Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9890424, 1 page
https://doi.org/10.1155/2023/9890424

Retraction

Retracted: Detection of Power Data Outliers Using Density Peaks Clustering Algorithm Based on *K*-Nearest Neighbors

Wireless Communications and Mobile Computing

Received 11 July 2023; Accepted 11 July 2023; Published 12 July 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Q. Li, L. Chen, and Y. Wang, "Detection of Power Data Outliers Using Density Peaks Clustering Algorithm Based on *K*-Nearest Neighbors," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 2203137, 7 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2203137, 7 pages
https://doi.org/10.1155/2022/2203137

Research Article

Detection of Power Data Outliers Using Density Peaks Clustering Algorithm Based on K-Nearest Neighbors

Qingpeng Li,[1] Lei Chen ,[2] and Yuhan Wang[1]

[1]*Nanchang Power Supply Company, State Grid Jiangxi Electric Power Company, Nanchang 330200, China*
[2]*School of Information Engineering, Nanchang Institute of Technology, Nanchang 330099, China*

Correspondence should be addressed to Lei Chen; 15161875401@163.com

Received 28 October 2021; Accepted 9 July 2022; Published 27 July 2022

Academic Editor: Chi-Hua Chen

As an important research branch in data mining, outlier detection has been widely used in equipment operation monitoring and system operation control. Power data outlier detection is playing an increasingly vital role in power systems. Density peak clustering (DPC) is a simple and efficient density-based clustering algorithm with a good application prospect. Nevertheless, the clustering results by the DPC algorithm can be greatly influenced by the cutoff distance, indicating that the results are highly sensitive to this parameter. To address the shortcomings of the DPC algorithm and take the characteristics of power data into consideration, we propose a DPC algorithm based on K-nearest neighbors for the detection of power data outliers. The proposed DPC algorithm introduces the idea of K-nearest neighbors and uses a unified definition of local density. In the DPC algorithm, only one parameter (K) needs to be determined, thus eliminating the influence of cutoff distance on the clustering result of the algorithm. The experimental results showed that the proposed algorithm can achieve accurate detection of power data outliers and has broad application prospects.

1. Introduction

With the construction and development of smart grids, power enterprises have accumulated a large amount of power data in various forms, from different sources, and with complex structures. With the rapid progression of artificial intelligence [1], effective power data mining can be achieved. Not only has it promoted the transformation of the power grid from the traditional physical model-based business model to a data-based one [2, 3] but also it has empowered the power grid enterprises to embrace the new digital economy.

Anomalies in the power industry mainly include loss of primary attributes of power data, inconsistent statistical caliber of power data, abnormal power consumption behaviors of customers, and power equipment failure. Since anomaly detection can detect abnormal power consumption behaviors in power operations, it is widely used in the power system [4]. In the early days, the method used to detect power consumption anomalies was relatively simple. In most cases, technicians should go onsite to diagnose the problem. For this method, technicians should have a great amount of experience. It would lead to the waste of human and material resources and the detected power fault time lagging behind the actual fault time so that the operation status of the power grid cannot be maintained in real time. Additionally, the results obtained using this method were not highly accurate and were highly correlated with the experience of technicians. When data-driven methods are used to detect power data outliers, the reliance on human resources can be reduced and the power grid operation status can be monitored in real time. Besides, when the power grid is in an alert or abnormal state, alarms can be issued or the power supply can be cut off to prevent the spread of faults and reduce the economic loss of the power grid. Given this, the data-driven methods will be the new trend and promising direction of power anomaly detection [5, 6].

Currently, algorithms for the detection of power data outliers include algorithms based on the probability statistical model, classification-based algorithms, distance-based algorithms, and clustering-based algorithms.

