Design of Online Monitoring Method for Distribution IoT Devices Based on DBSCAN Optimization Algorithm

Chaofan Hou*, Nan Xu, Siyu Liu
State Grid Beijing Electric Power Company Information and Communication Branch, Beijing, China
E-mail: 18612330569@163.com, xunan2024@126.com, julyand131@163.com
*Corresponding author

Keywords: distribution internet of things, clustering algorithm, abnormal data detection, time correlation, parameter adaptation

Received: June 14, 2024

In response to the data mutation problem caused by equipment failures in the distribution Internet of Things, this study proposes a density-based clustering optimization algorithm for online monitoring of equipment data anomalies. This method considers the local and global similarity of high-dimensional measurement data and constructs a composite time series similarity measurement criterion. Improvements are made to the density-based clustering algorithm, which combines with preprocessed device historical measurement data to adaptively generate global density parameters. Through clustering training, core data points are obtained to detect abnormal changes in data. The experiment showed that compared to traditional density-based clustering algorithms, the improved algorithm had good clustering performance, with standardized mutual information and adjusted mutual information increased by about 2%. Compared to anomaly detection algorithms, the density-based clustering optimization algorithm for anomaly detection of equipment data in the distribution Internet of Things increased the detection rate by 38% and reduced the false detection rate by 65%. Therefore, the proposed online monitoring method for data anomalies can improve the data detection rate and has high practical value for the reliable operation of distribution Internet of Things systems.

Povzetek: Prispevek predstavlja izboljšano metodo za zaznavanje anomalij podatkov v distribucijskih IoT napravah z optimiziranim algoritmom DBSCAN, ki povečuje natančnost zaznavanja in zmanjšuje lažne alarme.

1 Introduction

With the continuous development of the national economy, the scale of the power grid has grown rapidly, and the stable operation of the power grid directly affects economic development and social production construction. “Smart grid” is a new type of power network that integrates modern sensing and measurement technology, information technology, communication technology, and computer control technology, achieving digitalization and intelligence in power generation and transmission [1-3]. To achieve the interconnection of all things and human-machine interaction in various aspects of the power system, the ubiquitous power Internet of Things (IoT) is further designed built on the smart grid, and the important link in its construction is the Distribution Internet of Things (D-IoT). Lingda et al. proposed a trust evaluation method based on information entropy to address the internal network attack threat of D-IoT terminals. It used an exponential distribution reputation model to estimate direct trust values, and combined entropy theory to compensate for the inaccuracy of direct trust judgments. This method could effectively resist switching attacks and collusion attacks [4]. To improve the accuracy of fault location for feeder terminal units in D-IoT, Xu et al. proposed a fault location algorithm based on the Genetic Algorithm (GA) and improved the unified matrix algorithm, which combines GA to achieve fault location. This algorithm could effectively obtain the localization results of error information [5]. Huang et al. proposed a cloud edge collaborative processing method based on an attention mechanism to solve the delay problem in data calculation of D-IoT edge devices. They combined edge computing and cloud computing to solve the timeliness of data analysis and calculation. This algorithm could achieve efficient utilization of electricity resources and accelerate data iteration [6].

D-IoT collects and uploads information through numerous Low-voltage Terminal Units (LTUs) to achieve fault detection and safe maintenance of distribution lines. LTU may cause Abnormal Data (AD) upload due to communication module failure or battery depletion, which can be detected through device self-inspection or main station monitoring. However, when there is a malfunction in the LTU sampling circuit or data anomalies caused by abnormal storage modules, the AD

is difficult to detect. The false alarm of abnormal LTU can have an impact on the reliability of solving system faults. Therefore, LTU measurement Data Abnormal Detection (DAD) is one of the important links to improve the reliable operation of D-IoT systems [7-9]. DAD is a research hotspot in data mining, and in recent years, experts and scholars have conducted relevant research on DAD. Lei et al. designed an unsupervised time series DAD framework based on spectral analysis to detect abnormal samples in time series. This framework utilized spectral analysis to decompose time series and combined neural networks to predict trend and seasonal sequences. This method has been proven to effectively improve the detection rate of AD [10]. Gan et al. introduced utility-aware anomaly sequence rules into anomaly detection methods, combined with pruning strategies, to explore the upper bound of utility-aware anomaly sequence rules for anomaly value detection. This method had good effectiveness and scalability [11]. Jeong et al. proposed local outlier probability to identify seismic noise in the context of unsupervised anomaly detection, and utilized binary decision-making combined with moving windows to reduce the energy loss of useful signals. This method had good feasibility in denoising applications [12]. Yang et al. proposed a multivariate regression DAD and restoration based on spatio-temporal correlation to address the issue of insufficient feature extraction ability in drone flight data. It used correlation analysis methods to select parameters and designed a multiple regression model that integrates attention mechanisms to improve DAD performance, which could achieve data recovery [13].

At present, research on D-IoT mainly focuses on application solutions and system architecture design. The amount of data collected by D-IoT LTU nodes is large, and the spatio-temporal complexity of the above DAD methods is low, which has certain limitations for high-dimensional D-IoT LTU DAD. This makes the scalability of the detection method weaker and the practical application range smaller. At the same time, the above methods still lack adaptability research involving datasets. Therefore, this study proposes an improved clustering algorithm-based D-IoT LTU node online monitoring method. This study innovatively adopts a classification method based on global density parameters, expanding the adaptability

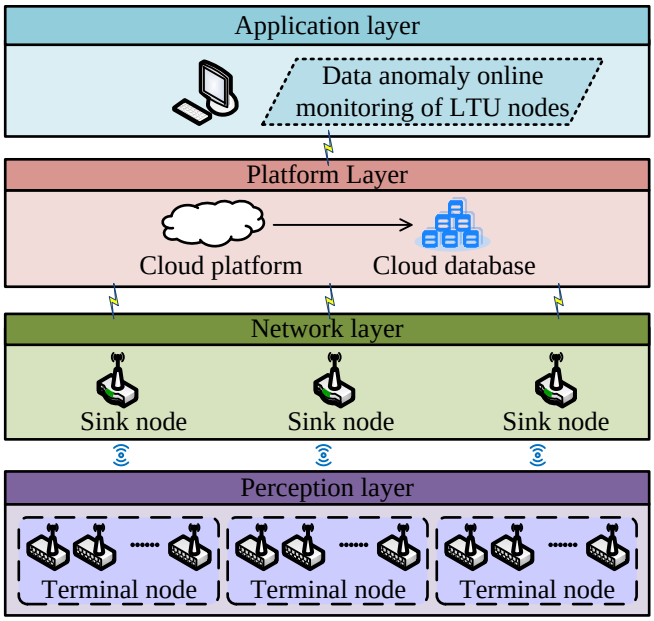


Figure 1: Architecture of D-IoT

of clustering algorithms to datasets. The research objective is to improve the accuracy of LTU node DAD and provide data foundation and technical reference for the stable operation and system management level of future D-IoT systems.

