Voltage sag analysis based on cluster analysis and correlation analysis

Hao Liu, Qinghui Zeng, Shaohui Liu, Zhuojun Wu, Lei Li and Rongrong Ma
Fenjiang South Road, Chancheng District, Foshan City, Guangdong Province

77163134@qq.com

Abstract. Based on data mining algorithm, a voltage sag analysis method is proposed, which first performs clustering analysis and then correlation analysis. Firstly, the climatic factors are used to cluster the data, and then the correlation analysis of each cluster is carried out according to the selected voltage sag feature dimensions, finally, the strong association rules are obtained. Through the analysis of the example, it is found that there are some certain correlations between the climatic factors, the voltage sag dimensions and the causes of the voltage sag. These correlations can provide a theoretical basis for the prevention and control of voltage sag.

1. Introduction
Voltage sag is a typical power quality problem[1]. According to statistics, over 80% of power quality problems are caused by voltage sags[2]. Voltage sag refers to "short-term voltage drop", which is caused by short circuit, lightning stroke, start-up of large motors, equipment failure, etc[3]. Once a voltage sag accident occurs, it will cause the motor to stop, data loss, etc., and the equipment will be damaged directly, which will seriously affect the normal life and operation of the users. Some surveys have shown that a voltage sag accident can cause a loss of 200-300 million yuan[4]. Therefore, the prevention and control of voltage sag has great practical significance.

At present, many scholars try to use data mining methods to study voltage sag. Luo Yan et al.[5] studied the historical events of voltage sag, selected seven dimensions related to voltage sag to analyze the association rules, and obtained strong association rules between voltage sag type and other dimensions. Xu Zhong et al.[6] used two-step clustering to study the voltage sag monitoring data in a certain area. Finally, the voltage sag was divided into four categories, and the characteristics of each cluster were given. Shen Xiang et al.[7] adopted the adaptive Gaussian cloud transform algorithm and gray target theory to establish a matching model between the actual scene and the association rules, and screened out the knowledge of the influence of the voltage sag on the nodes in the actual scene. The above research uses the data mining method to analyze the voltage sag historical data, but all of them are excavated from the perspective of some voltage sag data dimensions, which have certain one-sidedness.

This paper proposes a voltage sag analysis method that clusters first and then associates. Through this process the historical monitoring data and historical climate data accumulated by the power quality monitoring system will be analyzed, which proves that there is a certain relationship among the climate, load nature, voltage level, occurrence area, week, and time. That can provide a theoretical basis for the prevention and control of voltage sag.
2. Algorithm principle

2.1. Cluster analysis
As an important algorithm of data mining, cluster analysis mainly refers to the process of dividing a set into multiple classes which are composed of similar elements. First, select one point as the initial cluster center, and calculate the distance between the remaining points and the initial cluster center, and select the farthest point as the second cluster center. The distance is calculated as:

\[ d_{i,j} = \sqrt{(x'_1 - x_i)^2 + ... + (x'_n - x_i)^2} \] (1)

Where \(d_{i,j}\) represents the distance between the measurement day and clustering center \(i\), and \(x_1,...,x_n\) represents each cluster indicator. Then, according to the same method, select the cluster center \(i\) until there are no more cluster centers. Finally, calculate the distance from the remaining points to each cluster center, and assign each point to each class according to the principle of minimum distance.

2.2. Correlation Analysis
The purpose of correlation analysis is to mine the potential connections between data items which from the data. Correlation analysis can connect things that seem completely unrelated to data. According to the association rule, each sample data is called a "transaction", transaction database \(D\) consists of \(n\) transactions and each transaction is specified by multiple attributes, which are recorded as "items", and multiple items constitute an "item set". The probability \(P(C)\) of item set \(C\) is called support, and the item set exceeding the minimum support is called frequent item set.

