IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 June 2024, accepted 11 July 2024, date of publication 16 July 2024, date of current version 25 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3429304

RESEARCH ARTICLE

Abnormal Electricity Detection of Users Based on Improved Canopy-Kmeans and Isolation Forest Algorithms

JIANYUAN WANG AND XIAOYAO LI

Key Laboratory of Modern Power System Simulation and Control and Renewable Energy Technology Ministry of Education, Northeast Electric Power University, Jilin City, Jilin 132012, China

Corresponding author: Xiaoyao Li (369888300@qq.com)

ABSTRACT Aiming at the existing user abnormal electricity consumption detection methods that have the problem of difficult classification of user similar electricity consumption patterns, this paper proposes an unsupervised isolation forest abnormal electricity consumption detection model based on the Canopy-Kmeans algorithm with weighted density improvement. To start, we propose a composite parameter analysis method for user electricity consumption patterns, volatility, trends, and correlations using Irish smart meter data. This method involves joint data cleaning, interpolation, and feature construction. Additionally, principal component analysis is introduced to fuse features across layers and reduce dimensionality in user electricity consumption. Subsequently, we introduce the weighted density improvement Canopy-Kmeans clustering algorithm. This algorithm determines the K value and clustering centers using the maximum weight product method, based on definitions of sample density, average intra-class sample distance, and inter-class distance in the multilayer fusion feature data. Finally, we propose a fusion mechanism of weighted density improvement Canopy-Kmeans and isolation forest algorithms to jointly construct a model for detecting abnormal power usage based on multilayer fusion feature data analysis. The results demonstrate that multilayer fusion feature parameters vary in size and discretization among different user types, enabling classification of users with diverse electricity consumption patterns. Moreover, the anomaly detection model based on multilayer fusion feature data analysis improves accuracy rates, recall rates, and F1 scores compared to other algorithms.

INDEX TERMS Abnormal detection of electricity consumption by users, Canopy-Kmeans algorithm, isolation forest algorithm, unsupervised learning, weighted density.

I. INTRODUCTION

Non-technical loss (NTL) of electrical energy is that portion of electrical energy that is consumed by the consumer but not received by the electric utility [1]. Most of these losses, except for a small portion generated by the failure of the meter itself, are caused by the abnormal electricity consumption behavior of power users [2]. According to statistics, the losses due to electricity theft in China exceeds 6 percent of the total power supply losses [3]. In addition to economic losses, the presence of NTL can interfere with the measurement of power system network operating parameters, which in turn jeopardizing the reliability of power system operation [4], [5]. Abnormal electricity consumption detection work can identify faulty meters and electricity theft users, but early detection methods still rely on on-site screening, which is time-consuming and labor-intensive. The popularization of smart meters has collected a large amount of fine-grained electric load data for electric utilities [6], which provides a data source for data-driven anomalous power usage detection based on data.

According to whether data labeling is required or not, electricity load anomaly detection algorithms can be classified into 2 categories: supervised and unsupervised. Supervised

The associate editor coordinating the review of this manuscript and approving it for publication was Muammar Muhammad Kabir.

algorithms learn the features of anomalous users through the training set and thus obtain the ability to recognize anomalous users, such as support vector machines [7], [8], neural networks [9], [10], et al, which are more accurate but require a sufficient amount of labeled sample data as support. However, electricity load data have the problem of imbalance of data categories and a small proportion of abnormal users, which tends to make the training process underfitted [11], and may ultimately reduce the accuracy of the detection results. In contrast, unsupervised algorithms have the advantage that the detection process does not require data labeling and is less affected by data category imbalance due to the absence of a training process.

Currently, deep learning-based unsupervised algorithms are mostly used as auxiliary algorithms to supervised detection algorithms [12] to reduce the negative impact of category imbalance data on the detection results; and unsupervised algorithms based on outlier factors are not suitable for detecting electricity consumer loads with multiple electricity consumption patterns [13]. In contrast, clustering-based unsupervised algorithms can realize unsupervised anomaly detection in the whole process of anomaly detection (excluding model evaluation) and can realize anomaly detection for multiple categories of electric loads. K-means clustering [14], density-based spatial clustering of applications with noise (DBSCAN) [15], nearest neighbor propagation clustering [16], and outlier clustering [17] have shown good results in power load anomaly detection. However, the problem with clustering algorithms for anomaly detection is that parameter selection is often based on experience [18], and unsupervised algorithms lack anomalous user labels to assess the merits of the selected parameters, so it is difficult to obtain the optimal parameters.

Therefore, this paper proposes an unsupervised isolation forest anomalous electricity consumption detection model based on the weighted density improved Canopy-Kmeans algorithm. Firstly, based on Irish smart meter data, joint data cleaning, segmented linear interpolation, feature construction [19], [20] proposed a composite parameter analysis method of user electricity consumption pattern, volatility, trend, and correlation, and introduced the principal component analysis to carry out the feature layer fusion dimensionality reduction of user electricity consumption. Then, based on the difficulty of obtaining the optimal parameters of T_1 and T_2 in Canopy-Kmeans algorithm, the optimal parameter selection problem of T_1 and T_2 in Canopy-Kmeans algorithm with weighted density improvement is proposed, and the optimal K-value and initial clustering center are obtained by combining the multi-layer fusion feature data density calculation method and the maximum weight product method. Finally, the weighted density improvement Canopy-Kmeans and isolation forest algorithm fusion mechanism is proposed to jointly construct a multi-layer fusion feature data analysis of the user abnormal electricity consumption detection model. The results of this paper provide a new idea for the existing user abnormal electricity consumption detection methods that have the problem of difficult classification of user similar electricity consumption patterns. The results show that the multilayer fusion feature parameters of different types of users have different sizes and dispersion degrees, and the optimization of the optimal parameters of the Canopy-Kmeans algorithm can classify users with different electricity usage patterns; meanwhile, the anomaly detection model for calculating the anomaly scores based on the analysis of the multilayer fusion feature data of different types of users achieves an increase in the precision rate, the recall rate, and the F1 score.

II. RELATED WORK

Unsupervised methods are particularly popular in the field of power load anomaly detection because they do not require labeled data. This section analyzes and summarizes several unsupervised anomaly detection methods, including clustering techniques, density-based methods, and other unsupervised machine learning methods.

Xiao et al. [15] addressed the low efficiency of electricity theft detection by proposing a DBSCAN clustering and cubic smoothing exponential model. This method uses the residual term between the true value and the predicted value for clustering to detect electricity theft. While DBSCAN can efficiently handle noisy data, its performance in high-dimensional data is suboptimal, and parameter selection is crucial for effective clustering. Du et al [16] proposed a secondary screening method to address interference and false alarms when identifying low-power abnormal users based on derived indicator items. It uses the daily three-phase power of initially detected abnormal users as load characteristics, clusters daily power consumption, and determines power theft based on similarity to normal low power states. However, the screening features are limited and have certain constraints. Xie et al. [21] proposed a K-means clustering method using a graph attention self-encoder model for dimensionality reduction to address anomaly identification and user localization problems. By constructing a user electricity relationship network based on a probabilistic graph structure and the random walk algorithm, it can localize abnormal users and provide early warning for potential abnormal users. Although K-means clustering is widely used, it has issues with determining the optimal number of clusters and selecting the initial clustering centers Zheng et al. [22] employed fuzzy clustering combined with changes in the line loss rate of station areas as the state value. By analyzing the covariance matrix feature distribution, average spectral radius over time, and empirical spectral function, it accurately identifies periods of power theft. Fuzzy clustering is advantageous for handling uncertain data but has high computational complexity, making it unsuitable for large-scale real-time monitoring. Krishna et al. [23] presented a density-based spatial clustering anomaly detection method utilizing principal component analysis and applications with noise. This method, used by electric utilities to detect integrity attacks on smart meter communications in advanced metering infrastructures,

TABLE 1. Reference comparison.

