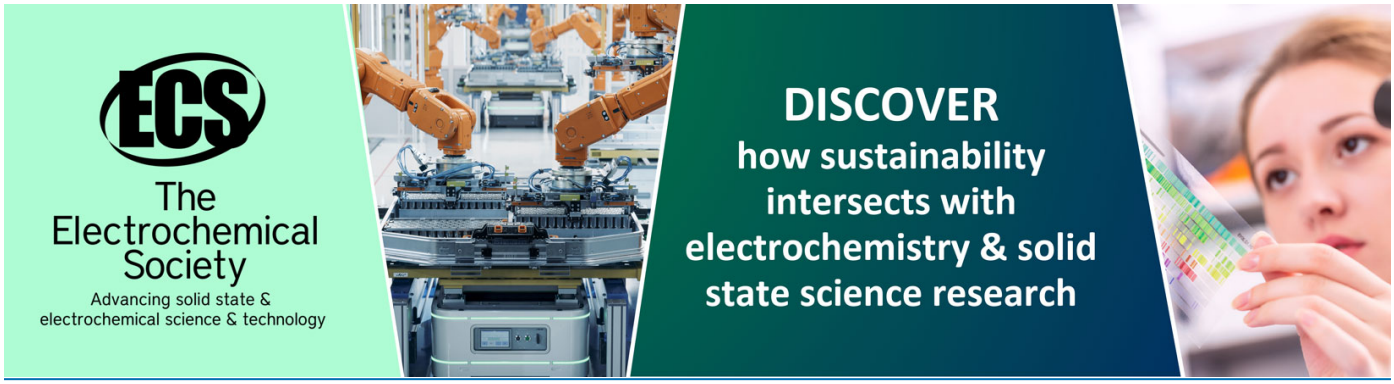
Research on Power Grid Over-voltage Anomaly Data Management Based on the Improved Clustering Algorithm



Research on Power Grid Over-voltage Anomaly Data Management Based on the Improved Clustering Algorithm

Lixia Jia[1*], Xiangming Zeng[1], Fangman Lin[1]

[1]Hainan Power Grid Co. Ltd, Haikou, Hainan, 570100, China

* Corresponding author: jialx@hn.csg.cn

Abstract. The relational database uses distributed storage for grid over-voltage anomaly data, which lacks the division of the anomaly data, resulting in a long query time for anomaly data management. For this reason, the research of grid over-voltage anomaly data management based on the clustering algorithm is proposed. The clustering algorithm is combined with the outlier detection to divide the anomaly data and improve the query efficiency. The data are classified according to their characteristics. Row storage is selected as the main storage method for grid over-voltage anomaly data, and a three-dimensional model library is used to build out the management framework of the anomaly data to realize the efficient management of the anomaly data. In the experiment, the query time consumption of the proposed method is tested, and the analysis of the experimental results shows that the proposed method has a high query efficiency in managing the grid over-voltage anomaly data.

1.Introduction

The existing method of grid over-voltage anomaly data management mainly adopts local centralized storage, storing the anomaly data in a relational database to realize the query and invocation of the anomaly data [1]. As the power grid scale expands continuously, the total data volume of historical data keeps rising, resulting in a large storage load for a relational database, and its data access performance is also affected. The traditional relational database lacks the processing capability of massive data, and it can no longer meet the management needs of large-scale abnormal data. The limitation of storage capacity causes the access efficiency of the relational database to the grid over-voltage anomaly data to be affected. Beyond the initial time, the traditional method of grid over-voltage anomaly data management also has the problem of poor scalability. To solve this problem, the operational performance can be improved by upgrading the hardware equipment of the database, or by distributing the relational database into data clusters through horizontal partitioning to reduce the inefficiency of management caused by poor scalability. However, the technology of this method is not mature enough. It lacks the existing technical framework, takes a long time cycle, and is costly. Therefore, this improvement method does not apply to the current urgent demand for abnormal data management. As the scale of the power grid continues to grow, new requirements are being placed on the management and storage of over-voltage anomaly data. The main requirement is to improve the storage space of the database to ensure data access performance. At the same time, the improved abnormal data management method needs to support efficient read and write access functions, to achieve fast data queries within the minimum read and write time frame, and improve the response speed of data operations [2-3].