If item set \(X \cap Y \neq \emptyset\) and both of them are all included in \(D\), then \(X \rightarrow Y\) is called the association rule. The probability that \(D\) contains the union of \(X\) and \(Y\) is called support, and the degree of support represents the importance and number of occurrences of association rules. The support degree of the association rule \(X \rightarrow Y\) is:

\[ S = P(Y \mid X) = \frac{\|\{t \in D \mid X \cup Y \in t\}\|}{\|D\|} \] (2)

When the confidence of the rule \(X \rightarrow Y\) 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 \(X \rightarrow Y\) is:

\[ P[X \rightarrow Y] = P[Y \mid X] = \frac{P(X \cap Y)}{P(X)} \] (3)

3. Voltage sag analysis algorithm based on cluster analysis and correlation analysis

3.1. Data preprocessing

3.1.1. Historical climate data collection. According to statistics, most of the voltage sag events are caused by lightning strikes. It can be seen that the climate has a certain impact on the voltage sag. This paper selects temperature, humidity, and wind speed as climatic factors, and obtains historical climate data from the Internet.

3.1.2. The value of voltage sag characteristic dimensions. The voltage sag data in this paper is mainly derived from the power quality detection system. In the power quality detection system, there are more than ten dimensions related to voltage sag. In order to simplify the calculation, this paper chooses the six dimensions of load nature, voltage level, occurrence area, week, time and sag cause as the characteristic dimension of voltage sag for correlation analysis. The results are shown in the table 1 below.
Table 1. The value of voltage sag characteristic dimensions

| Characteristic dimensions | Data value                                      |
|---------------------------|-------------------------------------------------|
| Load natures              | Ordinary load, heavy load, new energy, sensitive user |
| Voltage level             | 500kv and above voltage levels, 330kv, 220kv, 110kv, 35kv, 10kv and below voltage levels |
| Voltage sag area          | Geographical area of the voltage sag event      |
| Week                      | The nature of the voltage sag event occurring in week |
| Time                      | The time when the voltage sag event occurred in 1 day |
| Cause of voltage sag      | The main cause of voltage sags                  |

Link historical climate data to time and voltage sag data records and proceed to the next step.

3.1.3. Discretization of characteristic dimensions of voltage sag. Discrete the characteristic dimensions of voltage sag. Based on the classification in the power quality detection system, we divide the load properties into common load, heavy load, new energy, and sensitive users. According to China's voltage levels (220V, 380V, 10kV, 24kV, 35kV, 110kV, 220kV, 500kV, 800kV and 1000kV, respectively), the voltage levels are reclassified into five categories: 500 kV and above, 330 kV, 220 kV, 110 kV, 35 kV, 10 kV and below. In order to facilitate the processing, the voltage will be temporarily numbered according to the area. According to the date attributes and the household electricity time rules, the week of voltage sag will be divided into working days and holidays, and the time of voltage sag is divided into 00:00- 8:00, 8:00-18:00, 18:00-24:00. According to the classification of the main cause of voltage dip, the reasons will be temporarily divided into short circuit, heavy load, new energy and others. The discretization data for each dimension is shown in the following table 2.

Table 2. Voltage sag characteristic dimensions discretized data items

| Characteristic dimensions | Identifier | Data values                                      |
|---------------------------|------------|-------------------------------------------------|
| Load nature               | U          | Ordinary load, heavy load, new energy, sensitive user |
| Voltage level             | L          | 500kV and above voltage levels, 330kV, 220kV, 110kV, 35kV, 10kV and below voltage levels |
| Voltage sag area          | R          | 01, 02, 03, ..., 10                              |
| Week                      | W          | Working days(0), holidays(1)                     |
| Time                      | T          | 00:00-8:00, 8:00-18:00, 18:00-24:00              |
| Cause of voltage sag      | O          | Short circuit, heavy load, new energy and others |

3.2. Cluster analysis
This article classifies voltage pause records based on temperature, humidity, and wind. The steps are as follows:

1) Arbitrarily select a point as the first cluster center $Z_1$.
2) Calculate the distance $d_{x_i, z_1}$ between the remaining points $\{X_1, X_2, \ldots, X_n\}$ ($i = 1, 2, \ldots, n$) to $Z_1$, select the point farthest from $Z_1$, which is $\max\{d_{x_1, z_1}, \ldots, d_{x_n, z_1}\}$, as the new cluster center $Z_2$.
3) Calculate the distances $d_{x_i, z_1}, d_{x_i, z_2}$ from the remaining points to $Z_1, Z_2$, and take the minimum values $\min\{d_{x_i, z_1}, d_{x_i, z_2}\}$ of the two distances, if the value is greater than the distance between the two cluster centers, i.e.:

$$\min\{d_{x_i, z_1}, d_{x_i, z_2}\} > \sqrt{(Z_1 - Z_2)}$$

(4)

