

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Electricity Theft Detection in AMI Based on Clustering and Local Outlier Factor

Yanlin Peng¹, Yining Yang², Yuejie Xu¹, Yang Xue², Runan Song², Jinping Kang¹, *Member, IEEE*, and Haisen Zhao¹, *Senior Member, IEEE*

¹State Key Laboratory of Alternate Electricity Power System with Renewable Energy Sources, North China Electric Power University, Changping District, Beijing 102206, China

²China Electric Power Research Institute, Haidian District, Beijing 100192, China

Corresponding author: Haisen Zhao (e-mail: zhaohaisen@163.com).

This work was supported in part by the Key Project of “research and application of generic technology of national quality foundation” (2016YFF0201200) and the State Technology Project (JLB17201900137).

ABSTRACT As one of the key components of smart grid, advanced metering infrastructure (AMI) provides an immense number of data, making technologies such as data mining more suitable for electricity theft detection. However, due to the unbalanced dataset in the field of electricity theft, many AI-based methods such as deep learning are prone to under-fitting. To evade this problem and to detect as many types of theft attacks as possible, an outlier detection method based on clustering and local outlier factor (LOF) is proposed in this study. We firstly analyze the load profiles with k -means. Then, customers whose load profiles are far from their cluster centers are selected as outlier candidates. After that, the LOF is utilized to calculate the anomaly degrees of outlier candidates. Corresponding framework for practical application is then designed. Finally, numerical experiments based on realistic dataset show the good performance of the presented method.

INDEX TERMS Clustering, data mining, electricity theft detection, local outlier factor.

NOMENCLATURE

<i>Sets</i>		$\tilde{\mathbf{u}}_{i,d}^*$	Normalized recorded load vector for user i on day d .
\mathcal{A}	Set of all users in an area.	m	Number of load profiles for each user.
\mathcal{B}	Set of benign users in the area.	\mathbf{x}	A data sample in dataset \mathcal{D} .
\mathcal{C}	Set of fraudulent users in the area.	\mathbf{y}	A data sample in dataset \mathcal{D} .
\mathcal{D}	A dataset.	\mathbf{x}^j	Center of cluster j in dataset \mathcal{D} .
\mathcal{D}^j	The j -th cluster of dataset \mathcal{D} .	n	Number of nearest samples to \mathbf{x} .
\mathcal{O}	Set of outlier candidates for \mathcal{D} .	k	Number of clusters.
\mathcal{O}^j	Set of outlier candidates for \mathcal{D}^j .	ε	Ratio of outliers account in \mathcal{D} .
<i>Indices</i>		d_c^j	Cut-off distance of cluster j .
t	Index of time interval.	$\text{LOF}_{i,d}$	Value of LOF for user i on day d .
i	Index of user and data sample.	$\text{rank}_{i,d}$	Rank of user i on day d based on descending order of $\text{LOF}_{i,d}$.
j	Index of cluster.	$\overline{\text{rank}}_i$	Average rank of user i during m -days.
d	Index of day.	<i>Functions</i>	
<i>Variables and Parameters</i>		$ \cdot $	Size of a set.
$u_{i,t}$	Ground truth load for user i at time interval t .	$f(\cdot)$	Attack function.
$\tilde{u}_{i,t}$	Recorded load for user i at time interval t .	$\text{dist}(\cdot, \cdot)$	Euclidean distance between two data samples
$\tilde{\mathbf{u}}_i$	Recorded load vector for user i .	$\text{dist}_n(\cdot)$	The n -th nearest distance of a data sample.
$\tilde{\mathbf{u}}_i^*$	Normalized recorded load vector for user i .		

- $N_n(\cdot)$ The n -nearest objects of a data sample.
- $Rd(\cdot, \cdot)$ Reachability distance between two data samples.
- $\rho_n(\cdot)$ Local reachability density of a data sample.
- $\sigma(\cdot)$ Standard deviation of a vector.
- $\text{mean}(\cdot)$ Arithmetic average of a vector.

I. INTRODUCTION

Electricity theft in power system is that customers adopt certain techniques and devices to illegally tamper with the meters or intrude into the information flow of grid, resulting in the electricity consumption or the bills being lower than the actual amount [1]. Electricity theft seriously damages the economic benefits of power utilities and also lays down potential safety hazards such as power outages, equipment damage, and casualties. According to the report conducted by Northeast Group, the annual cost caused by electricity theft in the USA had reached \$10 billion in 2017 [2]. In China, State Grid Corporation has also retrieved the theft bills of nearly 13 billion yuan in the past three years.

