IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 21, 2020, accepted March 7, 2020, date of publication March 11, 2020, date of current version March 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2980079

Unsupervised Detection of Abnormal Electricity Consumption Behavior Based on Feature Engineering

WEI ZHANG, XIAOWEI DONG, HUAIBAO LI, JIN XU, AND DAN WANG
Department of Electrical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
Corresponding author: Xiaowei Dong (182540359@st.usst.edu.cn)

ABSTRACT The detection of abnormal electricity consumption behavior has been of great importance in recent years. However, existing research often focuses on algorithm improvement and ignores the process of obtaining features. The optimal feature set, which reflects customers' electricity consumption behavior, has a significant influence on the final detection results. Moreover, it is not straightforward to obtain datasets with label information. In this paper, a method based on feature engineering for unsupervised detection of abnormal electricity consumption behavior is proposed. First, the original feature set is constructed by brainstorming in the feature engineering step. Then, the optimal feature set, which reflects the customers' electricity consumption behavior, is obtained by features selected based on the variance and similarity between them. After that, in the abnormal detection step, a density-based clustering algorithm, in which the best clustering parameters are selected through iteration and evaluation, combined with unsupervised clustering evaluation indexes, is used to detect abnormal electricity consumption behaviors. Finally, using the load dataset of an industrial park, several typical feature strategies are applied for comparison with the feature engineering proposed in this paper. To perform the evaluation, the label information of abnormal behaviors is obtained by combining the original electricity consumption behavior detection results with abnormal data injections. The abnormal detection method proposed has given good results and outperformed typical feature strategies in an effective and generalizable way.

INDEX TERMS Abnormal detection, electricity consumption behavior, feature engineering, maximal information coefficient, unsupervised learning.

I. INTRODUCTION

With the widespread use of smart meters and the continuous development of advanced metering infrastructures (AMI), utilities are able to acquire fine-grained data about the real electricity consumption of end-users [1], [2]. Moreover, research on anomaly detection for electricity consumption behavior based on data from smart meters has gained extensive attention [3], [4]. Anomaly detection can regulate customers' electricity consumption behavior, reduce the losses of power utilities, and maintain the security of the smart grid [4], [5]. The data from smart meters indirectly reflect customers' electricity consumption behavior, and the essence of anomalous behavior detection is to distinguish abnormal data in the electricity consumption dataset [6], [7]. Generally, the data in abnormal detection include bad data and non-technical loss (NTL) data, which are directly reflected in the readings of smart meters. Nevertheless, there are fundamental differences between bad data and NTL data. Bad data include missing data as well as abnormal electricity consumption patterns caused by force majeure in the process of data collection, transmission, temporary power outages, or business rectification [3]. The generation of bad data is inevitable, objective, and temporary. The data in NTL are generated by electricity theft under specific strategies, including meter tampering, network intrusion, and measurement interruption, which have the characteristics of continuity, subjectivity, and illegality [8].

The problems with bad data, such as missing data, repeated data, and exceptions, are easy to find and handle in the process of data cleaning [9]. However, because of the large amount of data, distinguishing bad data manually without a unified

The associate editor coordinating the review of this manuscript and approving it for publication was Canbing Li.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

standard is subjective, and abnormal electricity consumption patterns in bad data cannot be distinguished effectively from normal patterns. At the same time, unplanned electricity consumption in bad data and electricity theft in NTL have similar performance on load curves and similar detection methods. Therefore, the abnormal electricity consumption patterns in bad data and electricity theft in NTL all belong to the category of abnormal electricity consumption behavior.

Current research on abnormal electricity consumption behavior detection generally follows the process of ''feature acquisition—abnormal detection'' [10]–[12]. A feature, also called an attribute or variable, represents a property of a process or system that has been measured or constructed from the original input variables [13]. Through different feature strategies, a series of electricity consumption behavior features are obtained from the data set and are used to train classifiers or to perform cluster analysis. However, existing research mainly focuses on the improvement of classification and clustering algorithms, while research on feature strategies is not a popular issue. ''Data and features determine the upper limit of machine learning, while models and algorithms only approach this upper limit'' [14]. The feature set, which can reflect customers' electricity consumption behavior based on load data, has played a decisive role in follow-up research [15]. In fact, in two recent research reviews [3], [16], feature engineering is considered to be a challenging field that has been ignored in previous studies.

Feature engineering mainly focuses on the mining and analysis of electricity consumption and load data in the existing research. Features can be divided into two categories according to whether they are interpretable. One is the category of interpretable features, such as the variability index, the daily average load, and the peak-to-valley time. These features generally rely on professional knowledge to execute feature construction or adopt specific strategies to select features from the commonly used electricity consumption feature set. The other is the category of noninterpretable features generated by machine learning or deep learning algorithms, such as a deep confidence network [9], an autoencoder [17], or a convolutional neural network [18]. Obtaining a feature set that can reflect customers' electricity consumption is the first step in detecting abnormal electricity consumption. The features obtained by different feature strategies lead to vast differences in the evaluation of anomaly detection algorithms [19]–[21]. At present, there are several problems in the research of customer electricity use features. The process of constructing the electricity behavior feature set relies heavily on professional knowledge. Features selected by experience have the disadvantages of subjectivity and one-sidedness. Additionally, customers have inherent electricity consumption characteristics, and unified selection of features ignores the differences between customers and can lead to the loss of crucial information. Deep learning algorithms are a popular research topic in feature engineering and are very effective on feature engineering datasets. However, almost all of them are based on a sample data set and require the dataset with labels for model training in supervised learning. The number of layers and parameters of deep learning algorithms means that researchers must carry out experiments continuously, and the extracted features are not interpretable. The deep learning algorithms cannot be adjusted adaptively and have some limitations when facing a new dataset.

After feature acquisition, the methods of detection can be divided into supervised and unsupervised methods according to whether label information is available in the dataset. Supervised methods include various classifiers and neural network models, such as support vector machines [17], [22], extreme learning machines [23], random forests [24], and deep learning algorithms. Unsupervised methods primarily include a variety of clustering algorithms, e.g., k-means clustering [25], fuzzy clustering [26], and other improved clustering methods [2], [27], [28]. However, as mentioned above, the labels of abnormal electricity consumption data in most datasets are difficult to obtain. Moreover, there is no standard to judge whether the customer data represent normal electricity consumption behavior and manual labeling is difficult to work that lacks reliability [3]. Therefore, supervised methods have certain limitations in practice, and unsupervised methods are more suitable for actual needs.

In order to solve the above problems of features acquisition in abnormal electricity consumption behavior detection, an unsupervised abnormal detection method based on feature engineering for electricity consumption behaviors is proposed. The proposed method consists of three parts: data preparation, feature engineering, and anomaly detection. First, data preprocessing is carried out to clean the data, and it is a vital and indispensable step. Next, through feature engineering, the optimal feature subset that reflects consumers' energy consumption behavior is obtained. Then, the best clustering parameters are found by iterative evaluation and cluster evaluation. Finally, anomaly detection is performed by a density-based clustering algorithm. The proposed method is based on unsupervised learning and does not rely on subjective experience and data. The optimal feature set obtained by the proposed method can comprehensively and objectively reflect the user's electricity consumption behavior, and realize abnormal detection.

ORGANIZATION OF THE PAPER

Section II discusses the related concepts of feature engineering and provides an overview of the existing studies related to electricity consumption feature strategies in anomaly detection. Section III proposed the abnormal detection method and elaborates on each component of the method in detail. Section IV shows the experimental results and compares them with previous works. Finally, the conclusion and discussion are discussed in Section V.

A. ABBREVIATIONS AND ACRONYMS

Abnormal data injection (ADI), advanced metering infrastructure (AMI), clustering evaluation score (CES), common feature construction (CFC), deep belief network (DBN),

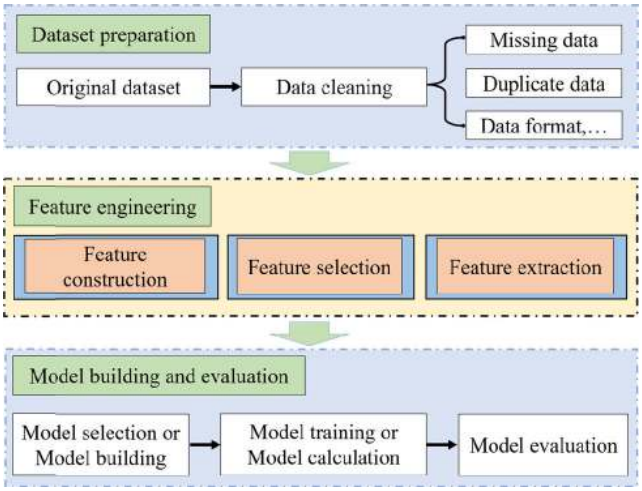


FIGURE 1. The position of feature engineering in the process of data mining.

feature construction (FC), feature extraction (FE), feature selection (FS), false positive rate (FPR), independent component analysis (ICA), linear discriminant analysis (LDA), maximum information coefficient (MIC), minimum number of points (MinPts), maximum-relevance-minimum-redundancy (mRMR), mutual information (MI), non-technical loss (NTL), principal component analysis (PCA), positive predictive value (PPV), restricted boltzmann machine (RBM), true positive rate (TPR).

II. RELATED WORK

This section discusses the related concepts about feature engineering and provides a brief review of the existing studies of feature engineering in abnormal electricity consumption detection.

A. FEATURE ENGINEERING

Feature engineering is the process of using data science knowledge to create feature sets that enable machine learning algorithms to achieve the best performance. Generally, the essence of feature engineering is to transform the preprocessed data, which includes three subproblems: feature construction (FC), feature extraction (FE), and feature selection (FS). The place of feature engineering in the process of data mining is shown in Fig. 1.

FC is based on the original dataset and relies on professional experience to build new features, which are generally interpretable. Usually, the process of FC requires much time to study data samples as well as particular abilities of insight and analysis. For power data, FC needs background information on the power industry, which relies on experience and subjective judgment. The task of FS is to select a feature subset from the original feature set based on specific evaluation criteria [29], [30]. The number of elements in the feature subset should be less than that in the original feature set. In essence, FS is a dimensionality reduction process, and the features themselves do not change. A good FS algorithm can effectively reduce the original feature set dimension, has low computational complexity, and can improve the effectiveness of subsequent classification and clustering [31]. Similar to FS, FE also reduces the dimension of the original feature set. Through dimension reduction, mapping, and other methods, FE keeps the original data information to as great an extent as possible and obtains more abstract and concise feature representation. Common FE methods include PCA, independent component analysis (ICA) and linear discriminant analysis (LDA). Unlike FS, the data themselves will change in the process of FE. Let $\boldsymbol{F} \in \mathbb{R}^{d\times n}$ be a feature set with n features as $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_i, \ldots, \boldsymbol{f}_n\}$, where $\boldsymbol{f}_i \in \mathbb{R}^{d\times 1}$. The process of FS is as follows:

$$\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_i, \ldots, \boldsymbol{f}_n\} \xrightarrow{\text{FS}} \left\{\boldsymbol{f}_{i_1}, \ldots, \boldsymbol{f}_{i_j}, \ldots, \boldsymbol{f}_{i_m}\right\} \quad (1)$$

where $i_j \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}$, $a = b$ can be deduced when $i_a = i_b$ and $a, b \in \{1, \ldots, m\}$. The process of FE is as follows:

$$\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_i, \ldots \boldsymbol{f}_n\} \xrightarrow{\text{FE}} \{g_1(\boldsymbol{F}), \ldots, g_i(\boldsymbol{F}), \ldots, g_m(\boldsymbol{F})\} \quad (2)$$

where $g_i(\boldsymbol{F})$ represents the transformation of feature set $\boldsymbol{F}$ by a series of FE methods, such as PCA, ICA, and LDA. In most cases, not all three subproblems of feature engineering are carried out, but one or two of them are selected according to the needs of model building and analysis.