Then use this point as the cluster center $Z_3$.
4) And so on, find all cluster centers until the number of cluster centers does not increase.
5) Calculate the distance between the remaining points and each cluster center, and assign them to each class according to the principle of minimum distance.
6) Repeat the previous step until all points have been divided.

The algorithm flow chart of the cluster analysis provided in this paper is shown in the figure 1 below.

**Figure 1.** Clustering algorithm flow chart

### 3.3. Correlation Analysis

According to the clustering result of 3.2, the correlation analysis is carried out in each cluster, and it is proposed to find out the association rules of the causes of voltage sag and the other factors in each cluster.
3.3.1. Building a dimension matrix. Traditional correlation analysis requires frequent scanning of the database. In order to reduce the scanning time and speed up the algorithm, this paper generates an array of all the information after a database scan, and then scans the database to scan the array. The dimension matrix $A$ is established according to discrete data items of each feature dimension.

$$
A = \begin{bmatrix}
1 & U & L_1 & R_1 & Q_1 & W_1 & T_1 & O_1 \\
2 & U & L_2 & R_2 & Q_2 & W_2 & T_2 & O_2 \\
& \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
n & U & L_n & R_n & Q_n & W_n & T_n & O_n \\
\end{bmatrix}
$$

(5)

3.3.2. Looking for frequent item sets. Next, using the dimension matrix $A$ for correlation analysis, then a frequent item set is obtained.

First, each value in the dimension matrix $A$ is used as a candidate 1-item set; then, by scanning the matrix $A$, the statistical values in the row of each value are accumulated to obtain the minimum support degree; the candidate 1-item concentration support does not match the item set required for minimum support is culled, resulting in a frequent 1-item set.

Next, the rows in the matrix $A$ that do not contain the frequent 1-item set are deleted, and the dimension matrix $A_2$ is obtained. At the same time, the item sets in the frequent 1-item set are connected to obtain the candidate 2-item set, and the candidate 2-item centralized support is less than the minimum support degree, and the frequent 2-item set is obtained.

By analogy, the dimension matrix $A_{k+1}$ is obtained by deleting the rows of the dimension matrix $A_k$ that do not contain the frequent $k$-item set, and the candidate $(k+1)$-item set is obtained according to the frequent $k$-item set self-joining, and then delete the items in candidate $(k+1)$-items set that do not meet the minimum support, the frequent $(k+1)$-item set is obtained. When a new frequent item set cannot be generated, the operation ends.

3.2.3. Output strong association rules. According to frequent item sets, the non-empty subset is used to get multiple association rules $X \rightarrow Y$. Among them, item set $Y$ is the cause of voltage sag, and item set $X$ is one or more dimensions of the other six dimensions, namely load nature, voltage level, occurrence area, occurrence quarter, week and time. Calculate the confidence of the association rule and compare it with the set minimum confidence. If it is not less than the minimum confidence, it is a strong association rule.

4. Analysis of results

Taking the historical power quality detection record as the data source, and 6060 voltage sag records of 70 detection points in 10 areas were selected for mining. Clustering by climatic factors yielded three clusters, and the characteristics are shown in the Table 3.

| Cluster | I | II | III |
|---------|---|----|-----|
| Proportion | 39% | 44% | 17% |
| Temperature | ![Temperature Graph](chart1) | ![Temperature Graph](chart2) | ![Temperature Graph](chart3) |

Table3. Clustering result feature distribution
Assuming that the minimum support is 10% and the minimum confidence is 50%, the obtained partial strong association rules are as follows.