Shortcomings	Reference	Improvements
Poor performance on high-dimensional data	[15]	Introduced principal component analysis for feature layer fusion and dimensionality reduction of user electricity consumption, effectively reducing data dimensionality while retaining key features.
Parameters greatly impact results and are difficult to select	[12] [13] [15] [21] [23] [25] [26]	Used maximum density and maximum weight to select clustering centers, making the entire clustering process parameter-free.
Secondary screening feature selection is single and has certain limitations	[16]	Proposed a composite feature analysis method for user electricity consumption patterns, volatility, trends, and correlations.
High computational complexity	[22] [24]	The clustering algorithm selects the clustering centers sequentially based on maximum density and maximum weight, which reduces the clustering time and reduces the complexity by performing anomaly detection after user classification.

provides quantitative parameters and design choices, and is quantitatively evaluated against average detectors proposed in related works. Parameter selection impacts the results. Cui et al. [24] introduced a two-step power theft detection strategy. In the first step, a convolutional auto encoder neural network model is proposed for power theft identification, using convolutional layers to extract and identify anomalies in stealing users based on the uniformity and periodicity of normal power consumption characteristics. In the second step, the anomalies of each identified power theft user are predicted using a modified regression algorithm combining extreme gradient boosting and transfer adaptive boosting training strategies. However, the introduction of convolutional layers increases computational complexity.

Additionally, unsupervised learning methods often clustered users first to classify different power users, followed by outlier detection for each class of users [13], [25], eliminating the need to train on power theft users. Other studies started with feature extraction, utilizing auto encoders to automatically extract user features [12], [26], and then combined clustering and outlier detection to detect power theft users. These methods achieved anomaly detection without training on power theft users, but the clustering effectiveness directly impacted the final detection results. Table 1 illustrates how the algorithm proposed in this paper addressed the aforementioned issues.

In this paper, the proposed algorithm as well as improvements are as follows:

A. CANOPY ALGORITHM

Canopy algorithm is proposed by McCallum et al. [27], which is an unsupervised preclustering algorithm. The preprocessing step of divisive clustering algorithms and hierarchical clustering algorithms often adopts the Canopy algorithm, which requires setting 2 distance thresholds T_1 and T_2, and randomly selecting the initial clustering centers, and calculating the Euclidean distance between the sample objects and the initial clustering centers. The samples are classified into the corresponding class clusters according to the thresholds. Finally, the clustered dataset is divided into k class clusters.

The algorithmic flow of Canopy is as follows:

Step 1 Given the dataset D $=\{x_1, x_2, \ldots, x_n\}$, set the thresholds T_1 and $T_2 (T_1 > T_2)$.

Step 2 Take a random sample point S from the dataset D and calculate the Euclidean distance d between the remaining sample points in the dataset and the point S respectively. If $d< T_1$, the sample point will be added to the current Canopy layer.

Step 3 Then the distance d is compared with T_2, if $d< T_2$, the sample point is removed from the dataset D and is not added to the other Canopy layers.

Step 4 Repeat Steps 2 and 3 until the dataset D is empty.

The Canopy algorithm is difficult to determine the thresholds T_1 and T_2, and the value of the thresholds has a great impact on the clustering results. If the threshold is too large, data belonging to different classes will be grouped into the same class; if the threshold is too small, data belonging to the same class will be divided into several classes. Therefore, a weighted density improved threshold parameter selection is proposed to solve this problem.

B. WEIGHTED DENSITY IMPROVED CANOPY ALGORITHM

1) RELATED CONCEPTS

Let the dataset D $=\{x_1, x_2, \ldots, x_n\}$ be a collection of data containing n sample objects, each with d-dimensional feature attributes. Where x_i denotes the p-th dimension attribute of the i-th data object ($i = 1, 2, \ldots, n$; $p = 1, 2, \ldots, d$).

Definition 1 The Euclidean distance between data objects x_i and x_j is d_{ij}:

$$d_{ij} = \sqrt{\sum_{p=1}^{d} \left(x_{ip} - x_{jp}\right)^2} \tag{1}$$

where i, j is from 1 to n.

Definition 2 The average distance of all sample elements in dataset D is:

$$AverDis\left(D\right) = \frac{2}{n\left(n-1\right)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d_{ij} \tag{2}$$

Definition 3 The cardinality ρ_i of a sample object x_i in dataset D is expressed in as the number of sample objects falling within a region centered on the sample object x_i with $AverDis(D)$ as the radius.

Definition 4 According to Definition 3, ρ_i is the number of sample objects that satisfy the condition that the distance from other sample objects to sample object x_i is less than $AverDis(D)$. The samples that satisfy the condition form a class cluster where the average distance between sample objects is:

$$\alpha_i = \frac{2}{\rho_i(\rho_i - 1)} \sum_{i=1}^{\rho_i} \sum_{j=i+1}^{\rho_j} d_{ij} \tag{3}$$

Definition 5 The class cluster distance s_i denotes the distance between a sample object x_i and another sample object x_j. The definition of s_i depends on the density value of x_i:

If x_i has the maximum density value, then s_i is defined as the maximum distance to any other sample object, expressed as $max\{d_{ij}\}$.

If x_i does not have the maximum density value, then s_i is defined as the minimum distance to any other sample object, expressed as $min\{d_{ij}\}$, which is defined as follows:

$$s_i = \begin{cases} \underset{j:\rho_j>\rho_i}{min} \{d_{ij}\}, \rho_j > \rho_i \\ max\{d_{ij}\}, \rho_i > \rho_j \end{cases} \tag{4}$$

2) MAXIMUM WEIGHT PRODUCT METHOD

Definition 6 The density weights for the selected clustering centers are defined as follows:

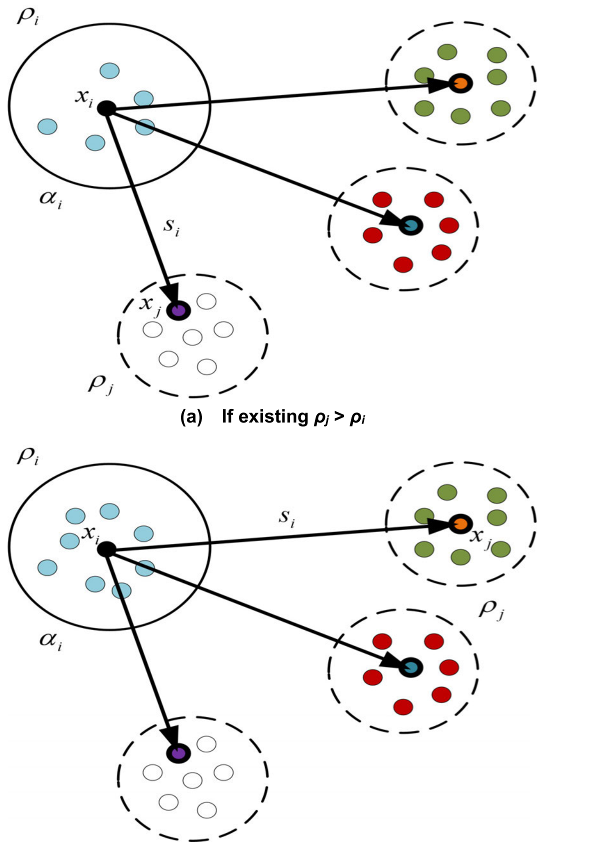
$$\omega_i = \rho_i * \frac{1}{\alpha_i} * s_i \tag{5}$$

The threshold value is randomly selected in the traditional Canopy algorithm, which has a large impact on the clustering results. This paper proposes the maximum weight product method, which can reduce the uncertainty caused by randomness and can improve the clustering accuracy.

The maximum weight product method is shown in Fig. 1. First, the density of the samples is calculated according to Definition 3, and the maximum value of the density is taken as the 1st clustering center, and all the sample points that satisfy the condition that the distance of the samples from the initial clustering centers is less than the $AverDis(D)$ condition in Definition 2 are added to the current clusters, and these sample points are removed from the dataset. Then, the weight product of the remaining samples is calculated according to Definition 3 and refer to (3)-(6), the maximum value is found and the corresponding sample is selected as the 2nd clustering center, and finally, the above steps are repeated until the dataset is empty.

In addition, a larger ρ_i indicates more element points around the sample element x_i and a greater concentration of sample elements. A smaller α_i means a larger $1/\alpha_i$, indicating closer proximity of elements in the class clusters. A larger s_i means that the two class clusters are further apart from each other, leading to a greater degree of dissimilarity. Therefore, the optimal clustering center can be found using the maximum weight product method.

The most suitable number of clusters in the optimal division is determined according to the maximum weight product method proposed in this paper, and the specific improvement steps are as follows:



(a) If existing $\rho_j > \rho_i$

(b) If not existing $\rho_j > \rho_i$

FIGURE 1. Diagram of the maximum weight product method.

Step 1 Calculate the Euclidean distances of all elements in the power dataset, the average distance between samples, and the local density with reference to equations (1), (2), and Definition 3.

Step 2 Select the sample element with the highest density as the first initial cluster center, add it to the Canopy center set, and remove all sample elements within the neighborhood of this sample from the dataset.

Step 3 Calculate the density, average distance between sample objects, and inter-cluster distance of the remaining sample elements. Determine the second cluster center using the maximum weight product method, add this center to the Canopy center set, and remove all sample elements from the dataset that meet the specified conditions.

Step 4 Determine whether the dataset is empty: if it is not empty, loop the execution of steps 3 until the dataset does not contain sample points; if it is empty, output the Canopy center set, which contains K initial clustering centers, as the initial parameters of the output of K-means.

C. WEIGHT DENSITY IMPROVED CANOPY-KMEANS ALGORITHM

In this paper, the algorithm of weighted density improved Canopy-Kmeans clustering is proposed, and it is proposed to use the weighted density Canopy algorithm to precluster the

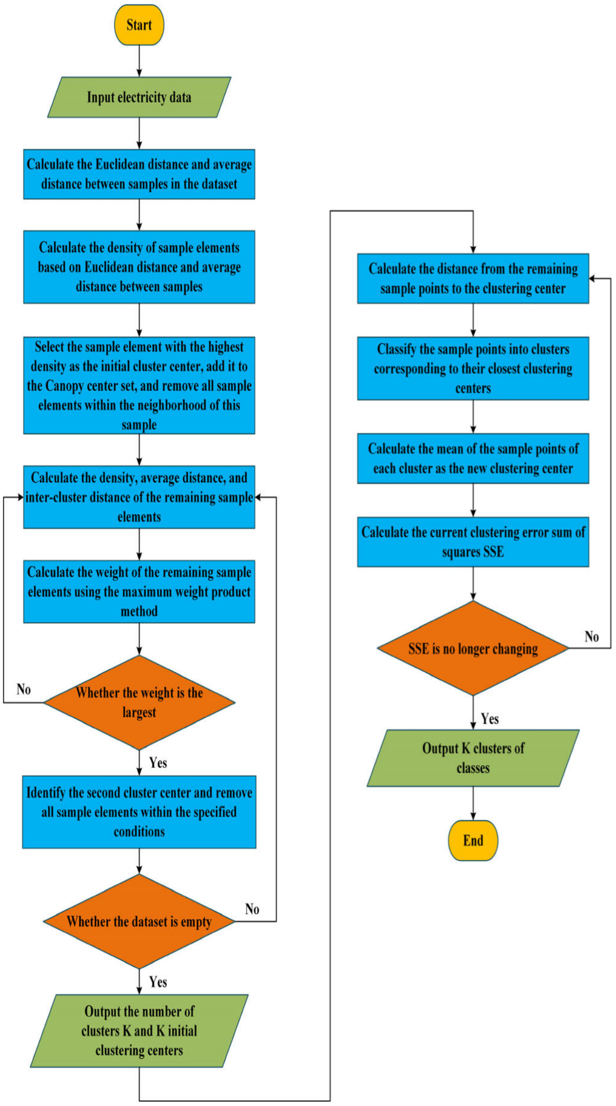

Start
Input electricity data
Calculate the Euclidean distance and average distance between samples in the dataset
Calculate the density of sample elements based on Euclidean distance and average distance between samples
Select the sample element with the highest density as the initial cluster center, add it to the Canopy center set, and remove all sample elements within the neighborhood of this sample
Calculate the density, average distance, and inter-cluster distance of the remaining sample elements
Calculate the weight of the remaining sample elements using the maximum weight product method
Whether the weight is the largest
No
Yes
Identify the second cluster center and remove all sample elements within the specified conditions
Whether the dataset is empty
No
Yes
Output the number of clusters K and K initial clustering centers
Calculate the distance from the remaining sample points to the clustering center
Classify the sample points into clusters corresponding to their closest clustering centers
Calculate the mean of the sample points of each cluster as the new clustering center
Calculate the current clustering error sum of squares SSE
SSE is no longer changing
No
Yes
Output K clusters of classes
End

FIGURE 2. Flowchart of Canopy-Kmeans algorithm for weighted density improvement.

electricity data, and use the weighted density as the basis for selecting the Canopy centers, and abolish the setting of the distance threshold, to avoid the problem of difficulty in the selection of the distance threshold, and then use the number of clusters K obtained by the Canopy algorithm and the K initial clustering centers as the initialization input parameters of K-means algorithm. This improved method avoids the instability caused by randomness and can effectively improve the accuracy of clustering. The flow chart of the algorithm is shown in Fig. 2.