B. FEATURE CONSTRUCTION AND EXTRACTION

Feature construction and feature extraction are used to build feature vectors of consumers in [32]. An unsupervised algorithm for detecting abnormal electricity consumption patterns is proposed. They use monthly average load data as the dataset and obtain trend index, variability index, and volatility index as feature sets. Then, PCA is used in FE of the feature sets to two dimensions, and the local outlier factor is calculated to find anomalous power consumption patterns. Sun *et al.* [33] proposed an improved outlier detection method based on the Gauss kernel function. After consumers are classified by clustering, the feature set of the electricity consumption behavior of each type of consumer is constructed, such as trend indicators, the standard deviation of daily load series. Then, the feature set is reduced to 2 dimensions by PCA, and the improved Gaussian kernel function detects the outliers. This method is similar to [32] in its feature strategy. An optimum-path forest (OPF) clustering algorithm is proposed in [34]. They take the problem of NTL recognition as an anomaly detection task to analyze and use eight features to represent each electricity customer for clustering analysis.

As mentioned above, several of features constructed are subjective and cannot reflect the inherent electricity consumption behavior of customers. Additionally, the features extracted lack of interpretability.

C. FEATURES OBTAINED BY DEEP LEARNING ALGORITHMS

Zhang *et al.* [9] used a deep belief network based on real-valued (RDBN) to detect electricity theft. The electricity consumption data are divided into a training set and a test set, and dimension reduction is performed by factor analysis. The RDBN is built and trained to obtain electricity consumption behavior features and to realize anomaly detection. Yu *et al.* [10] introduced a feature extraction method based on deep learning theory to detect electricity theft. The stacking uncorrelation autoencoder (SUAE) is proposed and used to extract features from electricity consumption data. It can extract highly abstract and concise features due to its deep structure and high nonlinearity. A convolutional neural network (CNN) based deep learning method is proposed in [18] and used to extract features from massive load profiles automatically.

The features generated by deep learning algorithms lack interpretability and rely on data with labels. At the same time, it makes model building and parameter selection challenging.

D. FEATURE ENGINEERING TO OBTAIN THE OPTIMAL FEATURE SET

R. Razavi *et al.* introduced the concept of feature engineering into anomaly detection applications for the first time [35]. The feature engineering framework proposed was used to create a set of features that could best express the dynamics of a load over time. It was easier to detect abnormalities and fraud behaviors of anomalous households in comparison with similar households. In [36], Lu Jun *et al.* proposed a strategy based on feature information quantity to select the optimized feature set of customers' electricity consumption behavior. According to the mutual information (MI) and the correlation coefficient between features, the optimal feature subset was selected based on the common feature set. Aydin and Gungor [21] and Toma and Li [22] presented two feature construction techniques for NTL characterization. Compared with algorithm improvement, they were more concerned about finding the set of features that best discriminate legal and illegal profiles.

In summary, the problem of how to acquire the optimal feature set, which can reflect users' electricity consumption behavior, is becoming a research hotspot. A series of methods (i.e., feature strategies) of feature engineering in machine learning have gradually become more prevalent in abnormal electricity consumption behavior detection.

III. METHODOLOGY

The abnormal detection method of electrical consumption behavior is proposed firstly in this section, and the conducted of the research is also discussed. After that, the details of the method are discussed in the rest of this section.

A. PROCESS OF THE METHOD

Based on the related concepts of unsupervised learning, feature engineering and the general process of data mining, the abnormal detection method of electricity consumption behavior is proposed, and it is shown in Fig. 2. The method includes three parts: data preparation, feature engineering, and anomaly detection.

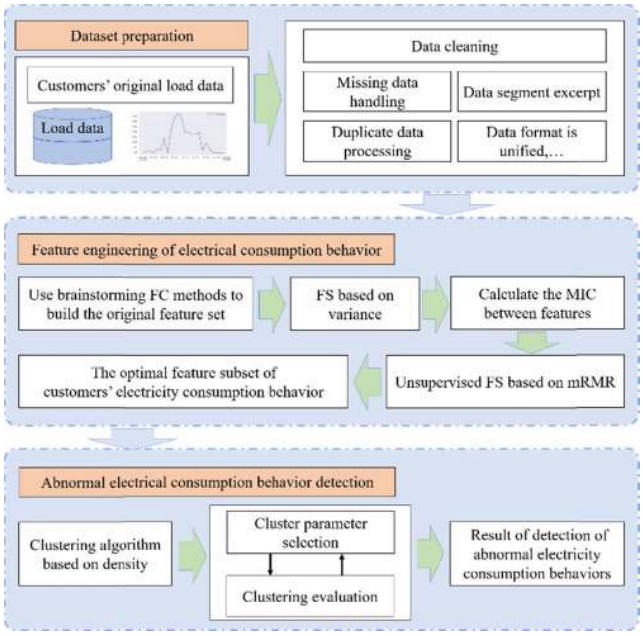


FIGURE 2. The unsupervised method of anomaly detection for electricity consumption based on feature engineering.

In the process of dataset preparation, different methods are selected to address specific problems in the dataset. Feature engineering includes the original feature set construction of electricity consumption behavior, FS based on variance, and FS based on improved maximum relevance minimum redundancy (mRMR). The problem of missing crucial features missing, caused by subjective selection, can be avoided by using the brainstorming method in building the original feature set of electricity consumption behavior. The maximum information coefficient (MIC) is used to replace the MI to measure the correlation and redundancy between features based on the supervised FS algorithm of mRMR. Compared with MI, MIC has the advantages of generality and equitability. In addition, the concept of relevance in mRMR is redefined to apply in unsupervised cases. In the anomaly detection step, a density-based clustering algorithm is adopted, and the best clustering parameters are obtained by iteration based on a certain clustering evaluation standard. Finally, customers' abnormal electricity consumption behavior is detected; the load data of an industrial park are selected for experimental analysis.

B. FC BASED ON TSFRESH

To overcome the problems in the FC process mentioned earlier, a Python package named tsfresh (time series feature extraction based on scalable hypothesis tests) was used to construct the original feature set from the load data set. tsfresh is a Python-based development package that can quickly calculate a large number of time series characteristics, namely,

features [24], [25]. The features calculated by tsfresh can be used in subsequent research, e.g., classification, regression, and clustering. At present, tsfresh provides 63 time-series feature methods, and 794 time-series features can be calculated by changing the methods' parameters [40].

Customers' load data are essentially a time series [3], [26]. Many aspects of the information on electricity consumption behavior can be obtained by constructing the original feature set based on tsfresh. Problems such as missing vital features due to insufficient consideration can be avoided. At present, tsfresh can extract 794 features from the customers' daily load data, including the commonly used electricity consumption features in the existing research. However, when constructing the feature set through tsfresh, the correlation between the features and the labels cannot be considered because of the lack of label information. Unavoidably, there are a large number of redundant features in the original electricity consumption behavior feature set (referred to as the original feature set). The dimension of the original feature set is higher due to the existence of a large number of redundant features, which reduces the efficiency of the algorithm and leads to the ''curse of dimensionality'' [11], [18], [42]. Therefore, to ensure the interpretability of features, it is necessary to carry out the FS process for the original electricity consumption feature set.

C. FS BASED ON VARIANCE

FS based on variance, which belongs to the filtering method in FS, looks only at the features, not the desired outputs, and can thus be used for unsupervised learning. After the variance of features is calculated, the variance of a feature being minimal means that the feature has a low difference in the sample set and a weak ability to distinguish samples. In contrast, if the variance of a feature is maximal, the feature has a higher impact on the overall sample set, including more information that can reflect the differences between samples. For the feature matrix $\boldsymbol{F}$ with d samples and n features

$$\boldsymbol{F} = \begin{bmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & f_{ij} & \vdots \\ f_{d1} & \cdots & f_{dn} \end{bmatrix}_{d\times n} \tag{3}$$

The equation of the ith feature variance is

$$V_j = \frac{\sum_{i=1}^{d} \left(f_{ij} - \mu_j\right)^2}{d} \tag{4}$$

where μ_j is the average value of the j-th feature, and the equation is as follows:

$$\mu_j = \frac{\sum_{i=1}^{d} f_{ij}}{d} \tag{5}$$

To improve the efficiency of the method and reduce information loss, it is necessary to reduce the number of features in the original feature set to a reasonable number. Therefore, the variance of each feature in the original feature set is calculated first. After that, only features with variance equal to zero are removed, and the features with nonzero variance are retained.

D. MAXIMUM INFORMATION COEFFICIENT

The relationship coefficient is usually used to measure the similarity between features. In statistics, the Pearson correlation coefficient and Spearman correlation coefficient are the most well-known measures to calculate the correlation between feature vectors. However, the scope of its application is limited because only linear and simple nonlinear associations can be identified. The correlation between most features is complex and nonlinear, and the traditional correlation measurement method cannot reflect sophisticated associations. The MIC proposed by Reshef *et al.* [43] based on MI can not only find linear or nonlinear correlations but can also widely mine the nonfunctional dependence between vectors. It has been proven that in many cases, the MIC has a better performance than MI. The MIC has been applied in the power industry. For example, in [44], K. Zheng *et al.* used the MIC to find the correlations between the NTL and tampered load profiles of consumers to find abnormal users. The scale of the MIC is [0,1], and a larger value indicates a stronger correlation between two feature vectors. The MIC has the characteristics of generation, equitability, and symmetry. Generation means that when the sample size is sufficient, the MIC can capture a wide range of interesting associations that are not limited to specific function types. Equitability means that the MIC can give similar scores to different types of correlation with equal noise levels. The definition of symmetry is as follows:

$$MIC\left(\boldsymbol{a};\boldsymbol{b}\right) = MIC\left(\boldsymbol{b};\boldsymbol{a}\right) \tag{6}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are two vectors. Given a finite set D, the elements of which are data points with two dimensions of x-values and y-values. An $x-by-y$ grid G can be created in finite two-dimensional space based on the supposition that the x-values of D can be partitioned into x bins, and the y-values of D can be partitioned into y bins. Let all elements in D be arranged in the grid G. Let $D|_G$ represent the distribution of D divided by one of $x-by-y$ grids as G. Calculate the MI under each grid division as follows:

$$I^*\left(D, x, y\right) = \max_{G} I\left(D|_G\right) \tag{7}$$

where $I\left(D|_G\right)$ denotes the MI of $D|_G$, $D \in \mathbb{R}^2$, and $x, y \in \mathbb{N}^*$. Because the division of grid G is infinite, the number $I\left(D|_G\right)$ is also infinite, and the largest representation is $I^*\left(D, x, y\right)$. The characteristic matrix $\boldsymbol{M}$ is composed of the maximum normalized MI obtained under different grid divisions, which is defined as follows:

$$\boldsymbol{M}\left(D\right)_{x,y} = \frac{I^*\left(D, x, y\right)}{\log\min\left\{x, y\right\}} \tag{8}$$

where the $\log\min\left\{x, y\right\}$ is the maximal possible MI of a distribution on an $x-by-y$ grid [45]. Normalizing by the $\log\min\left\{x, y\right\}$ creates a score that can be compared across

grids with different dimensions and therefore across different distributions. It also ensures that almost all noiseless functions receive perfect scores and the entries of the characteristic matrix $\boldsymbol{M}$ range from zero to one [43]. Furthermore, the MIC of a finite set D of two-variable data with sample size n is given by

$$MIC\,(D) = \max_{xy<B(n)} \left\{ \boldsymbol{M}\,(D)_{x,y} \right\} \tag{9}$$

where $B\,(n)$ is the upper bound of the grid size. This paper uses $B\,(n) = n^{0.6}$ because it has better performance in practice [43], [44], [46], [47].