In outlier detection algorithms based on probability statistical models, it is assumed that the detected data fit a statistical model, such as a parametric model like the Gaussian mixture model or a nonparametric model like kernel density estimation. The outliers are detected by fitting the detected data to the statistical model and then comparing the deviation of the detected data with the model to determine whether an outlier is present in the data or not. Methods for the detection of data outliers based on parametric statistical models [7] and nonparametric statistical models [8] have been introduced. The experimental results showed that these methods could obtain good detection results when the statistical laws of the data were simple and the data size was small. Nevertheless, power data shows a typical temporal coupling pattern and is high dimensional. For such datasets, these methods often fail to achieve good detection results. Although these methods have a good theoretical basis, the practical application of these methods is limited by the fact that the specific statistical model fitting the detected data cannot be known in advance, resulting in blind detection and uncertain results.

The classification-based outlier detection method is a semi-supervised learning method [9]. This type of method operates in a two-phase fashion, that is, the training phase and the testing phase. The training phase requires sufficient labeled samples to train the classifier, and then the classifier determines whether the detected data are normal or outliers. The drawback of this method is that it requires enough labeled samples to train the classifier, and the performance of the classifier directly affects the detection accuracy. The neural network is a classification-based outlier detection algorithm with good self-learning capability, but its derivation process lacks interpretability [10].

For distance-based outlier detection, it is assumed that a data object is an outlier if it is far away from other points. The distance-based outlier detection method is simpler than the statistical model-based method because it is easier to define a distance-based metric for a dataset than to determine the distribution of the dataset. Fan et al. [11] proposed the outlier detection algorithm with personalized K-nearest neighbors (PKNN). In this algorithm, the number of K-nearest neighbors of each sample is determined automatically by the algorithm without any human intervention, so that samples in dense areas have more nearest neighbors and samples in sparse areas have fewer, which is more consistent with the actual distribution of the dataset. The idea of this type of method is simple, but the time complexity is high because of the required calculation of the distance between two data points. Additionally, the method is sensitive to the parameter K, thereby obtaining highly variable detection results for different values of K.

The clustering-based algorithm for outlier detection is an unsupervised learning method. This type of algorithm assumes that normal data belong to one or more clusters, and samples that do not belong to any cluster are considered outliers. Outlier detection is the process of identifying outliers in a dataset through cluster analysis. Clustering algorithms that have been frequently used for outlier detection include DBSCAN, BIRCH, CLARANS, STING, CLIQUE, and KNN [12, 13]. The clustering-based algorithms for outlier detection can obtain good detection results, but their time complexity is generally high and the clustering results are greatly influenced by the parameters.

The density peak clustering (DPC) algorithm is a novel density-based clustering method proposed by Rodriguez and Laio [14] in 2014. The main idea of this algorithm lies in the portrayal of cluster centers. The authors considered that the cluster centers were composed of many samples with a higher density and larger relative distance. The DPC algorithm can automatically determine the number of clusters and achieve any status quo clustering, which is a hot research topic in cluster analysis. However, the DPC algorithm has some problems. First, the clustering results by this algorithm are dependent on the cutoff distance, which can be hardly determined. Second, the definition of the local density of the algorithm does not take the data size and its distribution into account, resulting in unideal clustering accuracy [15].

Based on the characteristics of power data, we proposed a DPC algorithm based on K-nearest neighbors for the detection of power data outliers. Currently, the DPC algorithm has been seldomly applied for outlier detection. In this paper, we redefined the local density of the DPC algorithm by integrating the K-nearest neighbors. By considering the local characteristics of data and using a unified definition of local density, the original algorithm was improved. Meanwhile, only one parameter (K) needs to be determined, which eliminates the influence of cutoff distance on clustering performance, and its value can be easily determined. The experimental results showed that the proposed algorithm can achieve accurate detection of power data outliers.

