Analysis and Investigation on Causes of Voltage Sag Based on A Novel Apriori Algorithm

Hao Liu*, Qinghui Zeng, Shaohui Liu, Zhuojun Wu, Lei Li, and Rongrong Ma

Foshan Power Supply Bureau of Guangdong Power Grid Co., Ltd, Foshan 528011, Guangdong, China

Abstract. As more and more power users are increasingly demanding the quality of electricity, the losses caused by voltage sags are becoming more and more serious. Therefore, it is very important to analyze the cause of the voltage sag to prevent the voltage sag in time. This paper proposes a new algorithm that combines Apriori correlation analysis algorithm and cluster analysis algorithm to analyze the causes of voltage sags. Because some typical climatic conditions also have an important influence on the cause of voltage dips, The data is initially processed using climate factors as clustering indicators, and then the correlation analysis between typical electrical characteristics and voltage sags is performed, and strong association rules are finally obtained. According to the calculation and analysis of examples, some factors with high correlation with the causes of voltage sag are found, which will provide theoretical support for the prevention of voltage sag and provide ideas for further research on the causes of voltage sag.

1 Introduction

With the high-speed development of smart electric power and energy systems [1]-[3], smart grids become the development trend of future power systems [4], and more and more artificial algorithms are gradually applied to power grids [5]. Under this background, high-end manufacturing industry becoming more and more sensitive to the unavoidable voltage sag events in power supply system, voltage sag has caused huge losses to users, so it has become the focus of industry and academia[6]. Some data show that among many power quality related problems, the power quality problems related to voltage sag account for more than 80% [7].

Voltage sag refers to short-term voltage drop, which is caused by short circuit, lightning stroke, start-up of large motors, equipment failure, etc. [8]. Voltage sag accidents may cause problems such as shutdown of various electrical equipment, loss of important data, and even direct damage to the equipment. The consequences may be very serious. An investigation has shown that a voltage sag accident can cause huge losses of more than 200 million yuan [9]. In conclusion, it is of great significance to take timely and effective measures to deal with voltage sag.

As data mining technology becomes more and more mature, more and more scholars are trying to find research methods in the field of data mining to deal with the related problems of voltage sag. Based on a large number of historical events of voltage sags, Luo et al. [10] selected 7 dimensions that have a greater correlation with voltage sags to analyze the association rules between the causes of voltage sags and characteristic dimensions, and obtained the strong relationship between the voltage sag types and other dimensions. Based on the monitoring data of voltage sags in a certain area, Xu Zhong et al. [11] used a two-step clustering method to divide voltage sags into four categories, and gave the characteristics of each cluster based on the clustering results. Shen et al. [12] applied the adaptive Gaussian cloud transform algorithm and the gray-scale target theory to the voltage sag study, and finally established a matching model between the actual scene and the association rules, and screened out the law of the voltage sag affecting the nodes in the actual scene. Many of the above studies use advanced data mining techniques to analyze a large amount of historical voltage sag data, and dig out the potential correlations and hidden information, but the research entry points are all one-sided.

In this paper, firstly, we use the K-means clustering algorithm to classify the data, and then use a novel Apriori algorithm to analyze the data association in each classification. The results show that climatic factors, the nature of the switching load, voltage level, area, time and other factors have a certain influence on the type of voltage sag. This analysis result provides a more specific theoretical basis for the prevention of voltage sag.

2 Algorithm Principle

2.1 K-means Cluster Analysis Algorithm

As a common clustering algorithm, K-means clustering algorithm has high efficiency for processing large data sets. The workflow of K-means clustering algorithm is as follows:

* Corresponding author: 930125316@qq.com

First, we randomly determine k initial points as centroids, then we find the nearest centroid for each point in the data set and assign it to the cluster corresponding to the centroid. After this step is completed, the centroid of each cluster is updated to the average of all points in the cluster. The Euclidean distance from each point in the data set to the centroid is calculated as:

$$d_{f,i} = \sqrt{(x_{f1} - x_{i1})^2 + ... + (x_{fn} - x_{in})^2} \quad (1)$$

In the formula, $d_{f,i}$ represents the Euclidean distance between the point f to be clustered and the cluster center i, and $x_{f1},...,x_{fn}$ represents the attribute of each point, that is, the feature used for clustering.

