DIANA) is also called a top-down method. First, all objects are regarded as a cluster, and then a cluster is continuously selected for decomposition until all objects are individually in a cluster or meet certain termination conditions.

![Figure 1. Schematic diagram of hierarchical clustering algorithm.](image-url)
The hierarchical clustering method must use certain standards to measure the similarity between clusters when merging or splitting clusters. In the text representation based on the vector space model, the distance between clusters is often used as the measurement standard. These distances are mainly Minimum distance, maximum distance, class center distance, class average distance, etc.

For a given document set $D = \{d_1, d_2, \cdots, d_n\}$, assuming that $C_i$ and $C_j$ are two clusters in the clustering process, the definitions of the above four distances are as follows:

1. **Minimum distance**: $\text{Dis}_{\text{min}}(C_i, C_j) = \min_{d_i \in C_i, d_j \in C_j} \|d_k - d_i\|

2. **Maximum distance**: $\text{Dis}_{\text{max}}(C_i, C_j) = \max_{d_i \in C_i, d_j \in C_j} \|d_k - d_i\|

3. **Class average distance**: $\text{Dis}_{\text{ave}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{d_i \in C_i} \sum_{d_j \in C_j} \|d_k - d_i\|

4. **The distance between cluster centers**: $\text{Dis}_{\text{cen}}(C_i, C_j) = \|m_i - m_j\|$, where $m_i = \frac{1}{n_i} \sum d_i$, $n_i$ is the number of documents in $C_i$.

In this type of method, the agglomerated hierarchical method is often used for text clustering. For document sets $D = \{d_1, d_2, \cdots, d_n\}$, the agglomerated hierarchical clustering algorithm based on the minimum distance is as follows:

**Input**: $k$ number of clusters; document set containing texts of $D$.

**Output**: A hierarchical cluster of $k$.

1. Initialize $l = n, C_i = \{d_i\}, i = 1, 2, \cdots, n$, get the initial cluster of $C = \{C_1, C_2, \cdots, C_n\}$;

2. According to $\text{Dis}_{\text{min}}(C_i, C_j) = \min_{d_i \in C_i, d_j \in C_j} \|d_k - d_i\|$ the distance between the clusters $C_i$ in the calculation;

3. find the cluster pair $\text{MinDis}(C_i', C_j') = \min_{C_i', C_j' \in C} \{\text{Dis}_{\text{min}}(C_i, C_j)\}$ with the smallest distance between clusters;

4. Combine the cluster pairs $C_i'$ and $C_j'$ into a new cluster $C_i$ and set $l = l - 1$ to get a new cluster $C = \{C_1, C_2, \cdots, C_l\}$;

5. If $l = k$ then terminate; otherwise, go to 2).

The hierarchical clustering method is simple, can find clusters of any shape, and has greater flexibility in clustering granularity.

**Advantages**: but in the clustering process, the distance between all clusters must be compared globally, which leads to a high complexity of the algorithm, which is not suitable for clustering large-scale data sets, and once the clusters are merged or split, it cannot be undone. Therefore, if a cluster merge or split is inappropriate, the quality of the clustering result will be poor. The most basic and representative algorithms among the hierarchical methods are the agglomerated hierarchical clustering method AGNE and the split hierarchical clustering method DLANA proposed by Kaufmann and Rousseeuw in 1990.

### 2.2. Association rules

The association rule is an implication of the form $X \rightarrow Y$, where $X$ and $Y$ are respectively called antecedent or left-hand-side (LHS) and successor (consequent or right-hand-side, RHS) of the association rule. Among them, the association rule $XY$ has support and trust.

The association rule mining process mainly includes two stages: the first stage must first find all the high-frequency item groups (Frequent Itemsets) from the data collection, and the second stage will generate association rules (Association Rules) from these high-frequency item groups.
2.2.1. **Apriori algorithm.** Apriori algorithm is the most influential algorithm for mining frequent itemsets of Boolean association rules. The core is a recursive algorithm based on the two-stage frequency set idea. The association rule is classified as a single-dimensional, single-layer, Boolean association rule. Here, all itemsets with support greater than the minimum support are called frequent itemsets, or frequency sets for short.