With the establishment of Advanced Metering Infrastructure (AMI) and application of smart meter, the massive amounts of electricity consumption data they provide make data mining technologies more suitable for electricity theft detection [3]. However, the software and communication technologies applied in AMI make it possible for malicious users to tamper with the smart meters and intrude into the information flow of grid via cyberattacks. Corresponding high-tech electricity theft cases were reported in Fujian Daily of China [4]. Unlike traditional physical attacks such as meter-bypassed or meter-tampered, cyberattacks modify the data in a more random way and leave little physical trace, making them more difficult to be detected. Because of such increasingly severe situation, corresponding detection methods become urgently needed to address the problems of electricity theft in AMI.

Current data-driven electricity theft detection methods (ETDMs) can be divided into three categories [3], as follows.

1) *Game theory based*. In this type of methods [5]-[6], electricity theft is described as a game between power suppliers and customers. The game equilibrium theory can be used to derive the difference in the distribution of electricity consumption between normal customers and abnormal ones. The game theory-based methods are useful for understanding the potential strategies and interactions among different players, but are hard to formulate a practically applicable model to involve all the players.

2) *System state based* [7]-[10]. The methods based on system state utilize the fact of data inconsistency caused by data tempering of fraudulent customers to realize theft detection. The physical model of a power network indicates that the system variables should satisfy certain mathematical equations, which derives the consistency of the variables. But the data tempering of fraudulent users will destroy this consistency and cause some anomalies such as non-technical

loss (NTL) and voltage limit violation. Works [7]-[8] on this direction perform distribution system state estimation to realize the detection for electricity theft detection. However, the state-estimation-based methods need precise information of network topology and parameter, which are not available at the end-user level. Thus, the practical applicability is limited in this situation.

3) *Power consumption pattern based*. It is widely believed that the consumption patterns of fraudulent users differ from those of benign users. Based on such characteristics, this type of ETDM utilizes logistic regression [11]-[12] or artificial intelligence such as classification [13]-[16] and clustering [17]-[21] to analyze the load profiles of customers for electricity theft detection. Specifically, classification methods usually involve vast labeled historical electricity usage data to train the detection models. Examples including support vector machines (SVM) [13], convolutional neural networks (CNNs) [14] and other artificial neural networks [15]-[16] have been tested in literature. In contrast, unsupervised methods like clustering, focus on the information without labels. For example, [17] adopted the density-based spatial clustering of applications with noise (DBSCAN) to calculate the anomaly degrees of users.

The existing data-driven ETDMs have some limitations. First, the game theory -based methods mainly focus on theoretical analysis with strong assumptions, thus are not competent in engineering practicality. Second, supervised methods need vast reliable theft samples to train the detection models. But the small proportion of theft users and the data poisoning (the false labeled samples) [18] limited their accuracies. Worse yet, they might not distinguish between electricity theft and non-malicious activities like meter reinstallation.

Since the amounts of fraudulent users account for a very little proportion in reality and their consumption patterns deviate from the normal ones, it's quite suitable for outlier detection methods [19] to be utilized in electricity theft detection. However, traditional outlier detection methods such as local outlier factor (LOF) can't detect the overlapping outliers. To handle this problem and to detect as many types of theft attack as possible, an improved outlier detection method based on clustering and local density is proposed in this study. This method adopts k -means to analyze the load profiles of users. And then, customers whose load profiles are far from their cluster centers are selected as outlier candidates. After that, the LOF is utilized to calculate the anomaly degrees of the outlier candidates. The main contributions of this paper are as follows.

1) *New techniques*: Combining k -means and LOF for electricity detection, which not only evade the problem of unbalanced dataset but also realize the detection for overlapping outliers.

2) *Extensive experiments*: We have conducted extensive and comprehensive experiments based on a realistic dataset.

The comparisons with some other detection methods validate the effectiveness and superiority of our method.

The rest of this paper is organized as follows. In Section II, we review the literature related to electricity theft detection in AMI. In Section III, the AMI model and the attack functions are pointed out. Section IV presents the theory of the two techniques and the framework of the detection method. Numerical experiments are conducted and evaluation results are shown in Section V. Finally, we conclude this paper in Section VI.