2.Grid over-voltage data division based on an improved clustering algorithm

The core concept of the clustering algorithm is to divide the data points in the form of clusters so that the data inside the clusters have the highest similarity and differ from the clusters outside. This method can be applied to the detection of over-voltage anomalies in power grids, where the clustering algorithm can divide the over-voltage anomalies from the common values and achieve effective management of over-voltage anomalies in power grids.

Outlier detection, the main tool of data mining, can be used to mine anomalous data that deviates significantly from other data to obtain anomalous data values, which can be helpful for the detection of over-voltage anomalies in power grids. Therefore, combining the clustering algorithm with the outlier detection method [4] and optimizing the traditional clustering algorithm can effectively improve the efficiency and accuracy of the grid over-voltage anomaly [5] data detection work, so as to meet the actual detection requirements.

The idea of grid over-voltage outlier detection based on the clustering algorithm is: since the distance between normal data and cluster center is closer than that between abnormal data and cluster center, the clustering algorithm can be used to cluster all data of grid over-voltage, and the abnormal data of grid over-voltage can be identified by comparing the distance between data and cluster center. The distance between the abnormal data and the cluster center is used as an outlier measure to divide the abnormal data and the normal data. As a typical clustering algorithm, K-means has the advantages of simple steps and fast calculation speed. It is suitable for clustering large-scale data and is widely used in the management of abnormal data. Therefore, the K-means clustering algorithm is selected as the core algorithm for detecting the data above the grid over-voltage in this paper to divide the data of the grid over-voltage. The specific steps are as follows:

All data of grid over-voltage, including abnormal data, are taken as the main modeling object. It is assumed that the modeling sample is $\{x_1, x_2, x_3, \dots x_m\}$, where $x_i \in R_n$. Firstly, the data points of grid over-voltage are randomly selected. Assuming that the number of selected data points is k, the data points are taken as the primary clustering center, and the data points are $\{\mu_1, \mu_2, \mu_3, \dots \mu_k\}$, where $\mu_k \in R_n$. Secondly, the distance between each grid over-voltage data point and the cluster center is calculated, and the data points are allocated according to the size of the distance, and the data points are allocated to the nearest cluster center. The cluster center c_j corresponding to the sample j can be calculated. The specific formula is as follows:

$$c_j = arg\ min\|x - \mu_k\| \tag{1}$$

The secondary calculation is performed for the allocated cluster center. The specific calculation formula is as follows:

$$d = \sum_{i=1}^{m} \left(c_j - c_k\right) \tag{2}$$

Where, d represents the cluster center after the secondary allocation, and c_k represents the distance between the cluster center and the data point when the number of data points is k. Repeat the above steps until all the grid distribution voltage data are distributed. When there are no data points that can be allocated in the cluster center, it means that the K-means clustering algorithm has reached the termination condition, that is, the newly allocated data points cannot be received in different clusters. At this time, the sum of error squares is the minimum value. Generally, the degree of clustering is described by defining the distortion function [6]. The specific expression is as follows:

$$J(c, \mu) = \sum_{i=1}^{m} \|x - \mu_k\| \tag{3}$$

The J function represents the square difference in the distance between the data sample and the cluster center. The purpose of the clustering algorithm is to reduce the J value to the minimum range and ensure that all the data of the grid over-voltage are fully distributed. If the value of J is still less than the minimum value of the distance square difference, the centroid of the cluster center can be fixed, and the size of J can be changed by adjusting the category of the cluster center. Similarly, by fixing the

category of the cluster center and adjusting the centroid of the cluster center, the J purpose of function value size. When the value of J function is at the minimum, it indicates that all the data of power grid over-voltage has been divided at this time. And the divided abnormal data and normal data are obtained, which can provide support for the subsequent abnormal data management.

3.Characteristic analysis of abnormal grid over-voltage data

Since the statistical workload of the grid over-voltage abnormal data is large and there are many types of data, it is necessary to analyze the characteristics of the abnormal data for scientific management, and manage the abnormal data according to the characteristics. The causes of abnormal data are complex. For example, the low initial data accuracy causes abnormal data in subsequent processing. Or the coarse-grained data aggregation method [7] is used in the statistical process of some over-voltage data, resulting in fluctuations in the abnormal values. Therefore, it is necessary to classify and manage the grid over-voltage anomaly data according to its specific causes and characteristics. The specific classification results of the grid over-voltage anomaly data characteristics are shown in the following table.

