Anomaly Detection of Big Data Based on Improved Fast Density Peak Clustering Algorithm

Fulong Zhong and Tongxi Lin(✉)

Guangzhou Huashang Vocational College, Guangzhou 511300, China
Linxixi12021@163.com

Abstract. Aiming at the problems of traditional clustering algorithms in the process of large amount of big data anomaly monitoring under the background of big data, a big data anomaly detection method based on improved fast density peak clustering algorithm is proposed. Set up a big data anomaly clustering framework, automatically select parameters and clustering centers, evaluate big data outliers through standardized local density and distance, and can obtain outliers, extract and calculate thresholds according to features, complete anomaly scheduling, and achieve anomaly detection. The clustering algorithm designed in this paper is used for example analysis, which shows that the algorithm designed in this paper can meet the needs of actual users and improve the detection accuracy of power big data outliers.

Keywords: Improved Density Peak · Fast Peak · Peak Clustering Algorithm · Big Data Exception · Abnormal Detection

1 Introduction

Due to different data sources, different statistical caliber of the same data, front-line personnel data entry, abnormal behavior and other problems, as well as the lack of corresponding data quality control system, abnormal data is often generated. The abnormal data contains relevant information about the occurrence of system abnormal conditions, so there is a huge research value hidden behind the abnormal data, which can provide help for practical applications, such as equipment failure and abnormal power consumption detection.

In the early days, the anomaly detection was mostly conducted by technicians on the spot, and some scholars carried out relevant research. Literature [1] proposed to improve the power big data outlier detection and analysis technology in the high-speed density peak space clustering algorithm, analyze the traditional density peak space clustering method, obtain the effect of the traditional algorithm in the actual application process, improve the fast density peak value clustering algorithm, and use the automatic selection of adaptive parameters and aggregation center, The normalized local density coefficient

B. Wang et al. (Eds.): ICMTEL 2023, LNICST 534, pp. 332–348, 2024.
https://doi.org/10.1007/978-3-031-50577-5_24

and spacing can be used to evaluate the outliers of the maximum data, but the algorithm of this method is computationally expensive and time-consuming in operation. Literature [2] proposed a warning scheme for abnormal detection of power generation operation data based on big data analysis technology. It used big data analysis technology to estimate the maximum likelihood value of power generation operation data and other signals, carried out clustering analysis on sub sequences in the asynchronous window, obtained the abnormal information in each time window, and completed the data abnormal detection. However, this method has poor stability in the detection process, and the system often crashes.

In order to make up for the problems of traditional methods, this paper proposes an outlier detection method for power big data based on multidimensional time series analysis. This paper redefines the local density and distance with the idea of KNN, which improves the shortcomings of the original algorithm that does not consider the local characteristics of the data and depends on the truncation distance. The clustering algorithm avoids the clustering process, reduces the time complexity of the outlier detection algorithm based on clustering, and absorbs the advantage that the outlier detection based on neighborhood is insensitive to the data dimension. In order to ensure the scheduling effect, the scheduling exception data source is refined twice and the scheduling exception sample set is obtained. The sequential leveling method is used to complete the data processing, and the scheduler is designed according to the established optimal clustering scheduling model. The clustering algorithm is used to classify the data, and finally the category of unknown data is predicted. Simulation results based on the daily load curves of a single transformer and multiple transformers in a province show that the improved algorithm can effectively detect outliers in the data.

2 Big Data Anomaly Clustering Framework

According to different usage scenarios and types, clustering algorithms can be divided into partition clustering, pattern clustering, hierarchical clustering, and grid clustering. Different clustering algorithms are applicable to different application fields. The clustering algorithm is shown in Fig. 1:

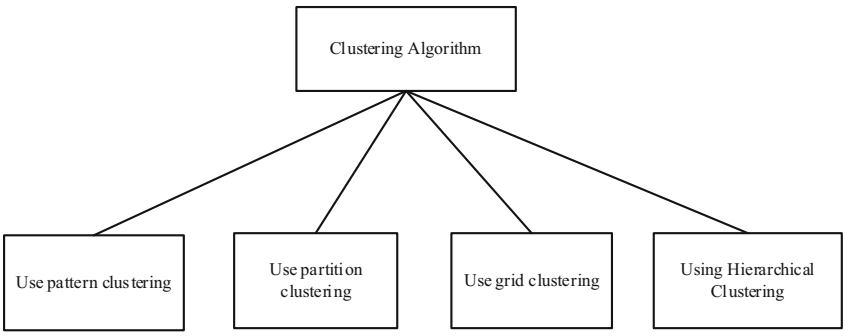

Fig. 1. Classification of clustering algorithms

Parallel processing uses multiple processing units to process input information at the same time, so as to shorten the time spent in executing tasks and effectively improve the system processing speed. In the parallel processing control system, the system acceleration ratio is closely related to the number of processing units and the task parallelism. Specifically, the task parallelism will restrict the performance of the parallel processing system. When the task parallelism is a certain value, the acceleration ratio obtained by increasing the number of processing units will reach a limit value. Parallel processing control technology includes multi processor tight coupling technology, shared memory technology, direct memory technology, cache memory technology, etc.