2 Methods and materials

This study proposes a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for adaptive generation of global density parameters, which optimizes the global density parameters of DBSCAN through the Composite Time Series Similarity Measure Criterion (CTSSMC). On the basis of optimizing the DBSCAN and combining the time correlation of D-IoT LTU Measurement Data (LTU-MD), a D-IoT LTU node data anomaly online monitoring method is proposed to detect LTU nodes with abnormal changes in data.

2.1 DBSCAN algorithm for adaptive generation of global density parameters

Cluster analysis refers to the classification of data points in a dataset according to a specific metric, where the differences between data points in the same class are minimized and the similarities between data points in different classes are minimized [14-16]. D-IoT data have continuity and similarity, and clustering analysis can distinguish between normal data and mutation data. Figure 1 shows the architecture of D-IoT, which includes the perception layer, network layer, platform layer, and application layer. The perception layer is the end of the D-IoT, and a large amount of data information collected by LTU nodes is transmitted to the application layer for detecting faults and monitoring loads. The data collected by LTU is time-series data with timestamps, so Time Series Similarity Measurement (TSSM) is very important for the clustering analysis of D-IoT data.

The existing TSSM methods can be divided into elasticity metric and lock step metric based on the time series comparison method [17-18]. Due to the temporal nature of D-IoT data, timing is crucial for TSSM. Therefore, a lock-step metric is used to analyze the similarity of D-IoT data. The distribution of each component in the sequence can lead to significant differences in measurement results [19]. Therefore, this study considers the distribution of each component in the sequence and proposes CTSSMC to analyze the global and local similarity of LTU-MD sequences. The measurement time series of two LTU nodes in D-IoT at a certain time is shown in equation (1).

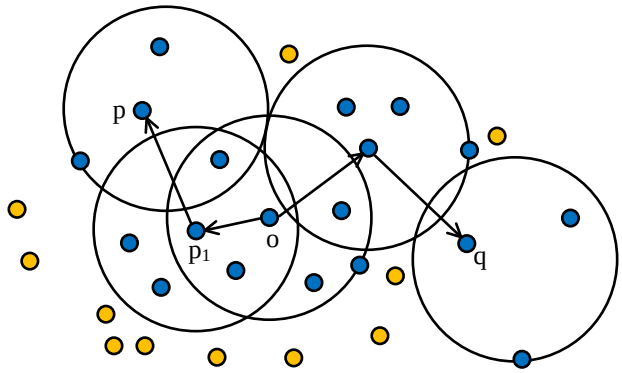


Figure 2: The relationship between data points

$$\begin{cases} X_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,t}\} \\ X_j = \{x_{j,1}, x_{j,2}, \cdots, x_{j,t}\} \end{cases} \quad (1)$$

In equation (1), X_i and X_j are LTU nodes. CTSSMC comprehensively considers the Jensen-Shannon (JS) divergence distance, Euclidean distance, and D-IoT error mode distance to measure the similarity of probability distributions, as shown in equation (2).

$$D_{ij} = dist(X_i, X_j) = \frac{1}{4} D_{jsd} + \frac{1}{2} D_{edc} + \frac{1}{4} D_{mis} \quad (2)$$

In equation (2), D_{ij}, D_{jsd}, and D_{edc} respectively represent the similarity between two sequences, the probability distributions between sequences, and the sequence amplitude. D_{mis} is the error mode distance of D-IoT. D_{jsd} is calculated based on the JS divergence, as shown in equation (3).

$$\begin{cases} D_{jsd} = \frac{1}{2} D_{kl}\left(P_i \| M\right) + \frac{1}{2} D_{kl}\left(P_j \| M\right) \\ M = \frac{P_i + P_j}{2} \end{cases} \quad (3)$$

In equation (3), $D_{kl}(\square)$ is the KL divergence. $P_i = p_i(f)$ and $P_j = p_j(f)$ represent the probability density function distribution of sensor data flow based on samples. M represents the average probability density function distribution of P_i and P_j based on samples. The calculation of $D_{kl}(\square)$ is equation (4).

$$D_{kl}\left(P_i \| P_j\right) = \sum_i p_i(f) \ln \frac{p_i(f)}{p_j(f)} \quad (4)$$

D_{edc} is calculated based on Euclidean distance, as shown in equation (5).

$$D_{edc} = \left\| X_i - X_j \right\| = \left(\sum_{n=1}^{t} \left(x_{i,n} - x_{j,n} \right)^2 \right)^{\frac{1}{2}} \quad (5)$$

The error modes of LTU-MD include erroneous measurement values with collected data streams less than zero and missing data, calculated as shown in equation (6).

$$D_{mis} = \begin{cases} 0, if\ \varphi_i = \varphi_j, \varphi_i = \begin{cases} 0, if\ \exists x \in \{1, \cdots t\},\ f_{i,x} < 0 \vee f_{i,x} = NaN \\ 1, other \end{cases} \\ 1, other \end{cases} \quad (6)$$

The JS divergence and Euclidean distance are calculated using preprocessed LTU-MD, while the error mode distance is obtained by unfilled and uncleaned LTU-MD. This study utilizes DBSCAN's global density parameters to classify the data in the dataset [20-21]. The relationship between data points in dataset D is shown in Figure 2.

The density p1 can be directly reached from data point o, the density q can be reached from data point o, and the density p can be connected to data point q. The DBSCAN algorithm classifies data points based on global density parameters and can detect noisy data. Therefore, the DBSCAN is often used in DAD. Global density parameters contain the Neighborhood Radius (NR) and Neighborhood Density Threshold (NDT) of the data, and the NR calculation is equation (7).

$$N_{Eps}(p) = \left\{ q \in D \middle| dist(p,q) \le Eps \right\} \quad (7)$$

In equation (7), $N_{Eps}(p)$ is the dataset, Eps represents NR, and $dist(p,q)$ is the distance measurement criterion between p and q.

The selection of global density parameters directly affects the clustering results, and inappropriate global density parameters can lead to false positives and missed detection in DBSCAN during DAD. In response to the shortcomings of DBSCAN, this study proposes the DBSCAN algorithm of Adaptive Generating Global Density Parameters (AGDP-DBSCAN). This algorithm distinguishes between normal and abnormal measurement data based on the distribution of AD in D-IoT LTU measurement data, combined with analysis of the distance distribution of LTU historical data points. Based on the discrete distribution characteristics of AD, NR and NDT are determined. The distance matrix of $n \times n$ is obtained from the measurement dataset with a data volume of n, calculated by equation (2). The distance curve in the distance matrix refers to the distance between each data point and other data points. Four distance curves are selected for polynomial fitting, and the fitted curves are shown in equation (8).

$$dist_i(x) = p_i x^4 + q_i x^3 + r_i x^2 + s_i x + t_i \quad (8)$$

In equation (8), $dist_i(x)$ is the i-th distance curve, x is the data point number. p_i, q_i, r_i, s_i, and t_i denote the fitting curve parameters. The second-order derivative of the fitted curve $dist_i''(x)$ is equation (9).

$$dist_i(x) = p_i x^4 + q_i x^3 + r_i x^2 + s_i x + t_i \quad (9)$$

In equation (9), assuming $dist_i''(x) = 0$, and the root of the equation is found.