2. DPC Algorithm

The basic idea of the DPC algorithm is as follows: (1) density peaks have a high local density and are surrounded by neighbors with lower density; (2) the density peaks are relatively far from each other. In the DPC algorithm, two variables are introduced to characterize the density and distance of Sample i, namely, the local density ρ_i and the relative distance δ_i to the nearest sample with a higher local density. The local density ρ_i can be calculated by:

$$\rho_i = \sum_j \chi(d_{ij} - d_c),$$
$$\chi(x) = \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases}, \tag{1}$$

where d_{ij} is the Euclidean distance between Samples i and j. When the size of the dataset is small, the local density is calculated by the Gaussian kernel function as follows.

$$\rho_i = \sum_j \exp\left(-\frac{d_{ij}^2}{d_c^2}\right). \tag{2}$$

The relative distance of Sample i to the nearest sample with higher local density is denoted by δ_i, which can be calculated by:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j>\rho_i} \left(d\left(x_i, x_j\right)\right), & \text{if } \exists \text{j s.t.} \rho_j > \rho_i \\ \max\limits_{j} \left(d\left(x_i, x_j\right)\right), & \text{otherwise} \end{cases}. \tag{3}$$

If Sample i has the maximum local density, the corresponding relative distance will also be the largest.

The DPC algorithm takes the samples with the larger ρ_i and δ_i as density peaks. To find the density peaks, the DPC algorithm draws a decision diagram with ρ_i on the horizontal axis and δ_i on the vertical axis to select the density peaks. To better represent the density peaks, the DPC algorithm defines a decision value γ_i.

$$\gamma_i = \rho_i \times \delta_i. \tag{4}$$

The DPC algorithm considers the samples with a high local density and large distance as the density peaks. It means that the points with high γ_i shall be selected as the density peaks. After the density peaks are found, the remaining samples shall be allocated to the cluster that the nearest samples with a higher density than them belong.

3. Detection of Power Data Outliers Using DPC Based on K-Nearest Neighbors

3.1. DPC Algorithm Based on K-Nearest Neighbors. The local density of the DPC algorithm using either the Gaussian kernel or the cutoff kernel is related to the cutoff distance, and the optimal cutoff distance varies greatly among different datasets [16]. Additionally, the local density of the DPC algorithm is mainly determined by the samples within the range of cutoff distance, and the samples beyond the range contribute little to the local density. Owing to this, density peaks are more likely to appear in the region with high local density. For data with uneven density distribution, the cluster centers are concentrated in dense regions and there are no cluster centers in sparse regions. After analysis, it can be known that the relative density of a sample and its K-nearest neighbors can truly reflect whether the sample is a density peak or not.

Definition 1. Local density based on K-nearest neighbors. The local density of new samples can be defined by:

$$\rho_i = \exp\left(-\frac{\sum_{j=knn(i)} {d_{ij}}^2}{\sum_{j=knn(i)}\sum_{v=knn(j)} {d_{vj}}^2}\right), \tag{5}$$

where d_{ij} is the Euclidean distance between Samples i and j, $knn(i)$ is the set of K-nearest neighbors of Sample i, $\sum_{j=knn(i)} {d_{ij}}^2$ is the sum of the Euclidean distance between Sample i and its K-nearest neighbors, indicating the degree of outlierness of Point i. The larger the value of $\sum_{j=knn(i)} {d_{ij}}^2$, the greater the degree of outlierness, the more locally sparse Sample i. $\sum_{j=knn(i)}\sum_{v=knn(j)} {d_{vj}}^2$ indicates the sum of the degree of outlierness of K-nearest neighbors of Sample i. The larger the value, the greater the local density of the point.

When the local density is defined in this way, the local density of the sample is only related to its K-nearest neighbors, thereby eliminating the interference of the distant irrelevant points. Also, the calculated local density is the relative density of the point and its K-nearest neighbors. It means that the local density of the sample in clusters with varying density distribution can be adjusted so that the local density of the sample in sparse clusters increases and the local density of the sample in dense clusters decreases. In doing so, the density peak in sparse clusters can be found more easily and thus improving the clustering performance for datasets with large differences in density.