E. FS BASED ON IMPROVED mRMR

This section proposes an improved algorithm based on the supervised FS algorithm mRMR [48], which can be used in the unsupervised case. The algorithm selects the MIC to replace the MI as the feature coefficient evaluation metric and defines relevance in FS without label information. The mRMR maximizes the relevance between features and category variables, i.e., class labels, and minimizes the redundancy between features in the constructed feature subset. The m-th feature vector is selected based on mRMR from the original feature set $\boldsymbol{F}$ and represented by $\boldsymbol{f}_{g_m}$ as follows:

$$\boldsymbol{f}_{g_m} = \max_{\boldsymbol{f}_i \epsilon \boldsymbol{F}-\boldsymbol{S}_{m-1}} \left\{ I\left(\boldsymbol{f}_i; \boldsymbol{c}\right) - \frac{1}{m-1} \sum_{\boldsymbol{f}_t \in \boldsymbol{S}_{m-1}} I\left(\boldsymbol{f}_i; \boldsymbol{f}_t\right) \right\} \tag{10}$$

where $\boldsymbol{S}$ represents the selected feature subset, $I\,(\cdot)$ denotes the MI between a pair of vectors, and $\boldsymbol{c}$ is the class label variable. Due to the difficulty in obtaining the class label $\boldsymbol{c}$, the unsupervised FS is based on the mRMR algorithm proposed, which should meet the following requirements:

1. The m-th feature selected should contain the largest amount of information, i.e., this feature can minimize the uncertainty of other features;

2. In the optimal feature subset $\boldsymbol{S}$, the relevance between features should be as small as possible, which will minimize the redundancy between the selected features.

The unsupervised FS algorithm starts from an empty set and uses a step-by-step method to select a feature from the original feature set $\boldsymbol{F}$ and add it to the feature subset $\boldsymbol{S}$. The first feature selected $\boldsymbol{f}_{g_1}$, which has the maximum average MIC value, found by calculating the MIC between features in the original feature set $\boldsymbol{F}$, is defined as follows:

$$\boldsymbol{f}_{g_1} = \max_{\boldsymbol{f}_i \epsilon \boldsymbol{F}} \left\{ \frac{1}{n} \sum_{t=1}^{n} MIC\left(\boldsymbol{f}_i; \boldsymbol{f}_t\right) \right\} \tag{11}$$

where $n = |\boldsymbol{F}|$ represents the number of features in $\boldsymbol{F}$. The first feature has the highest relevance to the remaining features, which can reduce the uncertainty of other features to the greatest extent and provide the most information compared with the others. Therefore, the relevance of a feature is defined as the average value of the MIC between the feature and the other features in the set.

The m-th feature selected mainly considers the following two aspects: containing the largest amount of information with other features in the set $\boldsymbol{F}$ and having the smallest degree of redundancy with other features in the subset $\boldsymbol{S}$. By using the incremental search, the search for the m-th feature is written as an optimization problem as follows:

$$\boldsymbol{f}_{g_m} = \max_{\boldsymbol{f}_i \epsilon \boldsymbol{F}-\boldsymbol{S}_{m-1}} \left\{ \frac{1}{n} \sum_{t=1}^{n} MIC(\boldsymbol{f}_i; \boldsymbol{f}_t) - \frac{1}{m-1} \sum_{j=1}^{m-1} MIC\left(\boldsymbol{f}_i; \boldsymbol{f}_j\right) \right\} \tag{12}$$

where $m-1$ represents the number of features in feature subset $\boldsymbol{S}$ before feature selection. To make the algorithm take both efficiency and effect into account, the termination condition of FS is defined as follows:

$$\frac{(m-1)\sum_{t=1}^{n} MIC\left(\boldsymbol{f}_{g_m}; \boldsymbol{f}_t\right)}{n\sum_{j=1}^{m-1} MIC\left(\boldsymbol{f}_{g_m}; \boldsymbol{f}_j\right)} \leq T \tag{13}$$

For the selected m-th feature, when the amount of information it contains has less redundancy than a certain threshold, the selection is stopped. The flow of the unsupervised FS algorithm based on improved mRMR is shown in Algorithm 1.

FS is usually in an independent part of the data mining process, which cannot benefit from the previous data exploration process. Information loss is inevitable with the feature dimension reduction in the FS process, but it can reduce the computational complexity of the process.

F. CLUSTERING ALGORITHM

Density-based spatial clustering of applications with noise (DBSCAN) is a classical clustering algorithm based on density that is widely used in anomaly detection [49]. Compared with k-mean and partition-based clustering algorithms, DBSCAN can find any shape of clustering without setting the number of clusters in advance, and it is not sensitive to the order of data points. At the same time, it can detect outliers in the process of clustering.

Given a set of points in a certain space, DBSCAN can divide the points in the high-density area into a group and mark the outliers in the low-density area, i.e., outliers. DBSCAN needs to set the following two neighborhood parameters: the ε-neighborhood and the minimum number of points (represented by MinPts) needed to form a high-density area. The ε-neighborhood describes the neighborhood distance threshold of a sample, and MinPts describes the threshold of the number of samples in the ε-neighborhood. In short, the basic idea of the algorithm is to explore the ε-neighborhood of any point that is not visited. If the number of points in the ε-neighborhood reaches MinPts, a new cluster is established. Otherwise, the point is labeled as an outlier.

Algorithm 1 Unsupervised FS based on improved mRMR

Input: Feature set $\boldsymbol{F} = \{f_1, \ldots, f_i, \ldots, f_n\}$ of customers' electricity consumption behaviors, where n represents the total number of features, which is determined by FC based on tsfresh and FS based on variance.

Output: The optimal feature subset $\boldsymbol{S}$, which reflects the customers' electricity consumption behavior.

Begin:
Initialization:
$\boldsymbol{S} = \emptyset$
$\boldsymbol{F} = \{f_1, \ldots, f_i, \ldots, f_n\}$
MIC_martix $= O_{n\times n}$
/* Calculate MIC between features in set $\boldsymbol{F}$*/
for all $f_i, f_j \in \boldsymbol{F}$ **do**
    MIC_martix[i][j] $= MIC(f_i; f_j)$
    MIC_martix[j][i] = MIC_martix[i][j]
end for
Find the first feature f_{g_1} according to (11)
$\boldsymbol{S} = \{f_{g_1}\}$
$\boldsymbol{F} = \boldsymbol{F} \setminus f_{g_1}$
while True:
    Find the feature f_{g_m} in $\boldsymbol{F}$ according to (12)
    if f_{g_m} satisfies the condition of (13) **do**
        break
    else:
        $\boldsymbol{S} = \boldsymbol{S} \cup \{f_{g_m}\}$
        $\boldsymbol{F} = \boldsymbol{F} \setminus f_{g_m}$
end while
Return S
End

G. EVALUATION INDEX

The unsupervised evaluation indexes are selected to evaluate clustering performance, e.g., the silhouette coefficient (SC), the Calinski Harabasz index (CHI), and the Davies-Bouldin index (DBI). The higher the scores of SC and CHI, the lower the scores of DBI, indicating that the clustering algorithm has better-defined clusters and better separation between the clusters. The scores of indexes under different clustering parameters are obtained by iteration. The ranking sum of the three index scores under different clustering parameters is used to evaluate clustering to avoid the problem of different index magnitudes. The clustering evaluation scores (CES) are defined as follows:

$$\mathrm{CES} = \mathrm{rank}_{\mathrm{SC}} + \mathrm{rank}_{\mathrm{SHI}} - \mathrm{rank}_{\mathrm{DBI}} \tag{14}$$

The clustering parameters with the highest CES will be used for clustering and obtain the final anomaly detection results.

In order to realize methods evaluation and comparison, the confusion matrix and several derived indexes are used. The confusion matrix is shown in Table 1. FN refers to the actual abnormal electricity consumption behavior that is detected as normal electricity consumption behavior, and FP refers to the actual normal electricity consumption behavior that is detected as abnormal. TP and TN indicate correct detection, i.e., the actual abnormal behavior is detected as abnormal, and the actual normal behavior is detected as normal. Several evaluation criteria can be derived from the confusion matrix and used to evaluate the results of different feature strategies methods. The true positive rate (TPR), also known as the sensitivity or recall, is defined as the proportion of detecting as abnormal in actual abnormal electricity consumption behavior. The false positive rate (FPR) is defined as the proportion of detecting as abnormal in actual normal electricity consumption behavior. The positive predictive values (PPV) also known as the precision is defined as the proportion of actual abnormal electricity consumption behavior in detected as abnormal electricity consumption. TPR, FPR, and PPV are defined as follows:

TABLE 1. Confusion matrix for abnormal electricity consumption behavior detection.

		Detection	
		Abnormal	Normal
Actual	Abnormal	TP (true positive)	FN (false negative)
	Normal	FP (false positive)	TN (true negative)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{15}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \times 100\% \tag{16}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\% \tag{17}$$

In many cases, there is an imbalance between positive and negative samples in the data set, and TPR (i.e. recall) and PPV (i.e. precision) are contradictory in most cases. If the method wants to detect more abnormal electricity consumption behavior, it will be possible to detect more normal. Therefore, there will be higher TPR and lower PVV. On the contrary, if the method is relatively conservative, only certain samples are detected. The method will have a higher PVV and a lower TPR. Therefore, the F1 Score (also called balanced F score) and FPR are selected as the evaluation criteria of the method in this paper. The F1 Score is defined as the harmonic mean of TPR and PPV, which can be used to evaluate the performance of unbalanced label data set. The F1 Score is defined as follows:

$$\mathrm{F1} = \frac{2 \times \mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} \tag{18}$$

F1 Score helps to compute TPR and PPV in one equation so that the problem to distinguish the models with low TPR and high PVV or vice versa could be solved.

IV. EXPERIMENT

The load dataset of an industrial park is used for analysis in this paper. First, inherent electricity consumption habits and abnormal electricity consumption behavior of different users are detected through the method proposed. Then, to evaluate the method, six methods given in Appendix A are used

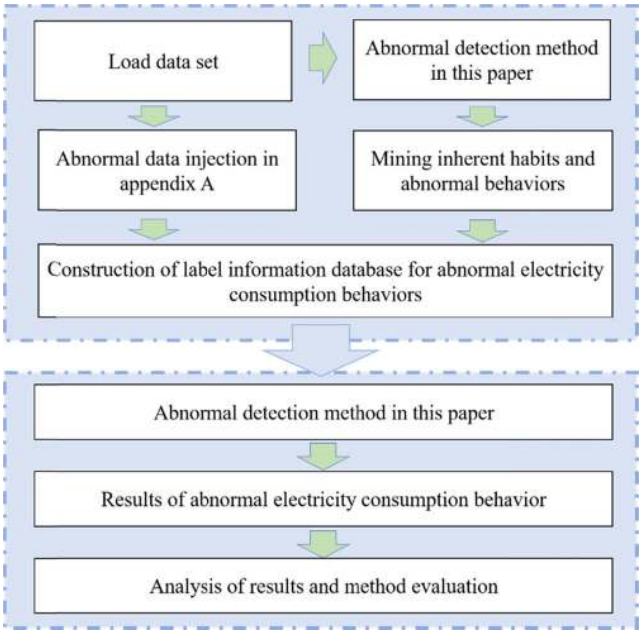


FIGURE 3. The flow chart of the experimental stage.

