J Inf Process Syst, Vol.21, No.3, pp.318~327, June 2025
https://doi.org/10.3745/JIPS.03.0205 ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Anomaly Detection of Power Load Based on Robust PCA and Improved K-Means Clustering Algorithm

Xinjian Zhao[1,*], Weiwei Miao[1], Song Zhang[1], Youjun Hu[2], and Shi Chen[1]

Abstract

The interaction of power load information provides reliable data support for accessing user-side electrical energy storage devices and distributed renewable energy sources. However, owing to the large volume of interactive information and the numerous security threats faced during the interaction, anomaly detection has become one of the most challenging problems in smart grids. To address this issue, an anomaly detection method was developed that consists of three stages. First, feature extraction is performed based on the power load information. Then, a robust principal component analysis method is used for the preliminary classification of the extracted features. Finally, an improved K-means clustering algorithm is employed to refine the classification results into completely non-overlapping groups and detect anomalies from the classified data. Experimental results demonstrate that the proposed method can effectively and accurately detect anomalies from power load data.

Keywords

Anomaly Detection, K-Means, Feature Extraction, Power Load, Robust Principal Component Analysis

1. Introduction

With the gradual increase of the construction of new power systems, an increasing number of user-side electrical energy storage devices and distributed renewable energy sources have been integrated into power systems through third-party aggregation platforms [1]. These multisource, heterogeneous, distributed resources generate massive amounts of power load information [2]. Efficient anomaly detection for power load interaction information is essential for providing reliable data support for the scheduling and control of new power systems.

In recent years, researchers have proposed various solutions to the problem of anomaly detection for power load data. Regarding specific implementation methods, the research results can be categorized into two types: traditional machine-learning-based and deep-learning-based anomaly detection. Anomaly detection algorithms based on deep learning have better detection abilities, but their performance often depends on the size of the dataset, and their computational overhead is large [3–5]. In practice, the power load has certain regional characteristics and imbalances, which makes the dataset always have similar

Manuscript received June 11, 2024; first revision October 14, 2024; accepted December 27, 2024.

* **Corresponding Author:** Xinjian Zhao (zhxj198708@126.com)
[1] Information & Telecommunication Branch, State Grid Jiangsu Electric Power Co. Ltd., Nanjing, China (mww196801@sgcc.cn, zhs199608@sgcc.cn, chs199509@sgcc.cn)
[2] Nari Information & Communication Technology Co. Ltd., Nanjing, China (hyj198111@sgepri.sgcc.cn)

characteristics to a small data volume. However, it is usually necessary to perform simple and fast anomaly detection at the edge of the network. Therefore, in comparison, traditional machine-learning-based anomaly detection methods can detect anomalous data on a limited dataset with a lower overhead, which is more suitable for the existing power grid environment. In anomaly detection for power data, the main methods based on machine learning include isolation forest [6], one-class support vector machine (SVM) [7], local outlier factor (LOF) [8], density-based spatial clustering of applications with noise (DBSCAN) [9], local correlation integral (LOCI) [10], connectivity-based outlier factor (COF) [11], and histogram-based outlier score (HBOS) [12]. Existing machine-learning-based anomaly detection methods typically use different combinations of algorithms to obtain anomaly detection results by classifying or clustering the original power load dataset. However, the actual accessible power load data often have uncertainties in such issues as data integrity and synchronization, and these uncertainties affect the data detection results significantly.

In this study, an anomaly detection scheme for power loads was developed based on robust principal component analysis (PCA) and clustering algorithms. First, user power load data are collected and analyzed to extract features. Then, a robust PCA algorithm is used for preliminary classification, categorizing the power load data into suspected abnormal and suspected normal groups. Based on this classification, an improved clustering algorithm is used to refine further and extract the results.

The main contributions of this study are as follows.

- A phased anomaly detection scheme for power load data was developed that realizes anomaly detection using a robust PCA algorithm and an improved K-means algorithm.
- The proposed scheme was tested with public datasets, and the results show that the scheme has a better detection performance than similar mainstream schemes.