3.2. Principle of Outlier Detection. Upon determination of (ρ_i, δ_i) of all sample points in the dataset, the decision diagram of ρ_i and δ_i shall be drawn. The points with both large ρ and δ are identified from the decision diagram, and these points are used as the cluster centers of the dataset. From the perspective of outlier detection, the points with smaller ρ and larger δ can also be visually seen in the decision diagram, and these points are identified as outliers.

Considering the characteristics of power data, we assumed that outliers should satisfy the following conditions: (1) the local density is less than the threshold of local density, that is, $\rho_i < \rho_f$; (2) the relative distance is greater than the threshold of relative distance, that is, $\delta_i > \delta_f$. Based on this, the outliers of the power data can be determined. The threshold of local density ρ_f can be calculated by:

$$\rho_f = \frac{1}{N}\sum_{i=1}^{N} \rho_i - \varepsilon_\rho. \tag{6}$$

The threshold of relative distance δ_f can be calculated by:

$$\delta_f = \frac{1}{N}\sum_{i=1}^{N} \delta_i - \varepsilon_\delta, \tag{7}$$

where N denotes the total number of samples in the power dataset, and ε_ρ and ε_δ denote the empirical parameters.

3.3. Procedures of Outlier Detection. Input: power load data X, number of neighbors K (the experimental result is optimal when K is 25).

Output: outliers.

Step 1: Data preprocessing. Preprocess the load data, such as replacing the missing values with the mean substitution method

Step 2: Calculate the Euclidean distance between samples and construct the distance matrix of samples

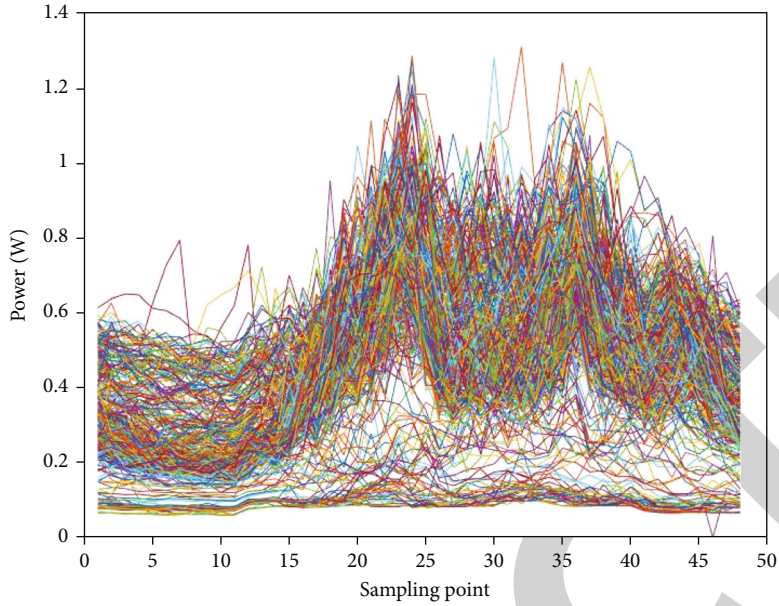

Figure 1: Daily load profile of single transformer.