In the face of abnormal data, traditional serial processing methods have limitations in computing resources and time constraints, and are difficult to adapt to massive data [3, 4]. The clustering algorithm is applied to the control system for parallel processing of abnormal data to enhance the flexibility of the system and maintain efficient clustering effect for both small data elements and large data element sets. The system can process both numerical and non numerical data, and has the ability to process different data types. The advantages of clustering algorithm are shown in Fig. 2:

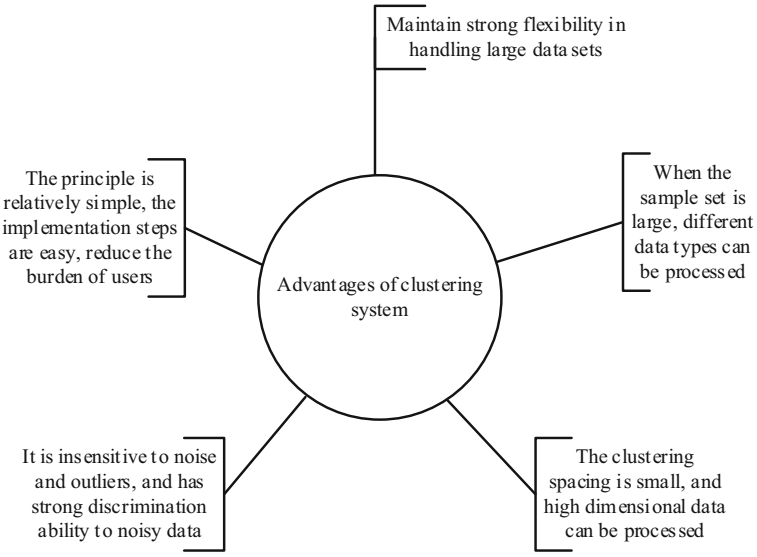

Fig. 2. Advantages of clustering algorithm

The traditional parallel processing control system manages multiple tasks by setting time sharing mode for single machine operation. After the parallel processing control system enters the running mode, any node can update the data information in the data stream. The system data is updated in the way of message on demand. The original data is partially or completely invalidated, and other nodes participating in the operation are notified. During the operation of the parallel processing control system, it is necessary to ensure that no resource livelocks occur while the system is updated. Therefore, this

paper uses the clustering algorithm to set the exception data processing framework, as shown in Fig. 3:

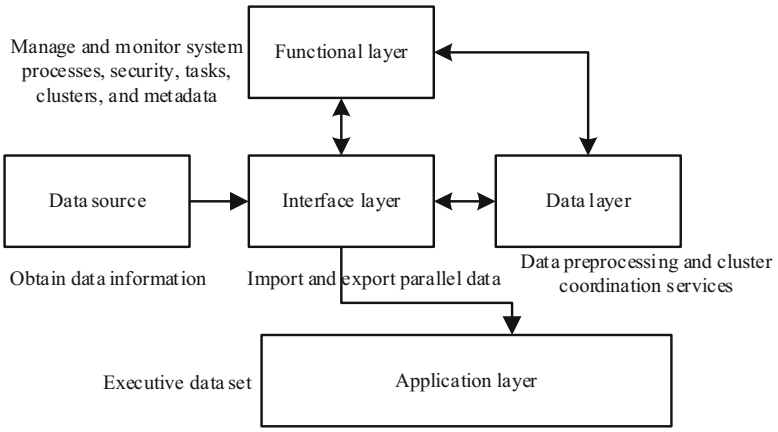

Fig. 3. Exception Data Parallel Processing Framework

As shown in Fig. 3, the control system for parallel processing of abnormal data based on clustering algorithm is mainly composed of data source, data layer, interface layer, resource layer, application layer and function layer. The system obtains data information from data sources, imports and exports parallel data through the interface layer, completes data preprocessing and cluster coordination services at the data layer, and performs clustering algorithm calculation and storage. The management and monitoring of system processes, security, tasks, clusters and metadata are completed at the functional level.

The data source and network data crawling in the interface layer constitute the network data acquisition module. The clustering module is to transfer the input data from the data source to the database in the data layer through the interface layer, convert it into data vectors for normalization, perform algorithm decomposition, and establish the clustering model [5, 6]. After cluster evaluation of the input data, the application layer of the dataset will be carried out, and the data will be stored and displayed in the foreground while completing the required functions. When the input data is abnormal, the system will automatically detect the abnormality, judge whether the system enters the fault mode, and maintain the normal operation of the system to the greatest extent.