The maximum in the root serves as the value of the inflection point position x_i, and the NR of the data is taken as the mean of the experimental evaluation curve, as shown in equation (10).

$$Eps = \frac{1}{n} \sum dist_i(x_i) \quad (10)$$

In equation (10), the derivative of the fitted curve is taken to obtain $dist_i'(x)$, and substituted into equation (8) to obtain x_i'. The specific calculation is equation (11).

$$dist_i(x_i') = Eps \quad (11)$$

In equation (11), x_i' is the minimum number of sample points for the i-th evaluation curve. NDT is the average of the minimum points on all curves, calculated as expressed in equation (12).

$$MinPts = \frac{1}{n} \sum x_i' \quad (12)$$

In equation (12), $MinPts$ is NDT.

2.2 D-IoT LTU node data anomaly online monitoring method

Based on the Temporal Correlation (TC) of D-IoT measurement data and combined with the AGDP-DBSCAN algorithm mentioned above, this study designs a D-IoT LTU node data anomaly online monitoring method. In D-IoT, the TC of LTU-MD nodes refers to the functional connection between data continuity

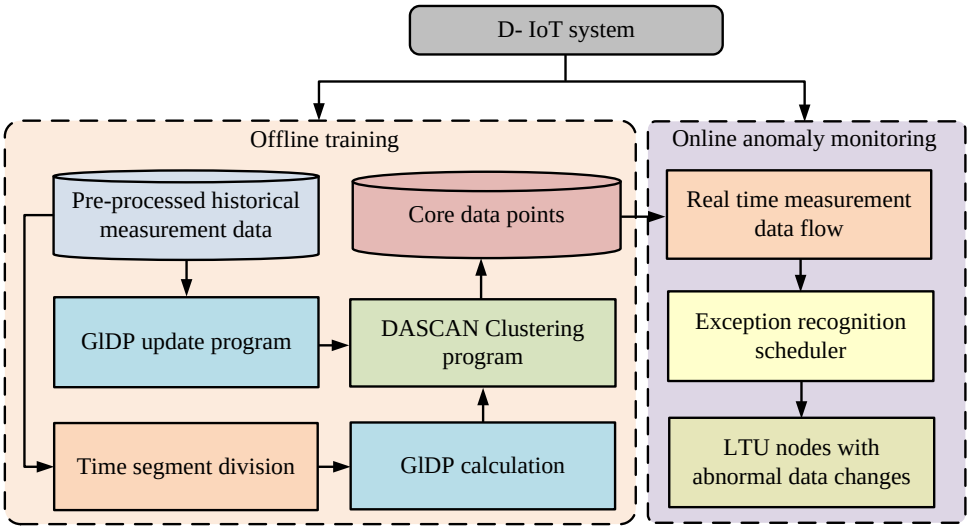


Figure 3: Data anomaly monitor program based on AGDP-DBSCAN

and similarity. The variation pattern of LTU-MD can be obtained through the TC of measurement data, mining historical data information, extracting measurement data change characteristics, and improving the efficiency of online monitoring of LTU node data anomalies. In traditional clustering data anomaly monitoring, the overall temporal data is used as the input for clustering analysis. When outliers appear in the clustering results, it is determined that the data of the LTU node has undergone abnormal changes. The data dimension of D-IoT is high. Directly inputting the overall data will increase the calculation time, and directly clustering high-dimensional data will cause dimension disasters, leading to a decrease in monitoring accuracy. When measurement data from new nodes is added, it is necessary to perform cluster analysis on the overall data again. Therefore, based on the TC of the LTU-MD of D-IoT, this study divides the measurement data into time segments with a length of L. This segment is further divided into sub time segments of l length. Each sub segment constitutes a dimension of the time segment, that is, the data dimension of the time segment is L/l. According to equation (2), the distance of input data points for LTU node clustering analysis is equation (13).

$$D_{m,n} = dist\left(F_m, F_n\right) \tag{13}$$

In equation (13), $D_{m,n}$ is the distance between the data points input for clustering analysis. F_m and F_n represent time segments.

All measurement data are divided into time segments based on the time series, and the dataset for clustering analysis input is calculated according to equation (13). Figure 3 shows the data anomaly monitor program framework for D-IoT LTU nodes based on the AGDP-DBSCAN algorithm. This framework consists of offline training and online anomaly monitoring. Offline training is based on the historical LTU-MD nodes to gain the global density parameters of Core Data Points (CDPs) and DBSCAN clustering. Online anomaly monitoring utilizes training results to detect measurement data.

In offline training, after preprocessing the historical LTU-MD nodes, time segments are divided and used as the training dataset. Each LTU node is trained to obtain CDPs that represent the changes in LTU-MD nodes from

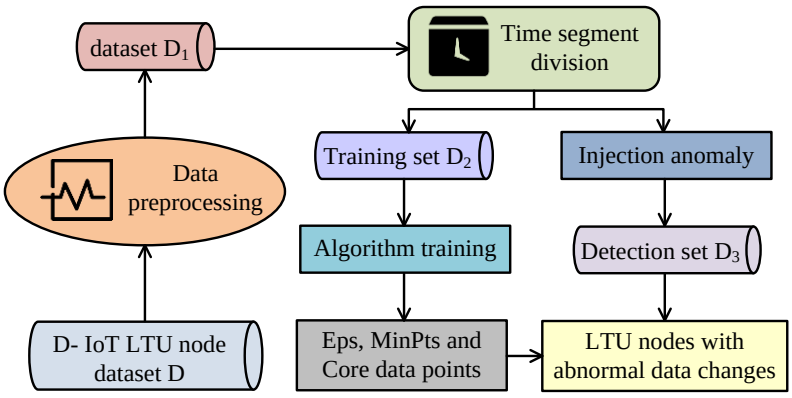


Figure 4: Data anomaly detection process

the training results, forming a CDP set. In online anomaly monitoring, when LTU measures new time period data (i.e. new data points, NDP), equation (2) is used to calculate the distance between NDP and all CDPs obtained from offline training to determine whether NDP is within the neighborhood range of the core point. The NDP is within the neighborhood range and represents normal measurement data. The NDP is not within the neighborhood range, indicating abnormal changes in the LTU node data. The randomly selected LTU-MD may not surely contain all types of AD. This study incorporates artificially simulated AD to obtain a dataset of abnormal nodes with D-IoT data characteristics. There are 5 types of simulated AD, including constant anomaly, drift anomaly, bias anomaly, shock anomaly, and periodic anomaly. The error function of drift anomaly is equation (14).

$$e = k\left(t - t_0\right) \tag{14}$$

In equation (14), e is the error function, k is the gain factor, and t_0 is the time when an anomaly occurs. The error function of periodic anomalies is equation (15).

$$e = a_0 + \sum_{n=0}^{\infty} a_n \sin\left(n\omega t + \varphi_n\right) \tag{15}$$

In equation (15), a_0, a_n, and φ_n are all constants, and $n = 1, 2, 3, \cdots$. 790 LTU terminal nodes in the D-IoT system of a certain university are selected, using a portion of the selected LTU-MD as dataset D for detection. Dataset D is subjected to DAD by the aforementioned data anomaly monitoring program, and the DAD process is shown in Figure 4.