2.2 Apriori Correlation Analysis Algorithm

Association analysis is an unsupervised learning algorithm that finds interesting relationships in large-scale data sets. These relationships can have two forms: frequent item sets or association rules. Frequent item sets are collections of things that often appear in one piece. Association rules imply that there may be a strong relationship between two things.

In the correlation algorithm, each piece of sample data is defined as a transaction, and the transaction set D is composed of n transactions, and each transaction has multiple attributes. These attributes are recorded as items, and multiple items constitute an item set. The support factor $S(C)$ of item C is defined as the proportion of records in the data set containing item set C. The itemsets whose support coefficient exceeds the preset minimum support coefficient are defined as frequent itemsets.

If item set A and item set B in transaction set D have intersection, then $A \rightarrow B$ is called the association rule. The probability that D contains the union of A and B is called support factor, and the degree of support factor represents the importance and number of occurrences of association rules. The support degree of the association rule $A \rightarrow B$ is

$$S = P\{B \mid A\} = \frac{\| \{t \in D \mid A \cup B \in t\} \|}{\| D \|} \quad (2)$$

When the confidence of the rule $A \rightarrow B$ is not less than the confidence of the set of the minimum support set by the artificial, the rule is a strong association rule. The confidence of the rule $A \rightarrow B$ is

$$P\{A \rightarrow B\} = P\{B \mid A\} = \frac{P\{A \cap B\}}{P\{A\}} \quad (3)$$

In order to reduce the time needed to find the correlation between item sets, Apriori principle was discovered. This principle shows that if a item set C is not a frequent itemset, then all item sets containing C are not frequent item sets. It can help us reduce item sets that may be of interest and save us time.

3 Analysis of Calculation Examples Based on Apriori Correlation Analysis

3.1. Data Preprocessing

According to the statistics of voltage sag events, most voltage sag events are caused by lightning strikes. In addition, other climatic factors such as wind also have an impact on voltage sag events. Therefore, this article selects temperature, humidity, and wind as climate factors to start research.

1) *Select the characteristic dimensions of voltage sag.* This article obtains the voltage sag data required for the experiment from the power quality detection system. According to the historical data of voltage sag, this paper selects the load characteristics, the voltage level where the sag occurs, and the area where the voltage sag occurs as the transaction attributes of this study, and studies the correlation between them and the cause of the voltage sag. The analysis results obtained are shown in Table 1 below. First associate historical climate data with time and voltage sag data records, and then continue to the next step.

Table 1. Results obtained from data preprocessing.

Characteristic dimensions	Data value
Load characteristics	Ordinary load, heavy load, new energy and sensitive user
Voltage level	the voltage level where the sag occurs
Voltage sag area	Geographical area of the voltage sag event
Cause of voltage sag	The main cause of voltage sags

2) *Discretize voltage sag characteristic dimensions.* Discrete the characteristic dimensions of voltage sag. According to the classification in the power quality detection system, we divide the load attributes into ordinary load, heavy load, new energy and sensitive users. According to China's voltage level, the voltage level is divided into four categories: 330kV and above, 220kV, 110kV, 35kV, 10kV and below. In order to better use the software for data analysis and processing, the sag area will be numbered once from 0 according to the number of sag area in the data. According to the main reasons of voltage sag, the reasons will be temporarily divided into short circuit, heavy load, new energy, etc. The discretization data of each dimension is shown in Table 2 below.