**Table 4. Cluster matching results**

| Cluster  | Strong association rules | confidence |
|----------|--------------------------|------------|
| Climate I | {35kV} → Short circuit   | 87%        |
|          | {Area 10, 220kV, Ordinary load} → Heavy load operation | 52%        |
|          | {35kV} → Short circuit   | 95%        |
| Climate II | {35kV} → Short circuit    | 82%        |
|          | {Area 10, 220kV, Ordinary load} → Heavy load operation | 50%        |
| Climate III | {35kV} → Short circuit   | 86%        |
|           | {Area 10, 220kV, Ordinary load} → Heavy load operation | 59%        |

Rule 1: Climate I ∨ 35kV → short circuit, confidence = 87%; Climate II ∨ 35kV → short circuit, confidence = 95%; Climate III ∨ 35kV → short circuit, confidence = 86%.

This rule shows that the voltage sag that is prone to short-circuit properties at 35kV, especially in climate II, the voltage sag accident at the 35kV monitoring point is almost always caused by a short circuit.

Inference 1 can be obtained by using rule 1 and proportion of sample size of clustering.

Inference 1: 35kV → short circuit, confidence = 90%.

Rule 2: Climate II ∨ area 10 ∨ 35kV ∨ ordinary load → short circuit, confidence = 82%.

This rule indicates that the voltage sag accidents occurring at the 35kV common load monitoring point in the climate region 2 and in the area 10 are mostly caused by short circuits.

Rule 3: Climate I ∨ area 10 ∨ 220kV ∨ ordinary load → heavy load operation, confidence = 52%.

Climate II ∨ area 10 ∨ 220kV ∨ ordinary load → heavy load operation, confidence = 50%.

Climate III ∨ area 10 ∨ 220kV ∨ ordinary load → heavy load operation, confidence = 59%.

The rule shows that the proportion of voltage sag accidents caused by heavy load operation of 220kV ordinary load in area 10 accounts for half, and when the climate is climate III, the proportion of heavy load operation is significantly higher than the other two clusters.

Rule 4: Climate I ∨ area 10 ∨ 220kV ∨ new energy load → new energy operation, confidence = 90%.

This rule shows that in the voltage sag accidents in the 220kV new energy load in the area 10 under the climate I, the accidents caused by the new energy operation account for 90%.
By analyzing all the strong rules, it can guide the voltage sag investigation to a certain extent, and provide data support for the prevention and treatment of voltage sag.

5. Conclusions
This paper divides the historical voltage sag data by cluster analysis, and obtains several categories of different climatic characteristics. Then, it combines the correlation analysis to analyze each cluster, and puts forward a voltage sag analysis method based on cluster analysis and correlation analysis. The results show that the climate has a certain impact on the voltage sag. In some climates, the voltage sag caused by the short circuit is significantly more than other climates. Moreover, the voltage level, region, load type, week, time and other dimensions in the voltage sag dimension also have a certain impact on the voltage sag event. Most of the voltage sag events at the 35kV monitoring point are caused by the short circuit. The research results can effectively guide the prevention and treatment of local voltage sag and reduce the occurrence of voltage sag.

Acknowledgments
Foundation item: Supported by the Technical Projects of Guangdong Power Grid Limited Liability Company (No. 030600KK52160006)

References
[1] Sun J.J. Voltage sag in modern distribution network and its operational toughness improvement. Industrial Engineering, 2019, 38(3): 4-7.
[2] Yang X.D. Study on stochastic estimation of voltage sag and its economic management method. North China Electric Power University, 2009.
[3] Zhang Y., Yin L.S., Ma R.Q., Liu D.M., Yang C.L. Voltage sag detection method based on complex wavelet transform and effective value algorithm. Electrical Measurement and Instrument, 2017, 54(10): 74-79.
[4] Tao S. Study on the influence of voltage sag on reliability of distribution system and its evaluation index. North China Electric Power University, 2005.
[5] Luo Y., Qi L.H. Application of association rules in voltage sag analysis. China Science and Technology Information, 2013, 23: 169-171.
[6] Xu Z., Mo W.X., Zhang Z., Zhong Q. Cluster analysis of urban power system voltage sag event. Power supply and consumption, 2018, 35(02): 31-35.
[7] Shen X., Yang H.G., Duan C. Data mining analysis method based on gray target theory and cloud model for voltage sag event. Power grid technology, 2019, 43(02): 722-731.