International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2024): 6.623

Voltage Sag Homology Detection Based on DBSCAN Algorithm

Meijing Jiang

[1]North China Electric Power University, School of Control and Computer Engineering, Changping District, Beijing, China
Email: *120222227098[at]ncepu.edu.cn*

Abstract: ***Inverters, AC contactors and other equipment used in high-tech manufacturing are very sensitive to voltage sags. Voltage sags may cause equipment failure, production interruption, data loss, damage to sensitive equipment and unstable energy supply. A short circuit fault may trigger multiple power quality monitoring devices to record voltage sag waveforms. The problem of voltage sag data redundancy seriously affects data application. Therefore, identifying the source of voltage sags is of great significance for scientifically and rationally evaluating the severity of regional power grid voltage sags. Therefore, this paper proposes a voltage sag source identification algorithm based on the DBSCAN algorithm. By adopting appropriate feature engineering, three-dimensional clustering features are selected, and then appropriate clustering algorithm parameters are selected through iterative method for clustering. Finally, the algorithm effect is evaluated through six clustering evaluation indicators. Experiments were conducted on the jupyter notebook programming platform using data provided by a provincial power company. The final results prove the effectiveness of the proposed algorithm.***

Keywords: voltage sag, clustering, DBSCAN, voltage sag homology detection

1. Introduction

Voltage sags cause production interruptions in precision processing industries such as microelectronics and intelligent control, resulting in huge economic losses for users and becoming the most complained power quality problem [1], [2] . A short circuit fault may trigger multiple power quality monitoring devices to record voltage sag waveforms. The redundancy of voltage sag data seriously affects data application [3], [4] and may cause overestimation of the severity of regional power grid voltage sags [5] . At the same time, repeated analysis of multiple data caused by the same voltage sag source will increase the computational intensity and complexity. Identifying multiple voltage sag events as the same voltage sag source is an urgent problem to be solved in the field of power quality monitoring. Identifying the same source of voltage sags can reduce the data redundancy of the power quality monitoring system of the power grid and avoid overestimation of the regional power quality level. It is a necessary prerequisite for clarifying the power quality level of the regional power grid and is of great significance for scientifically and rationally evaluating the severity of regional power grid voltage sags.

Identification of the source of voltage sags is to classify multiple voltage sag data monitored in a short period of time, and classify the voltage sag monitoring data triggered by the same voltage sag source into one category. In recent years, a large number of studies have been carried out on the source of voltage sags at home and abroad. Existing research mainly includes feature extraction and selection [6], data mining and machine learning algorithms [7], [8], [9], and algorithm fusion and integration [10].

In summary, this paper proposes a homology identification method based on the DBSCAN algorithm, and conducts a clustering experiment using 10049 temporary drop data provided by a provincial power company. Finally, six clustering evaluation indicators are evaluated on the clustering results to prove the accuracy and effectiveness of this experiment.

2. Related Concepts

2.1 Basic knowledge of voltage sag

2.1.1 Definition of voltage sag

Voltage sag (also known as voltage sag, dip or sag) refers to a transient disturbance phenomenon in which the voltage root mean square value temporarily drops to 90%~10% of the rated voltage amplitude and lasts for 0.5~30 cycles [11] . Voltage sag is extremely harmful, and more than 70% of power quality problems are caused by voltage sag. The classification of voltage sag sources is the premise for understanding the inherent properties and laws of voltage sag events.

2.1.2 Source of voltage sag

Voltage sags in power systems can be caused by a variety of factors [12], [13], the following are some of the most common ones:

1) Load change: When the load in the power system increases suddenly, such as when large mechanical equipment is started, air conditioning systems are put into operation, or there is a sudden large-scale power consumption, the system voltage will temporarily drop. This is because the power system needs to adapt to the load change in a short period of time, and it may take some time to return to normal voltage levels.

2) Short circuit fault: In the power system, a short circuit fault refers to a direct connection between two circuits or wires with different voltages. This will cause a sudden increase in current, resulting in a temporary drop in system voltage, and may trigger the action of protective equipment to isolate the fault point.

3) Action of overcurrent protection device: In the power system, overcurrent protection device is used to detect

abnormal current in the system to protect equipment and circuits from damage due to overload or short circuit fault. When the overcurrent protection device is activated, it may cut off the circuit or reduce the current, causing a temporary drop in system voltage.

4) System failure or fault recovery: Faults in the power system, such as generator failure, transformer failure, or transmission line failure, can cause voltage sags. When these faults are repaired or the system is restored, the voltage will gradually return to normal levels.