The specific steps of the improved algorithm are as follows:

Step 1 Calculate the Euclidean distances of all elements in the power dataset, the average distance between samples, and the local density with reference to equations (1), (2), and Definition 3.

Step 2 Select the sample element with the highest density as the first initial cluster center, add it to the Canopy center set, and remove all sample elements within the neighborhood of this sample from the dataset.

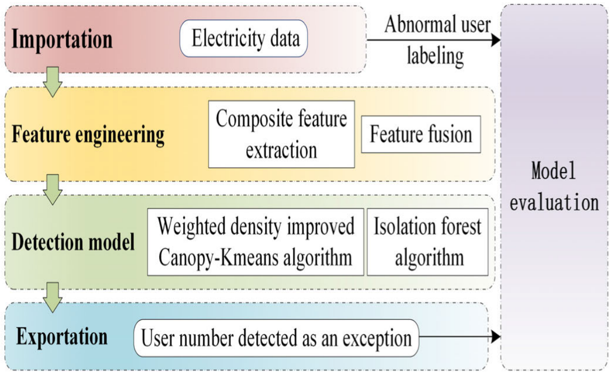

Importation
Electricity data
Abnormal user labeling
Feature engineering
Composite feature extraction
Feature fusion
Detection model
Weighted density improved Canopy-Kmeans algorithm
Isolation forest algorithm
Exportation
User number detected as an exception
Model evaluation

FIGURE 3. Overall framework of proposed model.

Step 3 Calculate the density, average distance between sample objects, and inter-cluster distance of the remaining sample elements. Determine the second cluster center using the maximum weight product method, add this center to the Canopy center set, and remove all sample elements from the dataset that meet the specified conditions.

Step 4 Determine whether the dataset is empty: if it is not empty, loop the execution of steps 3 until the dataset does not contain sample points; if it is empty, output the Canopy center set, which contains K initial clustering centers, as the initial parameters of the output of K-means.

Step 5 Calculate the distance between the clustering center and the remaining sample points in the electricity dataset, and classify the remaining sample points into clusters in which the clustering centers closest to them are located.

Step 6 Calculate the mean of the sample points of each cluster as the new clustering center.

Step 7 Calculate the error sum of squares for the current clustering as a judgment condition for convergence of the clusters.

Step 8 Determine whether the SSE changes or not: if it changes, loop the execution of steps 5-7 until the SSE no longer changes; if it remains unchanged, the algorithm ends and outputs the final division of the K class clusters.

III. ANOMALY DETECTION MODEL

The overall framework of the unsupervised power load anomaly detection model based on the weighted density improved Canopy-Kmeans algorithm and the isolation forest algorithm is shown in Fig. 3.

A. COMPOSITE FEATURE EXTRACTION

Composite feature extraction can mine the deep information of the original load data and improve the accuracy of the anomaly detection model. This model extracts the morphological, volatility, trend, and correlation indicators based on the daily and monthly electricity consumption data of users.

1) MORPHOLOGICAL INDICATORS

1) Average daily and monthly electricity consumption.

2) Daily and monthly electricity consumption rate, that is, the ratio of average electricity consumption to maximum electricity consumption.

3) Peak-to-valley ratio of monthly electricity consumption [28], that is, the ratio of the difference between the maximum and minimum electricity consumption to the maximum electricity consumption.

4) Quarterly electricity consumption as a percentage of annual electricity consumption.

2) VOLATILITY INDICATORS

1) Daily and monthly electricity consumption dispersion coefficients [29], which is the ratio of the standard deviation of daily and monthly electricity consumption to the mean of daily and monthly electricity consumption.

2) Ratio of daily and monthly electricity consumption dispersion coefficients to the daily and monthly electricity consumption dispersion coefficients of the industry [29] (using the average of all customer electricity consumption to represent the electricity consumption of the industry).

3) The first and last differences between electricity consumption in the first m months and the next m months.

3) TREND INDICATORS

1) The slope of the linear fit of the daily electricity consumption series k.

2) Upward and downward trends in monthly electricity consumption series.

4) CORRELATION INDICATORS

A daily series of electricity consumption of a customer with the Pearson correlation coefficient of a typical daily series of electricity consumption (expressed as a series of daily averages of all customers).

B. FEATURE FUSION

The number of extracted features is large and may include features with strong correlation. In order to reduce the complexity of the composite statistics extracted feature vector, it is necessary to reduce the dimensionality of the extracted composite features. The feature layer fusion method chosen in this paper is principal component analysis (PCA) to reduce the dimensionality of composite feature fusion, in order to determine the dimensionality of the fusion features, the cumulative contribution rate is used to reflect how much information of the original composite features is contained in the fusion features. The cumulative contribution rate usually has to reach 85% to express the original composite feature information more comprehensively [30]. The formulas for the contribution rate and cumulative contribution rate of individual fusion features are as follows:

$$\varepsilon_i = \frac{\lambda_i}{\sum_{j=1}^{P} \lambda_j} \tag{6}$$

$$\varepsilon_{ci} = \sum_{j=1}^{i} \varepsilon_j \tag{7}$$

where ε_i is the contribution rate of the i-th fusion feature; λ_i is the value of the i-th fusion feature; P is the total number of fusion features; and ε_{ci} is the cumulative contribution rate of the previous i fusion features.

C. ISOLATION FOREST ALGORITHM ANOMALY DETECTION

Clustering algorithms are mostly used for solving classification problems and for anomaly detection problems, isolation forest algorithm is chosen for anomaly detection. Isolation forest algorithm is an unsupervised anomaly detection algorithm for continuous type data. Unlike other anomaly detection algorithms that use quantitative indicators such as distance and density to characterize the degree of sparsity between samples, this algorithm uses an isolation tree structure to isolate the samples. Since the number of outliers is small and most of the samples are sparse, the outliers are isolated earlier, that is, the outliers are closer to the root node of the isolation tree. Also, the outlier score for each sample point can be calculated. The anomaly score of each sample can be used as an indicator of the anomaly of the sample. Compared with traditional algorithms such as local outlier detection algorithm and K-means, the isolation forest algorithm has better robustness to high dimensional data.

An isolation forest consists of multiple isolation trees, and the structure of an isolation tree is the same as that of a binary search tree, so the average path length of a leaf node is equivalent to the expectation of a binary search tree. Therefore, the isolation forest algorithm draws on methods related to analyzing binary search trees to predict the average path length of its leaf nodes.

Given a dataset containing n samples, the average path length of the tree is:

$$C\left(n\right) = 2H\left(n-1\right) - \frac{2\left(n-1\right)}{n} \tag{8}$$

where $H(i)$ is the harmonic number, which can be estimated as $ln(i)+c_0$, and c_0 is known as Euler constant (c_0 =0.5772156649). $C(n)$ is the average of the path lengths for a given number of samples n, which is used to normalize the path length $h(x)$ for sample x.