Table 1.Characteristic analysis of grid over-voltage abnormal data

Data category	Data characteristics	Read/write characteristics	Recommended technology
Model type data	Complex relationships	Multiple updates	Relational database
Real-time data	High real-time performance	Fast access speed	In-memory database
Historical data	Large volume and complex	Fast access speed	Distributed storage databases
Document data	Large volume and complex	Multiple queries	Distributed file system

Based on the above classification results, the grid over-voltage anomaly data can be classified according to characteristics, improving the management efficiency and providing the initial data set for pre-processing.

To improve the effectiveness of grid over-voltage anomaly data management, the extracted and classified data need to be pre-processed. Usually, the original data set of over-voltage anomalies includes noisy data and irrelevant data. The error value of noisy data is caused by the deviation of the over-voltage operation of the power grid, and the error size is within the normal range. Some of the abnormal data are not caused by the deviation from the normal operation of the grid. Therefore, these data belong to the error data. When pre-processing the abnormal data, the error data needs to be removed to improve the statistical accuracy.

The Schauwelle method is used to remove the error values from the over-voltage data of the power grid. The Schauwelle method is based on the principle that the number of measurements is n. Data with an error that does not occur half as many times is rejected. The approximation can also be performed by the following formula level.

$$x \leq 1 + 0.41n^2 \tag{4}$$

Where x represents the number of times the error occurs in the abnormal data. When x satisfies the above condition, it means that x belongs to the error data at this moment and needs to be rejected.

After rejecting the error data, normalization of the abnormal data can improve the execution speed, and the specific normalization formula is as follows.

$$P_i^* = \frac{P_i}{\sum_{i=1}^{96} P_i} \tag{5}$$

Where P_i represents the abnormal data value and P_i^* represents the normalized abnormal data.

After normalizing the abnormal data, the initial data set is obtained, and the initial abnormal data set is stored to ensure the timeliness and safety of the abnormal data.

4.Setting up a storage management mode for over-voltage anomaly data on the power grid

The purpose of managing the grid over-voltage anomaly data is to achieve efficient transmission of the anomaly data within the cloud platform. Therefore, a distributed data storage management model can be formed with the help of block-chain technology [8-9]. An analysis matrix based on the identification of data management risks [10] is first constructed.

$$Q = \left[f(a,b)_{ij}\right] \tag{6}$$

$$f(a,b)_{ij} = f(a,b)_{ji}^{-1} \tag{7}$$

Where a represents the growth rate of abnormal data storage, b represents the growth of abnormal data storage time, and ij represents the length of the time block chain. The consistency between the behavior of anomalous data management and the actual anomalous data management can be solved by the following equation.

$$\vartheta = \frac{\varepsilon-\omega}{\omega-1} \cdot W \tag{8}$$

Where, ϑ represents the data management consistency metric, ε represents the minimum eigenvalue of the identification matrix, ω represents the number of eigenvalues in the anomaly data set, and W represents the anomaly data storage speed. When the calculated consistency index is less than 1%, the management behavior of the matrix is proved to be consistent with the actual management behavior, and the management of the over-voltage anomaly data of the grid can be realized. In this regard, data aggregation and collation architecture can be conceived. The specific structure is shown in the figure below.