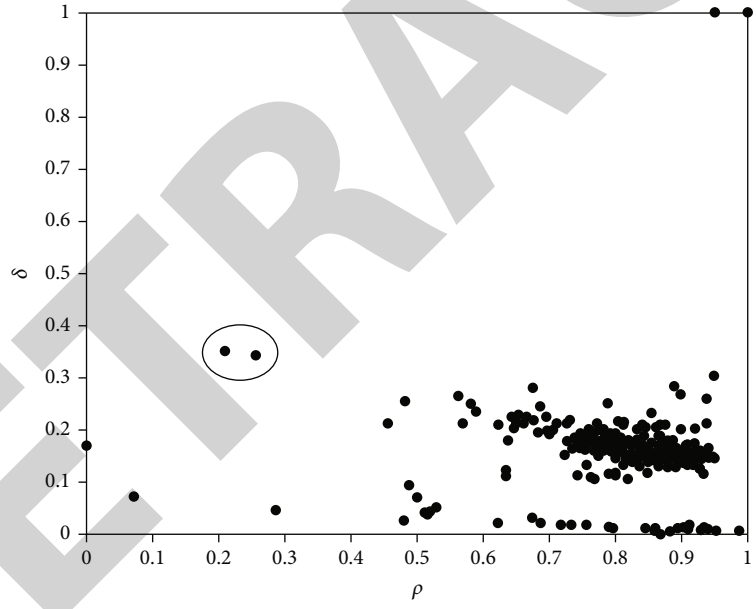

Figure 2: Outlier detection decision diagram of single transformer.

Step 3: Calculate the local density ρ_i and relative distance δ_i of samples according to Eqs. (5) and (3), respectively

Step 4: Set the threshold of local density and relative distance based on the domain expert's experience according to Eqs. (6) and (7)

Step 5: Identify Sample i with $\rho_i < \rho_f$ and $\delta_i > \delta_f$ as outliers and output outliers

4. Results and Analysis

4.1. Data Source. To verify the effectiveness of the DPC algorithm based on K-nearest neighbors for detection of power data outliers, we used the load data of AC 10 kV distribution transformers in a region for 366 days from January 1, 2020, to December 31, 2020. The power load data belongs to the storage sector, and its collection frequency is 0.5 h. The daily load profile has 48 data points. Case 1 is the data of a single transformer, and Case 2 is the daily load data of 10 transformers during 6 days from September 22, 2020, to September 27, 2020. In this paper, the mean substitution method was used, i.e., the mean of all the values except the missing point was used to replace the missing value of the attribute.

4.2. Outlier Detection of Load Profiles of Single Transformer. The load data profile of a single AC 10KV distribution

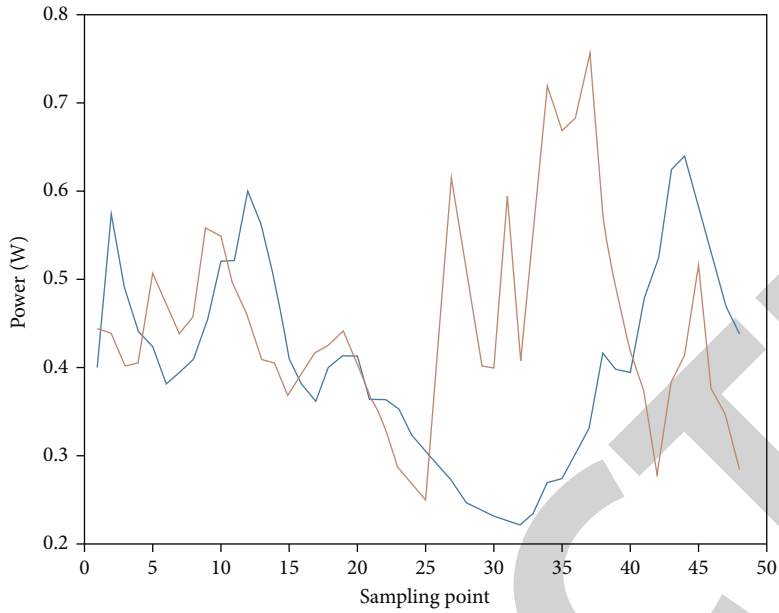

Figure 3: Outliers of single transformer.