2. Phased Anomaly Detection Method for Power Load Data

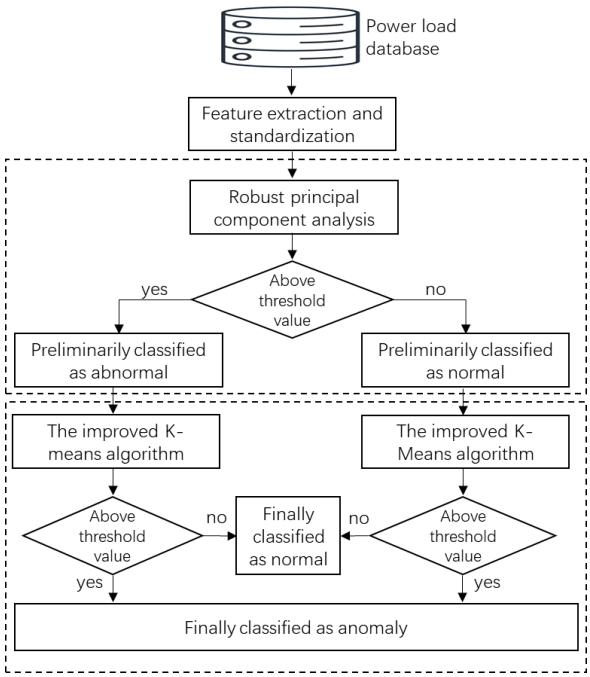


Fig. 1. Framework of the proposed scheme.

A phased anomaly detection method for power load data was developed, and its framework is shown in Fig. 1. First, intelligent terminals obtain the user power load data and extract features from these data, including Kullback–Leibler (KL) score, flat points, Canberra distance, and crossover points. Second, the robustness-enhanced PCA algorithm is used to analyze the features of power load data, preliminarily dividing them into abnormal and normal groups. Finally, the improved K-means algorithm is used to remove outliers from the classification results obtained from the preliminary classification in the previous stage and to obtain a classification of the power load data with clear boundaries.

2.1 Feature Extraction

In this study, the monthly power load was used to represent the electricity usage of each user, which is defined as the power load of the user every 30 days. Feature extraction is the process of extracting the intrinsic features of a dataset, which reduces the dimensionality of the data and saves computational resources. The extracted features include the average load (the average power load of each platform during this period), variance (the variance of the power load of each platform during this period), horizontal displacement difference (the maximum difference in average load between days), variance difference (the maximum difference in variances between months), KL score (the maximum difference in KL divergence between consecutive months), flat spot (the length of the largest flat interval within each month), and Canberra distance (a numerical measure of distance between points in vector space).

Two of the most essential features are the Canberra distance and flat point. The Canberra distance is considered a weighted version of the Manhattan distance and is calculated using Eq. (1):

$$\mathrm{CD} = \sum_{i}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}, \tag{1}$$

where x_i and y_i are different data points in the real value.

2.2 Preliminary Classification based on Robust Principal Component Analysis

A robust PCA algorithm [13] is used to preliminarily classify the users reflected by the power load information into suspected abnormal and normal groups with the following workflow.

Step 1: The feature matrix of the power load data, $Y_{n\times p}$, is input, where n is the number of users, and p is the number of features for each user. The algorithm first normalizes the given matrix and then transforms it into an affine space based on singular value decomposition.

Step 2: The anomaly index of each data point $y_i (i = 1,2,\ldots,n)$ is calculated, as shown in Eq. (2):

$$out_A(y_i) = max_{v\in B} \frac{|y_i^T v - a_{MCD}(y_i^T v)|}{s_{MCD}}, \tag{2}$$

where B contains all nonzero vectors, and a_{MCD} and s_{MCD} are the mean and standard deviation calculated using the minimum covariance determinant (MCD) method, respectively.