2.1.3 Voltage sag hazards

Voltage sag may cause some damage to power systems and related equipment [14], [15] , including:

1) Equipment failure: Voltage sags may cause equipment failure or damage. Low voltage levels may not provide sufficient power supply, causing the equipment to not operate normally or work unstably. In some cases, voltage sags may cause problems such as motor overload, equipment startup difficulties, and electronic equipment failure.

2) Production shutdown: In industrial production environments, voltage sags can cause equipment or production line shutdowns. Some equipment may require a stable voltage supply to operate properly. If a voltage sag causes equipment shutdown, it will cause production interruptions, resulting in production losses and increased downtime.

3) Data loss: For computer systems, servers, data centers and other equipment, voltage sags may cause data loss or damage. Unstable voltage supply may cause computer system crashes or storage device failures, resulting in the loss or irrecoverable loss of important data.

4) Unstable operation: Voltage sag may cause unstable operation of the power system as a whole. When the voltage drops, the frequency of the power system may also be affected, which may cause other equipment or power system components to fail to operate normally. This may trigger a chain reaction and affect the operational stability of the entire power system.

2.2 DBSCAN algorithm

DBSCAN is one of the most typical density-based clustering algorithms. Its main idea and implementation method are as follows: 1) Draw a circle with each data point as the center and the neighborhood radius (ε) as the radius. The area enclosed by the circle is the neighborhood of the data point; 2) Traverse the data, find high-density points, and gradually connect the high-density points in its neighborhood; 3) Find low-density points, connect them to the nearest high-density points in the neighborhood, and call them boundary points; if there are no high-density points in its neighborhood, the point is noise; 4) The connected points form a cluster and clustering is completed [16] . If the number of points in the neighborhood of a data point is greater than or equal to the density threshold (δ), then the point is a high-density point. Otherwise, it is a low-density point. The values of ε and δ need to be set manually.

This paper selects DBSCAN as the temporary drop clustering because DBSCAN, as a clustering algorithm, has the following advantages and can better cluster the temporary drop data of this experiment:

1) Density-based: The DBSCAN algorithm divides data points into core points, boundary points, and noise points through density-based clustering. Core points are data points with sufficient density within a given radius ε, boundary points are non-core points adjacent to core points within a given radius ε, and noise points are data points that are neither core points nor boundary points.

2) Clusters of any shape can be found: DBSCAN algorithm can find clusters of any shape, regardless of the distribution of data points. It forms clusters by connecting data points that are connected by density, without pre-specifying the number or shape of clusters.

3) Strong robustness: The DBSCAN algorithm is highly robust to noise data. Noise data points are marked as noise points in the clustering results and will not be classified into any valid clusters, thus reducing the impact of noise on the clustering results.

4) No need to set the number of clusters in advance: Unlike some traditional clustering algorithms (such as K-means), DBSCAN does not require the number of clusters to be specified in advance. It controls the compactness of clustering by setting two parameters, namely ε (neighborhood radius) and MinPts (minimum number of neighborhood points).

5) Sensitive to outliers: The DBSCAN algorithm can identify outliers and mark them as noise points. This is useful for tasks such as anomaly detection and data cleaning, and can filter out data points that do not conform to the clustering rules.

6) Efficiency: The DBSCAN algorithm has a low time complexity and can run efficiently on larger data sets. It utilizes the density information of data points and improves the efficiency of the algorithm by reducing the number of traversals of data points.

2.3 Voltage sag homology detection features

This paper selects three features of the data set, namely, the grid topology monitoring nodes, the sag start time and the sag amplitude, to perform homology identification.

1) Monitoring nodes: When naming monitoring nodes, the power grid topology often names monitoring nodes with similar topological distances. When a voltage sag event occurs, it usually spreads from one monitoring node to other nodes, so the sag events that occur at several nodes that are relatively close are likely to have the same source as the sag events that occur at the monitoring node. Therefore, using the monitoring node as a homologous identification feature can help group sag events of the same source together.

2) Sag start time: The voltage sag start time is associated with the propagation path of the fault. If the event occurs in a specific area and then voltage sags occur in other areas of the

power system, this may mean that the sag propagates in the power system. Therefore, it is meaningful to analyze the start time of the voltage sag event during the same source identification process.

3) Voltage sag amplitude: Voltage sag amplitude can provide clues to help determine the source type of the sag. Different types of sag sources (such as short circuit faults, load changes, equipment failures) usually cause voltage drops to varying degrees. By analyzing the sag amplitude, we can preliminarily infer the possible source type and identify the same source of the sag.