TABLE 2. Basic information on the data set.

Customer type	Number	Period	Days	Sampling interval	Data type
Industrial and commercial	15	2015-2017	547	1 hour	Load data

to inject abnormal electricity consumption data. After that, the label information database is established by combining the original abnormal electricity consumption behavior and the abnormal data injection information. Finally, with the label information database, the feature engineering method proposed and several common feature strategies are used for comparison and evaluation. The experiment shows that the method based on the feature engineering proposed is better than the others by comparing the clustering evaluation results. The flow chart of the experimental stage is shown in Fig. 3.

A. EXPERIMENTAL SETUP

All experiments are implemented on a single computer with a CPU of 2.6 GHz and a memory of 16 GB. Experiments and results are completed with the jupyter notebook compiling environment with Python version 3.6 and the relevant Python library.

B. DATA PREPARATION

The basic information about the data set is shown in Table 2; it includes 15 industrial and commercial electricity users, and the data were collected over a period of 547 days with a sampling interval of 1 hour. The anomaly detection work in this paper is accurate to the unit of days. Four typical users are selected from the data set for analysis, numbered User1 to User4.

TABLE 3. Main problems in the data cleaning process.

Problems	Processing method
Data loss in a long time span	Delete missing data
Data with duplicate date index	Delete duplicate data
Data of less than 24 sampling points in a day	Interpolation fill
Missing data one day	Delete corresponding day
Inconsistent time format	Unified time format

Data cleaning plays a vital role because the data set is of raw data that has not been processed. The main problems found in the process of data cleaning are shown in Table 3.

C. FEATURE ENGINEERING FOR USERS' ELECTRICITY CONSUMPTION BEHAVIOR

The tsfresh package is used to construct the original feature set in days. Seven hundred ninety-four electricity consumption behavior features are extracted for each customer, and the feature matrix of size 547*794 is generated. The original feature set of electricity consumption behavior on the ith day is expressed as follows:

$$\left[f_1^i, \cdots f_j^i, \cdots f_{794}^i\right] \tag{19}$$

The Gauss kernel function is used to estimate the variance distribution of the 794 column features, as shown in Fig. 4. According to Fig. 4, the logarithms of the feature variance in different customers' original feature sets have a similar distribution, which is mostly concentrated between [−20,20], and most values are equal to 0. Although the FS based on variance has dramatically reduced the dimension of the original feature set, there is still much redundant information between the features. Therefore, FS based on improved mRMR is needed. The MIC between feature sets is calculated after FS based on variance is performed, and the time complexity is $O\left(n^2\right)$. Therefore, only the upper triangle (or lower triangle) matrix of the MIC matrix is calculated to reduce calculation time based on the symmetry of MIC in (6). The MIC matrix is obtained, as shown in Fig. 5, where the parameters for the MIC are selected according to [43].

Different colors in Fig. 5 correspond to different values of the MIC. When there is no correlation between the two features, the MIC is 0, which is shown in red. When there is a certain correlation relationship between features, the MIC is 1, which is shown in blue. From the color density and the sum of all elements in the customers' MIC matrix, it can be seen that there are apparent differences in the feature sets between customers. The MIC of the User2 feature set is the smallest, and the proportion of red is the highest, which means that the correlation between features is the lowest. In contrast, the MIC of the User4 feature set is the largest, and the proportion of blue is the highest, which means that the correlation between features is the highest.

The FS based on mRMR is conducted based on the MIC matrix. A smaller number of features will result in missing crucial information; on the contrary, a large number of features will increase the running time and cause

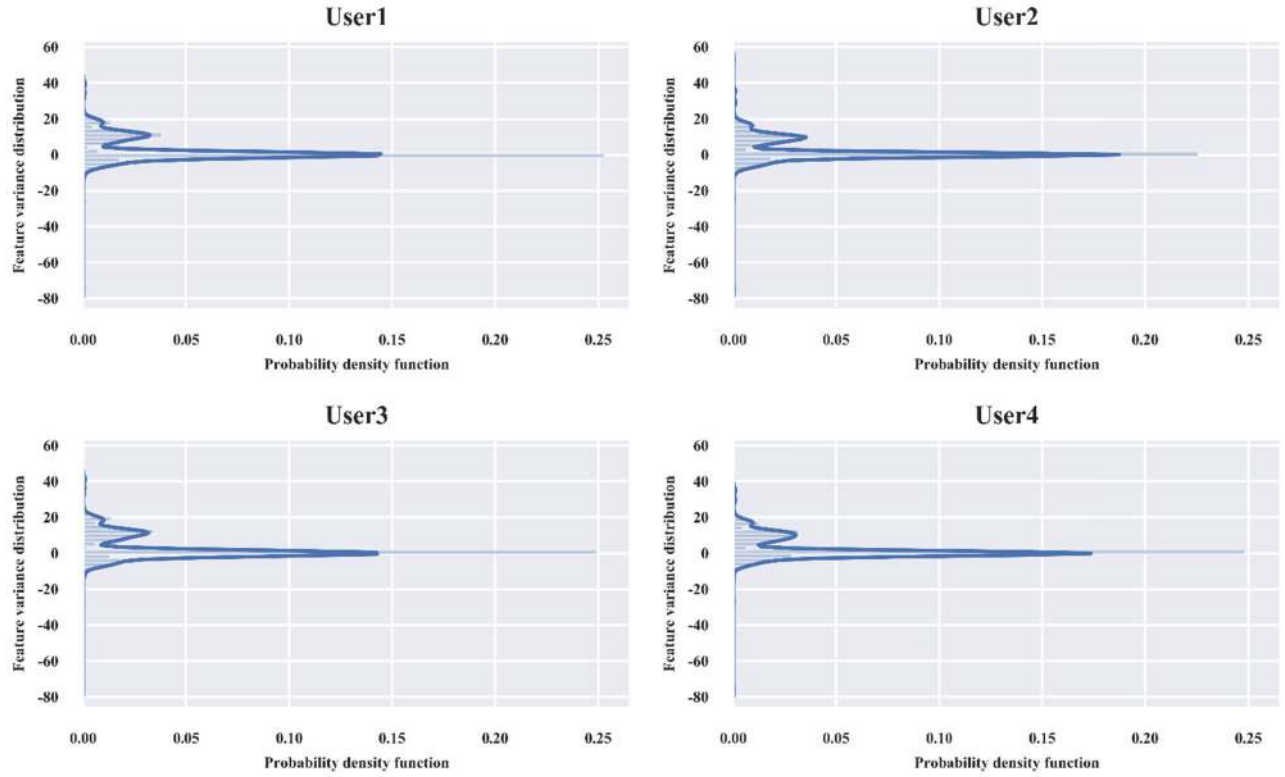


FIGURE 4. Estimation of the logarithm of the feature variance in the original feature set based on the Gaussian kernel function.

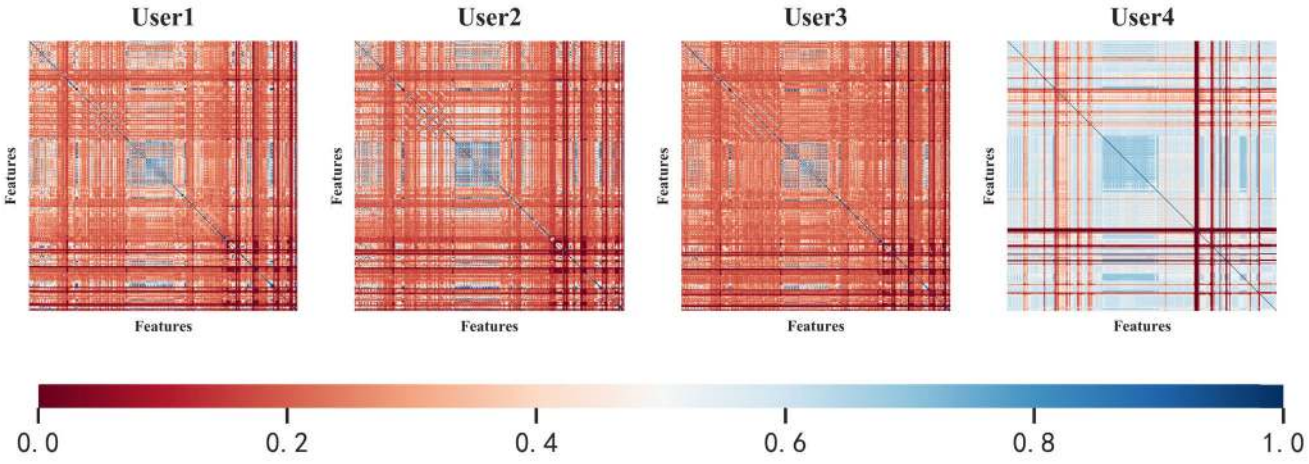


FIGURE 5. Heatmap of the MIC between features in the feature set after FS based on variance.

dimension disaster. Therefore, setting a reasonable value of the termination parameter T in (13) is very important. After many experiments and analyses, when $T = 1$, the proposed method can meet the requirements of accuracy and generalizability. That is, when the correlation of the m-th feature is less than its redundancy, the FS is stopped.

From Table 4 and Fig. 5, User2, with the lowest sum of the MIC matrix, has the highest number of features after FS. In contrast, User4, with the highest sum of the MIC matrix, has the lowest number of features ultimately obtained. Therefore, the conclusion is that more features are needed to reflect customers' electricity behavior when the correlation between features in the electricity behavior feature set is lower. A smaller number of features are needed when the correlation between features is higher. The violin plots of the features in the optimal feature set of customer electricity consumption behaviors are shown. Those of User1 and User4 are shown in Fig. 6, and those of User2 and User3 are shown in Appendix B.

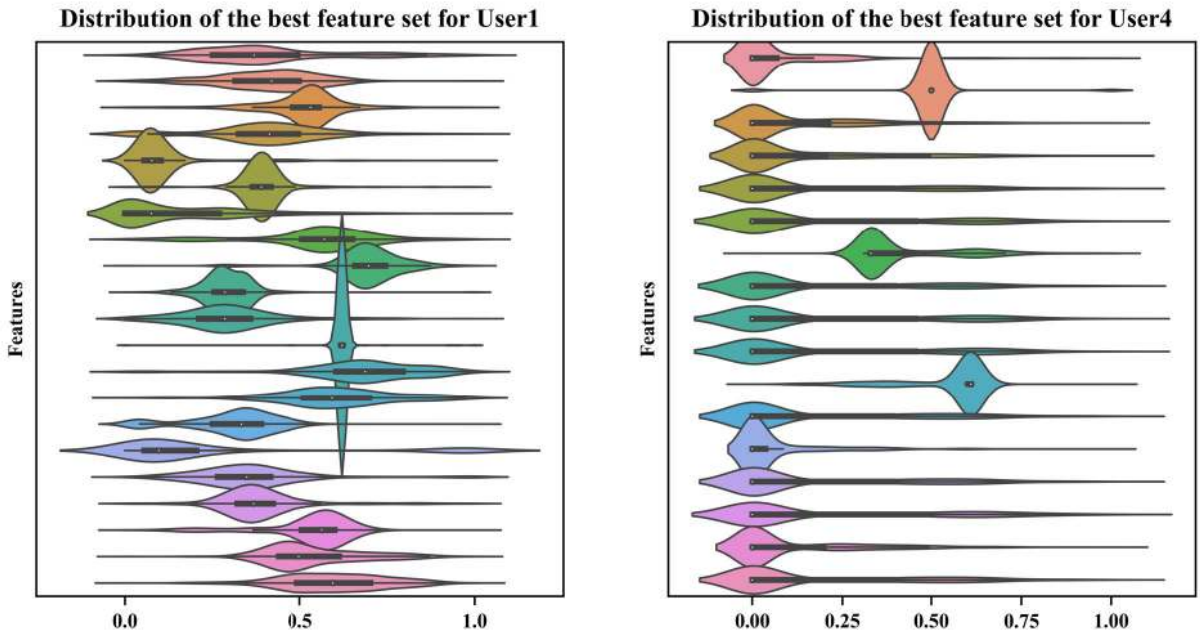


FIGURE 6. The violin plot of the features in the optimal feature sets of user1 and user4.