The anomaly score of the sample is defined as:

$$s\left(s, n\right) = 2^{-\frac{E(h(x))}{C(n)}} \tag{9}$$

where $E(h(x))$ is the expectation of the path length of sample x in a batch of isolated trees.

TABLE 2. Confusion matrix applied in anomaly detection for electricit load.

Users	Actual abnormal user	Actual normal user
Detected as abnormal user	TP	FP
Detected as normal user	TN	FN

D. MODEL EVALUTION INDICATORS

Anomaly detection of electricity load is essentially a binary classification problem with category imbalance, and evaluation metrics based on accuracy cannot be used because a high evaluation can be obtained even if the classifier recognizes all users as normal users. The merit of the electricity load anomaly detection model is commonly evaluated by the AUC metric of the receiver operating characteristic (ROC). The AUC is derived by first obtaining the confusion matrix of the binary classifier.

1) CONFUSION MATRIX

The confusion matrix contains all the possible classification results of a binary classifier, as shown in Table 2.

In Table 1, the letters T and F denote the correct and incorrect classification results of the classifier, and the letters P and N denote the classifier predictions as abnormal and normal, respectively.TP and TN denote the 2 correct classification results, and FP and FN denote the 2 incorrect classification results.

From the confusion matrix the following metrics can also be calculated:

1) Recall Rate:

$$R = \frac{A_{TP}}{A_{TP} + A_{FN}} \times 100\% \tag{10}$$

2) Accuracy Rate:

$$P = \frac{A_{TP}}{A_{TP} + A_{FP}} \times 100\% \tag{11}$$

where A_{FP} denotes the number of users that the classifier predicts as abnormal but are actually normal; P denotes the ratio of the number of correctly detected abnormal data to the number of all detected abnormal data. The higher the accuracy P the lower the false detection rate and the better the classifier performance.

3) F1 score:

$$F_1 = \frac{2PR}{P + R} \tag{12}$$

2) ROC CURVES AND AUC INDICATORS

Based on the confusion matrix, the true positive rate (TPR) and false positive rate (FPR) of the classifier can be calculated, which are used to reflect the detection rate and false detection rate, respectively, and different thresholds correspond to different values of TPR and FPR.The ROC curves, with FPR as the horizontal axis and TPR as the vertical axis, reflect the trade-off between detection rate and false detection rate under different thresholds [31].

The value range of the AUC indicator is [0, 1], the larger the AUC value, the closer the ROC curve is to the optimal classification point (0, 1), and the better the classification effect.

IV. ANALYSIS OF ABNORMAL DETECTION RESULTS

A. DATA INTRODUCTOPN

The data in this paper uses the Irish smart meter dataset [32], which contains 536 days of electricity consumption data (in kW·h) from 6445 electricity consumers, sampled at a frequency of every 30 min. Users with anomalous electricity consumption behavior have been labeled, totaling 205. The abnormal user labeling is only used as a basis for model evaluation and is not used in the detection process.

The daily electricity consumption data for 536 days is obtained by aggregating the electricity consumption of each user on a daily basis. To extract characteristics of electricity consumption reflecting a longer time span, the daily electricity consumption data for every 30 days is aggregated to obtain monthly electricity consumption data for 18 months. When cleaning the daily electricity consumption data, 10 users whose electricity consumption is 0 for all 536 days are removed. Additionally, the dataset contains data points where electricity consumption is 0 for an entire day, which is common in practice. However, these data points may render some feature values impossible to extract during the process of extracting composite features. In this experiment, all the data points with 0 electricity consumption for one day are modified to 0.01, and a very small value is used to indicate that the user has no electricity consumption in one day.

B. EXPERIMENTAL RESULTS AND ANALYSIS

The 18 composite features extracted from the daily and monthly electricity consumption sequences of the users are:

- Average daily and monthly electricity consumption (*F1* and *F2*).
- Discretization coefficients for daily and monthly electricity consumption series (*F3* and *F4*).
- Daily and monthly electricity consumption rates (*F5* and *F6*).
- Peak-to-valley difference in the monthly electricity consumption series (*F7*).
- Difference in electricity consumption between the first 3 months of the 1st year and the first 3 months of the 2nd year (*F8*).
- Ratios of the discrete coefficients of daily and monthly electricity consumption of the user to the discrete coefficients of daily and monthly electricity consumption of the industry (*F9* and *F10*).
- The ratio of the electricity consumption of each of the 4 quarters of the spring, summer, fall, and winter in the 1st year to the annual consumption of the year (*F11-F14*).

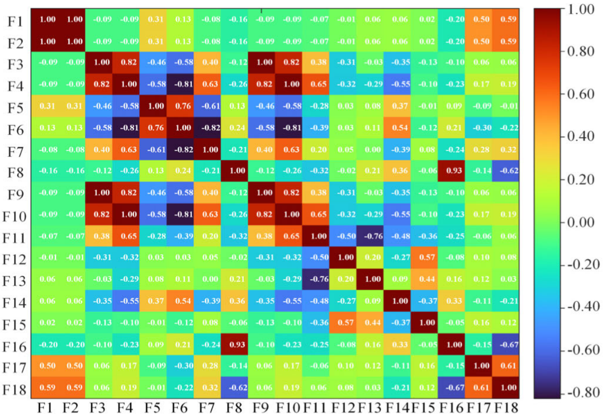


FIGURE 4. Correlation matrix of 18 features.

- Correlation coefficient between daily electricity consumption of consumers and typical daily electricity consumption (*F15*).
- Slope of the linear fit for the daily consumption series (*F16*).
- Upward and downward trend indexes (*F17* and *F18*).

Since the above composite features are not of the same order of magnitude, in order to balance the impact of composite features on the results, the above composite features are normalized with reference to (13).

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{13}$$

Considering that features *F11-F14* together represent the electricity consumption pattern throughout the year, they are multiplied by a weighting factor of 0.25.

The correlation analysis of all the extracted composite features is performed and the obtained correlation matrix of 18 composite features is shown in Fig. 4. Among them, *F1* is linearly correlated with *F2*, *F3* with *F9*, and *F4* with *F10*, so *F2*, *F9*, and *F10* are deleted and the remaining 15 features are retained.

Some composite features are still highly correlated, meaning they contain overlapping information. To address this, a PCA model is employed to perform dimensionality reduction on the 15 composite features, resulting in a set of fused features. These fused features are ordered based on their contribution rates, from largest to smallest. Table 3 intercepts the first 8 fusion features with a higher contribution rate. When the number of fusion features is 3, the cumulative contribution rate of PCA feature fusion dimensionality reduction method is 92.83%, which can represent the information of composite features better (the cumulative contribution rate of 90% can represent the information of composite features well). Therefore, the number of new features taken in this experiment is 3.

The weighted density improved Canopy-Kmeans clustering algorithm is used to classify the new feature data obtained from the PCA model, as shown in Fig. 5, which divides the new feature dataset into 5 classes.

TABLE 3. Contribution rate and cumulative contribution rate of fusion features obtained from PCA models.