Step 3: The covariance matrix $S_0 = P_0 L_0 P_0^T$ is calculated, where $L_0 = diag\left(\tilde{l}_1,\ldots,\tilde{l}_r\right), r < r_1$ is the feature value matrix, and P_0 is an orthogonal matrix of r_1 rows and r columns. The data points are projected on the subspace spanned by the first k_0 eigenvectors of S_0, as shown in Eq. (3):

$$Y_{n\times k_0}^* = \left(Y_{n\times k_0} - 1_n \hat{\mu}_1^T\right) p_{r_1\times k_0}, \tag{3}$$

where $p_{r_1\times k_0}$ consists of the first k_0 columns of P_0.

Step 4: The MCD estimator is used to estimate the scatter matrix of data points in of $Y^*_{n\times k0}$ robustly. The robust covariance matrix $S = P_{p\times k}L_{k\times k}P^T_{p\times k}$, so the robust principal component matrix can be rewritten as $T_{n\times k} = (Y_{n\times p} - 1_n\hat{\mu}^T)P_{p\times k}$.

Step 5: Orthogonal distance is defined as the distance between each observation and its projection onto the new subspace, as shown in Eq. (4):

$$OD_i = \left||y_i - \hat{y}_i\right||, \tag{4}$$

where y_i is the i-th data, and $\hat{y}_i$ is the projection data point on the k-dimensional subspace. The score distance is calculated using Eq. (5):

$$SD_i = \sqrt{t_i^T L^{-1} t_i} = \sqrt{\sum_{j=1}^{k}(\frac{t_{ij}^2}{l_j})}, \tag{5}$$

where l_j is the set of eigenvalues, and k is the number of principal components.

Step 6: Two threshold values are calculated for the sum of the dataset to separate normal observations from abnormal observations. These are calculated separately using Eqs. (6) and (7):

$$c_{OD} = (\hat{\mu} + \hat{\sigma}_{Y0.975})^{3/2}, \tag{6}$$

$$c_{SD} = \sqrt{X^2_{k,0.975}}, \tag{7}$$

where $\hat{\mu}$ is the average value, $\hat{\sigma}$ is the variance for the given data point, 0.975 means 97.5% quantile of the Gaussian distribution, and X_k^2 is the square of the Mahalanobis distance. Based on the threshold values, the power load data are preliminarily classified into two major groups.

2.3 Anomaly Detection based on the Improved K-Means Algorithm

Because the traditional K-means algorithm initializes cluster centers randomly [14], the results of the algorithm depend highly on the initial selection of cluster centers, and the final result is likely to have a high error. The method of selecting clustering centers based on the relative distance and density between data points makes a better separation of outliers possible. Based on this, an improved K-means clustering algorithm was developed to detect anomalies and generate distinct load types with apparent boundaries to address the slight overlap between the two categories of load data identified in the previous stage. The specific flow of the algorithm is shown in Algorithm 1.

First, the relative distances and densities of the data points are calculated. The $D(x_i)$ values are ranked in descending order, and the data point c_i with the maximum density is selected as the initial clustering center. The thresholds are then calculated separately for different cases based on the range of the center node density values. The clustering process of the nodes is completed by iteration to obtain the optimal number of clusters K. In the second stage, an outlier factor o_i is computed for each observation in the K clusters, which depends on its distance from the clustering center. The value of o_i ranges from 0 to 1, and the value of μ is 0.95 in this algorithm, which means the data with $o_i \geq 0.95$ are considered abnormal.

This algorithm is used to remove outliers for each of the two categories of users derived in Section 2.1, the data exceeding the thresholds among the suspected normal and suspected abnormal users are finally categorized as abnormal data, and all the rest are finally categorized as normal data.

Algorithm 1. Anomaly detection based on the improved K-means clustering algorithm