Dataset D is preprocessed to obtain dataset D_1. The TC of data is analyzed, and time segments are divided. The training set D_2 is selected from D_1, the remaining data of D_1 are injected into the labeled artificially simulated abnormal data to obtain the detection dataset D_3, and finally the D_3 is input into D_2 for DAD.

3 Results

This study verified the superiority of the proposed AGDP-DBSCAN algorithm by comparing the clustering results and dataset evaluation indicators of different algorithms. At the same time, the time correlation, recall rate, and F1-Score of AD in D-IoT LTU nodes based on AGDP-DBSCAN algorithm were analyzed.

3.1 Experiment and analysis of parameter selection model for DBSCAN optimization

To test the performance of the AGDP-DBSCAN algorithm, artificial datasets Aggregation and Compound were selected in

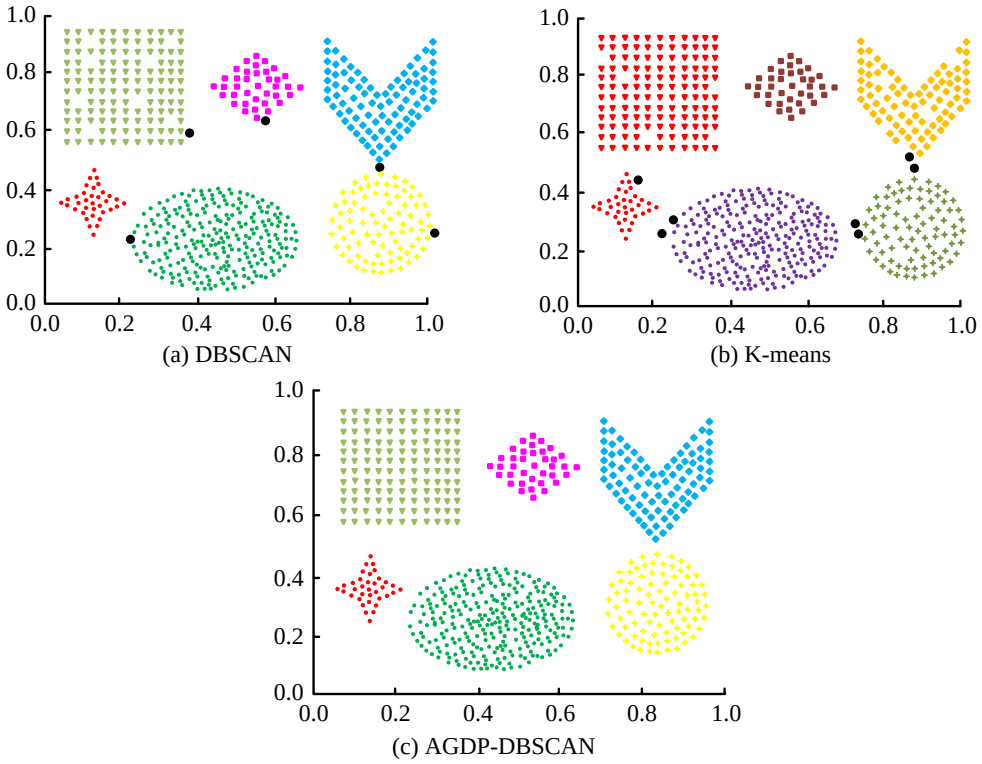


Figure 5: Clustering results on the Aggregation dataset

the experiment. The DBSCAN algorithm, partition-based K-means algorithm (K-means), and AGDP-DBSCAN algorithm were compared to verify the effectiveness of the AGDP-DBSCAN algorithm. The density distribution of data samples in Aggregation was relatively uniform, with connected data points between clusters. Figure 5 shows the clustering results of different algorithms on Aggregation.

In Figure 5 (a), DBSCAN performed better in clustering when the sample distribution density was uniform. In Figure 5 (b), there were a few clusters with connected samples, and the K-means algorithm performed well. In Figure 5 (c), AGDP-DBSCAN performed similarly to the first two clustering algorithms on samples with uniform distribution density, indicating that for samples with uniform distribution density in the dataset, the clustering effects of different clustering algorithms were basically the same. To further demonstrate the performance of AGDP-DBSCAN, experiments were conducted on datasets with uneven sample density distribution. The density distribution of data samples in Compound was uneven, and there were clusters with different densities. Figure 6 is the clustering results of various algorithms on Compound.

In Figure 6 (a), DBSCAN treated the data in the upper left corner of the dataset as noise points and did not achieve complete clustering of the data. In Figure 6 (b), the K-means also treated the data in the upper left corner as noise points and divided the cluster in the lower left corner into multiple small clusters. In Figure 6 (c), the AGDP-DBSCAN algorithm correctly identified two clusters and had good adaptability to datasets with uneven density. Due to the different densities of the two datasets, it directly affected the clustering