TABLE 4. Change in users' feature numbers in the process of feature engineering.

Feature engineering stage	User1	User2	User3	User4
FC based on tsfresh	794	794	794	794
FS based on variance	349	348	344	364
FS based on improved mRMR (termination parameter $T = 1$)	21	71	38	17

A violin plot plays a similar role as a box-and-whisker plot, which is used to visualize the distribution of the data and its probability density. A violin plot shows the distribution of quantitative data across several levels of more categorical variables so that those distributions can be compared [50]. The violin plot shows a kernel density estimation of the underlying distribution rather than actual data points. The violin plot in Fig. 6 indicates that the distributions of features in the optimal sets of electricity consumption features of different users are quite different. Each ''violin'' in Fig.6 and Appendix B. represents a feature in the optimal feature set of a user. The shape represents the density estimate of the feature, i.e. the more data points in a specific range, the larger the violin is for that range. Different colors are used to distinguish different features in a user's optimal feature set.

The violin plot in Fig. 6 and Appendix B is a combination of a box plot and a density plot that is rotated and placed on each side, to show the distribution shape of each feature. The white dot in the middle is the median value and the thick black bar in the center represents the interquartile range. The whiskers show a 95% confidence interval and the shape of the violin display frequencies of values. The features of Users 1 to 3 have a scattered density distribution, which indicates an excellent ability to distinguish the different electricity consumption behaviors. It means that the user's electricity consumption behavior can be well described by the features in the optimal feature set obtained by the feature engineering in this paper. That is, if the features extracted by a user have a similar distribution, i.e., they have a similar median and frequency, the different electricity consumption behaviors of the user cannot be reflected from the features. It shows that this feature set cannot effectively distinguish the abnormal electricity consumption behaviors from the normal. On the contrary, if the distribution of multiple features of the user is quite different, the extracted features can effectively describe the user's electricity consumption behavior, then it can deeply mine the user's electricity consumption law and realize abnormal detection. However, the violin plot of User4 is different from that of other users because of the characteristics of the load curve itself, and the details will be explained later. The results show the effectiveness of the feature engineering strategy proposed.

The optimal feature subset, which reflects the electricity consumption behavior of different users, is obtained through the feature engineering strategy proposed. Moreover, the features in the optimal feature subsets of users are different because users have different electricity consumption behaviors.

D. DETECTION OF ANOMALOUS ELECTRICITY CONSUMPTION BEHAVIOR

The feature vector based on days is standardized before clustering to eliminate the weight difference between different

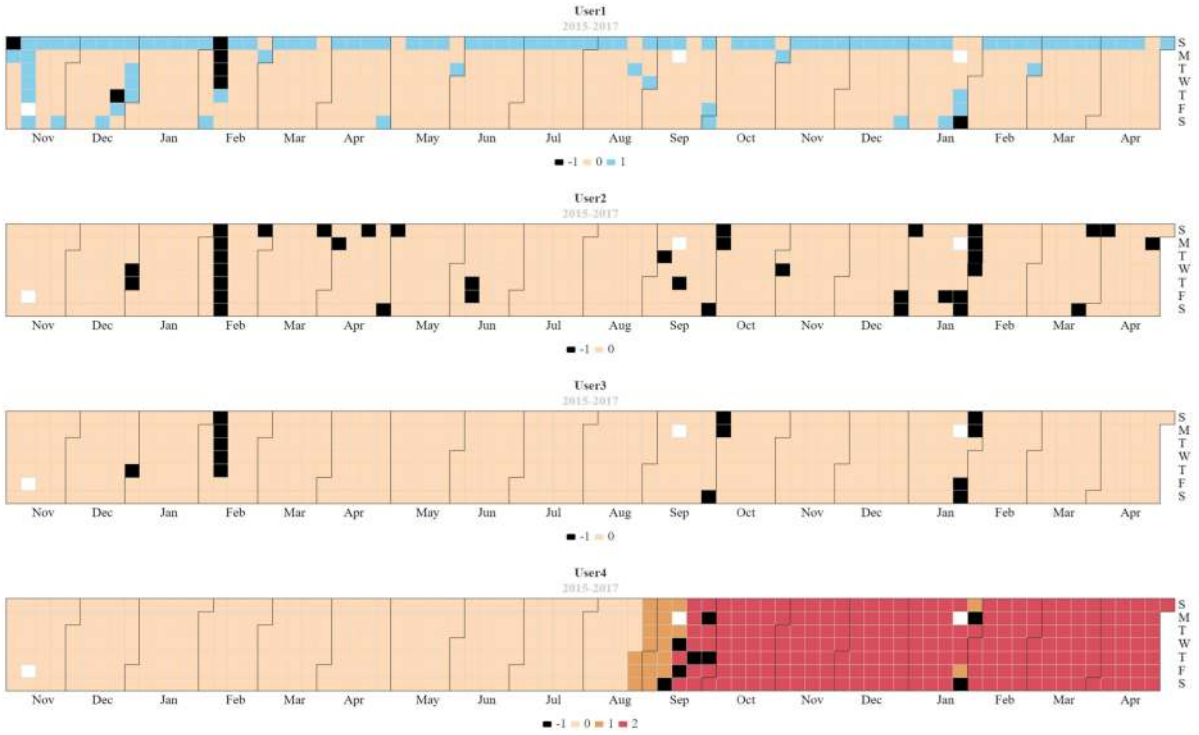


FIGURE 7. Calendar chart showing the electricity consumption behaviors of different users.

TABLE 5. Optimal DBSCAN clustering parameters for each user.

Clustering parameters	User1	User2	User3	User4
ε-neighborhood	0.501	0.901	0.001	0.651
MinPts	7	7	9	7

features. The standardization method is as follows:

$$f_j^i = \frac{f_j^i - \mu^i}{\sigma^i} \tag{20}$$

where f_j^i is the j-th feature on the i-th day, σ^i and μ^i are the standard deviation and mean value of the feature vector of the i-th day.

Through the iterative clustering algorithm parameters, multiple clustering results are obtained based on the CES, and the optimal clustering parameters are selected for different customers as shown in Table 5.

Fig. 7 shows the final clustering results reflected in the calendar chart. Colors represent different clustering results, i.e., different electricity consumption behaviors. The dates of detected abnormal electricity consumption behaviors are unified and marked in black with the category mark -1. The peach color is used to represent the most common kind of electricity consumption behavior of users, which has the category mark 0. The sky blue shown for User1 has category mark 0, and the orange and red for User4 have category marks of 1 and 2, respectively, representing other electricity consumption behaviors of the users. In addition, data loss is represented with a blank mark. As shown in Fig. 7, users have habitual electricity consumption behaviors. User1's electricity consumption behavior on Sundays is grouped into one group. Most of the abnormal electricity consumption behavior of User2 occurs at the beginning and the end of the month. User4's electricity use behaviors are grouped into three groups. During the holidays, User1, User2, and User3 all have abnormal electricity consumption behavior. For visual display, the clustering results are reflected in the original load curve, as shown in Appendix C.

Using the best clustering parameters, shown in Table 5, the normalized sum of the three indexes in CES are adopted to evaluate whether the termination parameter T is reasonable. The variable T directly affects the number of features in the optimal feature subset for clustering. With a change in the feature number for clustering, i.e., the variable T, the clustering result changes accordingly, and the results are shown in Fig. 8. In order to compare different users horizontally, the sum of three indexes in CES under different feature numbers are normalized and represented as the cluster evaluation, which is defined as follow:

$$\mathrm{CES}^* = \frac{\mathrm{CES}_i - \min(\mathrm{CES})}{\max(\mathrm{CES}) - \min(\mathrm{CES})} \tag{21}$$

where CES^* represents the cluster evaluation in Fig. 8, CES_i represents the CES under the number of features i. The normalized cluster evaluation changes as the number of features used for clustering increases from 1 to 100 in Fig. 8. When $T = 1$, the number of features used for clustering by different

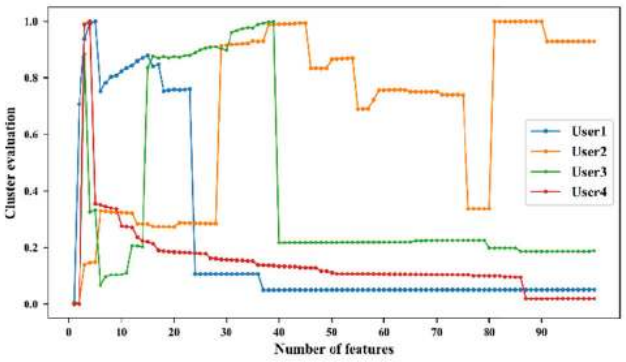


FIGURE 8. Changes in the cluster evaluation with the number of features.

users is shown in Table 4. Under the number of optimal feature subsets, the cluster evaluations of different users are all at a high level, which ensures that the algorithm has generalizability.

As shown in Fig. 8, when the number of features is in a small range (this means that the variable T is in a high range.), the clustering evaluations of Users 1 to 4 are all at a high level, such as User 1, User 3, and User 4 when the number of features ranging from 1 to 5, and User 2 when the number of features ranging from 5 to 10. Although it has a good cluster evaluation when there are few features, it is difficult to obtain an accurate value of T due to the higher cluster evaluations only exist in a small range of changes in the number of features. According to Fig. 8, when the number of features used for clustering analysis reaches a specific critical value, the cluster evaluations drop sharply due to dimension disaster, such as User1 when the number of features is 20 to 25, User2 when the number of features ranging from 70 to 80, User3 when the number of features ranging from 35 to 40, and User4 when the number of features ranging from 85 to 90.

Considering the generalizability and robustness of the method, the stage of feature number range where the cluster evaluation changes relatively smoothly and having a higher level is chosen. Therefore, it is reasonable to choose the termination parameter T in (13) as 1.

E. COMPARISON AND EVALUATION

For each user, 10% of the load data in units of days are randomly selected and changed by the six abnormal data injection (ADI) methods in Appendix A, which are represented as ADI1 to ADI6. Meanwhile, the dates of ADI are recorded, and the label information database is established by factoring in the original abnormal electricity consumption behavior dates. It is worth noting that the label of abnormal electricity consumption behaviors is only used for evaluation, and the label information is not included in the detection process.

Two feature strategies, FE and FC, which are relied on experience and selected as contrasts to the feature engineering method given in this paper. In FE, three feature extraction methods, PCA, restricted boltzmann machine (RBM) and deep belief network (DBN), are used for comparison. The PCA selects the parameters as 2 and 3, i.e., the original load data in days will be reduced to 2 and 3 dimensions after normalization, and the results are represented by PCA_2d and PCA_3d. The RBM is a generative stochastic artificial neural network that can learn a probability distribution through input data sets. The standard RBM consists of binary hidden layer and visible layer without connections between the units within the same layer. RBM can extract discriminative features from complex data set by the unsupervised way due to the introduction of hidden units [51]. The DBN is a kind of probability generating model including multiple hidden layers, which can be regarded as a composite model composed of multiple simple unsupervised learning models [52], [53]. The high flexibility of DBN makes it possible to learn discriminant features from the high-dimensional complex data set. RBM and DBN are all belonged to deep learning and can be used in unsupervised learning. They can reduce the dimension of the original data and achieve feature extraction. The RBN and DBN structures in this paper are shown in Appendix D.