1. Calculate the relative distance d_{ij} between data points x_i and x_j, then generate a distance matrix M.
2. For each data point x_i, calculate its density $D(x_i) = \frac{1}{\sum_{x_j \in G_t(x_i)} d_{ij}}$, where $G_t(x_i)$ is a set containing the t closest elements to x_i, and calculate the average density D_{ave}.
3. Select the data points with the largest $D(x_i)$ values as the initial clustering centers c_i.
4. If $D(c_i) \geq D_{ave}$,
   set the threshold for the first clustering center, which is selected as $\delta_i = \frac{n}{K} + \frac{D_i - D}{D_i + D} \times \frac{n}{K}, i = 1$.
5. Remove S_1 from the dataset and divide the remaining equally into K−1 clusters, where S_i is the neighboring node set of c_i, and the threshold is $\delta_i = \frac{n - \sum_{j=1}^{i-1} |S_j|}{K - i + 1} + \frac{D_i - D}{D_i + D} \times \frac{n - \sum_{j=1}^{i-1} S_j}{K - i + 1}, i = 2,3,\ldots,K$.
6. If $D(c_i) < D_{ave}$, set the threshold as $\delta_i = \begin{cases} \frac{n}{K} + \frac{D - D_i}{D + D_i} \times \frac{n}{K}, i = 1 \\ \frac{n - \sum_{j=1}^{i-1} |S_j|}{K - i + 1} - \frac{D - D_i}{D + D_i} \times \frac{n - \sum_{j=1}^{i-1} |S_j|}{K - i + 1}, i = 2,3,\ldots,K \end{cases}$.
7. Remove the corresponding data points of S_i continuously until the initial cluster centers are selected and apply the improved K-means algorithm based on the selected cluster centers until it converges to the optimal K value.
8. Calculate the maximum distance of all data points to their cluster centers, $d_{max} = \max\{\|x_i - c_{pi}\|\}$, where c_{pi} is the cluster centers.
9. Calculate the anomaly factor for each of the data $o_i = \frac{||x_i - c_{pi}||}{d_{max}}$; the data points for which $o_i \geq \mu$ are considered anomalous and separated from the rest of the data, where μ is the threshold of the anomaly factor.
10. Repeat the process until the clusters no longer overlap.

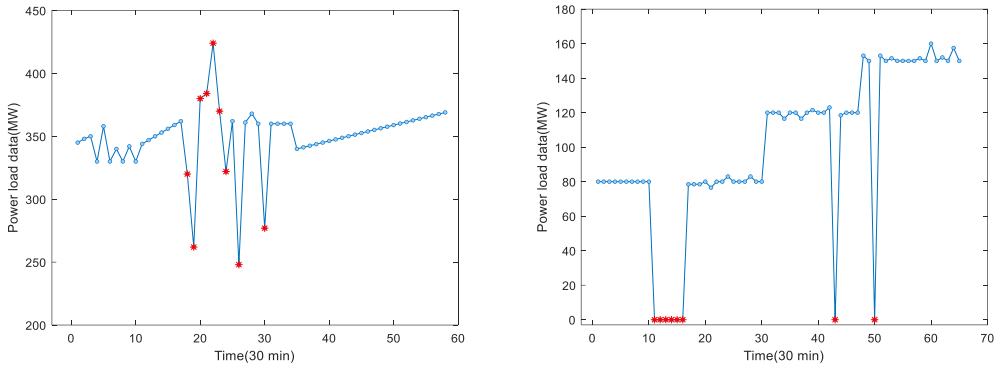
3. Experiment and Result Analysis

The experimental environment was based on the Ubuntu environment using Python. The CPU was Intel Xeon Silver 4208R with 16 GB RAM.

3.1 Experimental Dataset

This experiment used power consumption load data from the Los Alamos Public Utility Department in New Mexico, USA [15]. The data were collected using Landis+Gyr smart meter devices from 1,757 households in North Mesa, Los Alamos, NM, USA. The sampling rate was one observation every 15 minutes. For most customers, the data span approximately 6 years, from July 30, 2013 to December 30, 2019.

Affected by extreme weather, production, life, and equipment failure, and other uncertainties, the power load is characterized by randomness, volatility, and sudden changes. Power load data often show abnormal jumps, manifesting as fluctuation anomalies and extreme value anomalies. The fluctuation anomaly load curve shows a large number of burrs and frequent fluctuations in a short period compared with the normal load fluctuation law, which has a significant jump, as shown in Fig. 2(a). The extreme anomaly load curve is manifested as a load data extreme value abnormality in a certain period (duration is usually minutes) of load spike, valley, or significant peak–valley difference, destroying the curve similarity and periodicity, as shown in Fig. 2(b).