Table 2. Voltage sag characteristic dimensions discretized data items

Characteristic dimensions	Identifier	Data values
Load characteristics	L	Ordinary load, heavy load, new energy, sensitive users
Voltage level	V	330kV and above voltage levels, 220 kV, 110 kV, 35 kV, 10kV and below voltage levels
Voltage sag area	R	01, 02, 03, …, 10
Cause of voltage sag	O	Short circuit, heavy load, new energy and others

3.2 Clustering Analysis Based on K-means Algorithm

According to the preliminary observation of the climate data, this paper uses temperature, humidity, and wind force as the clustering index to divide all the voltage sag data into three categories. The steps are as follows:

a) Randomly select three points as initial centroids and record them as Z_1, Z_2, Z_3 respectively.

b) Calculate the Euclidean distance from each point to each centroid, find the nearest centroid for each point, and assign it to the cluster corresponding to the centroid.

c) After all points are assigned, the centroid of each cluster is updated to the average of all points in the cluster. The new centroid can be calculated by the following formula:

$$Z_i^{(k+1)} = \frac{\sum_{X_j \in C_i^{(k)}} X_j}{\left|C_i^{(k)}\right|}, i = 1,2,3 \tag{4}$$

where $C_i^{(k)}$ is a cluster centered on $Z_i^{(k)}$ and $Z_i^{(k)}$ represents the coordinate after the k-th update of the centroid Z_i, X_j represent a point in the cluster $C_i^{(k)}$, and $\left|C_i\right|$ represents the number of points of the cluster C_i where the centroid Z_i is located.

Concretely, repeat the third step until the result reaches the convergence condition. The algorithm flow chart of the K-means cluster analysis provided in this paper is shown in the Fig. 1 below.

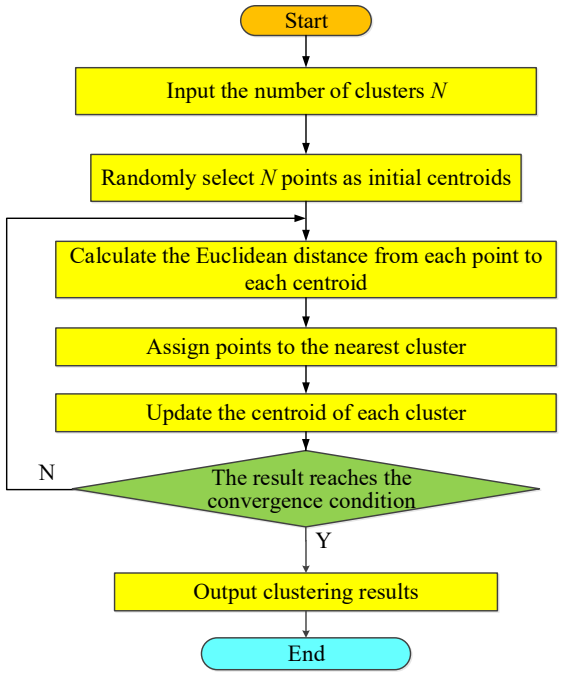

Fig. 1. K-means clustering algorithm flow chart

3.3 Correlation Analysis

In Section 3.2, we get the clustering results with climate factors as the clustering index. On this basis, the correlation analysis of each attribute in each climate type is carried out to find out the correlation rules between the voltage sag causes and other characteristics in each cluster.

1) *Building a dimension matrix.* We use Apriori algorithm to find frequent item sets. The two input parameters of Apriori algorithm are the minimum support and the data set. We need to construct the data set matrix D with the existing data.

$$D = \begin{bmatrix} d_1 \\ d_2 \\ \dots \\ d_n \end{bmatrix}, di = [L_i \quad V_i \quad R_i \quad O_i] \tag{5}$$

where d_i represents the i-th row vector of matrix D, L_i represents load nature, V_i represents voltage level, R_i represents voltage sag area, and O_i represents cause of voltage sag.