3 Abnormal Feature Extraction of Big Data

3.1 Feature Extraction

The abnormal characteristics of big data server include the characteristics of power grid communication network and big data abnormal characteristics. In the process of connecting big data with high-level services, the big data server equipment needs to be matched abnormally. After matching, SIFT algorithm is used to extract the service

features and network features in the process of big data transmission. The established digital has a strong digital level. In the process of extracting network features, the data transmission platform MSTP and wavelength division multiplexing WDM technology improve the extraction speed.

After the extraction is completed, the big data features are classified into backbone communication network and electric communication network in terms of big data transmission business and communication business. After the classification is completed, the dispatching data features and comprehensive data features in the backbone communication network are extracted, and network protection is achieved by SDH transmission and OTN transmission.

Because big data features have self similarity and long-range correlation, it is necessary to determine whether the state of the communication link is stable when extracting big data anomaly features [7, 8]. On the premise that the link operation state is stable, extract the similar features of the overall anomaly features of big data, and eliminate different data before the aggregation of big data anomaly features. At the same time, big data anomaly features have uneven distribution of anomalies. Therefore, in order to reduce the difference of anomaly features, big data anomaly features have periodic characteristics at different time scales. In order to reduce the periodic change amplitude of data anomalies, when extracting abnormal features, because big data flow belongs to low-speed message data, the extraction direction should be from the process layer to the bay layer of big data. If the data is abnormally large, it is necessary to use ADF for inspection first. After the inspection, the abnormal features are extracted when the transmission channel is not blocked.

3.2 Abnormal Threshold Detection

Big data has different abnormal features and patterns, which can be analyzed through FARIMA model. FARIMA model can supervise and learn big data network exceptions, find out the relationship between data input and output through big data sample information, and determine the threshold.

According to the above extracted big data server anomaly characteristics, the big data server anomaly threshold is detected. For different time series of big data, the measurement probe is used to collect the target grid anomaly data. Under different time series, the power communication network data and the grid anomaly data are aggregated. The aggregation formula is as follows:

$$Y_i(j) = \sum_{k-i(j-1)+1}^{ij} Y(k) \tag{1}$$

where, $Y(k)$ represents the characteristic sequence corresponding to the power communication network data; $Y_i(j)$ represents the characteristic sequence of aggregated abnormal data of power grid; i is the aggregation time; After the aggregation is completed, $Y_i(j)$ and $Y(k)$ need to be normalized, which is the first step to determine the big data server exception threshold. The normalized big data is preprocessed under the condition of retaining the original big data, so as to carry out the multifractality and periodicity analysis of the original data. The data preprocessing will not change the original abnormal

data. After the processing, the big data server exception threshold can be obtained by the autocorrelation function, The obtained formula is:

$$H = \frac{1}{3}\sqrt{(1-B)^d} \tag{2}$$

where, H represents the big data server exception threshold; B represents the time series corresponding to the FARIMA model; d represents the order difference of FARIMA model.

After the threshold calculation is completed, the threshold detection is carried out. The big data server exception threshold detection formula is:

$$X_a = -2\sum_{b=1}^{n} \ln f_a\left(\phi_a|\hat{\phi}_b\right) + 2p_b \tag{3}$$

where, X_a represents the abnormal output value of big data during threshold detection; f_a stands for stationary series; p_b represents the non-stationary sequence after parameter fitting. After threshold detection, big data server exceptions can be better identified under different confidence levels.

4 Optimal Scheduling Based on Improved Fast Density Peak Clustering Algorithm

4.1 Optimizing Clustering

Compared with classification algorithm, clustering algorithm is an unsupervised learning algorithm, which automatically divides similar samples into the same categories according to their similarity. Prior to data training, it was not known that the data would be divided into several clusters, and data objects in the same cluster would ensure maximum similarity. From the point of view of outlier detection, although this is a clustering idea, it avoids the clustering process, reduces the time complexity of the outlier detection algorithm based on clustering, and absorbs the advantage that the outlier detection based on neighborhood is insensitive to the data dimension. Determine the data nodes in different abnormal databases, use the abnormal data nodes to establish the scheduling network link in the system, optimize the clustering through node optimization, set the network where the databases of different abnormal data sources are located as R, the node is m, the side used by the network is d, and set the scheduling network set as $R = (d, m)$. Determine the number of internal nodes of the smart network, and express the number of node links through the abnormal data connectivity. The node degree calculation process is shown in Formula (4):

$$z = \frac{\wp \cdot d}{m} \tag{4}$$

where z represents the node degree of the scheduling network; $\wp$ represents the load of the scheduling process. Determine the scheduling density of multi-source data according to the node degree of the scheduling network obtained. The larger the z is, the greater the

scheduling density is, and the more data is scheduled. Analyze the degree of connection between data. The precision of different data sources is shown in Formula (5):

$$M = \frac{|\wp \cdot L_n|}{k_n(k_n - 1)} \tag{5}$$

where M represents the precision of the abnormal data source; L_n represents the clustered data set; k_n represents the node degree of data.