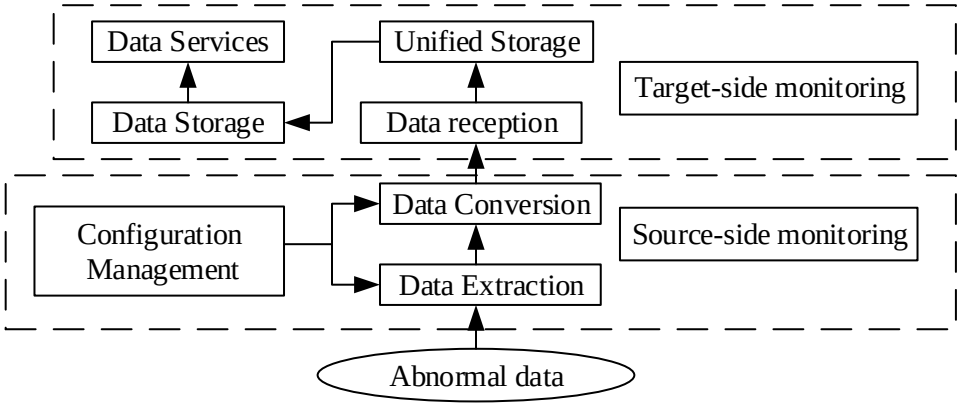


Figure 1.The architecture of anomaly data pooling and organizing

Following the abnormal data collection and collation architecture, the grid over-voltage abnormal data is extracted and converted, received through the cloud, and stored uniformly at the target end.

For the storage method, there are usually two ways to store the grid over-voltage data, namely row storage and column storage. Columnar storage stores the over-voltage data in columns, while row storage arranges the data in rows and stores them in cells. Since the characteristics of the over-voltage data are analyzed above, the abnormal data can be stored in the database in different categories, and the categories are indexed to facilitate the query of the data. In this regard, row storage is chosen as the storage method in the grid over-voltage anomaly data management system, which is conducive to improving the data reading performance and query efficiency, and is suitable for large-scale scanning queries.

5.Building a framework for grid over-voltage anomaly data management

To manage the grid over-voltage abnormal data visually [11], the abnormal data is managed through the 3D modeling platform to realize the visual query and invocation of the data. Since the total amount of

grid over-voltage abnormal data is large, a model library management scheme needs to be built, and the specific management model structure is shown in the figure below.

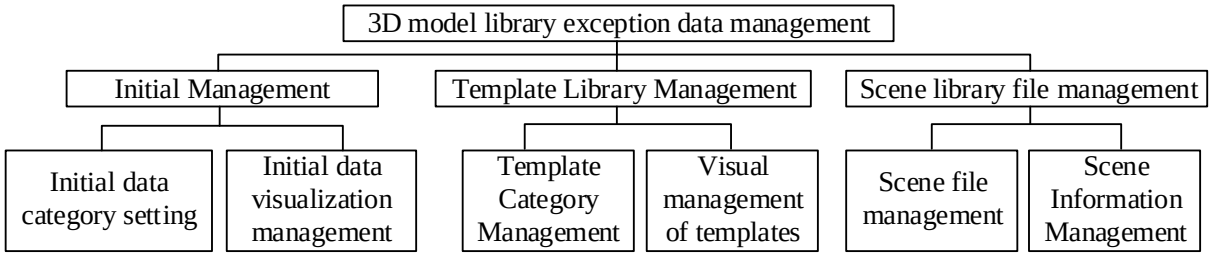


Figure 2.Abnormal data 3D model library management scheme

According to the above 3D model library management scheme, a 3D management model is constructed to manage the grid over-voltage anomaly data hierarchically through a five-layer structure: anomaly data collection layer, anomaly data cloud layer, anomaly data consensus layer, anomaly data contract layer, and anomaly data application layer. Each layer is interconnected with the other and each independent layer has its internal logic, which together forms the management model of grid over-voltage anomaly data.

The anomaly data collection layer integrates the anomaly data collected by the clustering algorithm [12-13] and extracts the key information for the features of the anomaly data. The anomaly data is transformed according to the mathematical binary of rated length for the characteristics of the data. After the anomaly data format is unified, the data is chained to the block master chain and stored as internal data of the block node.

The accuracy of the data is determined based on the data structure of the anomaly data and the characteristics of the address source, etc. If the anomaly data is accurate, the data is stored in the main body of the block through the anomaly data cloud fault. If there is an error in the anomaly data, the chain network will stop the transmission of the data; call the anomaly data back to the pre-processing step, and reject the error data to ensure that the error data will not be stored in the block body as anomaly data.

To ensure that the abnormal data of grid over-voltage can be authenticated quickly, the interest of the abnormal data block bookkeeping right is set through the data consensus layer to improve the authentication efficiency. The trusted block is authorized to enable block nodes to reach consensus quickly and improve the accuracy of data processing.