3. Experimental Results and Analysis

3.1 Dataset Introduction

This experiment uses 10049 sag data recorded by a provincial power company for DBSCAN algorithm clustering analysis. The sag data records the data of sag events from November 1, 2022 to November 30, 2022, including nine descriptive features: monitoring node, bus, branch, event type (voltage swell/voltage sag), occurrence time, duration, characteristic amplitude, phase difference, and analysis (ITIC tolerance zone event/ITIC over-upper limit event/ITIC over-lower limit event). The data is sorted in order from small to large according to the occurrence time. The experiment selects the DBSCAN clustering algorithm, and uses the three dimensions of monitoring node, occurrence time, and characteristic amplitude of the sag data as the characteristic variables to measure the same sag source event. The occurrence time of sag events of the same sag source is generally concentrated within a few minutes, and the characteristic amplitude values are similar. According to this clustering criterion, the 10049 data are first manually labeled, and a total of 257 homologous events are labeled, which are used as the true labels of the algorithm to calculate the accuracy of the algorithm.

3.2 Data preprocessing

The experiment first selected the three required features from the features provided by the data set: monitoring nodes, occurrence time, and feature amplitude, and then preprocessed the three features respectively. For string-type monitoring nodes, labelEncode encoding is used to convert the node name into a number starting from 0, and then the data is reduced by 100 times; for the occurrence time feature, it is first converted to a timestamp type, and then the maximum and minimum standardization MinMaxScaler is used to convert the data into a range of 0-500; for the feature amplitude, its data is reduced by 100 times. Finally, the three-dimensional features to be involved in the clustering algorithm are obtained. The flowchart of data preprocessing is shown below Figure 1.

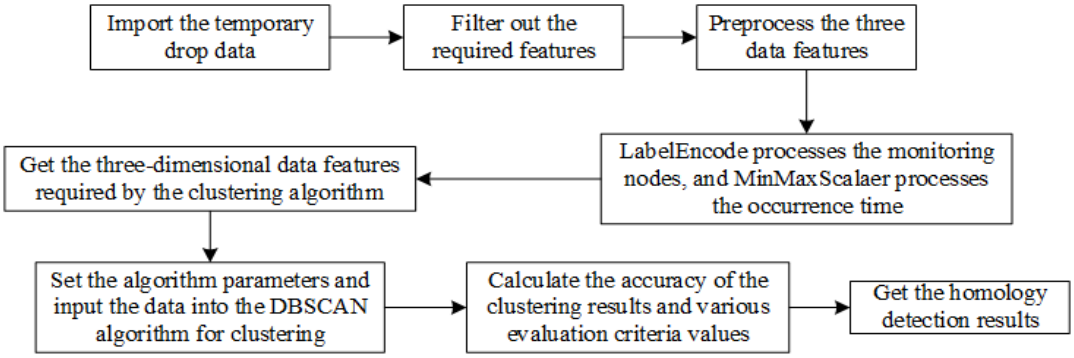


Figure 1: Voltage sag homologous detection flow diagram

3.3 Algorithm parameter settings

The DBSCAN algorithm needs to set two parameters, namely the neighborhood radius and the density threshold (i.e., the minimum number of sample points in a cluster). In order to ensure that the algorithm can distinguish the model outliers, the density threshold is selected as 2 in this experiment, and the parameter neighborhood radius should not be set too large. If the neighborhood radius is too large, all data will be clustered into one category, and if it is too small, the data will be clustered into multiple categories, resulting in inaccurate results. In order to ensure the accuracy of the experimental clustering results, this experiment uses accuracy and purity as reference functions for selecting neighborhood radius. Accuracy and purity both represent the ratio of the number of samples with accurate clustering to the total number of samples, but a higher accuracy does not mean a better clustering effect, because it ignores the differences between classes. Therefore, the two evaluation indicators of accuracy and purity are combined to iteratively select different neighborhood radii, calculate the accuracy and purity of the clustering results, and finally select the most suitable neighborhood radius for the final clustering effect evaluation and analysis.

3.4 Algorithm Evaluation Metrics

Since this experiment has manually annotated labels, six external clustering indicators with real labels are selected to detect the clustering effect, namely purity, normalized mutual information, adjusted mutual information, adjusted Rand coefficient, Fowlkes-Mallows index and accuracy [17].

1) Purity: Purity is an indicator used to measure the purity of clustering results. It calculates the sum of the most common true labels in each cluster in the clustering result and divides it by the total number of samples. The value of Purity ranges from 0 to 1. The larger the value, the purer the clustering result.