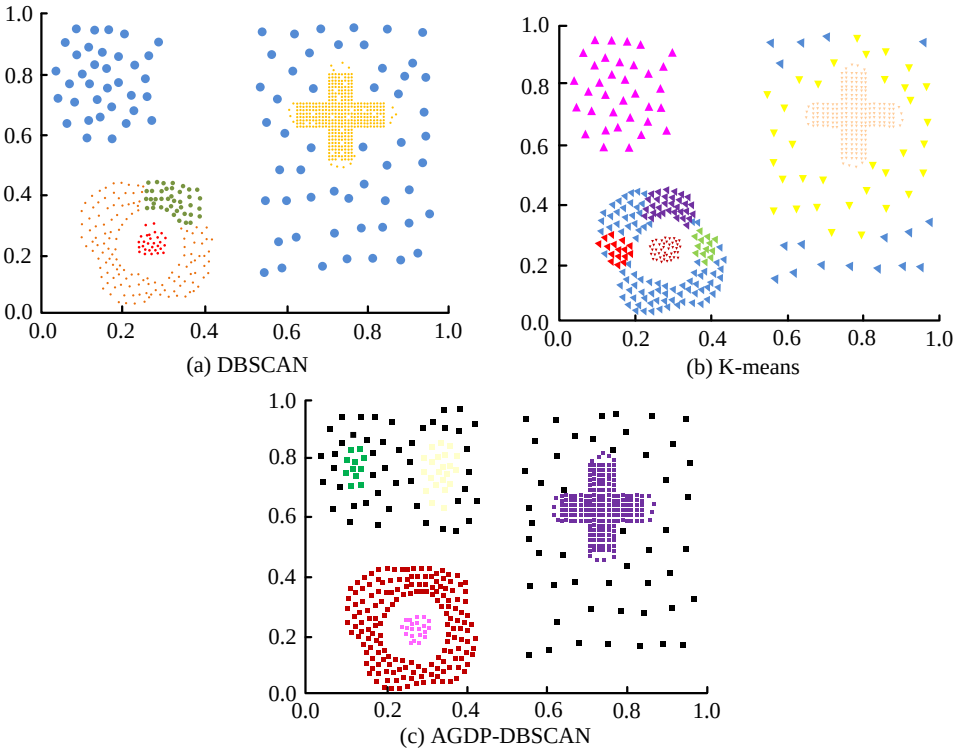
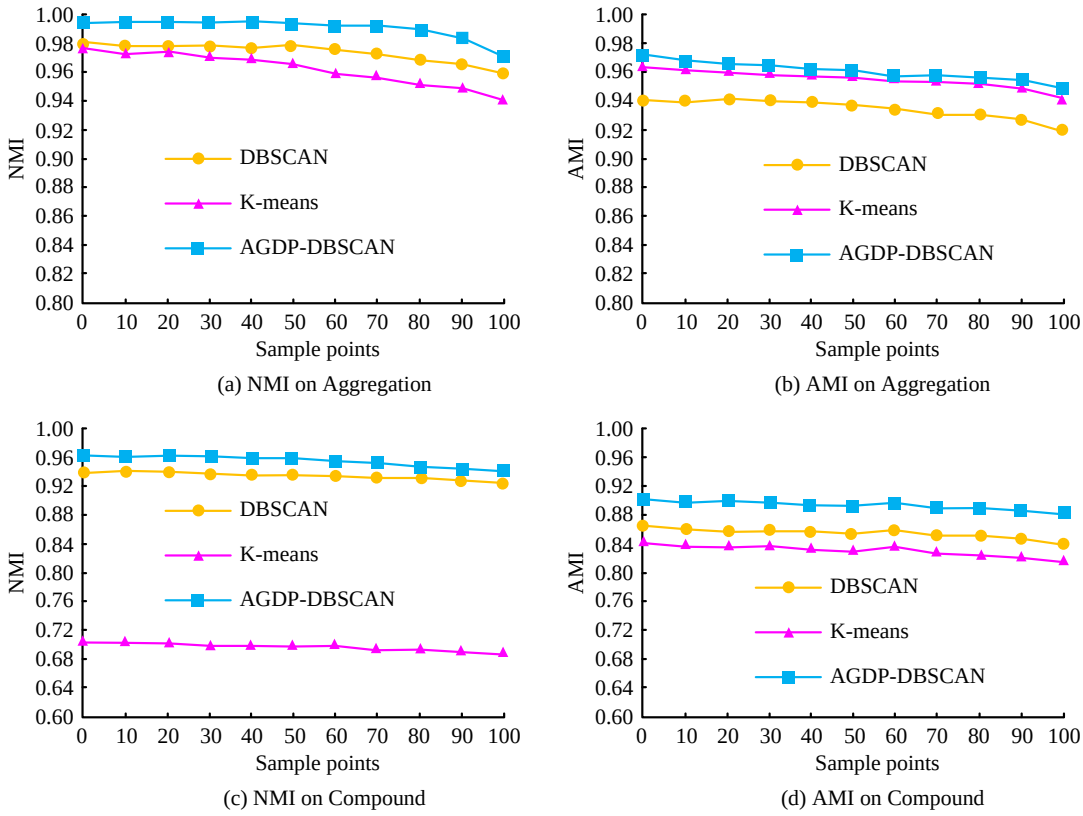


Figure 6: Clustering results on the compound dataset

performance of anomaly detection algorithms. Research on the use of similar duplicate record detection methods could optimize the DBSCAN algorithm and improve its clustering performance in different datasets. The experiment used Normalized Mutual Information (NMI) and Adjusted Mutual Information (AMI) to test the effectiveness of different clustering algorithms. Figure 7 shows the clustering performance of different clustering algorithms on artificial datasets.

In Figure 7 (a), the average NMI of DBSCAN, K-means, and AGDP-DBSCAN in the Aggregation were 0.974, 0.960, and 0.993, respectively. Compared to DBSCAN and K-means, the NMI of AGDP-DBSCAN increased by 2% and 3.4%, indicating that AGDP-DBSCAN had the best clustering effect. In Figure 7 (b), the average AMI of the three algorithms were 0.935, 0.948, and 0.955. The AMI of AGDP-DBSCAN was closest to 1, indicating that its clustering results were more consistent with the real clustering results. This was because DBSCAN randomly selected initial points and artificially sets NR and density thresholds, resulting in poor adaptability. K-means required an initial cluster center to determine the initial partition, continuously calculating and adjusting the cluster centers. As the amount of data increased, the clustering effect gradually deteriorated. In Figure 7 (c), the average NMI of DBSCAN, K-means, and AGDP-DBSCAN in Compound were 0.933, 0.695, and 0.952, respectively. The NMI of AGDP-DBSCAN has

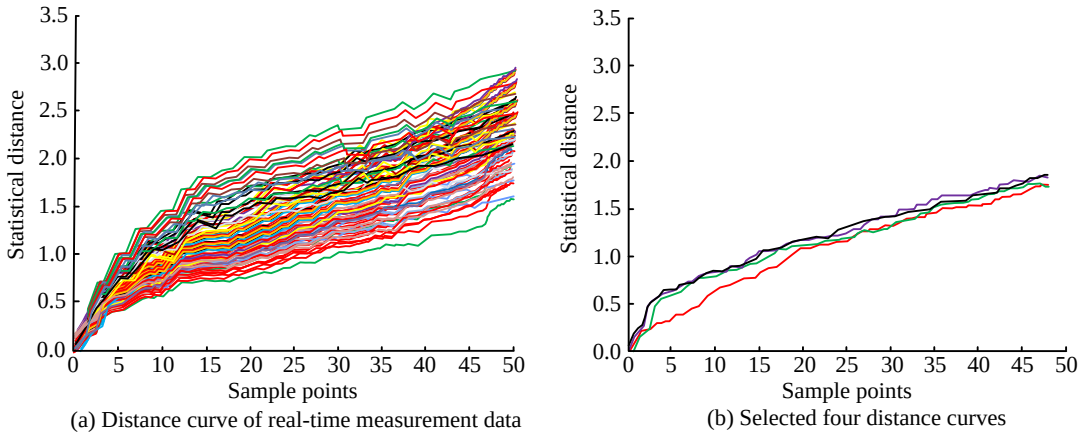
improved by 2% and 37% compared to the other two algorithms. In Figure 7 (b), the average AMI of each algorithm was 0.845, 0.828, and 0.898, respectively. AGDP-DBSCAN had the highest AMI, indicating its best clustering performance. This algorithm used CTSSMC to calculate distance



(a) NMI on Aggregation (b) AMI on Aggregation

(c) NMI on Compound (d) AMI on Compound

Figure 7: Clustering effect on the different dataset



(a) Distance curve of real-time measurement data (b) Selected four distance curves