Smart contracts [14-15] are adopted as the core part of the abnormal data contract layer to turn the abnormal data computer of grid over-voltage into an embedded contract layer and realize the automated management of abnormal data based on smart contracts. The activation conditions of the smart contract are formulated, and the smart contract will automatically monitor the code after the activation conditions are met to realize the management of the grid over-voltage abnormal data.

Through the above steps, the intelligent management framework of grid over-voltage anomaly data can be built, and this part will be combined with the data division, data feature analysis, and data management mode mentioned above, so that the design of grid over-voltage anomaly data management method based on improved clustering algorithm is completed.

6.Testing and Analysis

6.1.Test Preparation

For producing accurate test results, the grid over-voltage anomaly data management method in the text was tested. A traditional relational database was used as the comparison object to compare the query time of the two management methods. The two data management methods were tested using an anomaly

data cluster consisting of 8 nodes, with the specific nodes and related parameters shown in the table below.

Table 2.Test node data

IP Address	Host Name	Related Processes
195.150.1.01	sdh01	HMaster/Region server
195.150.1.02	sdh02	HMaster/Region server
195.150.1.03	sdh03	Region server
195.150.1.04	sdh04	Region server
195.150.1.05	sdh05	Region server
195.150.1.06	sdh06	Region server
195.150.1.07	sdh07	Region server
195.150.1.08	sdh08	Region server

The selected operating environment is AMD Opteron Processor 5640 processor with 64G of RAM and 5TB SATA server hard disk. The selected operating system is Windows 10 Professional (64-bit) and the distributed storage is Cloudera version 3.2.1. The block size for storing grid over-voltage data is 64MB, and each block contains 3 replica blocks. The performance difference between the two management methods is compared by reading the query time under the node on the client side.

6.2.Analysis of Test Results

Based on the above test results, it can be seen that the traditional relational database is less effective in query time consumption, and its query efficiency will be affected when managing data with different numbers of samples. And the average value of query time consumption is around 0.8s, which indicates that the traditional relational database cannot manage a large number of abnormal data efficiently. The query time is more stable with an average value of 0.3s, which can greatly improve the management efficiency of the grid over-voltage anomaly data, indicating that the anomaly data management method proposed in this paper is more advantageous in management performance and can meet the actual anomaly data management needs.