The abnormal data clustering center is established according to the data detection accuracy, and the obtained abnormal data clustering set is shown in Formula (6):

$$V = \sum_{i=1}^{M} \frac{|L_n|}{k_n(k_n - 1)} \tag{6}$$

where, V represents the clustering center of the obtained anomaly dataset. The data topology is optimized and analyzed in the obtained data clustering center to achieve optimal clustering.

4.2 Scheduling Model

By optimizing the data source information that needs to be scheduled through clustering analysis, set the total number of scheduling exception data sources to be n, classify the exception data samples, and obtain the scheduling exception data sample set $\mu = \{\mu_1, \mu_2, \ldots, \mu_n\}$. In order to ensure the scheduling effect, this paper refines the scheduled exception data sources twice, and the resulting scheduling exception sample set is $\mu' = \{\mu_{11}, \mu_{22}, \ldots, \mu_{in}\} (i = 1, 2, 3 \ldots\ldots n)$.

Detect the similarity of scheduling abnormal samples, obtain the similarity coefficient according to the detection results, and establish the sample optimization clustering matrix through the similarity coefficient of abnormal data, as shown in Formula (7):

$$\chi = \begin{bmatrix} \mu_{11} \cdots \mu_{1x} \\ \vdots \quad\quad \vdots \\ \mu_{n1} \ldots \mu_{nx} \end{bmatrix} \tag{7}$$

where, χ represents the obtained anomaly optimization clustering matrix.

After determining the abnormal data matrix, the characteristics of the scheduled abnormal data will be more clearly displayed. Analyze the data characteristics in the data matrix, screen the abnormal data, and mine the operation information of the normal data. This paper uses the successive square method to complete the data processing, and deepens the processing on the basis of matrix χ to obtain the closure relationship matrix ξ, and sets the clipping coefficients to $\tau, \tau \in [0, 1]$, The classification relationship and clustering relationship of the data are determined according to the cutting analysis results. The discrimination formula is as follows (8):

$$\xi r_{iy}^{*} = \begin{cases} 1, \ r_{iy}^{*} \geq \tau \\ 0, \ r_{iy}^{*} < \tau \end{cases} \tag{8}$$

where $\xi r_{iy}^{*} = 1$ represents abnormal data clustering; $\xi r_{iy}^{*} = 0$ represents data classification.

After judging the abnormal clustering relationship, calculate the scheduling clustering center of multi-source data,

$$\overline{M}_{v} = \frac{1}{p}\sum_{t=1}^{p}\xi_{k} \quad k = 1, 2, 3, \ldots, m \tag{9}$$

where, $\overline{M}_{v}$ represents the obtained anomaly data scheduling cluster center; p represents the number of samples scheduled; t represents the time spent in clustering; k represents the number of cluster scheduling.

According to the calculated cluster center and similarity detection results, the data source is optimally scheduled. The scheduling process is shown in Formula (10):

$$F(X, Y) = 1 - \frac{\sum_{i=1}^{\hbar} p^{2} x_{i} y_{i}}{\sqrt{\sum_{i=1}^{\hbar} p^{2} x_{i}^{2} \sum_{i=1}^{\hbar} p^{2} y_{i}^{2}}} \tag{10}$$

where, X and Y represent data samples that need to be scheduled within different data sources; $F(X, Y)$ represents the dispatching result; x_i represents the data to be scheduled in sample X; y_i represents the data to be scheduled in the sample Y; $\hbar$ represents the weight coefficient of attribute factors of different scheduling data. According to the data analysis, the scheduling model is established to realize the scheduling process.

5 Design of Multi Source Big Data Cross Source Scheduler

The scheduling program is designed according to the established optimization clustering scheduling model, and the scheduling process is shown in Fig. 4:

Step 1: Optimize the clustering model to cluster and integrate the data of different data sources to be scheduled, uniformly process the data, and calibrate the data after collecting all the data. The data scheduling work of different data sources puts forward higher requirements for information integration and analysis. Therefore, an optimized scheduling database is established to integrate data information and improve the scheduling process through the database,If the scheduling data is set to come from database X and database Y respectively, the calculation process of the established scheduling database is shown in formula (11):

$$\ell(x, y) = \sum_{n-1} \min\left(\frac{\phi_{x}\phi_{y}}{2}\right) \tag{11}$$

where, x represents the data to be scheduled from database X; y represents the data to be scheduled from the database Y; ϕ_x represents the restriction coefficient corresponding to the database; The limit coefficient corresponding to ϕ_y; n represents the total number of clustering data; $\ell(x, y)$ represents the established scheduling database.