(a) (b)

Fig. 2. Schematic diagram of anomalies of power load data: (a) fluctuation anomalies and (b) extreme value anomalies.

3.2 Evaluation Metrics

Owing to the abnormal detection of power load information, the interaction is a binary classification problem. For binary classification problems, the performance of a classification model can be evaluated using a confusion matrix, as shown in Table 1. The rows in the table represent the predicted classes, whereas the columns represent the actual classes. Based on the confusion matrix, one can obtain evaluation metrics, including accuracy (ACC), recall rate (RR), false positive rate (FPR), false negative rate (FNR), precision, F1-score, and Bayesian detection rate (BDR).

Table 1. Confusion matrix in anomaly detection

		Predicted label	
		Anomaly	**Normal**
True label	Anomaly	TP (True Positive)	FP (False Positive)
	Normal	FN (False Negative)	TN (True Negative)

3.3 Results and Discussion

Experiments were conducted using the proposed approach, and the results were analyzed and compared with some commonly used anomaly detection methods. In the performance metric scores shown in Table 2, the first six performance metrics are directly derived from the confusion matrix, and the last metric, BDR, is calculated using prior knowledge of fraud probability. The proposed approach achieved an accuracy of 91% and a recall rate of 81%, indicating that it could detect the types of user and most of the actual abnormalities.

ACC and recall are commonly used metrics in almost all abnormal electricity usage detection systems. However, these two metrics cannot be used as decisive indicators of scheme performance. The drawback of ACC is that, when the proportions of different categories of samples are highly unbalanced, the larger proportion category often becomes the main factor affecting ACC. For example, when defective samples account for 99% of the total, the classifier predicts that all samples are defective, resulting in an accuracy of 99%. Therefore, it is necessary to verify the performance of the scheme further by calculating and comparing the values of FNR and FPR. Under the same conditions, the FPR value of the proposed scheme is the lowest, and the FNR value is only slightly higher than that of one-class SVM. This indicates that

the proposed scheme can only classify a small portion of the normal electricity consumption as abnormal. Another critical performance metric used for qualitative analysis of anomaly detection methods is the F1-score. It helps strike a proper balance between precision and recall because these two metrics are contradictory. The higher the value, the better the predictability of the model, and vice versa. The F1-score obtained by the proposed method is 75%, the highest among the compared techniques. Because a reliable anomaly detection model has a high BDR value, the BDR score is used for the evaluation. In this study, the fraud probability was 16%, and the BDR value of the scheme was 63%, which were much higher than those of the other anomaly detection algorithms. Compared with the deep-learning-based detection model [5], the proposed method still has some advantages for most metrics because of the limited dataset volume.

Table 2. Performance metrics of proposed approach compared with commonly used anomaly detection algorithms

	ACC	RR	FPR	FNR	PR	F1-score	BDR
Proposed method	**0.91**	**0.81**	**0.07**	**0.16**	0.7	**0.75**	0.63
Isolation forest [6]	0.86	0.73	0.12	0.31	0.62	0.74	0.55
One-class SVM [7]	0.84	**0.81**	0.32	**0.16**	0.45	0.62	0.41
LOF [8]	0.81	0.67	0.22	0.32	0.49	0.59	0.48
DBSCAN [9]	0.77	0.64	0.26	0.47	0.43	0.58	0.43
LCI [10]	0.75	0.62	0.25	0.52	0.45	0.51	0.32
COF [11]	0.69	0.57	0.39	0.5	0.36	0.48	0.26
HBOS [12]	0.55	0.45	0.44	0.61	0.39	0.37	0.21
Bi-LSTM-AE [5]	0.89	0.8	0.09	0.18	**0.72**	0.73	**0.65**

The bold font indicates the best performance in each test.