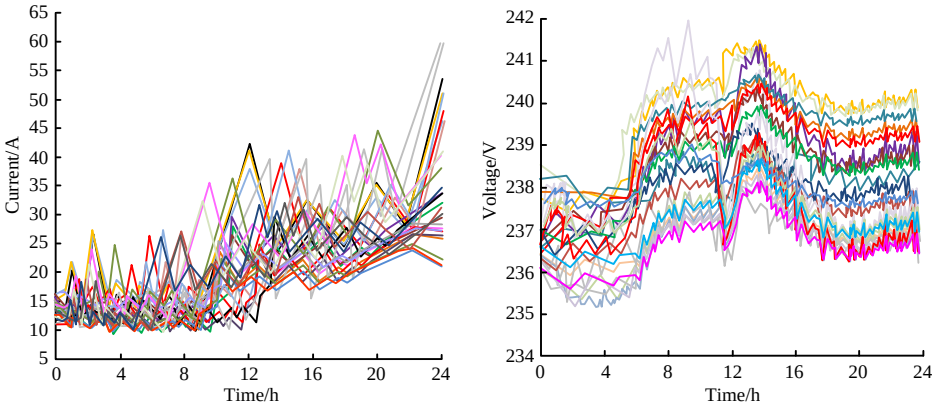
Figure 8: Distance curve chart

and adaptively generated NR and NDT based on LTU historical measurement data, which had good adaptability. Therefore, parameter optimization and dataset annotation were achieved through clustering analysis of datasets with different densities and artificial datasets, as well as adaptive neighborhood density methods.

Moreover, the results of the evaluation indicators also reflected the excellent clustering performance of the optimization algorithm on real datasets. Multiple experiments have demonstrated the robustness and superiority of the AGDP-DBSCAN algorithm.

3.2 Experimental and analysis of online monitoring of AD in D-IoT LTU Nodes

To test the performance of online monitoring



(a) Daily current timing chart (b) Daily Voltage timing chart

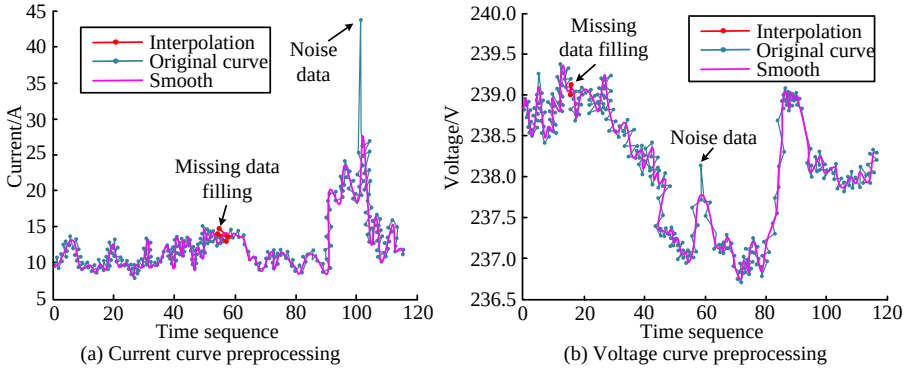
Figure 9: The variation pattern of daily LTU-MD nodes

of AD in D-IoT LTU nodes, real-time LTU-MD in a certain university's D-IoT was selected in the experiment. Figure 8 shows the distance curve drawn from the real-time LTU-MD.

In Figure 8 (a), in the real-time LTU-MD, the distance between normal and cluster data points was small and stable, showing a smooth front end of the distance curve. The proportion of AD points in the real-time entire LTU-MD was relatively low, and the distance from normal data points was relatively large, manifested as a steep distance from the end of the curve. In Figure 8 (b), at the position of sample point $n/2$, the location of the steep point was calculated by selecting four distance curves in ascending order of distance. The distance between steep points was used to calculate Eps, and $MinPts$ was calculated based on the amount of data points at the steep point. For testing the TC of D-IoT data, the daily variation pattern of LTU-MD nodes after normal operation of the system is shown in Figure 9.

Observing Figure 9 (a), when the system operated normally once, the daily current variation range was between 10A and 60A. The variation pattern of LTU node measured current over time showed similarity, and the same variation pattern was also observed for different periods, indicating that the current data measured by LTU had periodicity. In Figure 9 (b), after a normal operation of the system, the daily voltage changed over time, with a range of 235V to 242V. In addition, the daily voltage LTU-MD nodes exhibited significant similarity. Due to external signal interference, LTU-MD might experience data loss during measurement, transmission, and storage processes. This study used Z-Score standardization to standardize data and clean measurement data. Linear interpolation was used to fill in the cross-sectional data at a certain time, combined with Gaussian filtering for noise reduction and smoothing processing. The data preprocessing results of LTU nodes for one day are displayed in Figure 10.

In Figure 10 (a), the average values of the original current curve and the preprocessed current curve were 22.3A and 14.5A, respectively. Compared with the original current curve, the preprocessed current was reduced by about 35%. This was because the LTU current measurement data were interpolated and smoothed to remove the influence of noise, fill in missing data, and make the current curve relatively flat and smooth, ensuring the validity and



(a) Current curve preprocessing (b) Voltage curve preprocessing

Figure 10: Data pre-processing results

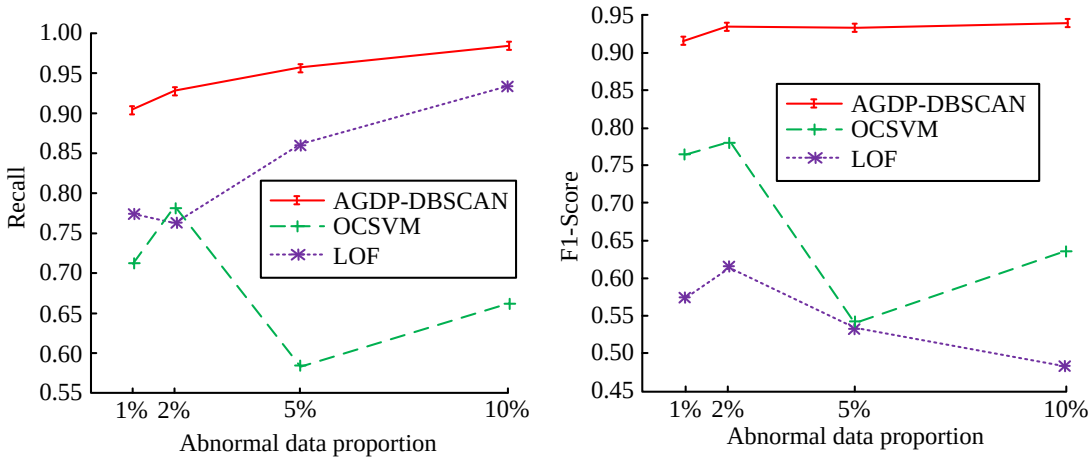
Table 1: Datasets with different proportions of anomalies

Datasets	Train sample days	Train sample samples	Test sample days	Test samples	Abnormal proportion
D3-I	25	8650	12	4030	1%
D3-II	25	8650	12	4030	2%
D3-III	25	8650	12	4030	5%
D3-IV	25	8650	12	4030	10%

completeness of the data. In Figure 10 (b), the average voltage before and after pretreatment was 238.8V and 237.5V, respectively. Compared to untreated, the voltage after pretreatment was reduced by about 0.5%. This was because noise had a relatively small impact on voltage. After filling in missing data and removing noise, the measured voltage data remained complete, making it easier for subsequent data analysis. To further test the superiority of AGDP-DBSCAN's D-IoT device online monitoring method, the experiment added artificially simulated LTU anomaly data with different proportions to the detection set D3, and obtained detection sets with different proportions of anomalies. Table 1 shows the dataset with different proportions of anomalies.