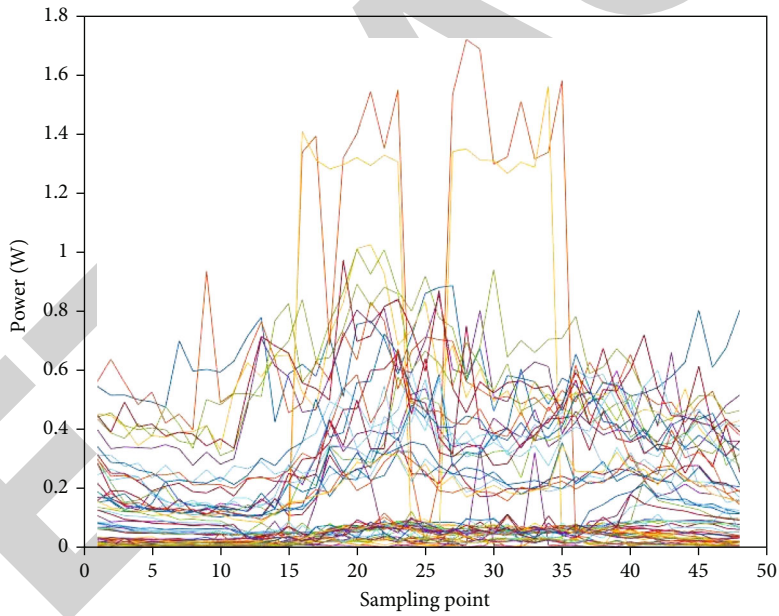

Figure 4: Daily load profile of 10 transformers during 6 days.

transformer during 366 days from January 1, 2020, to December 31, 2020, is shown in Figure 1.

As shown in Figure 1, the daily load trend of this transformer was more or less the same, but few profiles deviated from the normal operation pattern to a large extent. According to the steps of the DPC algorithm based on K-nearest neighbors for detection of power data outliers, the outlier detection decision diagram of this transformer was drawn, as shown in Figure 2.

As shown in Figure 2, the local density and relative distance of most of the samples fell in the region where the local density was higher than 0.4 and the relative distance was less than 0.3, and only very few samples had local density and relative distance fall beyond the above region. According to the principle that outliers should have low local density and large relative distance, the distribution of outliers was identified. The outliers were marked with hollow circles in Figure 2.

The empirical parameters were set as $\varepsilon_\rho = 0.4$ and $\varepsilon_\delta = 0.14$. Then, the outliers in the power data were found according to Eqs. (6) and (7), which were the points circled in Figure 2. The outliers of the load profile of a single transformer are shown in Figure 3.

As shown in Figures 1 and 3, the DPC algorithm based on K-nearest neighbors for detection of power data outliers

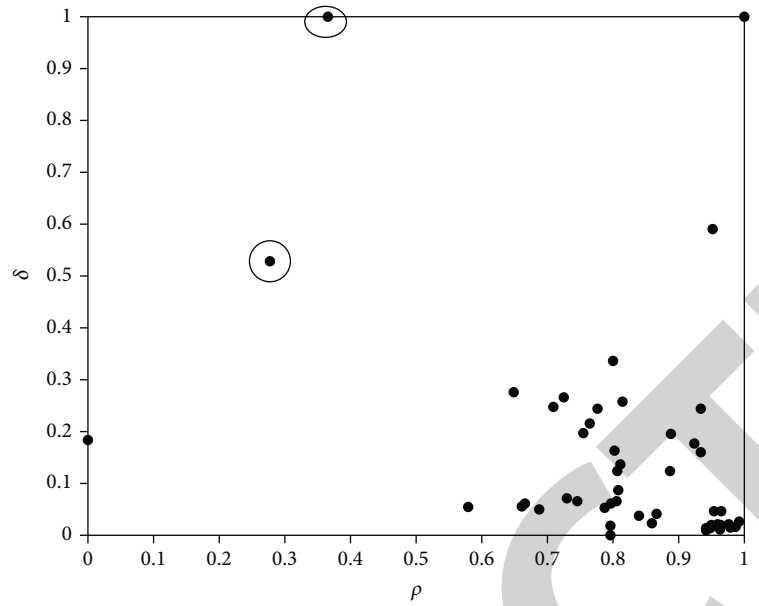

Figure 5: Outlier detection decision diagram of 10 transformers.