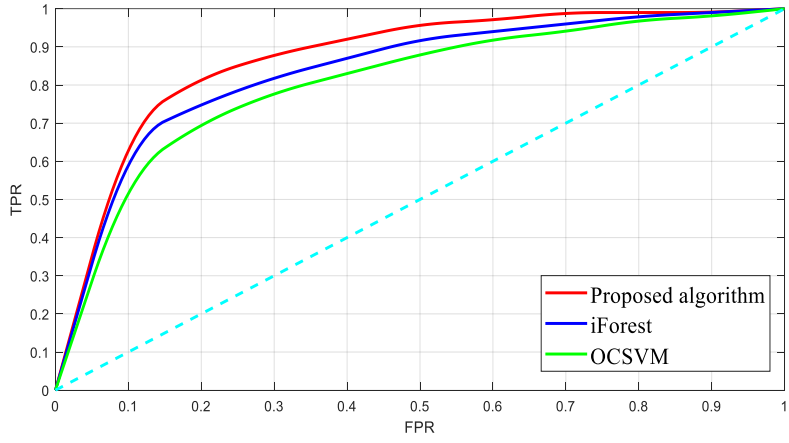


Fig. 3. ROC curves for different schemes.

Fig. 3 shows the receiver operating characteristic (ROC) curves of the proposed scheme, isolation forest, and one-class SVM. The ROC curve of the proposed scheme is closer to the upper left corner. The area under the curve (AUC) scores of these three schemes are provided to give a more comprehensive view of the performance of the method. The AUC values of the three algorithms are 0.81, 0.75, and 0.72, respectively. This implies that the detection effect of the scheme described in this section is better than that of the isolation forest and one-class SVM.

4. Conclusion

A phased anomaly detection method was developed to improve the efficiency of anomaly detection in power load data. First, the most important features of the power load were extracted from the data. Then, the power load was preliminarily classified into normal and abnormal groups based on the robustness-enhanced PCA algorithm. Based on the preliminary classification, an improved K-means clustering algorithm was used to obtain the final classification results. The experimental results show that the proposed method can improve the shortcomings of a single machine-learning method in terms of anomaly detection performance to some extent and provide more reliable anomaly detection results.

However, limited by the K-means algorithm itself, the clustering process cannot reflect the temporal characteristics of the data, which can affect the detection performance. Therefore, in future work, more attention should be paid to the temporal characteristics of power load data to improve the anomaly detection performance.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This work was supported by the Science and Technology Project of State Grid Jiangsu Electric Power Company Ltd. (Grant No. J2023124).

References

[1] S. Abbas, I. Bouazzi, S. Ojo, G. A. Sampedro, A. S. Almadhor, A. Al Hejaili, and Z. Stolicna, "Improving Smart Grids Security: an active learning approach for smart grid-based energy theft detection," *IEEE Access*, vol. 12, pp. 1706-1717, 2024. https://doi.org/10.1109/ACCESS.2023.3346327

[2] G. Bedi, G. K. Venayagamoorthy, R. Singh, R. R. Brooks, and K. C. Wang, "Review of Internet of Things (IoT) in electric power and energy systems," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 847-870, 2018. https://doi.org/10.1109/JIOT.2018.2802704

[3] A. Al-Khateeb, N. F. Kamal, H. Alnuweiri, S. Bayhan, and M. B. Shadmand, "Scalable light-weight anomaly detection for data of individual smart meters," in *Proceedings of 2024 4th International Conference on Smart Grid and Renewable Energy (SGRE)*, Doha, Qatar, 2024, pp. 1-6. https://doi.org/10.1109/SGRE59715.2024.10429010

[4] X. Zhang, C. Zheng, X. Wu, T. Wang, H. Gao, and J. Guo, "Anomaly detection method for interactive data of third-party load aggregation platform based on multidimensional feature information fusion," in *Proceedings of 2022 IEEE 22nd International Conference on Communication Technology (ICCT)*, Nanjing, China, 2022, pp. 1893-1897. https://doi.org/10.1109/ICCT56141.2022.10072826

[5] F. Zhu, M. Li, Y. Liu, J. Liu, Z. Wang, Y. Chen, and X. Zhang, "Anomaly detection in commercial load data using bidirectional LSTM and autoencoders," in *Proceedings of 2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, Changchun, China, 2024, pp. 902-904.