II. RELATED WORK

In this section we review existing data-driven ETDMs in literature. One direction for electricity theft detection is game-theory-based techniques. Cárdenas *et al.* [5]-[6] studied the use of game theory in energy theft behaviors. In [6], electricity theft and combat losses are modeled as non-zero sum Stakelberg game. The distribution deploy AMI to maximize the likelihood of detecting energy thieves, while the attackers schedule their electricity theft behaviors so that the probability of being caught is minimized.

Another solution for electricity theft pinpointing is state estimation. Leite *et al.* [7], adopted a state estimator to monitor the bias between the estimated and measured voltages. Once the bias is detected, the sources of NTL are located by a pathfinding procedure based on the A-Star algorithm. Their method, however, is only effective when precise information of network topology and parameter is available. If this is not the case, Salinas *et al.* [9] introduced a peer-to-peer ETDM. In their approach, a central meter is deployed in each neighborhood to measure the NTL of this area at each time instance. By solving the sparsest solution of a group of underdetermined linear equations between NTL and load vectors, the fraudulent users can be found. But the algorithms to get a solution of low percentage of sparsity are still in their infancy. To handle this problem, Zheng *et al.* [10] adopted the maximum information coefficient (MIX) to measure the correlation between NTL and load data of users. The stronger the correlation, the more suspicious the user is. Nevertheless, this correlation-based method could but detect linear false data injection.

Recently, with the booming of artificial intelligence (AI), techniques such as classification and clustering are utilized to analyze the load profiles of customers for energy thief locating. For example, Jokar *et al.* [13] summarized several modes of FDI to artificially generate fraudulent consumption data. Then, a support vector machine (SVM) was trained to detect whether a new sample of load profiles is normal or not. In [14], it was observed that the load curves of abnormal users have poor periodicity compared with those of normal users. And a conventional neural networks (CNNs) was trained to detect such abnormal users. However, these classification-based methods only can work if verified cases of theft samples are available. If this is not the case, then the unsupervised clustering which do not use electricity theft labels, must be

used. Examples include fuzzy C-means (FCM) [20] and optimum-path forests [21].

Perhaps, the most relevant in [22] presented the performance comparison for various existing outlier detection algorithms on real dataset. The results show the feasibility of outlier algorithms for electricity theft detection. Compared to [22], the proposed method analyzes the accuracy for detecting different attack functions and realizes the detection for overlap outliers.

III. PROBLEM STATEMENT

A. AMI SYSTEM MODEL

The architecture of AMI is shown in Figure 1. AMI is composed of smart meters, communication networks and data management system. Under the structure of AMI, each customer is equipped with a smart meter to record his electricity data. A concentrator is installed in an area with a group of neighborhood users to collect the data from smart meters in this area. Due to the stable topology within the area and the fine security of concentrators, the electricity consumption W_t recorded by concentrator is the sum of ground truth consumption of all customers in one area, i.e.,

$$W_t = \sum_{i \in \mathcal{A}} u_{i,t} \quad (1)$$

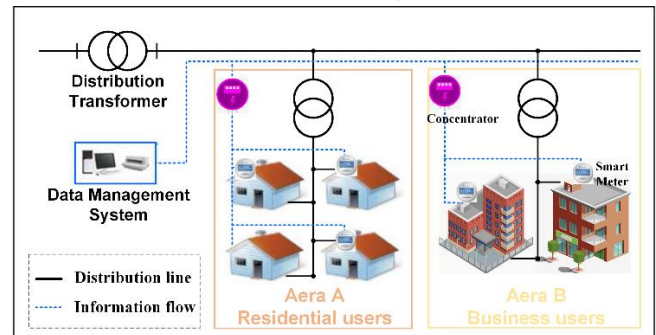


FIGURE 1. Illustration of AMI system model

where $u_{i,t}$ is the ground truth load of user i at time instance t , and \mathcal{A} is the set of all users in area A. If there are several fraudulent users in area A, the set of fraudulent users are denoted as \mathcal{C} whose size is $|\mathcal{C}|$ and remnant benign users are denoted as \mathcal{B} whose size is $|\mathcal{B}|$. The tampering behavior of the electricity thieves is to transform the ground truth data $u_{i,t}$ into modified data $\tilde{u}_{i,t}$, i.e.

$$\tilde{u}_{i,t} \leftarrow f(u_{i,t}) \quad (2)$$

where $f(\bullet)$ is an attack function to simulate the modification of fraudulent users. The problem this study focuses on is how to find the fraudulent users in \mathcal{C} with different types of $f(\bullet)$.