In FC, the common features are constructed by experience and are represented by CFC (common feature construction). The features in CFC are commonly used in current research to reflect electricity consumption behaviors. A total of ten features are included in CFC, including maximum and minimum values of the daily load, the daily peak-valley difference of the load, the difference of load variation, the daily average power consumption, the linear degree of the load curve, and the degree of load deviation. All of these can be obtained from the original feature set generated by tsfresh. The MI is also used in the process of FS based on improved mRMR to prove that the MIC has better performance than the MI. The F1 Score and FPR are selected as indicators to verify the detection ability of several feature strategies for six kinds of ADI.

Fig. 9 consists of six columns, each of which represents an ADI method and includes two subplots, one for F1 Score and the other for FPR, under different feature strategies. F1 Score and FPR of different users under different ADI are displayed in the form of histogram. Different colors are used to distinguish the results of different feature strategies.

Furthermore, the experimental results were analyzed as follows. For the feature strategy of PCA, the difference of the parameters (referring to the final feature dimension) has a significant influence on the result evaluation indexes F1 Score and FPR. The experimental results show the process of FE by PCA. When the parameter of PCA is selected as 3, the evaluation results are better than they are when the parameter is selected as 2, in most cases. Reducing to 2 dimensions means losing more information on the original load data than reducing to 3 dimensions, which leads to the loss of more crucial information that can distinguish abnormal electricity consumption behavior. Therefore, the follow-up clustering is affected, leading to the above results. Note that although the

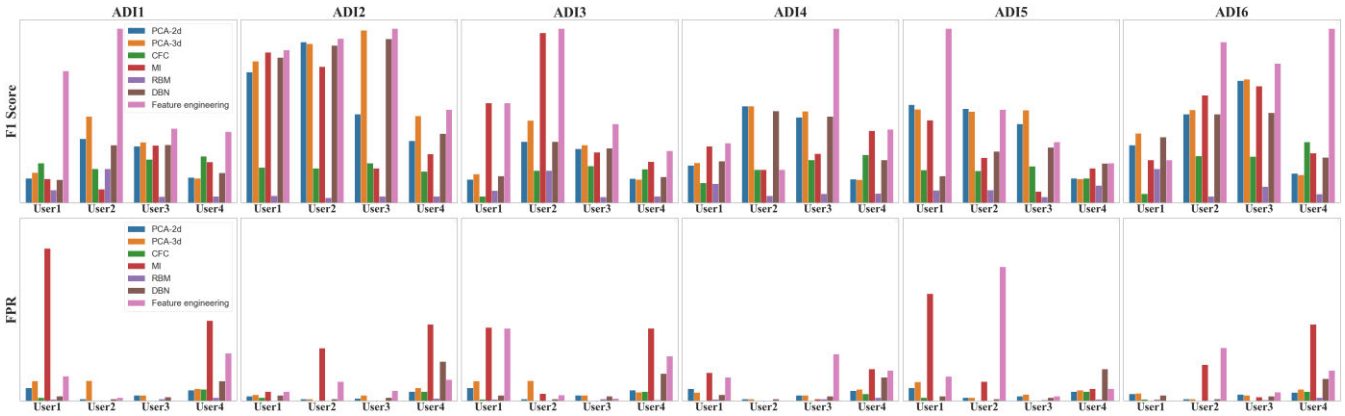


FIGURE 9. Values of F_1 score and FPR for each user under ADI1 to ADI6.

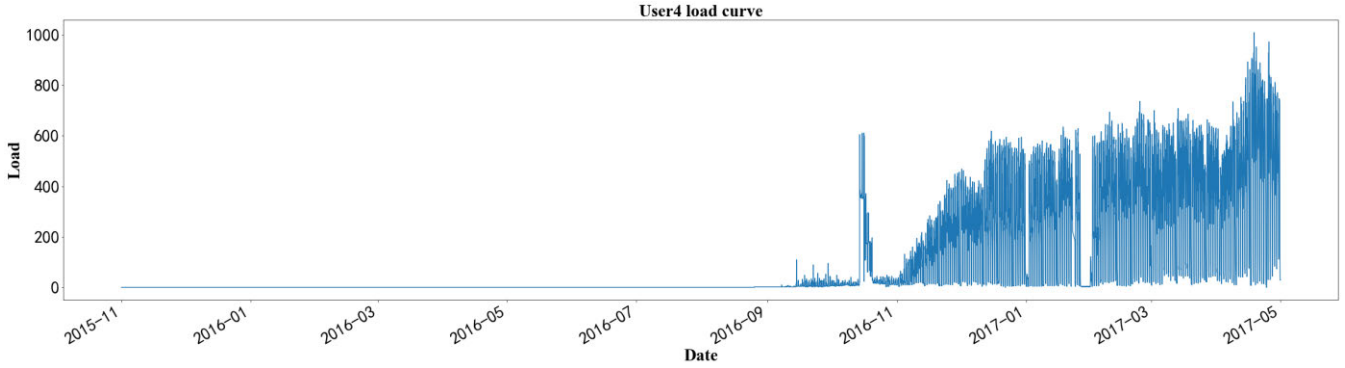


FIGURE 10. The original load curve of user4.

FPR of the PCA method is at a low level in all cases, the level of the F1 Score is also low, and so it cannot detect anomalies effectively.

In most cases, it is found that in the results of CFC, all samples are easily clustered into one group because of the lack of differences in features. This means that CFC has a lower F1 Score and a low-level FPR in the same case. CFC performs well only in a few cases, such as User 3 under ADI1, User 4 under ADI3, etc. The reason is that the consumption behaviors of the users and the results of performing ADI on the load curve do not perfectly correlate. Because CFC uses only experience and professional knowledge to build the feature set, it is unable to fully reflect the different electricity consumption behaviors of users. Therefore, CFC is only applicable to specific users or under specific ADI method and has limited generalizability.

The experimental results in Fig. 9 show that DBN has a higher F1 score and better performance than RBM in all cases. The reason is that compared with the RBM two-layer network, DBN deep-layer neural network structure can extract more abstract and concise electricity consumption behavior features of users. Even in some cases, for example, User2 under ADI4 and User1 under ADI6 are better than the feature engineering method proposed in this paper. However, the structure and parameter selection of RBM, DBN and other networks depend on experience and data itself to a great extent and need to adjust parameters many times for experiments. Different data sets often correspond to different neural network structures. Moreover, in the process of RBM and DBN network training, the acquired features are usually not interpretable.

Experimental results show that compared with MI, using the MIC to evaluate the degree of correlation between features can better mine the correlations between features. In most cases, the experimental results show that the F1 Score with MIC is higher than with MI and in a few cases, they have similar performance, such as those of User1 under ADI2, ADI3, AID4 and ADI6, User2 under ADI3 and ADI4, and User4 under ADI4 and ADI5. Although MI and MIC have similar F1 Score results sometimes, MI has the worst FPR result in most cases.

In some ADI modes, such as ADI1-ADI5, the F1 Score of User4 is lower than that of other users. As shown in Fig. 10, through the analysis of the original load curve, it is found that User4 has a large number of cases with a daily load of 0. The unusual characteristics of User4 also lead to the difference in the violin plot mentioned above. The ADI process has little impact on the original load data of User4,

so not all kinds of feature engineering strategies can detect the abnormal behavior well.

The results show that the detection of abnormal behaviors based on the feature engineering proposed can identify abnormal electricity consumption behavior well compared with detection based on other feature strategies. The proposed method has a high F1 score under different ADI methods and good generalizability for users with different electricity consumption behaviors. In a few cases, such as User2 under ADI4, User3 under ADI5, User1 under ADI6, the method proposed does not perform well compared with other feature strategies. The reason is that there are similarities between different users' inherent electricity consumption behavior load curve and the load curve injected by abnormal data, which makes the optimal feature set obtained by the proposed method unable to distinguish the abnormal electricity consumption behavior and normal effectively. The original feature set obtained by the proposed method is based on tsfresh, which is directly obtained from the user's electricity load curve. Other feature strategies, such as PCA and DBN, do not rely on the load data curve directly but acquire features through dimension reduction and training the neural network. Therefore, those feature strategies will not be affected by the above reason and have good performance compared with the proposed method in a few cases. However, the above phenomenon belongs to the existence of a specific situation. In practice, the abnormal electricity consumption behavior of users is generally different from their inherent electricity use habits. In addition, the better performance of other feature strategies also depends on the parameters detected through repeated experiments, but the selection of parameters largely depends on the data itself and experience. Although in a few cases the proposed method is inferior to other feature strategies, on the whole, it can effectively detect the abnormal electricity consumption behavior of most users under different ADI methods and can meet the requirements of unsupervised detection.

In order to ensure that the method proposed can detect all abnormal electricity consumption behaviors of users to the maximum extent, it is inevitable that a small number of normal electricity consumption behavior data are judged as abnormal data by the method in the clustering process. As a result, the proposed method has a higher FPR compared with other feature strategies in these cases, which are those of User1 under ADI3, User3 under ADI4, and User2 under ADI5 and ADI6. In some cases, the higher FPR is also affected by the user's own electricity consumption behavior. That is, if the user's minority normal electricity behavior and the abnormal electricity usage behavior are similar in some aspects, there is a possibility of misjudgment as abnormal electricity consumption. Because the proposed method detects abnormal electricity consumption behavior of users as much as possible, there are inevitably a few cases with high FPR. However, considering the generalization of the method and avoiding experience influence, in most cases, the proposed method can ensure that it has a higher F1 Score and lower FPR, which can meet the requirements of unsupervised anomaly detection.

TABLE 6. Six kinds of abnormal data injection methods.

Types		Modification
ADI1	$\hat{x}_t^m = \alpha x_t^m$	Where $\alpha = \text{random}(0.1,0.8)$
ADI2	$\hat{x}_t^m = f(t) \cdot x_t^m$	Where $f(t) = \begin{cases} 0 & t_{start} < t < t_{end} \\ 1 & \text{otherwise} \end{cases}$
ADI3	$\hat{x}_t^m = \alpha_t x_t^m$	Where $\alpha_t = \text{random}(0.1,0.8)$
ADI4	$\hat{x}_t^m = \alpha_t \text{mean}(x_t^m)$	Where $\alpha_t = \text{random}(0.1,1)$
ADI5	$\hat{x}_t^m = x_{24-t}^m$	When the sampling interval is 1 hour
ADI6	$\hat{x}_t^m = \begin{cases} x_t^m & x_t^m \le \gamma \\ \gamma & x_t^m > \gamma \end{cases}$	Where $\gamma = \text{random}(0, \max(X^m))$

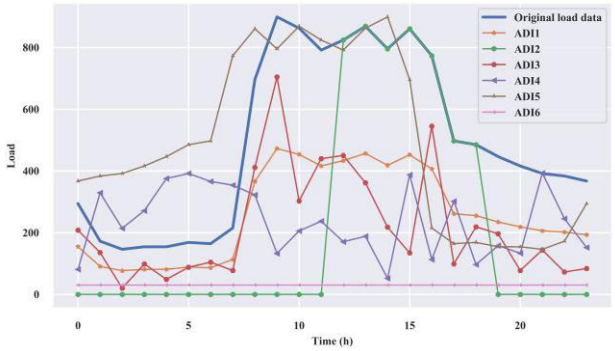


FIGURE 11. An instance of abnormal data injection for Table 6.