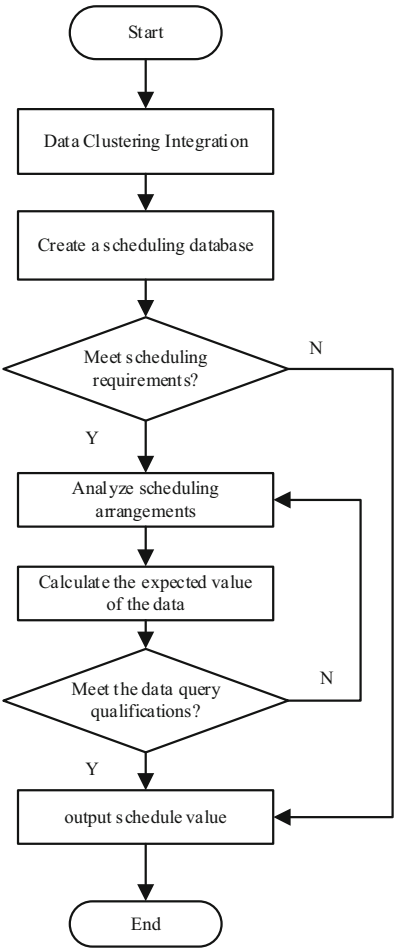

Fig. 4. Multi source big data cross source scheduling process

Step 2: Analyze the arrangement of the scheduled data sequence, and introduce Moran index to determine the expected value of the scheduled data. The time sequence of the scheduled data includes horizontal and vertical forms. In the scheduling process, the nearest field analysis method and Moran index should be introduced simultaneously to arrange the data to determine the expected value of the data. The data expectation analysis process is shown in Formula (12):

$$\varepsilon = \frac{\ell(x, y) \cdot \psi}{F(X, Y)} \tag{12}$$

where, ψ represents Moran index; ε represents the expected value of the data obtained. Transfer data in the scheduling database.

Step 3: Determine the data query conditions, schedule the data according to the query conditions, and establish the scheduling function. Data scheduling is directly related to the amount of data information. The greater the amount of data information, the greater the scheduling density. Therefore, the established transfer function should meet multiple scheduling requirements. The scheduling function is shown in Formula (13):

$$f(\ell) = \frac{\varepsilon\sqrt{\sum (F_{\max} - F_{\min})}}{n} \tag{13}$$

where $f(\ell)$ represents the established scheduling function; n represents the amount of scheduled data; $F_{\max}$ represents the maximum value of dispatching data; $F_{\min}$ represents the minimum value of scheduling data.

Step 4: Schedule the data through data separation and data integration, uniformly cluster the data information in the data scheduling set, analyze the data weight value and data expectation value, establish the data sequence, arrange the data according to the weight value, introduce the threshold to propose abnormal values, and schedule the data points using the clustering cycle mode to achieve the scheduling program. The data value after scheduling is shown in Formula (14):

$$F = \frac{f(\ell)}{N} \cdot \int k_n \cdot t + \xi r_{iy}^{*} \tag{14}$$

where, F represents the data value after scheduling; N represents the number of scheduling times. When scheduling data, the internal recording function will also be started to analyze the data transmission status, make data judgments, and ensure that the scheduling process can be completely recorded.

The parallel processing control system can process the input massive data quickly and efficiently. Generally, it can be divided into four steps: data acquisition, pre-processing, clustering and category prediction. After the collected data enters the data preprocessing stage, the missing values in the data will be supplemented and the data will be normalized. Then we use clustering algorithm to classify the data, and finally predict the category of unknown data.

Clustering algorithm will divide all data elements into clusters. There are many methods to measure similarity, such as Euclidean distance, cosine distance, Guben distance, Pearson similarity, etc. This paper uses cosine distance to calculate.Two n-dimensional vectors $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$ are established. The cosine distance is not sensitive to the absolute value, but distinguishes the difference of vectors in the direction. The cosine distance d between the two n-dimensional vectors AB is calculated by Formula (15). When d is 0, the distance is the nearest; The distance increases with the increase of d.

$$d = 1 - \frac{a_1 b_1 + \ldots + a_n b_n}{\sqrt{a_1^2 + \ldots + a_n^2}\sqrt{b_1^2 + \ldots + b_n^2}} \tag{15}$$

At the beginning of the application of clustering algorithm, it is necessary to manually select the cluster center, and the distance between the initial cluster center should be as

far as possible. First, randomly select a point in the dataset as the first cluster center, and calculate the distance di between the data object and the cluster center according to Formula (15) for each data object i in the dataset. Select a new cluster center from the data object i with a large distance di, repeatedly calculate the distance di between the data object and the cluster center, and select a new cluster center. Stop selecting when the set k cluster centers are reached, and complete the clustering classification of data.

Clustering algorithm analysis needs to be evaluated. The evaluation of clustering is usually divided into internal and external evaluation. The internal evaluation index is the sum of variance within the cluster to evaluate the data that represents the training in the use process. Formula (16) is used to complete the calculation. The external evaluation adopts F-measure index, combines the basic idea of recall ratio R and precision ratio P, and uses formula (17) to calculate and use data other than training data.