B. ATTACK FUNCTIONS

There are many known techniques for electricity theft in AMI, which can be categorized into three groups [13].

1) Physical Attacks: Fraudulent users manipulate smart meters physically to lower meter readings, such as meter-bypassed and meter-tempered.

2) Cyberattacks: Fraudulent users compromise meter readings remotely or modify the firmware on smart meters using communication technologies.

3) Data Attack: Fraudulent users inject bad data into the data management system or smart meters, which reduces their electricity bills and meter readings.

TABLE I
SEVEN TYPES OF THE ATTACK FUNCTIONS

Types	Attack Functions
Type 1	$\tilde{u}_{i,t} = \alpha u_{i,t}, 0.2 < \alpha < 0.8$
Type 2	$\tilde{u}_{i,t} = \begin{cases} u_{i,t} & \text{if } u_{i,t} < \sigma_i \\ \sigma_i & \text{if } u_{i,t} \geq \sigma_i \end{cases} \sigma_i < \max(u)$
Type 3	$\tilde{u}_{i,t} = \max(u_{i,t} - \sigma_i, 0), \sigma_i < \max(u)$
Type 4	$\tilde{u}_{i,t} = \alpha_i u_{i,t}, 0.2 < \alpha_i < 0.8$
Type 5	$\tilde{u}_{i,t} = \alpha_i \bar{u}_i, 0.2 < \alpha_i < 0.8$
Type 6	$\tilde{u}_{i,t} = \begin{cases} 0 & \text{if } t_1 < t < t_2 \\ u_{i,t} & \text{otherwise} \end{cases} t_2 - t_1 > 4\text{hours}$
Type 7	$\tilde{u}_{i,t} = \bar{u}_i$

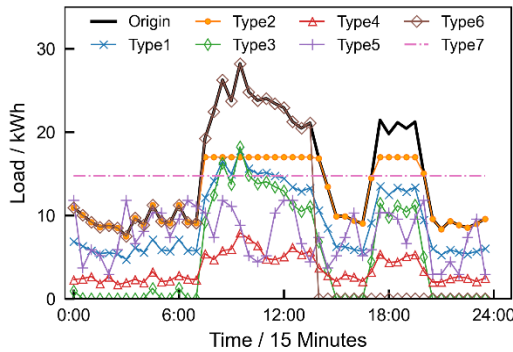


FIGURE 2. An example of different attack types

To simulate the tampering behaviors of above attacks, Jokar *et al.* [13] summarized several attack functions. Table 1 gives the details of these attack functions, and Figure 2 shows an example of the tampered load profiles. As shown in Table 1, Type 1 reduces the $u_{i,t}$ in a constant percentage throughout the entire fraudulent period. Type 2 means that $u_{i,t}$ above a threshold are clipped. In type 3, a cut off value is subtracted from all $u_{i,t}$. Type 4 modifies every $u_{i,t}$ in different ratios. Type 5 generates $\tilde{u}_{i,t}$ by multiplying the average consumption of this day by a random percentage defined for each user. In type 6, $u_{i,t}$ during a random period longer than 4 hours each day are replaced by zero. Finally, type 7 modifies all $u_{i,t}$ by the average consumption of this day to represent attacks against load control mechanisms in which the price of electricity varies over different hours of the day; while the total

amount of electricity usage stays the same. We utilize these 7 types of attack functions to generate fraudulent data to conduct numerical experiments for evaluation. There are many other theft attack functions in [32]. However, a characteristic can be generalized based on their definitions: An attack function either keeps the features and fluctuations of the original curve, or creates new patterns. This is the same for other attack functions, so our method can handle them as well.

III. METHODOLOGY

A. LOCAL OUTLIER FACTOR

The outliers are a sort of special data objects, which occupy a very little proportion and deviate from overall normal model. The outlier detection aims to find out these abnormal objects. Electricity thieves account for a very little proportion in reality, and their consumption patterns differ from normal ones. Thus, outlier detection methods are quite suitable for electricity identification. Local outlier factor (LOF) [23] is an outlier detection method based on local density, and has been proven to be very powerful in the field of fraud detection [25] and fault diagnosis [26].