At the same time, some cluster algorithms, such as k-means and fuzzy c-means are used to compare with the DBSACN clustering algorithm. The above clustering algorithms need to set the number of clusters and other parameters, but in practice, it is difficult to know the user's inherent electricity consumption behaviors in advance. Although the user's electricity consumption data can be divided into several clusters, it is not advisable to only rely on subjective experience to regard a certain cluster as abnormal electricity consumption behavior. Therefore, the above clustering algorithm is not suitable for unsupervised anomaly detection. However, in the case of supervision, the user's inherent electricity consumption behavior can be divided into several categories based on the user's various information, so the above clustering algorithm will have a good performance. DBSCAN clustering algorithm, as mentioned above, can detect outliers in the process of clustering without determining the number of clusters in advance, and it is not sensitive to the order of data points, so it is more suitable for unsupervised anomaly detection.

V. CONCLUSION

Aiming to solve the problems of the features obtained in the detection of abnormal electricity consumption

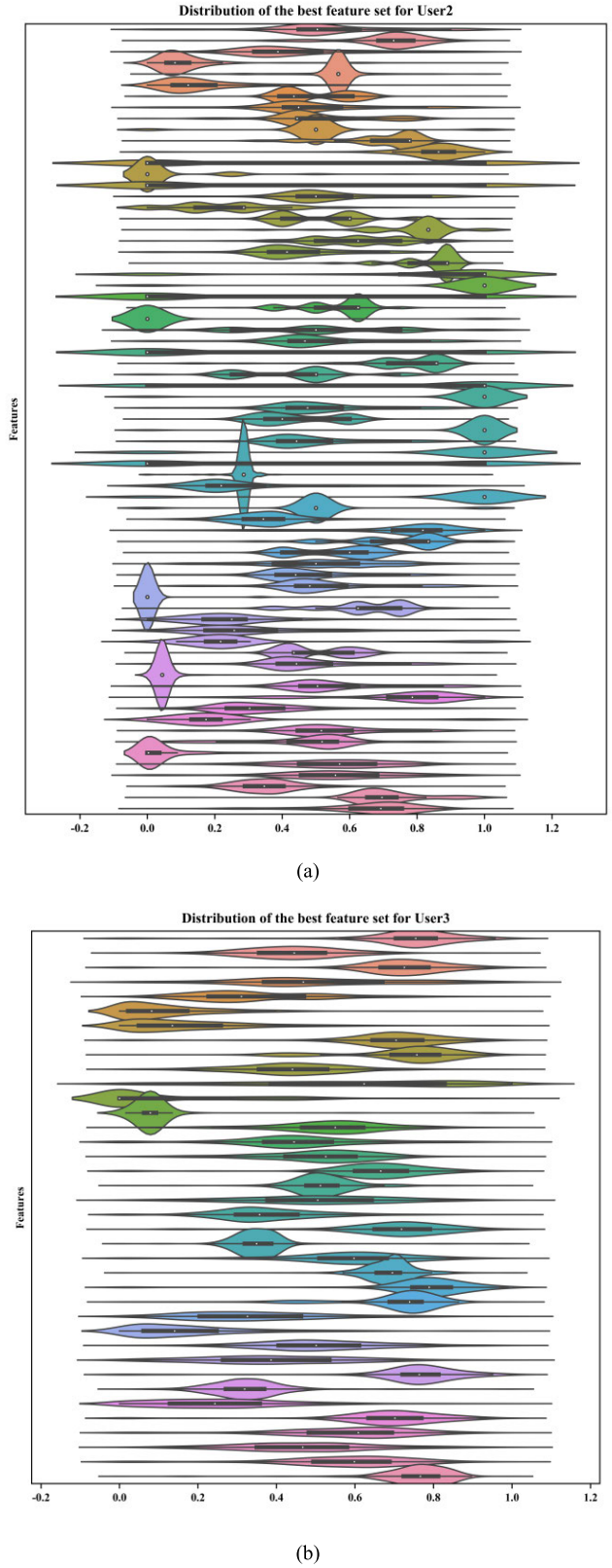


FIGURE 12. (a) The violin plot of the features in the optimal feature set of User2. (b) The violin plot of the features in the optimal feature set of User3.

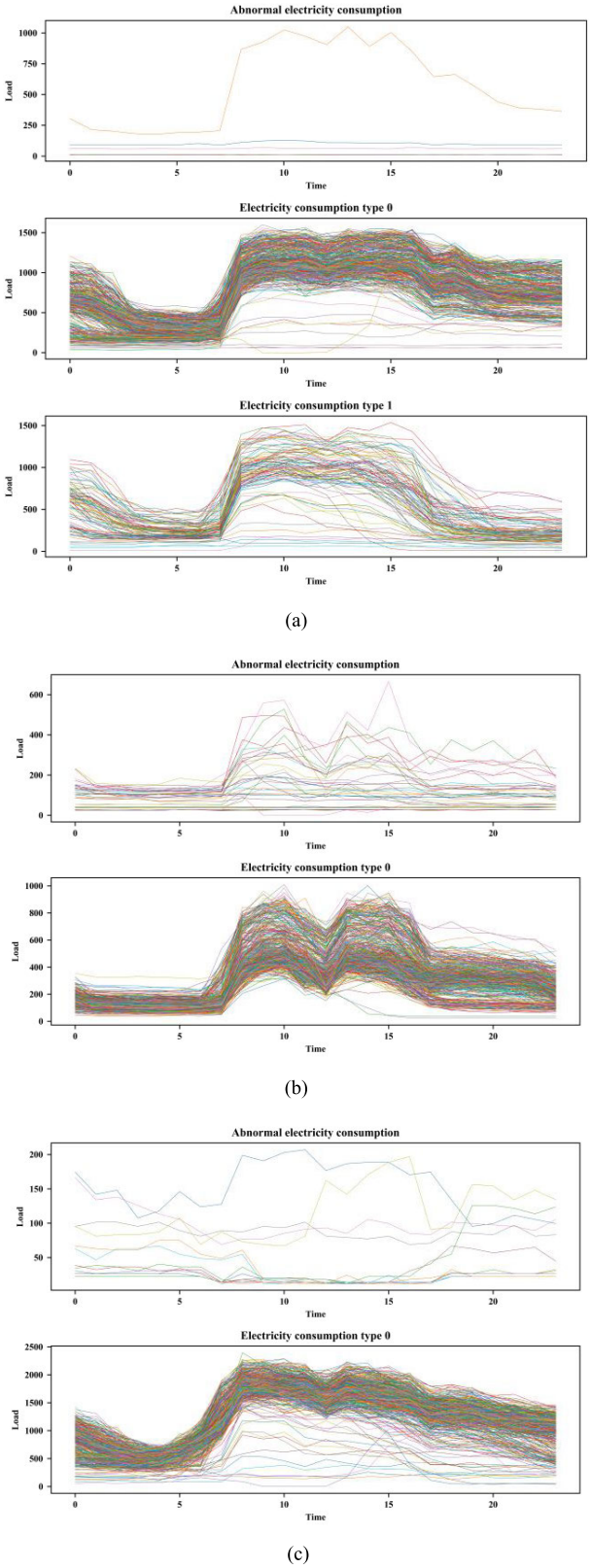
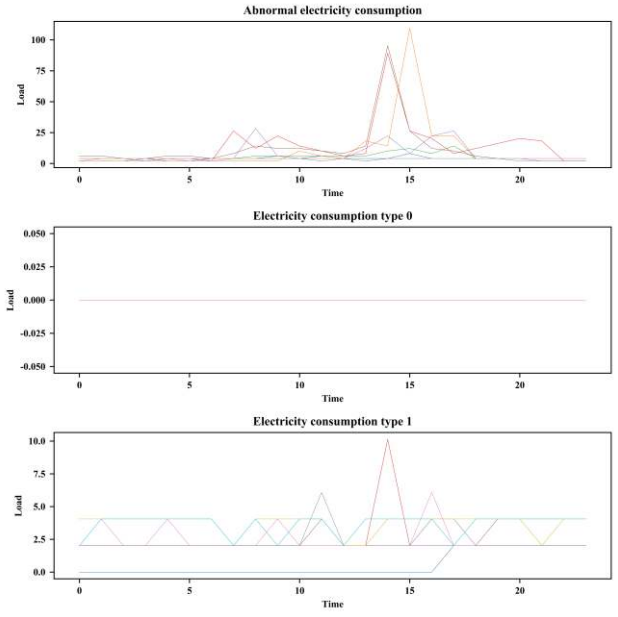


FIGURE 13. (a) The original load curve of User1 after clustering. (b) The original load curve of User2 after clustering. (c) The original load curve of User3 after clustering.



(d)

FIGURE 13. *(Continued.)* **(d) The original load curve of User4 after clustering.**

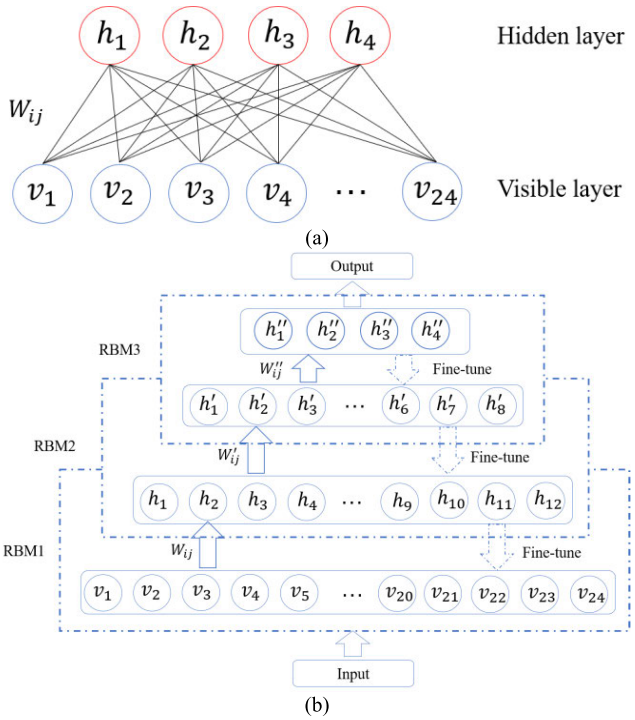


FIGURE 14. **(a) Architecture of the restricted boltzmann machine. (b) Architecture of the deep belief network.**

behavior, an unsupervised abnormal detection method based on feature engineering is proposed. The proposed method is based on unsupervised learning with feature engineering as the core, combined with a density clustering algorithm to detect abnormal electricity consumption behavior. In the proposed method, the feature engineering part does not rely on experience, and can comprehensively and objectively obtain the optimal feature subset reflecting the user's electricity consumption behavior, and the obtained features are interpretable. The abnormal detection part can avoid subjectivity through parameter iteration and clustering evaluation. In addition, the proposed method does not depend on the data and labels information, which has better generalization and practicability. At the same time, the proposed method involves customer portrait analysis, which can fully mine the intrinsic value of electricity consumption data.

Although the result of the anomaly detection method based on feature engineering is satisfactory, there is still room for improvement in terms of clustering algorithms based on density. In addition, because this paper studies small-scale data sets, in future works, when faced with the massive amount of data of the power grid, the problem of how to detect massive users in parallel still needs further research. At the same time, expansion of the application scenarios of the proposed method to cases such as demand-side response and power fault detection remains to be considered.

APPENDIX A

See Table 6 and Figure 11.

APPENDIX B

See Figure 12.