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| X_{\mathrm{ij}} - m_i \right\|^2 \tag{16}$$

$$F_{(i)} = \frac{2PR}{P+R} \tag{17}$$

In Formula (16), E is the square error function, Xij is the jth sample in the ith cluster, i is the i-th cluster from 1 to k, j is the jth sample from 1 to n, mi is the cluster center in the i-th cluster, and ni is the number of samples in the i-th cluster. In Formula (17), P is the precision ratio, R is the recall ratio, and the calculated F (i) is the external evaluation F-measure index.

6 Experimental Study

In order to verify the practical application effect of the big data anomaly detection method based on the improved fast density peak clustering algorithm, this anomaly detection method was used for specific data analysis experiments. The traditional big data anomaly detection method based on correlation analysis and the big data anomaly detection method based on information mining were also used to process the same group of data samples at the same time, and the data results from different systems were compared, Evaluate the application effect of the data detection system studied. Set the experimental parameters as shown in Table 1 below:

Table 1. Experimental Parameters

Project parameters	Project parameters
Operating voltage/V 150	Operating voltage/V 150
Working current/A 200	Working current/A 200
Operating frequency/Hz 220	Operating frequency/Hz 220
Operating system Windows10	Operating system Windows10

Three big data anomaly detection methods are used in the experiment, and the same group of unbalanced data sets are trained for multiple anomaly detection. T-sne algorithm is used to reduce the dimension of the data, and the data distribution result after two-dimensional space mapping is obtained; Then, the GAN algorithm is used to expand the original data. After multiple training, the expanded data associated with the original data is obtained. By comparing the distribution, concentration and distribution of the expanded data and the original data, the system's data detection and classification F1 value can be analyzed. The F1 value can be used as an evaluation index to measure the system's classification performance.

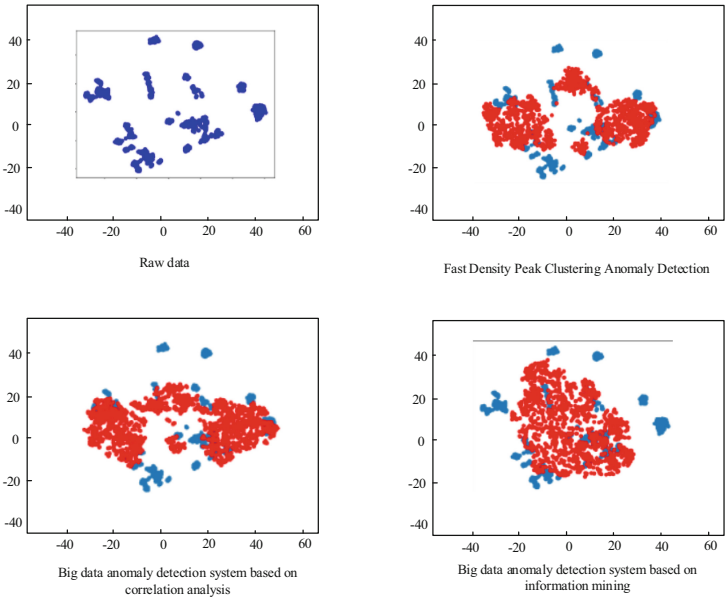

Fig. 5. Distribution of data generated after 5200 trainings

As can be seen from the above Fig. 5, compared with the other two traditional detection methods, the big data anomaly detection method studied in this paper, after 200 times of training, has a low coincidence rate between the expanded data and the original data, and the distribution is relatively scattered, indicating that the data analysis and classification detection capabilities of this method are low. In order to further test the classification performance of the data set of the method, continue to deepen the training degree of the data, record the similarity between the expanded data and the original data produced at different levels by increasing the training times, and then calculate the F1 value of the method data set classifier, which is compared with the F1 value of the original data, as shown in Fig. 6:

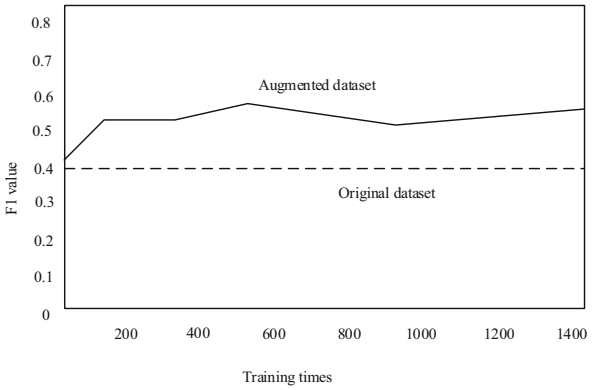

Fig. 6. Comparison Results of F1 Values between Generated Data and Original Data

It can be seen intuitively from the above Fig. 6 that the F1 value of the original data of the F1 range of the expanded data obtained after the method processing has a large difference, indicating that there is a large difference between the generated data and the original data, and the data set classification performance of the detection method is poor. And with the continuous increase of training times, the F1 value of the expanded dataset is still at a higher level, and does not decrease with the deepening of the processing degree.