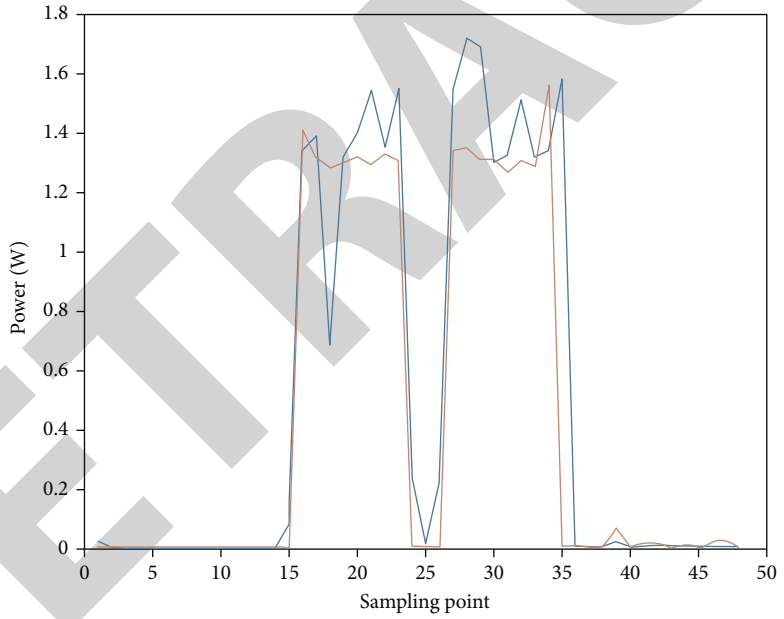

Figure 6: Outliers of 10 transformers.

can detect the profiles that are different from the conventional electricity consumption pattern from the load data. Among the total 366 daily load data, two power data outliers were detected, one on September 26, 2020, and the other on September 27, 2020. The blue and red profiles in Figure 3 represented the daily load profiles on September 26, 2020, and September 27, 2020, respectively. As shown in Figures 1 and 3, an electricity consumption peak should have appeared at sampling point 25 under normal conditions. However, sampling point 25 on September 27 reached the minimum electricity consumption in the day, and the electricity consumption at sampling point 25 on September 26 was also very low, and no peak of electricity consumption appeared between sampling points 25 and 35, which was inconsistent with the normal electricity consumption pattern of this industry. Therefore, this point was identified as an outlier.

4.3. Outlier Detection of Load Profiles of Multiple Transformers. The outlier detection was conducted simultaneously for the daily load data of 10 transformers, and the steps are the same as in Case 1. The daily load profiles of 10 transformers during 6 days from September 22, 2020, to September 27, 2020, are shown in Figure 4. As observed, the outliers were found in the data of few days.

Based on the proposed algorithm, the outlier decision diagram of the above data was drawn, as shown in Figure 5. It can be observed that the local density and relative distance of most of the samples fell in the region where the local density was higher than 0.5 and the relative distance was less than 0.4. The empirical parameters were set as $\varepsilon_\rho = 0.4$ and $\varepsilon_\delta = 0.3$. According to Eqs. (6) and (7), the outliers in the power data were identified, which were the two data points marked with hollow circles in Figure 5.

Figure 6 shows the detected outliers in the 10 transformers. Both outliers were the daily load data of the fifth transformer, which were the daily load data of the transformer on September 24, 2020, and September 25, 2020, respectively. As shown in Figure 6, the profiles of both outliners showed an abnormal increase and decrease in power consumption from sampling point 15 to sampling point 35, which was different from the power consumption of other transformers and other dates.

In summary, the DPC algorithm based on K-nearest neighbors for detection of power data outliers can not only detect the load data outliers of a single transformer but also get good results when being used for detecting the daily load data of multiple transformers, which indicates the applicability of the algorithm.