APPENDIX C

See Figure 13.

APPENDIX D

See Figure 14.

REFERENCES

[1] G. Micheli, E. Soda, M. T. Vespucci, M. Gobbi, and A. Bertani, "Big data analytics: An aid to detection of non-technical losses in power utilities," *Comput. Manage. Sci.*, vol. 16, nos. 1–2, pp. 329–343, Feb. 2019.

[2] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using Customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016.

[3] 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.

[4] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gomez-Exposito, "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.

[5] E. Villar-Rodriguez, J. Del Ser, I. Oregi, M. N. Bilbao, and S. Gil-Lopez, "Detection of non-technical losses in smart meter data based on load curve profiling and time series analysis," *Energy*, vol. 137, no. 7, pp. 118–128, Oct. 2017.

[6] C. C. Huang, Y. T. Tsao, and J. Y. J. Hsu, "Behavior-based detection of abnormal power consumption for power saving," *Appl. Mech. Mater.*, vols. 291–294, pp. 674–678, Feb. 2013.

[7] S.-C. Yip, K. Wong, W.-P. Hew, M.-T. Gan, R. C.-W. Phan, and S.-W. Tan, "Detection of energy theft and defective smart meters in smart grids using linear regression," *Int. J. Electr. Power Energy Syst.*, vol. 91, no. 10, pp. 230–240, Oct. 2017.

[8] W. Han and Y. Xiao, "Non-technical loss fraud in advanced metering infrastructure in smart grid," in *Proc. 2nd Int. Conf. Cloud Comput. Secur. (ICCCS)*, 2016, pp. 163–172.

[9] C. Z. Zhang, X. X. Yong, and Z. Z. Huang, "Electricity theft detection for customers in power utility based on real-valued deep belief network," *Power Syst. Technol.*, vol. 43, no. 3, pp. 1083–1091, 2019.

[10] H. T. Yu, G. Q. Lai, and S. Hongbin, "Electricity fraud detection based on stack uncorrelation autoencoder and support vector machine," *Autom. Electr. Power Syst.*, vol. 43, no. 1, pp. 119–125, Jan. 2019.

[11] 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.

[12] J. Mulongo, M. Atemkeng, T. Ansah-Narh, R. Rockefeller, G. M. Nguegnang, and M. A. Garuti, "Anomaly detection in power generation plants using machine learning and neural networks," *Appl. Artif. Intell.*, vol. 34, no. 1, pp. 64–79, Jan. 2020.

[13] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, Jan. 2014.

[14] G. De. *The Application of Machine Learning in the Starting Point of Gaode*. Dveloppaper. Accessed: Nov. 1, 2019. [Online]. Available: https://developpaper.com/the-application-of-machine-learning-in-the-starting-point-of-gaode/

[15] C. C. O. Ramos, A. N. de Souza, A. X. Falcao, and J. P. Papa, "New insights on nontechnical losses characterization through evolutionary-based feature selection," *IEEE Trans. Power Del.*, vol. 27, no. 1, pp. 140–146, Jan. 2012.

[16] P. Glauner, J. A. Meira, P. Valtchev, R. State, and F. Bettinger, "The challenge of non-technical loss detection using artificial intelligence: A survey," *Int. J. Comput. Intell. Syst.*, vol. 10, no. 1, pp. 760–775, 2017.

[17] T. Hu, Q. Guo, X. Shen, H. Sun, R. Wu, and H. Xi, "Utilizing unlabeled data to detect electricity fraud in AMI: A semisupervised deep learning approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3287–3299, Nov. 2019.

[18] Y. Wang, Q. Chen, D. Gan, J. Yang, D. S. Kirschen, and C. Kang, "Deep learning-based socio-demographic information identification from smart meter data," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2593–2602, May 2019.

[19] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[20] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 907–948, Feb. 2020.

[21] Z. Aydin and V. C. Gungor, "A novel feature design and stacking approach for non-technical electricity loss detection," in *Proc. IEEE Innov. Smart Grid Technol. Asia (ISGT Asia)*, May 2018, pp. 867–872.

[22] R. N. Toma, M. N. Hasan, A.-A. Nahid, and B. Li, "Electricity theft detection to reduce non-technical loss using support vector machine in smart grid," in *Proc. 1st Int. Conf. Adv. Sci., Eng. Robot. Technol. (ICASERT)*, May 2019, pp. 1–6.

[23] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946–955, Aug. 2008.

[24] J. A. Meira, P. Glauner, R. State, P. Valtchev, L. Dolberg, F. Bettinger, and D. Duarte, "Distilling provider-independent data for general detection of non-technical losses," in *Proc. IEEE Power Energy Conf. Illinois (PECI)*, Feb. 2017, pp. 1–5.

[25] E. W. S. Angelos, O. R. Saavedra, O. A. C. Cortés, and A. N. de Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," *IEEE Trans. Power Del.*, vol. 26, no. 4, pp. 2436–2442, Oct. 2011.

[26] J. L. Viegas, P. R. Esteves, and S. M. Vieira, "Clustering-based novelty detection for identification of non-technical losses," *Int. J. Electr. Power Energy Syst.*, vol. 101, pp. 301–310, Oct. 2018.

[27] T. Ahmad, H. Chen, J. Wang, and Y. Guo, "Review of various modeling techniques for the detection of electricity theft in smart grid environment," *Renew. Sustain. Energy Rev.*, vol. 82, no. 3, pp. 2916–2933, Feb. 2018.

[28] K. Zheng, Y. Wang, Q. Chen, and Y. Li, "Electricity theft detecting based on density-clustering method," in *Proc. IEEE Innov. Smart Grid Technol. Asia (ISGT-Asia)*, Dec. 2017, pp. 1–6.

[29] G. S. Thejas, S. R. Joshi, S. S. Iyengar, N. R. Sunitha, and P. Badrinath, "Mini-batch normalized mutual information: A hybrid feature selection method," *IEEE Access*, vol. 7, pp. 116875–116885, 2019.

[30] L. Zhao and X. Dong, "An industrial Internet of Things feature selection method based on potential entropy evaluation criteria," *IEEE Access*, vol. 6, pp. 4608–4617, 2018.

[31] W. Zhou, C. Wu, Y. Yi, and G. Luo, "Structure preserving non-negative feature self-representation for unsupervised feature selection," *IEEE Access*, vol. 5, pp. 8792–8803, 2017.

[32] C. J. Zhuang, B. Zhang, J. Hu, Q. Li, and R. Zeng, "Anomaly detection for power consumption patterns based on unsupervised learning," in *Proc. CSEE*, 2016, vol. 36, no. 2, pp. 379–387.

[33] Y. Sun, S. Li, C. Cui, B. Li, S. Chen, and G. Cui, "Improved outlier detection method of power consumer data based on Gaussian kernel function," *Power Syst. Technol.*, vol. 42, no. 5, pp. 1595–1604, 2018.

[34] L. A. Passos, Jr., C. C. O. Ramos, D. Rodrigues, D. R. Pereira, A. N. de Souza, K. A. P. da Costa, and J. P. Papa, "Unsupervised non-technical losses identification through optimum-path forest," *Electr. Power Syst. Res.*, vol. 140, pp. 413–423, Nov. 2016.

[35] R. Razavi, A. Gharipour, M. Fleury, and I. J. Akpan, "A practical feature-engineering framework for electricity theft detection in smart grids," *Appl. Energy*, vol. 238, no. 1, pp. 481–494, Mar. 2019.

[36] L. Jun, Z. Y. Ping, and S. Yi, "Feature optimization strategy for behavior analysis of intelligent power users," *Autom. Electr. Power Syst. Power Syst.*, vol. 41, no. 5, pp. 58–63, 2017.

[37] C. C. O. Ramos, A. N. Souza, G. Chiachia, A. X. Falcão, and J. P. Papa, "A novel algorithm for feature selection using Harmony search and its application for non-technical losses detection," *Comput. Electr. Eng.*, vol. 37, no. 6, pp. 886–894, Nov. 2011.

[38] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series FeatuRe extraction on basis of scalable hypothesis tests (tsfresh—A Python package)," *Neurocomputing*, vol. 307, no. 3, pp. 72–77, 2018.

[39] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," 2016, *arXiv:1610.07717*. [Online]. Available: http://arxiv.org/abs/1610.07717

[40] M. Christ, N. Braun, and J. Neuffer. *The Documentation of tsfresh*. Accessed: Oct. 25, 2019. [Online]. Available: https://tsfresh.readthedocs.io/en/latest/index.html

[41] X. Zhao, W. Deng, and Y. Shi, "Feature selection with attributes clustering by maximal information coefficient," *Procedia Comput. Sci.*, vol. 17, no. 5, pp. 70–79, 2013.

[42] R. Chen, N. Sun, X. Chen, M. Yang, and Q. Wu, "Supervised feature selection with a stratified feature weighting method," *IEEE Access*, vol. 6, pp. 15087–15098, 2018.

[43] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011.

[44] K. Zheng, Q. Chen, Y. Wang, C. Kang, and Q. Xia, "A novel combined data-driven approach for electricity theft detection," *IEEE Trans Ind. Informat.*, vol. 15, no. 3, pp. 1809–1819, Mar. 2019.

[45] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 2006.

[46] J. Liu, F. Qu, X. Hong, and H. Zhang, "A small-sample wind turbine fault detection method with synthetic fault data using generative adversarial nets," *IEEE Trans Ind. Informat.*, vol. 15, no. 7, pp. 3877–3888, Jul. 2019.

[47] Y. You, J. Sun, B. Ge, D. Zhao, and J. Jiang, "A data-driven M2 approach for evidential network structure learning," *Knowl.-Based Syst.*, vol. 187, Jan. 2020, Art. no. 104810.

[48] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1–6, Aug. 2005.

[49] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Knowl. Discovery Data Mining (KDD)*, 1996, pp. 226–231.

[50] W. Michael. *Seaborn: Statistical Data Visualization*. Seaborn. Accessed: Oct. 30, 2019. [Online]. Available: http://seaborn.pydata.org

[51] X. Cai, S. Hu, and X. Lin, "Feature extraction using restricted Boltzmann machine for stock price prediction," in *Proc. IEEE Int. Conf. Comput. Sci. Autom. Eng. (CSAE)*, May 2012, pp. 80–83.

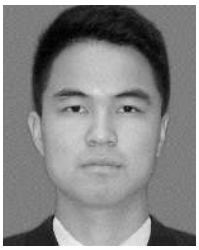
[52] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[53] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

WEI ZHANG received the Ph.D. degree in electrical engineering and its automation from Wuhan University, Wuhan, China, in 2015. He is currently a Lecturer with the Department of Electrical Engineering, University of Shanghai for Science and Technology, Shanghai, China. His current research interests include interdependent networks, the dispatch of integrated energy systems, and big data.

XIAOWEI DONG is currently pursuing the degree in electrical engineering with the Department of Electrical Engineering, University of Shanghai for Science and Technology, Shanghai, China. His research interests include data mining and machine learning in power systems.

HUAIBAO LI is currently pursuing the degree in electrical engineering with the Department of Electrical Engineering, University of Shanghai for Science and Technology, Shanghai, China. His research focuses on new energy and electricity markets.



JIN XU is currently pursuing the degree in electrical engineering with the Department of Electrical Engineering, University of Shanghai for Science and Technology, Shanghai, China. His researches focus on machine learning and distribution network topology recognition.

DAN WANG is currently pursuing the degree in electrical engineering with the Department of Electrical Engineering, University of Shanghai for Science and Technology, Shanghai, China. Her research interests include low-voltage topology identification and deep learning algorithms in power systems.