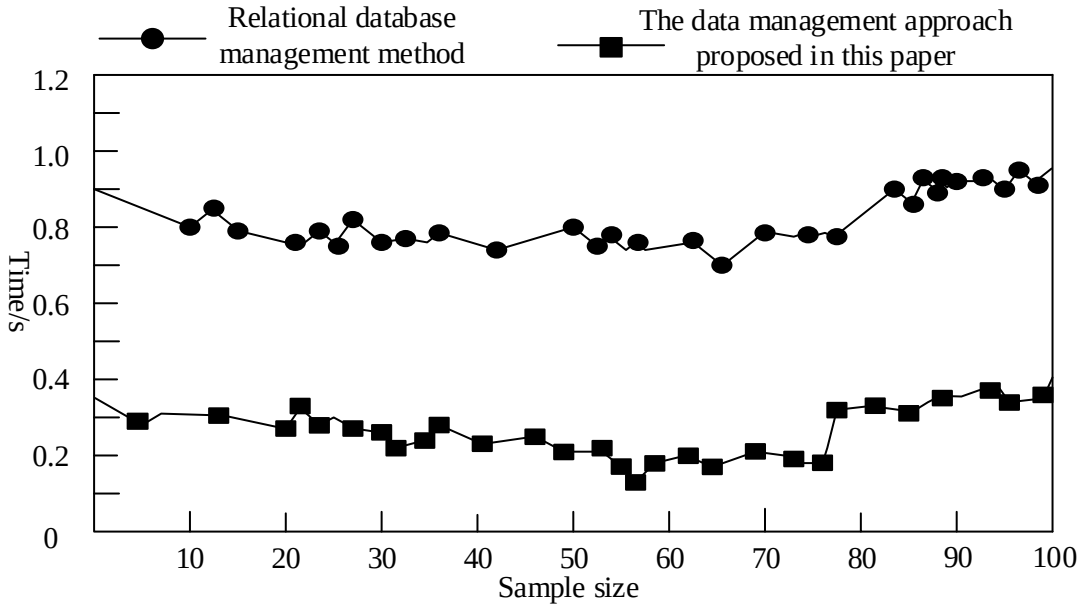


Figure 3.Management method query time consumption comparison

7.Conclusion

Based on the improved clustering algorithm technology, the article has designed a more efficient method for grid over-voltage anomaly data management, which opens up a new research idea for anomaly data management research. The improved clustering algorithm is used to divide the abnormal data, which improves management efficiency. In future research, the application scenarios of the management method still need to be explored to improve the practicality of the management method.

References

[1] Kreuwel, F. P., Mol, W. B., De Arellano, J. V. G., & Van Heerwaarden, C. C. (2021). Characterizing solar PV grid overvoltages by data blending advanced metering infrastructure with meteorology. Solar Energy, 227, 312-320.

[2] Rouindej, K., Samadani, E., & Fraser, R. A. (2020). A comprehensive data-driven study of the electrical power grid and its implications for the design, performance, and operational requirements of adiabatic compressed air energy storage systems. Applied Energy, 257, 113990.

[3] Backe, S., Kara, G., & Tomasgard, A. (2020). Comparing individual and coordinated demand response with dynamic and static power grid tariffs. Energy, 201, 117619.

[4] Gabbay, I., Shapira, B., & Rokach, L. (2021). Isolation forests and landmarking-based representations for clustering algorithm recommendation using meta-learning. Information Sciences, 574, 473-489.

[5] Jones, P. J., James, M. K., Davies, M. J., Khunti, K., Catt, M., Yates, T., ... & Mirkes, E. M. (2020). Filter K: A new outlier detection method for k-means clustering of physical activity. Journal of biomedical informatics, 104, 103397.

[6] Nedović, L., Pap, E., & Dragić, Đ. (2021). Aggregation of the triangle of distortion functions. Information Sciences, 563, 401-417.

[7] Caceres-Delpiano, J., Wang, L. P., & Essex, J. W. (2021). The automated optimization of a coarse-grained force field using free energy data. Physical Chemistry Chemical Physics, 23(43), 24842-24851.

[8] Tsao, Y. C., & Thanh, V. V. (2021). Toward sustainable microgrids with blockchain technology-based peer-to-peer energy trading mechanism: A fuzzy meta-heuristic approach. Renewable and Sustainable Energy Reviews, 136, 110452.

[9] Foti, M., Mavromatis, C., & Vavalis, M. (2021). Decentralized blockchain-based consensus for Optimal Power Flow solutions. Applied Energy, 283, 116100.

[10] Chen, F., Wang, H., Xu, G., Ji, H., Ding, S., & Wei, Y. (2020). Data-driven and safety-enhancing strategies for risk networks in construction engineering. Reliability Engineering & System Safety, 197, 106806.

[11] Manogaran, G., Shakeel, P. M., Baskar, S., Hsu, C. H., Kadry, S. N., Sundarasekar, R., ... & Muthu, B. A. (2020). FDM: Fuzzy-optimized data management technique for improving big data analytics. IEEE Transactions on Fuzzy Systems, 29(1), 177-185.

[12] Bas, E., & Egrioglu, E. (2022). A fuzzy regression functions approach based on the Gustafson-Kessel clustering algorithm. Information Sciences, 592, 206-214.

[13] Rezaee, M. J., Eshkevari, M., Saberi, M., & Hussain, O. (2021). GBK-means clustering algorithm: An improvement to the K-means algorithm based on the bargaining game. Knowledge-Based Systems, 213, 106672.

[14] Gourisetti, S. N. G., Sebastian-Cardenas, D. J., Bhattarai, B., Wang, P., Widergren, S., Borkum, M., & Randall, A. (2021). Blockchain smart contract reference framework and program logic architecture for transactive energy systems. Applied Energy, 304, 117860.

[15] Tolmach, P., Li, Y., Lin, S. W., Liu, Y., & Li, Z. (2021). A survey of smart contract formal specification and verification. ACM Computing Surveys (CSUR), 54(7), 1-38.