Fusion	PCA	
features	Contribution rate/%	cumulative contribution rate/%
1	61.18	61.18
2	25.44	86.62
3	6.21	92.83
4	3.45	96.28
5	2.12	98.40
6	0.77	99.18
7	0.55	99.73
8	0.12	99.85

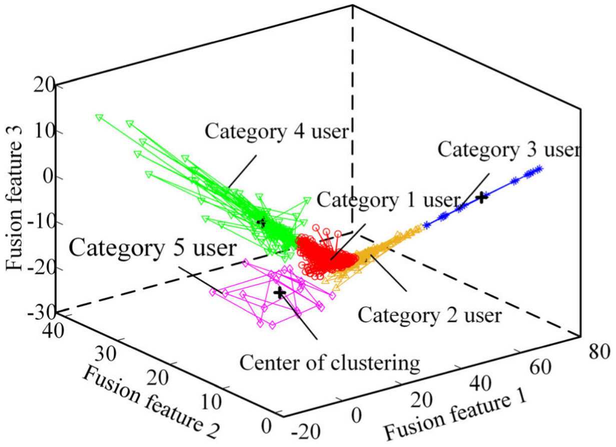


FIGURE 5. Plot of clustering results.

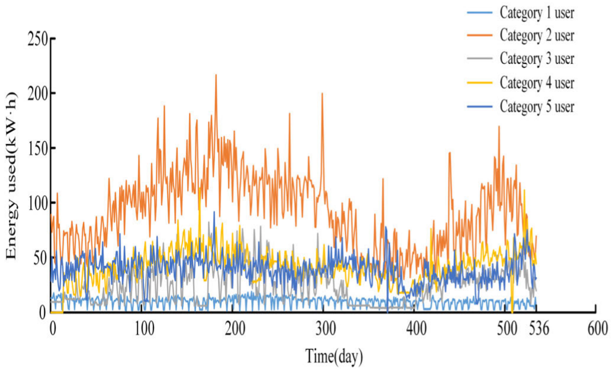
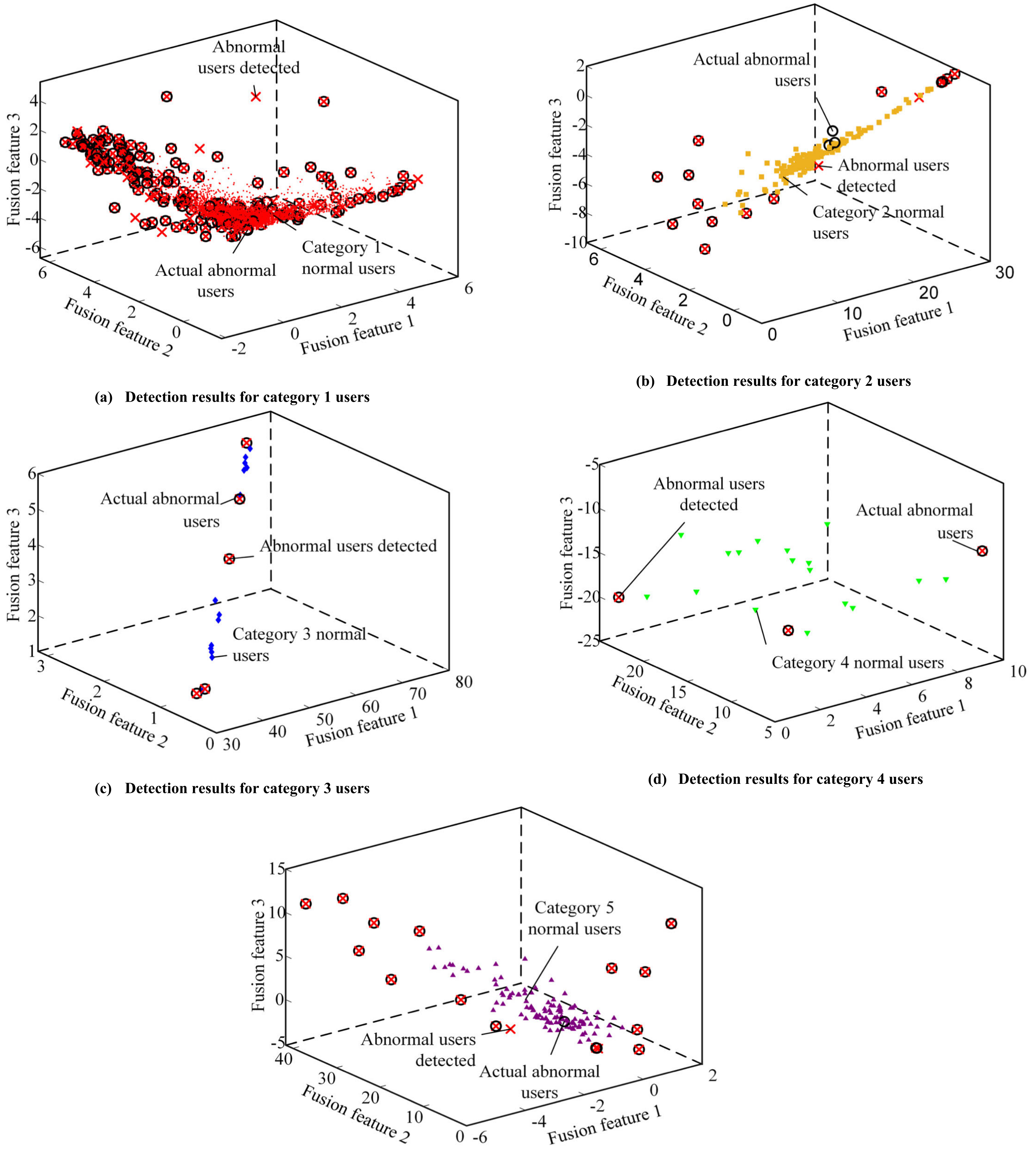


FIGURE 6. Electricity consumption curves for different categories of typical users.

Fig. 5 shows the weighted density improved Canopy-Kmeans algorithm to categorize the users and count the number of users in each category. There are 6097 users in category 1, 169 users in category 2, 19 users in category 3, 20 users in category 4, and 130 users in category 5. The cluster center of a cluster (if the cluster center is not included in the dataset, the user closest to the cluster center is selected) reflects the overall characteristics of all the samples in the category, so the electricity consumption curves of the users in the cluster centers can be used as



(a) Detection results for category 1 users

(b) Detection results for category 2 users

(c) Detection results for category 3 users

(d) Detection results for category 4 users

(e) Detection results for category 5 users

FIGURE 7. Detection results of the PCA model.

the typical daily electricity consumption curves of the corresponding categories, and the daily electricity consumption data of the users in the cluster centers determined by the weighted density-improvement Canopy-Kmeans algorithm are retrieved. The typical electricity consumption curves of different categories of users are shown in Fig. 6. Based on the weighted density improvement Canopy-Kmeans algorithm, the typical daily electricity consumption curves of category

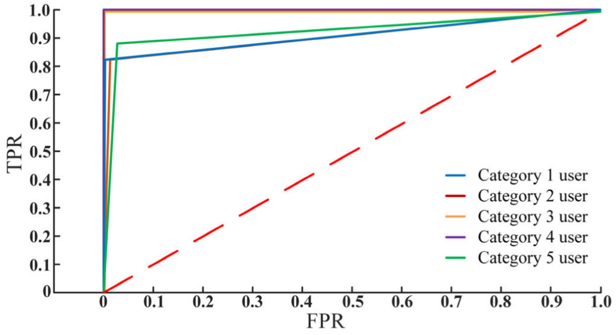


FIGURE 8. ROC curve for 5 types of users.