The experiment compared three DAD algorithms of AGDP-DBSCAN algorithm's D-IoT device online monitoring method (research method), One Class Support Vector Method (OCSVM), and Local Outlier Factor (LOF) algorithm. The comparison results of recall and F1-Score on datasets with different proportions of anomalies are shown in Figure 11.

In Figure 11 (a), the average recall rates of the research method, OCSVM, and LOF of the AGDP-DBSCAN algorithm were 0.94, 0.68, and 0.83, respectively. Compared with OCSVM algorithm and LOF, research method has improved recall rates by 38% and 13%, respectively, indicating a high detection rate for AD. The research method utilized time correlation and incorporated historical measurement data to generate distance curves during algorithm training, which could more precisely differentiate between normal data and AD. In Figure 11 (b), the mean F1-Score of the three algorithms was 0.92, 0.68, and 0.55, respectively. The F1-Score of oresearch method has improved by 35% and 67% compared to the other two algorithms. This was because the training dataset of OCSVM did not contain all kinds of AD, causing incomplete extracted AD features and a high false detection rate of the algorithm. The uncertainty of the distribution of current and voltage measurement data in LOF affected the



(a) Recall of different algorithms

(b) F1-Score of different algorithms

Figure 11: Recall rate and F1-Score of data with different proportions of anomalies

Table 2: Scalability verification results of different algorithms

Performance	Algorithms	Dataset size				
		2×104	4×104	6×104	8×104	10×104
Precision	LOF	0.64	0.70	0.75	0.83	0.85
	OCSVM	0.76	0.84	0.87	0.92	0.95
	K-Nearest Neighbor	0.73	0.84	0.86	0.91	0.94
	AGDP-DBSCAN	0.96	0.97	0.97	0.98	0.98
Running time/s	LOF	8s	33s	50s	64s	100s
	OCSVM	5s	9s	40s	58s	94s
	K-Nearest Neighbor	9s	34s	52s	67s	100s
	AGDP-DBSCAN	3s	5s	4s	5s	4s

calculation of LOF values, resulting in the highest false detection rate, especially as the proportion of anomalies increased, the false detection rate increased. The research method achieved high recall and F1-Score during DAD, and could accurately detect LTU-MD with abnormal changes.

To verify the scalability of the anomaly data detection algorithm, different algorithms were applied to different sizes of the same dataset, and their performance was compared. The results are shown in Table 2.

From Table 2, the running time of LOF algorithm and K-Nearest Neighbor algorithm varied greatly, and the time consumption increased with the increase of the dataset. The highest accuracy of the LOF algorithm was 0.85, while the highest accuracy of the K-Nearest Neighbor algorithm was 0.94. However, the proposed AGDP-DBSCAN algorithm performed the best in accuracy and operational efficiency, with the highest accuracy and operational efficiency of 0.98 and 3s, respectively. Therefore, it indicated that the online monitoring method of the AGDP-DBSCAN algorithm for D-IoT devices had good scalability in AD detection.

4 Discussion

As a technological goal for the development of a new type of power network, the construction of the power IoT and the D-IoT is a key factor for the stable development of the power grid. This study uses cluster analysis to partition normal data and AD through the architecture of the D-IoT system. Due to the time series nature of its data, a density-based clustering optimization algorithm is used to perform clustering analysis on the centralized data of the D-IoT. Through performance testing and analysis of different algorithms on the Aggregation dataset and Compound dataset, it can be concluded that the average NMI of the AGDP-DBSCAN algorithm is 0.993. The AMI of the AGDP-DBSCAN algorithm is closest to 1, indicating that the clustering results of the AGDP-DBSCAN algorithm are more consistent with the real clustering results. The maximum average AMI value of the AGDP-DBSCAN algorithm is 0.898, which is due to its use of CTSSMC to calculate distance and adapt to datasets of different densities, resulting in the best clustering performance on the dataset. Another method for monitoring LTU node data anomalies in distribution networks is designed and compared with the anomaly detection method proposed by Gan W et al. [11]. By utilizing the anomaly detection framework and utility-aware anomaly sequence rules, the algorithm's efficiency is evaluated, but its accuracy is 80%. However, the method proposed in this study combines the density of abnormal data sets, with a maximum accuracy of 98%. This is because the proposed algorithm has good adaptability to the proportion of AD in the dataset, which can reduce the influence of current and voltage, thereby improving the detection accuracy of AD and providing technical reference for the future development of smart grids.

5 Conclusion

To improve the reliable operation and equipment management level of the D-IoT system, this study proposed the D-IoT end device online monitoring method. This method analyzed the characteristics of D-IoT LTU-MD and proposed CTSSMC to calculate the distance matrix of the measurement sequence. Based on the TC of LTU-MD and the improved DBSCAN algorithm, this study designed a DAD method based on DBSCAN optimization algorithm. It improved the accuracy of DAD by fully mining LTU historical measurement data and adaptively generating the required global density parameters. The simulation results showed that the AGDP-DBSCAN could demonstrate good adaptability and clustering performance in datasets with uneven density distribution. In Compound dataset, compared to the traditional DBSCAN and K-means, AGDP-DBSCAN had a 2% and 37% increase in NMI. In detection sets with different proportions of anomalies, compared to the LOF algorithm, AGDP-DBSCAN's F1-Score and recall rates have increased by 67% and 13%, respectively. Therefore, AGDP-DBSCAN method

can improve dynamism, reduce clustering data dimensions, and be more stable. There are still shortcomings in identifying the sources of data anomalies, as well as the lack of deep-level monitoring functions for terminal nodes in the distribution logistics network architecture. In the data anomaly monitoring system, it is impossible to display the types and processing capabilities of data anomalies in real-time. Therefore, in the future, the sources of abnormal changes in LTU node data, the computational complexity of node data, and the influencing factors of monitoring models can be analyzed to achieve the intelligent application and development of the IoT in the distribution network.