2) NMI is an indicator based on information theory that measures the similarity between clustering results and true

labels. It measures the mutual information between clustering results and true labels, while taking into account the entropy of clustering results and true labels. The value of NMI ranges from 0 to 1. The larger the value, the more consistent the clustering results are with the true labels.

3) AMI (Adjusted Rand index), adjusted mutual information, is an improved NMI. AMI is calculated based on mutual information in information theory, but is adjusted to address the deviation caused by different numbers of clustering results or random label assignment. The calculation method of AMI takes into account the pairing between two clustering results and the entropy of each clustering result to measure the similarity between them. The value range of AMI is between 0 and 1. The closer the value is to 1, the more consistent the clustering result is with the true label.

4) ARI (Adjusted Rand index), ARI is an indicator used to measure the similarity between clustering results and true labels. It takes into account all pairs of classification results and adjusts them according to their consistency. The value range of ARI is between -1 and 1. The closer the value is to 1, the more consistent the clustering result is with the true label.

5) FMI (Fowlkes-Mallows index): FMI is an indicator based on the precision and recall of clustering results, which is used to evaluate the similarity between clustering results and true labels. It calculates the ratio between the number of pairs between members of the same category in the clustering results and the number of pairs between members of the same category in the true labels. The value range of FMI is between 0 and 1. The larger the value, the more consistent the clustering results are with the true labels.

6) ACC (Accuracy, ACC), clustering accuracy, is used to compare the obtained labels with the true labels provided by the data. The value range is between 0 and 1. The larger the value, the more consistent the clustering result is with the true label.

3.5 Training process

The entire experiment was implemented in the Jupyter notebook platform. The training environment is as follows: The processor model is Intel(R) Core(TM) i5-8300H CPU @ 2.30GHz, the processor frequency is 2.30 GHz, and the memory capacity is 16 GB.

Considering that the data feature values are small, the initial value of the function iteration radius is set to 0.01, the step size is 0.01, and the iteration end radius is 0.2. The iteration results are shown in Figure 2. As the neighborhood radius increases, the clustering accuracy continues to increase, while the purity shows a trend of first increasing and then decreasing. It is easy to see from the figure that when the radius reaches 0.05, the two curves intersect, so the experiment sets the neighborhood radius of DBSCAN to 0.05.

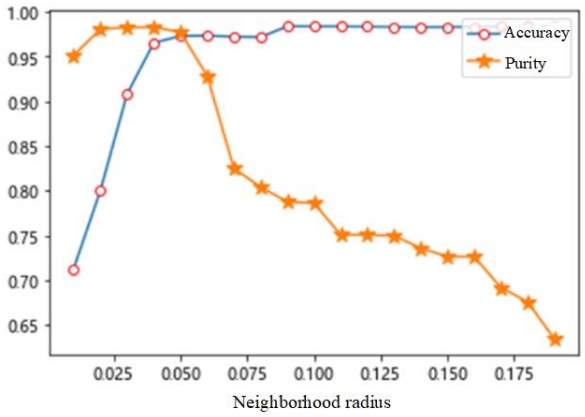
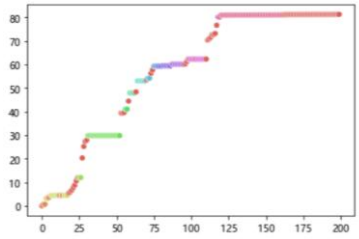


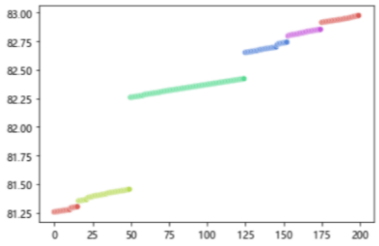
Figure 2: Neighborhood radius optimization line chart

3.6 Training Results and Analysis

After selecting the neighborhood radius, the experimental data is clustered again, and the occurrence time dimension of the first 400 data is selected for two-dimensional scatter display, as shown in Figures 3-2 and 3-3. As can be seen from the Figure 3, the algorithm clusters the temporary drop event data with an occurrence time of less than two minutes into one category, and the clustering effect is highly correlated in the time dimension, which is consistent with the identification criterion of the same source of temporary drops.



(a) 1-200 effect distribution diagram



(b) 200-400 effect distribution diagram

Figure 3: Homologous detection effect distribution diagram

At the end of the experiment, six external clustering indicators with real labels were selected as the measure of clustering effect, namely purity, normalized mutual information, adjusted mutual information, adjusted Rand coefficient, Fowlkes-Mallows index and accuracy. The six clustering indicator values of this experiment are shown in Table 1. According to the clustering indicator results, all clustering indicators have reached more than 97%, indicating that the distribution of predicted labels is basically consistent with that of real labels, the clustering effect is excellent, and the requirements of clustering homologous events are met.