On the basis of the above data set classification processing results, the data set features and anomalies are detected and analyzed. By comparing the number of detected feature points and the corresponding classification F1 value, the data detection accuracy of the three classification algorithms is analyzed.

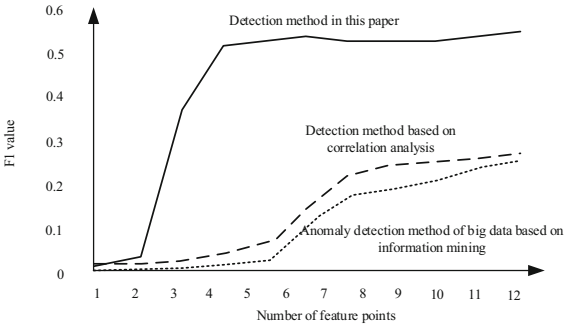

Fig. 7. Comparison Results of F1 Values Classified by Three Algorithms

It can be seen from the above Fig. 7 that when the number of feature points in the classification algorithm of the detection method studied in this paper is 2, the F1 value has reached about 0.5, while when the number of feature points in the traditional algorithm exceeds 5, the classification F1 value gradually rises, which indicates that the detection accuracy of the method studied in this paper is higher than that of the traditional method.

According to the above, the classification performance of the research method is high, so the detection and extraction ability of data feature points is also high. The accuracy of data classification detection and feature extraction for the same dataset is lower than that of traditional detection methods, which not only makes the analysis and processing of data information more difficult, but also reduces the overall working efficiency of detection methods, thereby affecting the progress of data detection. In conclusion, the large data anomaly detection method based on the improved fast density peak clustering algorithm studied in this paper has good data classification performance and high accuracy of data detection, which is conducive to the identification and detection of large-scale data anomaly information.

In order to further verify the effectiveness of the methods proposed in this paper, the detection methods proposed in this paper are selected and compared with the traditional data mining based method for identifying abnormal flow of power grid data servers and the width learning based method for identifying abnormal flow of power grid data servers. The average detection delay experimental results of the three identification methods are shown in Table 2 below:

Table 2. Average Time Delay Test Results

Load/kW	Average delay/ms		
	Data Mining Methods	Width learning method	Methods in this paper
0.1	2.25	2.15	0.12
0.2	2.46	2.37	0.14
0.3	2.69	2.54	0.19
0.4	2.88	2.68	0.23
0.5	3.01	2.79	0.26
0.6	3.23	2.97	0.27
0.7	3.66	3.09	0.31
0.8	3.98	3.24	0.34

According to the above Table 2, with the increase of load, the average delay of the anomaly identification method is also increasing. The average delay of the detection method proposed in this paper is always lower than that of the traditional identification method. The reason for this phenomenon is that the large data anomaly detection method proposed in this paper based on the improved fast density peak clustering algorithm has low mean square error, which can quickly extract the traffic characteristics and judge the average delay results according to the traffic characteristics. The threshold model proposed in this paper can well identify abnormal traffic, analyze the abnormal conditions of different networks, and complete the identification according to the abnormal conditions. However, the traditional anomaly detection method verifies the internal traffic through the network switch. The verification time is too long, and the parameter

coefficients need to be calculated repeatedly during the verification, resulting in a long verification delay.

The packet loss rate test results are shown in Table 3 below:

Table 3. Test Results of Packet Loss Rate

Load/kW	Packet loss rate/%		
	Data Mining Methods	Width learning method	Methods in this paper
0.1	10–6	10–6	10–6
0.2	10–6	10–6	10–6
0.3	10–5	10–6	10–6
0.4	10–5	10–5	10–6
0.5	10–5	10–5	10–6
0.6	10–4	10–5	10–5
0.7	10–4	10–4	10–5
0.8	10–4	10–4	10–4

It can be seen from the above Table 3 that the packet loss rate of the identification method proposed in this paper is always lower than that of traditional identification methods, and with the increase of load time, the packet loss growth rate is lower than that of traditional methods, which has better practical application effect.

Through experimental testing, observe the stability of the three methods during operation. The experimental results are shown in Fig. 8:

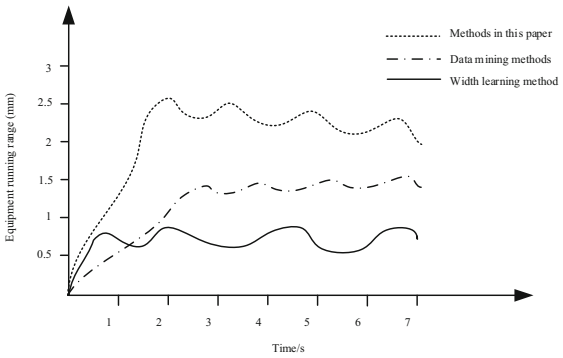

Fig. 8. Comparison Diagram of Equipment Stability

As shown in the Fig. 8, the stability performance of the equipment in this method is much higher than that in traditional methods. The hardware equipment and application program design in this paper cooperate with each other to make the operation of the

equipment more smooth. The equipment used in this paper is also highly organized. The operation of equipment is connected with each other, which does not cause the equipment to get stuck. Figure 9 shows the comparison of anomaly identification frequency between this method and traditional methods:

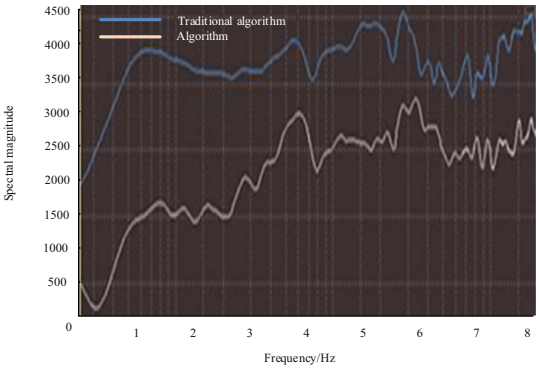

Fig. 9. Abnormal Identification Frequency Chart

As shown in Fig. 9, the anomaly identification frequency of the method in this paper is high, and it can quickly find exceptions. The SKTA protocol is referenced in the program designed in this paper, which greatly reduces the workload and speeds up the anomaly identification efficiency of the method in this paper. However, the data operation workload of traditional methods is huge, and the frequency of anomaly identification is low.

To sum up, compared with the traditional anomaly detection methods, the stability of the anomaly detection method studied in this paper is improved by 15.2%, and the anomaly frequency recognition ability is improved by 19.8%. The overall performance is better than the traditional performance, and has stronger practical operability.

In order to further verify the convergence of the proposed algorithm, it is compared with the above abnormal flow identification methods of power grid data server based on traditional data mining and width learning. The convergence results of the three identification methods are shown in Fig. 10.

In summary, compared with the two traditional anomaly detection methods, the anomaly detection method studied in this paper has stable operation and strong convergence. The overall convergence performance is better than that of the traditional method and has strong practicability.

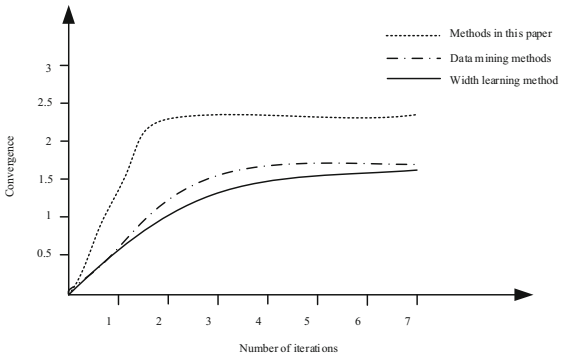

Fig. 10. Comparison results of convergence of the three algorithms

7 Conclusion

This paper proposes an improved fast density peak clustering algorithm, which defines the local density and distance through a new idea, improves the problems in the traditional algorithm, defines the rules for outlier judgment, and optimizes the algorithm based on outlier detection angle. This method has good performance when used in the simulation experiment of a transformer's daily load curve. After the detection of abnormal values, it can realize the analysis of abnormal power consumption and the monitoring of equipment status in combination with the actual business. It can also correct the abnormal values with business rules, so that the data quality can be improved.

References

1. Shao, M., Qi, D., Xue, H.: Big data outlier detection model based on improved density peak algorithm. J. Intell. Fuzzy Syst. **40**(9), 1–10 (2020)
2. Guo, L.: Research on anomaly detection in massive multimedia data transmission network based on improved PSO algorithm. IEEE Access **8**, 95368–95377 (2020)
3. Tu, B., Yang, X., Li, N., et al.: Hyperspectral anomaly detection via density peak clustering. Pattern Recogn. Lett. **129**, 144–149 (2020)
4. Zhang, G., Li, N., Tu, B., Liao, Z., Peng, Y.: Hyperspectral anomaly detection via dual collaborative representation. IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing **13**, 4881–4894 (2020)
5. Shi, Y., Shen, H.: Anomaly detection for network flow using immune network and density peak. Int. J. Netw. Secur. **22**(2), 337–346 (2020)
6. Xu, Z., Fu, N., Fu, Z., Liu, L.: FDTD Numerical simulation and detection capability analysis of small loop transient electromagnetic method in shallow water. IOP Conf. Ser.: Earth Environ. Sci. **660**(1), 012090 (2021)
7. Huang, S., Guo, Y., Yang, N., Zha, S., Liu, D., Fang, W.: A weighted fuzzy C-means clustering method with density peak for anomaly detection in IoT-enabled manufacturing process. J. Intell. Manuf. **32**(7), 1845–1861 (2020)
8. Guorui, C., Xuhua, Y.: Real-time detection algorithm of abnormal data based on HDFS open source architecture. Comput. Simul. **38**(8), 445–449 (2021)