2) *Looking for frequent item sets.* Next, using the dimension matrix D for correlation analysis, then a frequent item set is obtained. The process of discovering frequent item sets is shown in the following figure:

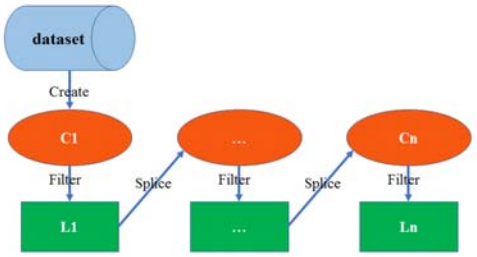

Fig. 2. Schematic diagram of frequent itemsets searching process

The specific process is shown in the figure above:

a) Generate candidate item set C_1 from the data set (1 means that each candidate item has only one data item).

b) Then filter out items with support less than the minimum support from C_1, and generate a frequent item set L_1 (1 means each frequent item Only one data item).

c) Combine the data items in L_1 pairwise to generate candidate item set C_2 (2 means that each candidate item has 2 data item).

d) Repeat the first three steps to generate larger frequent item sets until there is only one or no data item in L_n.

3) Obtain strong correlation characteristics of voltage sag. We can get multiple association rules $A \rightarrow B$ according to the many frequent itemsets filtered out. Among them, item set B is the cause of the voltage sag, and item set A is one or more of the other three dimensions, namely the nature of the load, the voltage level, and the occurrence area. We respectively calculate the confidence of each association rule and compare it with the preset minimum confidence. If it is greater than or equal to the minimum confidence, then the rule is artificially a strong association rule.

4. Case analysis

The example in this article uses 6060 voltage sag records obtained from the power quality detection system as the data source, and clusters three clusters according to the clustering indicators of temperature, humidity and wind. The characteristics are shown in Table 3.

Table 3. Clustering result feature distribution

Cluster	Ⅰ	Ⅱ	Ⅲ
Proportion	39%	44%	17%
Temperature	The temperature range is mainly 20-25 ℃.	The temperature range is mainly 10-20℃ and 25-35 ℃.	Almost no data on temperatures between 20 and 30 ℃ are included.
Humidity	The humidity range is mainly 20%-60%. A small part of the humidity range is below 20%.	The humidity range is mainly 20%-60%. A part of the humidity range is below 20%	The humidity range is almost all concentrated at below 20%.
Wind-force	The wind range is mainly concentrated in the 0-5 level.	The wind range is mainly concentrated in the 0-5 level.	The wind range is mainly concentrated in the 0-3 level.

According to the experience of correlation analysis, it may be assumed that our acceptable minimum support is 15% and the minimum confidence is 60%. Based on the above assumptions, we can analyze some strong association rules shown in Table 4.

Table 4. Cluster matching results

Cluster	Strong association rules	Confidence occurrence rate
Climate Ⅰ	{35kV} →Short circuit	87%
Climate Ⅱ	{35kV} →Short circuit	95%
	{Area 10，35kV，Ordinary load} →Short circuit	82%
Climate Ⅲ	{35kV} →Short circuit	86%

We can see that in the climate Ⅰ cluster, the confidence level associated with 35kV and short circuit is 87%. And in the climate Ⅱ, the confidence level associated with 35kV and short circuit is 95%. In the climate Ⅲ cluster, the confidence level associated with 35kV and short circuit is 86%.

From the results in the table, it is not difficult to infer that short-circuit type voltage sag accidents are prone to occur under 35kV voltage level, especially when the climate type belongs to climate Ⅱ, we can safely assert that voltage sag accidents at 35kV monitoring points are almost all caused by short circuits.

In addition, we can also get another strong association rule: when the climate is climate Ⅱ and the sag occurs in area 10 and the voltage level is 35kV, the confidence that the voltage sag type is short circuit is 82%.