https://doi.org/10.1109/EEBDA60612.2024.10485863

[6] L. Zhang and L. Liu, "Data anomaly detection based on isolation forest algorithm," in *Proceedings* of *2022 International Conference on Computation, Big-Data and Engineering (ICCBE)*, Yunlin, Taiwan, 2022, pp. 87-89. https://doi.org/10.1109/ICCBE56101.2022.9888169

[7] Y. Ji and H. Lee, "Event-based anomaly detection using a one-class SVM for a hybrid electric vehicle," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6032-6043, 2022. https://doi.org/10.1109/TVT.2022.3165526

[8] S. Yu, X. Li, L. Zhao, and J. Wang, "Hyperspectral anomaly detection based on low-rank representation using local outlier factor," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 7, pp. 1279-1283, 2021. https://doi.org/10.1109/LGRS.2020.2994745

[9] D. Deng, "Research on anomaly detection method based on DBSCAN clustering algorithm," in *Proceedings of 2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, Shenyang, China, 2020, pp. 439-442. https://doi.org/10.1109/ISCTT51595.2020.00083

[10] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: fast outlier detection using the local correlation integral," in *Proceedings 19th International Conference on Data Engineering (Cat. No. 03CH37405)*, Bangalore, India, 2003, pp. 315-326. https://doi.org/10.1109/ICDE.2003.1260802

[11] K. Feasel, "Connectivity-based outlier factor (COF)," in *Finding Ghosts in Your Data.* Berkeley, CA: Apress, 2022, pp. 185-201. https://doi.org/10.1007/978-1-4842-8870-2_10

[12] F. Pei, Z. Miao, and J. Wang, "Dynamic SLAM system using histogram-based outlier score to improve anomaly detection," in *Proceedings of 2021 China Automation Congress (CAC)*, Beijing, China, 2021, pp. 4909-4913. https://doi.org/10.1109/CAC53003.2021.9728124

[13] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden, "ROBPCA: a new approach to robust principal component analysis," *Technometrics*, vol. 47, no. 1, pp. 64-79, 2005. https://doi.org/10.1198/004017004000000563

[14] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-means and alternative clustering methods in modern power systems," *IEEE Access*, vol. 11, pp. 119596-119633, 2023. https://doi.org/10.1109/ACCESS.2023.3327640

[15] V. Souza, T. Estrada, A. Bashir, and A. Mueen, "LADPU smart meter data," 2020 [Online]. Available: https://datadryad.org/dataset/doi:10.5061/dryad.m0cfxpp2c.

Xinjian Zhao **https://orcid.org/0009-0005-2453-6752**

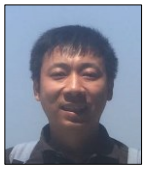
He received the M.S. degree in computer science and technology from Peking University, in 2010. He is currently a senior engineer with the Information and Telecommunication Branch, State Grid Jiangsu Electric Power Co. Ltd. His current research interests include big data and network security research.

Weiwei Miao **https://orcid.org/0009-0006-9071-6142**

He received the M.S. degree in automation theory and application from Southeast University, Nanjing, China, in 1992. He is currently a principal researcher with the Information and Telecommunication Branch, State Grid Jiangsu Electric Power Co. Ltd. His research interests include power communication systems, wireless access networks, and communication network management.

Song Zhang **https://orcid.org/0009-0002-1750-8153**

He received B.S. and M.S. degree in computer science from Southeast University, China, in 2018 and 2021, respectively. He is currently an engineer with the Information and Telecommunication Branch, Jiangsu Electric Power Co. Ltd. His current research interests include power informatization and network security.



Youjun Hu **https://orcid.org/0009-0007-4607-5912**

He received M.S. degree in computer science from Nanjing University, in 2009. He is currently an engineer in Nari Information & Communication Technology Co. Ltd. His current research interests include power informatization and network security.

Shi Chen **https://orcid.org/0009-0002-8389-7433**

He received the M.S. degree in electronic and communication engineering from Nanjing University, in 2020. He is currently an assistant engineer with the Information and Telecommunication Branch, Jiangsu Electric Power Co. Ltd. His current research interests include power informatization and network security.