Suppose that \mathbf{x} and \mathbf{y} are two data objects of dataset \mathcal{D} . Let us denote their Euclidean distance as $dist(\mathbf{x}, \mathbf{y})$. To calculate LOF, the reachability distance (RD) and the local reachability density (LRD) need to be defined. The n -objects in \mathcal{D} closest to \mathbf{x} are called n -nearest neighbors of \mathbf{x} and are denoted as $N_n(\mathbf{x})$. The reachability distance from \mathbf{x} to \mathbf{y} can be calculated as follows:

$$Rd(\mathbf{x}, \mathbf{y}) = \max \{ dist(\mathbf{x}, \mathbf{y}), dist_n(\mathbf{y}) \} \quad (3)$$

where $dist_n(\mathbf{y})$ is the n -th nearest distance between the objects in \mathcal{D} to \mathbf{y} . It is worthwhile to mention that the reachability distance $Rd(\mathbf{x}, \mathbf{y})$ from \mathbf{x} to \mathbf{y} may not equal to the reachability distance $Rd(\mathbf{y}, \mathbf{x})$ from \mathbf{y} to \mathbf{x} . As shown in Figure 3, when \mathbf{y} is in $N_n(\mathbf{x})$ but \mathbf{x} is not in $N_n(\mathbf{y})$, the $Rd(\mathbf{x}, \mathbf{y})$ is equal to $dist(\mathbf{x}, \mathbf{y})$ while the $Rd(\mathbf{y}, \mathbf{x})$ is equal to $dist_n(\mathbf{x})$.

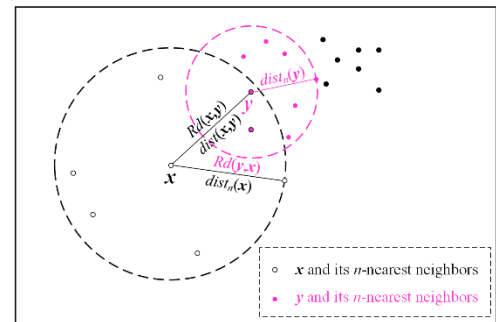


FIGURE 3. Illustration of the reachability distance

The local reachability density of \mathbf{x} is defined as the reciprocal of the average value of $Rd(\mathbf{x}, \mathbf{y})$ when $\mathbf{y} \in N_n(\mathbf{x})$, i.e.,

$$\rho_n(\mathbf{x}) = \frac{n}{\sum_{y \in N_n(\mathbf{x})} Rd(\mathbf{x}, \mathbf{y})} \quad (4)$$

$\rho_n(\mathbf{x})$ is able to measure how close \mathbf{x} is to its n -nearest neighbors $N_n(\mathbf{x})$. And a higher $\rho_n(\mathbf{x})$ indicates a closer distance between \mathbf{x} and $N_n(\mathbf{x})$. Finally, the LOF of \mathbf{x} is defined as the average of the specific value between $\rho_n(\mathbf{y})$ and $\rho_n(\mathbf{x})$ when $\mathbf{y} \in N_n(\mathbf{x})$, i.e.,

$$\text{LOF}_n(\mathbf{x}) = \frac{1}{n} \sum_{y \in N_n(\mathbf{x})} \frac{\rho_n(\mathbf{y})}{\rho_n(\mathbf{x})} \quad (5)$$

From (5), it can be seen that, LOF is a sort of density comparison which could represent the density contrast between \mathbf{x} and its n -nearest neighbors $N_n(\mathbf{x})$. The essence of LOF is to quantify the outlier degree of \mathbf{x} by its nearest neighbors. If \mathbf{x} is not a neighbor in the view of $N_n(\mathbf{x})$, which means that \mathbf{x} is isolated and separate from its neighbors, the value of LOF would be much higher than 1. Vice versa, when \mathbf{x} is a neighbor in the view of $N_n(\mathbf{x})$, which means that \mathbf{x} stays close to its n -nearest neighbors, the value of LOF would be close to 1.