Through this rule, we can get the following conclusions: the voltage sag accident that occurred at the 35kV public load monitoring point of area 10 in climate Ⅱ is mainly caused by a short circuit.

These strong rules obtained above can guide the related research on the causes of voltage sag to a certain extent, and then provide data and theoretical support for the prevention and control of voltage sag.

5. Conclusions

Through correlation analysis, the hidden features in the data that are strongly related to the occurrence of voltage sag can be discovered, which can provide theoretical support for studying the occurrence of voltage sag and contribute to the improvement of power quality.

This article first divides a large number of historical voltage sag data into three categories based on climate factors through cluster analysis. Then, combine the Apriori correlation analysis algorithm to analyze the data in each cluster. The results show that climate has a certain influence on voltage sags. Under certain climatic conditions, the voltage sag caused by a short circuit is much more serious than other climatic conditions. In addition, the voltage level, area location, and load type also have a certain impact on the occurrence of voltage sag events.

In addition, according to the experimental results, we can also get a highly reliable conclusion: most voltage sag events at the 35kV monitoring point are caused by short-circuit faults in the line.

Acknowledgment

This work is supported by the Technical Projects of Guangdong Power Grid Limited Liability Company (Grant. 030600KK52160006)

References

1. L.F. Cheng, T. Yu, X.S. Zhang, L.F. Yin, and K.Q. Qu, “Cyber-physical-social systems based smart energy robotic dispatcher and its knowledge automation: framework, techniques and challenges,” *Proceedings of the CSEE*, vol. 38, no. 1, pp. 25-40, Jan. 2018.
2. L.F. Cheng and T. Yu, “Typical scenario analysis of equilibrium stability of multi-group asymmetric evolutionary games in the open and ever-growing electricity market,” *Proceedings of the CSEE*, vol. 38, no. 19, pp. 5687-5703, Oct. 2018.
3. L.F. Cheng and T. Yu, “Game-theoretic approaches applied to transactions in the open and ever-growing electricity markets from the perspective of power demand response: An overview,” *IEEE Access*, vol. 7, no. 1, pp. 25727-25762, Mar. 2019.

4. T. Yu, L.F. Cheng, and X.S. Zhang, “The weakly-centralized Web-of-Cells based on cyber-physical-social systems integration and group machine learning: Theoretical investigations and key scientific issues analysis,” *Scientia Sinica(Technologica)*, vol. 49, pp. 1541–1569, 2019.
5. L.F. Cheng, T. Yu, X.S. Zhang, and L.F. Yin, “Machine learning for energy and electric power systems: state of the art and prospects,” *Automation of Electric Power Systems*, vol. 43, no. 1, pp. 15-31, Jan. 2019.
6. X.N. Xiao. *Analysis and control of power qua*lity. Beijing, China: China Electric Power Press, 2010: 124-138.
7. X.D. Yang. *Study on stochastic estimation of voltage sag and its economic management method*. Master's thesis, North China Electric Power University, Baoding, China, 2009.
8. Y. Zhang, L.S. Yin, R.Q. Ma, D.M. Liu, C.L. Yang Voltage sag detection method based on complex wavelet transform and effective value algorithm. *Electrical Measurement and Instrument*, 2017, 54(10): 74-79.
9. S. Tao. *Study on the influence of voltage sag on reliability of distribution system and its evaluation index*. Master's thesis, North China Electric Power University, Baoding, China, 2005.
10. Y. Luo, L.H. Qi Application of association rules in voltage sag analysis. *China Science and Technology Information*, 2013, 23: 169-171.
11. Z. Xu, W.X. Mo, Z. Zhang, Q. Zhong Cluster analysis of urban power system voltage sag event. *Power Supply and Consumption*, 2018, 35(02): 31-35.
12. X. Shen, H.G. Yang, C. Duan. Data mining analysis method based on gray target theory and cloud model for voltage sag event. *Power Grid Technology*, 2019, 43(02): 722-731.