Table 1: Homologous detection indicator results

	Purity	NMI	AMI	ARI	FMI	ACC
Result	0.977	0.982	0.98	0.974	0.975	0.973

4. Conclusion

This paper applies the DBSCAN algorithm to the process of identifying the source of voltage sag. The DBSCAN algorithm can not only identify outliers, but also does not require the number of clusters to be set in advance, which can well identify the source of sag data. During the experiment, the sag data was reasonably preprocessed through feature engineering, and the DBSCAN neighborhood radius parameter with the best clustering effect was selected through iteration, so that the clustering effect is even better than the result of manual labeling, achieving the goal of using the DBSCAN algorithm to identify the source of sag, while providing a basis for the prevention and control of subsequent voltage sags and improving the stability of the power system. Set your page as A4, width 210, height 297 and margins as follows:

References

[1] X. Xiangning, "Power Quality Analysis and Control[M]," Beijing: China Electric Power Press, 2010. (journal style)
[2] X. Xianyong, H. Hanyang, and W. Ying, "Analytical model of AC contactors for studying response mechanism to multi-dimensional voltage sag characteristics and its novel applications," IET Generation, Transmission & Distribution, pp. 3910-3920, 2019. (journal style)
[3] X. Yonghai, "Sensitivity of programmable logic controllers to voltage sags[J]," IEEE Transactions on Power Delivery, pp. 2-10, 2019. (journal style)
[4] G. Chun, Z. Luowei, L. Weiguo, "Three-phase power quality data compression method [J]," Power System Technology, pp. 130-134, 2011. (journal style)
[5] Z. Yi, Y. Honggeng, Y. Maoqing, "Massive power quality monitoring data management scheme based on distributed file system [J]," Automation of Electric Power Systems, pp. 102-108, 2014. (journal style)
[6] W. Ying, X. Jiani, D. Lingfeng, "Identification method of voltage sag homologous source based on typical waveform characteristics and improved DBSCAN[J]," Automation of Electric Power Systems, pp. 126-135, 2021. (journal style)
[7] X. Xianyong, G. Liangyu, L. Chengxin, "Detection method of multiple voltage sag events based on Wasserstein distance［J］," Power System Technology, pp. 4684-4693, 2020. (journal style)
[8] W. Ying, "Calculation of the point-on-wave for voltage dips in three-phase systems［J］," IEEE Transactions on Power Delivery, pp. 2068-2079, 2020. (journal style)
[9] L. Shunfu, X. Chao, T. Bo, "Application of data mining in power quality monitoring data analysis［J］," Electrical Measurement and Instrumentation, pp. 46-51, 2017. (journal style)
[10] S. Haoyuan, M. Fei, L. Danqi, "Research on voltage sag event type identification based on improved generative adversarial network [J/OL]," Proceedings of the CSEE, pp. 1-13, 2021. (journal style)
[11] S. Xuezhen, L. Qionglin, Y. Jiali, "Analysis of voltage sag characteristics based on measured data [J]," Electric Power Automation Equipment, pp. 144-149, 2017. (journal style)
[12] S. Mingming, "Practice of voltage sag monitoring and prevention [J]," Electrical Applications, pp. 4-9. 2018. (journal style)
[13] H. Wenxi, X. Xianyong, J. Yunling, Voltage sag waveform data analysis method and its application in monitoring system [J]. Power System Technology, 2019, 43(11): 4193-4199. (journal style)
[14] X. Xianyong, Z. Heqi, L. Chengxin, "Park-level and equipment-level voltage sag collaborative management optimization scheme and its investment and financing strategy [J]," Electric Power Automation Equipment, pp. 157-165, 2020. (journal style)
[15] W. Jianxun, Z. Yi, Z. Yan, "Comprehensive prevention and control scheme of voltage sag for modern industrial parks [J]," Automation of Electric Power Systems, pp. 156-163, 2020. (journal style)
[16] X. Lili, "Algorithms and applications of cluster analysis [D]," Changchun: Jilin University, 2010. (journal style)
[17] Z. Yuqin, L. Li, Z. Jianliang, F. Xiangdong, "Big data clustering method based on improved K-means++ and DBSCAN[J]," Foreign Electronic Measurement Technology, pp. 40-46, 2022. (journal style)