1 and 2 users do not cross most of the time, and the category attributes of category 2 users are clearer, which is conducive to the sub-clustering of the clustering algorithm; the typical daily electricity consumption curves of category 3, 4, and 5 users cross most of the time, but compared to category 4 and 5 users, the typical daily electricity consumption curves of category 3 users fluctuate most frequently in peaks and valleys; the trends of the typical daily electricity consumption curves of category 4 and 5 users are also different, and the attributes of the categories are clearer.

The fusion features of five types of users after PCA dimensionality reduction are sequentially used as the input data for the anomaly detection of the isolated forest algorithm, and the comprehensive anomaly score of each user is calculated, and it is used as the criterion for the anomaly detection of the electricity consumption of the user.

By referencing the abnormal user labels in the dataset and comparing the detection outcomes with the actual abnormal users, as illustrated in Fig. 7, it is evident that there are detection errors and instances of undetected abnormal users within the results for categories 1, 2, and 5. However, owing to the disparity in category numbers, the detection outcomes for categories 3 and 4 accurately identify the abnormal power users. Overall, the detection rate of 5 user categories under this model increases and the false detection rate decreases.

The ROC curves for outlier detection of the five classes of users are presented in Fig. 8. The difference in the area under the ROC curves (AUC) for category 1, category 2, and category 5 users is minimal, while the AUC for category 3 and category 5 users is the largest.

As shown in Table 4, the evaluation indexes are different for different user types, and the AUC indexes for all five types of users are greater than 0.9, which verifies that the model proposed in this paper can well detect abnormal power users. Owing to the considerable variance in user counts, the evaluation metrics for categories 1 and 2 exhibit slightly lower values compared to the remaining three user types. Nonetheless, the algorithmic enhancements contribute to improved accuracy rates, recall rates, F1 scores, and AUC indices across all user categories in the anomaly detection model.

Table 5 is a comparison table of accuracy rates, recall rates, F1 scores and AUC value of each algorithm. The AUC value of the elbow method improved K-means algorithm is slightly larger than that of the DBCSAN algorithm and the weighted density based improved Canopy-Kmeans and Isolated Forest anomaly detection model proposed in this paper, but the accuracy rates, recall rates, and F1 score of this paper's algorithm are slightly higher than that of the elbow method improved K-means algorithm, and significantly higher than that of the DBCSAN algorithm. Additionally, this paper's algorithm outperforms the SVM algorithm in all metrics. Considering the evaluation indexes of each algorithm comprehensively, this paper's algorithm is excellent in accuracy rates, recall rates, AUC value and F1 value, and has better anomaly detection effect.

TABLE 4. Comparison of model performance based on different types of users.

Users	Accuracy rate/%	Recall rate/%	F1 score	AUC
Category 1 user	88.24	82.32	0.8517	0.9101
Category 2 user	87.50	82.35	0.8485	0.9052
Category 3 user	100	100	1	1
Category 4 user	100	100	1	1
Category 5 user	87.50	87.50	0.8750	0.9287
Total users	88.60	83.42	0.8593	0.9153

TABLE 5. Comparison of model performance based on different types of algorithms.

Algorithm	Accuracy rate/%	Recall rate/%	F1 score	AUC
DBSCAN Algorithm	53.32	52.75	0.5303	0.8884
The SVM algorithm	61.07	69.42	0.6477	0.6144
The elbow method improved K-means algorithm	86.80	82.20	0.8440	0.9225
The algorithm of this paper	88.60	83.42	0.8593	0.9153

To better understand the computational complexity and efficiency of the algorithms in this paper, we compare them with the DBSCAN algorithm and the elbow method improved K-means algorithm, presenting the time consumption of the clustering phase for each algorithm in Table 6. Overall, the clustering runtime of our algorithm is 7.241 seconds, which is slightly longer than that of the DBSCAN algorithm. However, our algorithm reduces the clustering time by incorporating preclustering and achieves more efficient clustering and anomaly detection compared to the elbow method improved K-means algorithm and the DBSCAN algorithm. Nevertheless, our method spends more time on weighted density computation and does not significantly reduce the average runtime compared to the elbow method improved K-means algorithm.

TABLE 6. Comparison of clustering time and average runtime based on different types of algorithms.

Algorithm	Clustering time/s	Average runtime/s
DBSCAN Algorithm	3.491	3.491
The elbow method improved K-means algorithm	8.841	8.841
The algorithm of this paper	3.241	7.241

V. CONCLUSION AND DISCUSSION

Considering the challenge of classifying similar electricity consumption patterns among users, this paper proposes an unsupervised isolation forest anomalous electricity consumption detection model based on weighted density improved Canopy-Kmeans algorithm. First of all, joint composite parametric analysis and principal component analysis are used to carry out feature layer fusion and dimensionality reduction of user electricity consumption, which not only retains the characteristics of the data but also reduces the complexity of the data, and shortens the operation time of the model. Then, in order to avoid the effect of random selection of the two thresholds of T1 and T2, the optimal K-value and initial clustering centers are obtained through the joint multi-layer fusion of feature data density calculation and the maximum weight product method, and the results can clearly characterize various types of electricity users; and the clustering results can clearly characterize the abnormal electricity consumption of various categories. The clustering results can clearly characterize the power users of each category. Finally, based on the fusion mechanism of weighted density improvement Canopy-Kmeans and isolation forest algorithm, jointly construct a multi-layer fusion feature data analysis of user abnormal electricity use detection model. Under the premise of fine user classification, the model detection results are more comprehensive and accurate; for the impact of different user types on the detection effect of the model, the evaluation indexes of various types of users and the complete dataset are compared and analyzed. Additionally, the complete dataset is applied to different algorithms. The results demonstrate that the evaluation metrics of the model proposed in this paper have shown significant improvement.

The limitation of the model proposed in this paper is that the weighted density improved Canopy-Kmeans clustering algorithm has a high computational complexity when the dataset is transformed into values such as sample density, and the computation time is still slightly long in the face of massive load data, and the computational efficiency of the clustering algorithm is expected to be improved in subsequent iterations.

REFERENCES

[1] A. A. Esmael, H. H. da Silva, T. Ji, and R. da Silva Torres, "Non-technical loss detection in power grid using information retrieval approaches: A comparative study," *IEEE Access*, vol. 9, pp. 40635–40648, 2021.

[2] P. Massaferro, J. M. D. Martino, and A. Fernández, "Fraud detection in electric power distribution: An approach that maximizes the economic return," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 703–710, Jan. 2020.

[3] Z. Lin, X. Cui, W. Jin, S. Liu, X. Feng, and H. Feng, "Key technologies of electricity theft detection at consumer side," *Autom. Electr. Power Syst.*, vol. 46, no. 5, pp. 188–199, 2022.