5. Conclusions

A DPC algorithm based on K-nearest neighbors for the detection of power data outliers was proposed. This algorithm redefined the local density using the K-nearest neighbors of the samples, unified the definition of the local density of the samples, and eliminated the influence of the cutoff distance on clustering performance. Also, the rules for determining the outliers were defined and optimized from the perspective of outlier detection. The proposed algorithm performed well in the simulations of daily load profiles of single and multiple transformers, which verified the effectiveness and applicability of the proposed algorithm.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

[1] B. Saboury, M. Morris, and E. Siegel, "Future directions in artificial intelligence," *Radiologic Clinics of North America*, vol. 59, no. 6, pp. 1085–1095, 2021.

[2] Z. Dong, Z. Junhua, W. Fushuan, and Y. Xue, "rom smart grid to energy internet: basic concept and research framework," *Automation of Electric Power Systems*, vol. 38, no. 15, pp. 1–11, 2014.

[3] C. Chao, "Research for electric power big data quality evaluation model and dynamic exploration technology," *Modern Electronics Technique*, vol. 37, no. 4, pp. 153–155, 2014.

[4] G. Wang, Z. Guoliang, Z. Hongshan, and Z. Mi, "Fast clustering and anomaly detection technique for large-scale power data stream," *Automation of Electric Power Systems*, vol. 40, no. 24, pp. 27–33, 2016.

[5] H. Guo, J. Wang, Z. Li, and Y. Jin, "A multivariable hybrid prediction system of wind power based on outlier test and innovative multi-objective optimization," *Energy*, vol. 239, p. 122333, 2022.

[6] M. Farrokhifard, M. Hatami, and M. Parniani, "Novel approaches for online modal estimation of power systems using PMUs data contaminated with outliers," *Electric Power Systems Research*, vol. 124, pp. 74–84, 2015.

[7] W. Wei, L. Yujiang, and Z. Ye, "Anomaly detection for characteristics of power IT monitoring objects based on probability statistic mode," *Journal of Shandong Agricultural University (Natural Science Edition)*, vol. 50, no. 4, pp. 612–618, 2019.

[8] K. Zheng, C. Qixin, Y. Wang, C. Kang, and Q. Xia, "A novel combined data-driven approach for electricity theft detection," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1809–1819, 2018.

[9] S. C. Lee and R. Nevatia, "Hierarchical abnormal event detection by real time and semi-real time multi-tasking video surveillance system," *Machine Vision & Applications*, vol. 25, no. 1, pp. 133–143, 2014.

[10] H. A. Dau, V. Ciesielski, and A. Song, "Anomaly detection using replicator neural networks trained on examples of one class," *Lecture Notes in Computer Science*, vol. 8886, no. 1, pp. 311–322, 2014.

[11] F. Ruixuan, J. Gaoxia, and W. Wang, "Outlier detection algorithm with personalized k-nearest neighbor," *Journal of Chinese Computer Systems*, vol. 41, no. 4, pp. 752–757, 2020.

[12] J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "ensity-based clustering in spatial databases: the algorithm GDBSCAN and its applications," *Data Mining & Knowledge Discovery*, vol. 2, no. 2, pp. 169–194, 1998.

[13] H. J. Jia, S. F. Ding, and Z. Z. Shi, "Approximate weighted kernel k-means for large-scale spectral clustering," *Journal of Software*, vol. 23, no. 1, pp. 12–17, 2015.

[14] A. Rodriguez and A. Laio, "Machine learning. Clustering by fast search and find of density peaks," *Science*, vol. 34, no. 1, pp. 14–22, 2014.

[15] R. Bie, R. Mehmood, S. Ruan, and H. Dawood, "Adaptive fuzzy clustering by fast search and find of density peaks," *Personal and Ubiquitous Computing*, vol. 20, pp. 785–793, 2016.

[16] L. Lv, J. Wang, R. Wu, H. Wang, and I. Lee, "Density peaks clustering based on geodetic distance and dynamic neighbourhood," *International Journal of Bio-Inspired Computation*, vol. 17, no. 1, pp. 24–33, 2021.