The basic idea of the algorithm is: first find out all frequency sets, and the frequency of these itemsets is at least the same as the predefined minimum support. Then the frequency set generates strong association rules, these rules must meet the minimum support and minimum reliability. Then use the frequency set found in step 1 to generate the desired rules, and generate all the rules that only contain the items of the set, where there is only one item on the right of each rule, and the definition of the middle rule is used here. Once these rules are generated, only those rules that are greater than the minimum credibility given by the user are left. In order to generate all frequency sets, a recursive method is used.

The Apriori algorithm uses an iterative method of searching layer by layer. The algorithm is simple and clear, without complicated theoretical derivation, and easy to implement. But it has some insurmountable disadvantages:

1. Too many scans of the database.
2. Apriori algorithm will generate a large number of intermediate itemsets.
3. Use unique support.
4. The adaptability of the algorithm is narrow.

2.2.2. **Algorithm based on partition.** The partition-based algorithm first logically divides the database into several disjoint blocks, considers a block separately each time and generates all frequency sets for it, and then combines the generated frequency sets to generate all possible frequency set, and finally calculate the support of these itemsets. The size of the blocks is chosen so that each block can be placed in the main memory, and each stage only needs to be scanned once. The correctness of the algorithm is guaranteed by every possible frequency set at least in a certain block. The algorithm is highly parallel, and each block can be assigned to a certain processor to generate a frequency set. After each cycle of generating the frequency set ends, the processors communicate to generate a global candidate k-item set. Usually, the communication process here is the main bottleneck of the algorithm execution time; on the other hand, the time for each independent processor to generate the frequency set is also a bottleneck.

3. **Suspected familial defect analysis model**
Obtain the original data such as the defect records of the same batch of substation equipment of the power company and the equipment records of the asset system through the PMS system, take the defect data as an example, and mine the defect data of substation equipment based on cluster analysis and association rule algorithms. Further analysis of the obtained association rules can result in guiding results for actual production.

3.1. **Sample information**
After simple cleaning of the defect records and equipment record data of the same batch of substation equipment of the electric power company, a total of 30,000 sample data of equipment data, 150,000 sample data of component data, and 11224 sample data of defect records were obtained. Among them, four types of equipment are mainly selected as transformers, circuit breakers, isolating switches, and combined electrical appliances. The causes of equipment defects include technical reasons and liability reasons. The severity of defects includes three types: critical, serious, and general. Technical difficulties When clearing up the collected sample data, the following technical difficulties were found:

1. The sample data size of confirmed familial defects is too small to meet the training sample size required for machine learning.
2. Many information in equipment files, component files, and defect records are manually entered. Therefore, there are many data quality problems in data such as manufacturers and models. How to standardize the data of manufacturers and models with more errors.
3. Familial defects may occur in multiple manufacturers in a group, how to identify manufacturers in the same group.

4. Familial defects may occur in multiple models in one series, how to identify the models in the same series.

5. Familial defects may occur in multiple batches shipped within a period of time, and the length of the time interval that needs to be considered is variable. How to accurately locate the interval of the factory time.

6. The defective parts, defect descriptions, and reasons for defects in the defect record are manually entered, and the filling is random. Most of the defective parts are defective or misfilled. How to identify which part the defect occurs on.

7. As the amount of data increases, the various combinations of manufacturers, models, and delivery dates are increasing geometrically. How to calculate various combinations to avoid the influence of accidental factors and find the most likely suspected family defects.