[4] W. O. Amolo, P. Musau, and A. Nyete, "Non-technical power loss reduction and transients stability: Optimal placement of reclosers," in *Proc. IEEE PES/IAS PowerAfrica*, Nairobi, Kenya, Aug. 2020, pp. 1–5.

[5] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2661–2670, May 2019.

[6] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019.

[7] J. Tang, Y. Yang, S. Liu, Y. Zhang, Q. Li, and Y. Yi, "On-line optimization method for phase sequence in station area based on improved support vector machine and non-dominated sorting genetic algorithm-III," *Autom. Electr. Power Syst.*, vol. 46, no. 3, pp. 50–58, 2022.

[8] G. M. Messinis, A. E. Rigas, and N. D. Hatziargyriou, "A hybrid method for non-technical loss detection in smart distribution grids," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6080–6091, Nov. 2019.

[9] D. Wang and K. Yang, "A data generation method for electricity theft detection using generative adversarial network," *Power Syst. Technol.*, vol. 44, no. 2, pp. 775–782, 2020.

[10] M.-M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Hybrid deep neural networks for detection of non-technical losses in electricity smart meters," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1254–1263, Mar. 2020.

[11] N. Zhao, X. Zhang, and L. Zhang, "Overview of imbalanced data classification," *Comput. Sci.*, vol. 45, no. B06, pp. 22–27, 2018.

[12] T. Hu, Q. Guo, and H. Sun, "Nontechnical loss detection based on stacked uncorrelating autoencoder and support vector machine," *Autom. Electr. Power Syst.*, vol. 43, no. 1, pp. 119–125, 2019.

[13] Z. Chijie, Z. Bin, H. Jun, L. Qiushuo, and Z. Rong, "Anomaly detection for power consumption patterns based on unsupervised learning," *Proc. CSEE*, vol. 36, no. 2, pp. 379–387, 2016.

[14] J. Yang, K. Fei, F. Ren, J. Li, Y. Duan, and L. Dong, "Non-technical loss detection using missing values' pattern," in *Proc. ICSGCE*, Kuching, Malaysia, 2020, pp. 149–154.

[15] Y. Xiao, K. Zheng, Z. Yu, M. Zhou, S. Li, and Q. Ma, "Power data anomaly detection based on holt-winters model and DBSCAN clustering," *Power Syst. Technol.*, vol. 44, no. 3, pp. 1099–1104, 2020.

[16] Z. Du, S. Su, Z. Liu, Y. Xue, Y. Yang, and S. Liu, "Second inspection method for electricity theft detection with low false alarm rate based on identification of production and operation status," *Power Syst. Technol.*, vol. 45, no. 2, pp. 97–104, 2021.

[17] S. Hussain, M. Mustafa, T. Jumani, S. K. Baloch, and M. S. Saeed, "A novel unsupervised feature-based approach for electricity theft detection using robust PCA and outlier removal clustering algorithm," *Int. Trans. Electr. Energy Syst.*, vol. 30, no. 11, pp. 4425–4436, 2020.

[18] Y. Zhang and Y. Zhou, "Review of clustering algorithms," *J. Comput. Appl.*, vol. 38, no. 7, pp. 1869–1882, 2019.

[19] S. Rajendran, W. Meert, V. Lenders, and S. Pollin, "Unsupervised wireless spectrum anomaly detection with interpretable features," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 3, pp. 637–647, Sep. 2019.

[20] W. Zhang, X. Dong, H. Li, J. Xu, and D. Wang, "Unsupervised detection of abnormal electricity consumption behavior based on feature engineering," *IEEE Access*, vol. 8, pp. 55483–55500, 2020.

[21] L. Xie, H. Zhang, L. Yang, Y. Yang, L. Wang, and H. Li, "K-means traceability study of abnormal electricity based on graph neural network," in *Proc. 3rd Int. Conf. Electr. Eng. Control Sci. (IC2ECS)*, Hangzhou, China, Dec. 2023, pp. 1609–1613.

[22] S. Zheng, Q. Liang, X. Peng, W. Zhang, and H. Wang, "Research on abnormal electricity consumption behavior identification based on fuzzy clustering," *Electr. Meas. Instrum.*, vol. 57, no. 19, pp. 40–44, 2020.

[23] V. B. Krishna, G. A. Weaver, and W. H. Sanders, "PCA-based method for detecting integrity attacks on advanced metering infrastructure," *Quantum Eval. Syst.*, vol. 2015, no. 1, pp. 70–85, 2015.

[24] X. Cui, S. Liu, Z. Lin, J. Ma, F. Wen, Y. Ding, L. Yang, W. Guo, and X. Feng, "Two-step electricity theft detection strategy considering economic return based on convolutional autoencoder and improved regression algorithm," *IEEE Trans. Power Syst.*, vol. 37, no. 3, pp. 2346–2359, May 2022.

[25] Y. Sun, S. Li, C. Cui, B. Li, S. Chen, and G. Cui, "Outlier detection method for power users' electricity consumption data based on improved Gaussian kernel function," *Power Syst. Technol.*, vol. 42, no. 5, pp. 1595–1606, 2018.

[26] C. Pang, J. Yu, C. Feng, Y. Liu, and Y. Jiang, "Electric load clustering modeling and characteristic analysis based on LSTM autoencoder," *Autom. Electr. Power Syst.*, vol. 44, no. 23, pp. 57–63, 2020.

[27] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2000, pp. 169–178.

[28] J. Li, J. Zhang, G. Mu, Y. Ge, G. Yan, and S. Shi, "Day-ahead optimal scheduling strategy of peak regulation for energy storage considering peak and valley characteristics of load," *Electr. Power Autom. Equip.*, vol. 40, no. 7, pp. 128–133, 2020.

[29] M. Deng, Z. Yu, Z. Deng, Z. Zhang, and Y. Chi, "Study on stealing recognition model based on multi feature fusion," *Comput. Digit. Eng.*, vol. 45, no. 12, pp. 2398–2401, 2017.

[30] X. Yu, L. Sun, Y. Yan, and G. Liu, "A short-term traffic flow prediction method based on spatial–temporal correlation using edge computing," *Comput. Electr. Eng.*, vol. 93, Jul. 2021, Art. no. 107219.

[31] Y. Huang and Q. Xu, "Electricity theft detection based on stacked sparse denoising autoencoder," *Int. J. Electr. Power Energy Syst.*, vol. 125, Feb. 2021, Art. no. 106448.

[32] (May 16, 2011). *Data From the Commission for Energy Regulation(CER)—Smart Metering Project [EB/OL]*. [Online]. Available: https://www.ucd.ie/issda/data/commissionforenergyregulationcer

JIANYUAN WANG received the Ph.D. degree in electrical engineering from Harbin Institute of Technology University, Harbin, China. He is currently a Professor with Northeast Electric Power University.

His research interests include the application of computers in power systems, the application of power electronics technology in power systems, and power quality control technology.

XIAOYAO LI is currently pursuing the M.S. degree in electrical engineering with Northeast Electric Power University, Jilin, China.

His research interest includes abnormal electricity detection of users.