Table 3: Summary table of relevant literature

Reference number	Research methods	Research results and significance
Lingda et al [4]	Propose a trust evaluation method based on information entropy, using an exponential distribution reputation model to estimate the direct trust value, and combining entropy theory to compensate for the inaccuracy of direct trust judgments.	This method can effectively resist switching attacks and collusion attacks.
Xu et al [5]	Propose a fault location algorithm based on genetic algorithm, improve the unified matrix algorithm, and combine genetic algorithm to achieve fault ranging.	This algorithm can effectively obtain the localization results of erroneous information.
Huang et al [6]	A cloud edge collaborative processing method based on attention mechanism is proposed, which combines edge computing and cloud computing to solve the timeliness of data analysis and calculation.	This algorithm can achieve efficient utilization of electrical energy resources and accelerate data iteration.
Lei et al [10]	Design an unsupervised time series data anomaly detection framework based on spectral analysis, which decomposes the time series using spectral analysis and combines neural networks to predict trend and seasonal sequences.	This method can effectively improve the detection rate of abnormal data.
Gan et al [11]	Introducing utility aware anomaly sequence rules into anomaly detection methods, combined with pruning strategies, to mine the upper bound of utility aware anomaly sequence rules for anomaly detection.	This method has good effectiveness and scalability.
Jeong et al [12]	Propose the probability of local outliers to identify seismic noise, utilize binary decision-making, and combine with moving windows to reduce useful signal energy loss.	This method has good feasibility in denoising applications.
Yang et al [13]	Propose anomaly detection and recovery of multiple regression data based on spatio-temporal correlation, select parameters using correlation analysis methods, and design a multiple regression model that integrates attention mechanism.	This method can achieve data recovery and improve the performance of anomaly data detection.

References

[1] J. Shi, Y. Gong, D. Guang, C. Zuo, X. Wu, L. Lu, G. Zhang, S. Li, R. Wang, and B. Yu, "Improved topography measurement with a high dynamic range using phase difference sensing technology," Optics Letters, vol. 48, no. 17, pp. 4657-4660, 2023. https://doi.org/10.1364/OL.495680

[2] G. Bandewad, K. P. Datta, B. W. Gawali, and S. N. Pawar, "Review on discrimination of hazardous gases by smart sensing technology," Artificial Intelligence and Applications, vol. 1, no. 2, pp. 86-97, 2023. https://doi.org/10.47852/bonviewAIA3202434

[3] S. Rostampour, N. Bagheri, B. Ghavami, Y. Bendavid, S. Kumari, H. Martin, and C. Camara, "Using a privacy-enhanced authentication process to secure IoT-based smart grid infrastructures," Journal

of Supercomputing, vol. 80, no. 2, pp. 1668-1693, 2024. https://doi.org/10.1007/s11227-023-05555-y

[4] L. Kong, F. Zhai, Y. Zhao, N. Qin, D. Li, and S. Cai, "Evaluation method of trust degree of distribution IoT terminal equipment based on information entropy," Journal of Physics: Conference Series, vol. 1754, no. 1, pp. 12108-12115, 2021. https://doi.org/10.1088/1742-6596/1754/1/012108

[5] Y. Xu, Q. Song, and Z. Chen, "Fault-tolerant ability of fault location on distribution internet of things in electricity using genetic algorithm," Journal of Physics: Conference Series, vol. 2033, no. 1, pp. 12077-12083, 2021. https://doi.org/10.1088/1742-6596/2033/1/012077

[6] X. Huang, P. Zhang, R. Feng, C. Huang, K. Zhang, and K. Jin, "Research on cloud-edge collaborative processing method of distribution internet of things based on attention-LSTM," 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture, vol. 2022, no. 3, pp. 52-56, 2021. https://doi.org/10.1145/3495018.3495029

[7] M. Gheisari, H. Hamidpour, Y. Liu, P. Saedi, A. Raza, A. Jalili, H. Rokhsati, and R. Amin, "Mining techniques for web mining: A survey," Artificial Intelligence and Applications, vol. 1, no. 1, pp. 3-10, 2023. https://doi.org/10.47852/bonviewAIA2202290

[8] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A review on outlier/anomaly detection in time series data," ACM Computing Surveys, vol. 54, no. 3, pp. 56-89, 2021. https://doi.org/10.1145/3444690

[9] W. Jeong, M. S. Almubarak, and C. Tsingas, "Seismic erratic noise attenuation using unsupervised anomaly detection," Geophysical Prospecting, vol. 69, no. 7, pp. 1473-1486, 2021. https://doi.org/10.1111/1365-2478.13123

[10] T. Lei, C. Gong, G. Chen, M. Ou, K. Yang, and J. Li, "A novel unsupervised framework for time series data anomaly detection via spectrum decomposition," Knowledge-based Systems, vol. 280, no. 11, pp. 111002-111015, 2023. https://doi.org/10.1016/j.knosys.2023.111002

[11] W. Gan, L. Chen, S. Wan, J. Chen, and C. Chen, "Anomaly rule detection in sequence data," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 12, pp. 12095-12108, 2023. https://doi.org/10.1109/TKDE.2021.3139086

[12] W. Jeong, M. S. Almubarak, and C. Tsingas, "Seismic erratic noise attenuation using unsupervised anomaly detection," Geophysical Prospecting, vol. 69, no. 7, pp. 1473-1486, 2021. https://doi.org/10.1111/1365-2478.13123

[13] L. Yang, S. Li, C. Zhu, A. Zhang, and Z. Liao, "Spatio-temporal correlation-based multiple regression for anomaly detection and recovery of unmanned aerial vehicle flight data," Advanced Engineering Informatics, vol. 60, no. 4, pp. 102440-102454, 2024. https://doi.org/10.1016/j.aei.2024.102440

[14] S. Singh, and K. Kumar, "A study of lean construction and visual management tools through cluster analysis," Ain Shams Engineering Journal, vol. 12, no. 1, pp. 1153-1162, 2021. https://doi.org/10.1016/j.asej.2020.04.019

[15] K. Z. Takahashi, "Molecular cluster analysis using local order parameters selected by machine learning," Physical Chemistry Chemical Physics, vol. 25, no. 1, pp. 658-672, 2023. https://doi.org/10.1039/d2cp90240k

[16] L. Chen, and A. K. Aklikokou, "Relating e-government development to government effectiveness and control of corruption: a cluster analysis," Journal of Chinese Governance, vol. 6, no. 1, pp. 155-173, 2021. https://doi.org/10.1080/23812346.2019.1698693

[17] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A review on outlier/anomaly detection in time series data," ACM Computing Surveys, vol. 54, no. 3, pp. 1-33, 2021. https://doi.org/10.48550/arXiv.2002.04236

[18] A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall, "The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances," Data Mining and Knowledge Discovery, vol. 35, no. 2, pp. 401-449, 2021. https://doi.org/10.1007/s10618-020-00727-3

[19] R. Balaji, R. B. Bapat, and S. Goel, "Generalized euclidean distance matrices," Linear and Multilinear Algebra, vol. 70, no. 21, pp. 6908-6929, 2022. https://doi.org/10.1080/03081087.2021.1972083

[20] N. Gholizadeh, H. Saadatfar, and N. Hanafi, "K-DBSCAN: An improved DBSCAN algorithm for big data," The Journal of Supercomputing, vol. 77, no. 6, pp. 6214-6235, 2021. https://doi.org/10.1007/s11227-020-03524-3

[21] K. Sabor, D. Jougnot, R. Guerin, B. Steck, J. M. Henault, L. Apffel, and D. Vautrin, "A data mining approach for improved interpretation of ERT inverted sections using the DBSCAN clustering algorithm," Geophysical Journal International, vol. 225, no. 2, pp. 1304-1318, 2021. https://doi.org/10.1093/gji/ggab023