3.2. Model principle

Familial defects are a hidden state, which will leave information in data such as inspections, tests, and monitoring. The suspected familial defect model is to capture this information from the data through algorithms. For the mining of suspected familial defects, based on the manufacturer, model, and time of delivery of the equipment, the manufacturer, model, and time of delivery of the component, combined with the defect record, the mining algorithm is used to find the suspected familial defect. The model process is as follows:

Firstly, build the manufacturer standardization model and model standardization model to unify the manufacturers and models; then build the manufacturer group recognition model and model series recognition model to identify the equipment and components of the same group manufacturer and the same model series; then build an adaptive Discrete model of delivery date to locate the date interval of suspected familial defects; then construct a defective part identification model to identify the part where the defect occurred; on this basis, analyze the law of occurrence of defects, and discover the manufacturers and models of equipment and parts. The probability of suspected familial defects under various combination conditions on the date of delivery; finally, through the combined condition cleaning algorithm, the repeated combinations and false combinations in the mining results are removed, and the analysis results of suspected familial defects with business guidance significance are output, and partial results As shown in Table 1.

| Manufacturer, model and batch combination of equipment and components | Support | Confidence | Number of defects involved | Explanation |
|---------------------------------------------------------------|--------|------------|----------------------------|-------------|
| 'PARTS_CLASSIFY_ID': '13146', 'PARTS_MODEL_group': 'BRLW', 'PARTS_LEAVE_FACTORY': '1978-2004', 'PARTS_FACTORY': 'Nanjing Electric Porcelain Factory' | 0.010  | 0.5        | 21                         | The main transformer bushing whose component manufacturer is Nanjing Electric Porcelain General Factory, component model is BRLW-110/630-3, and component production date is 1978-2004, has a suspected family defect. |
| 'DEVICE_FACTORY': Shandong Taikai Transformer company, 'PARTS_CLASSIFY_ID': '13146' | 0.021  | 0.577      | 18                         | The manufacturer of the equipment is Shandong Taikai Transformer which has a suspected family defect. |

Table 1. Model running results
4. Conclusions and recommendations

Based on cluster analysis and Apriori algorithm, this paper mines and analyzes the equipment data, component data and defect data of the same batch of substation equipment. By analyzing the results of defect data mining, the following conclusions are obtained:

1. This method can effectively analyze the weak links of the same batch of substation equipment, and can find the causes of the weak links, and provide a reference for the operation, maintenance and control of substation equipment.

2. This method can analyze the family defects of substation equipment manufacturers, and has guiding significance for the equipment acceptance and operation and maintenance of the same manufacturer.

3. This method can analyze the problems of various manufacturers in the production of equipment, making the improvement plan of equipment quality more targeted.

However, due to some differences in expression or incomplete information in the equipment defect data derived from the PMS system, data cleaning is required before analyzing the defects of substation equipment. However, due to the low efficiency of data cleaning and fewer useful defect records obtained, this limits the ability to analyze defect data to a certain extent. The next step will be to solve the problems encountered in order to obtain safe and stable operation of the power system. More meaningful conclusions.

References

[1] Jingbin Li, Tian Xia, Juan Chen, Analysis and treatment of a partial discharge family defect in HGIS equipment, <<High Voltage Electrical Appliances>>, 2017, Issue 5

[2] Huiming Wang, Ping Zhao, Jun Zeng, Youming Zheng, Analysis of Familial Defects of 220kV Cable GIS Oil Injection Terminal, The Ninth National Power Cable Operation Experience Exchange Conference, 2013-10-21.

[3] Xiaodong Shen, Operation risk assessment of substation equipment based on data mining, <<New Industrialization>>, 2018, Issue 8

[4] Hancong Huangfu, Zhaodi Xiao, Research on Data Mining and Analysis of Defects of Power System Secondary Equipment Based on Apriori Algorithm, <<Electronic Design Engineering>>, 2019, Issue 5

[5] Xiliang Xiong, Research and application of power transformer fault prediction based on streaming data mining, <<Nanchang University>>, 2016

[6] Cheng Xie, Zhangjie Cao, Dian Wen, Yongtao Jin, State Evaluation of Distribution Transformer Based on Real-time Operation Data Mining, <<Zhejiang Electric Power>>, 2017-08