Compared with other outlier detection method such as DBSCAN and variogram cloud, LOF can consider clusters with an arbitrary shape and requires only one parameter n (We set n as 5% of the total number of \mathcal{D} in this paper because it is found to work well in practice). However, the definition of LOF also suggests that, when some outliers overlap together, the LOFs of these overlapping outliers could be close to 1. And attack type 7 modifies the load curves to straight curves which will overlap together after normalization. Thus, this problem of LOF need to be tackled to detect type 7.

B. CLUSTERING AND LOCAL OUTLIER FACTOR

To handle the problem of LOF's failure to overlapping outliers, we proposed an improved outlier detection method based on clustering and local outlier factor (CLOF). The idea of CLOF is to cluster dataset \mathcal{D} with k -means and select the objects which deviate from their cluster centers as the outlier candidates set \mathcal{O} . And then, LOF is adopted to measure the outlier degrees of the objects in \mathcal{O} .

In CLOF, the k -means is firstly utilized to classify the objects in \mathcal{D} . The cluster number could be easily chosen according to elbow method [20]. For every object in \mathcal{D} , we calculate its Euclidean distance to its cluster center, i.e.,

$$\text{dist}(\mathbf{x}_i^j, \mathbf{x}^j) = \|\mathbf{x}_i^j - \mathbf{x}^j\|_2 \quad (6)$$

where, \mathbf{x}_i^j is the i -th object in j -th cluster \mathcal{D}^j ; \mathbf{x}^j is the cluster center of \mathcal{D}^j , $\|\cdot\|_2$ is the 2-norm of a vector. For \mathcal{D}^j , the objects are chosen as its outlier candidates set \mathcal{O}^j according to the following equation.

$$\mathcal{O}^j = \left\{ \mathbf{x}_i^j \mid \text{dist}(\mathbf{x}_i^j, \mathbf{x}^j) > d_c^j \text{ or } |\mathcal{D}^j| < \varepsilon |\mathcal{D}| \right\} \quad (7)$$

where, d_c^j is the cut off distance of \mathcal{D}^j , $|\mathcal{D}^j|$ is size of \mathcal{D}^j ; $|\mathcal{D}|$ is the size of dataset \mathcal{D} , ε is the ratio of outliers. In this paper, d_c^j is defined as triple standard deviation of \mathcal{D}^j based on "three-sigma rule of thumb", and ε is set as 5% according to the parameter n in LOF. Finally, we can get the overall outlier candidates set $\mathcal{O} = \mathcal{O}^1 \cup \mathcal{O}^2 \cup \dots \cup \mathcal{O}^k$.

From (7), it can be concluded that, the outlier candidates in \mathcal{O} is composed of two kinds of objects: one is the objects that deviate from their cluster centers, the other is the clusters that account for very little proportion of the whole dataset. After \mathcal{O} is obtained, the LOFs are calculated for the outlier candidates in \mathcal{O} .

Algorithm of CLOF

Input:	Dataset \mathcal{D} , and parameters n, ε
Output:	Rank of every sample in \mathcal{D}
Step1:	Get the cluster number k according to elbow method.
Step2:	Analyze the data samples in \mathcal{D} with k -means and divide \mathcal{D} into k -clusters $\mathcal{D} = \mathcal{D}^1 \cup \mathcal{D}^2 \cup \dots \cup \mathcal{D}^k$
Step3:	For each cluster \mathcal{D}^j : Calculate the Euclidean distance $\text{dist}(\mathbf{x}_i^j, \mathbf{x}^j)$ between every data sample \mathbf{x}_i^j in this cluster and the cluster center \mathbf{x}^j ; Calculate the cut-off distance d_c^j of \mathcal{D}^j ; Get the outlier candidate set \mathcal{O}^j of \mathcal{D}^j according to (7);
Step4:	Get the outlier candidate set $\mathcal{O} = \mathcal{O}^1 \cup \mathcal{O}^2 \cup \dots \cup \mathcal{O}^k$
Step5:	For every data sample \mathbf{x} in \mathcal{D} : Get the n -nearest $N_n(\mathbf{x})$ neighbors of \mathbf{x} ; Calculate the n -th nearest distance $\text{dist}_n(\mathbf{x})$ of \mathbf{x}
Step6:	For every data sample \mathbf{x} in \mathcal{D} : Calculate the reachability distance $Rd(\mathbf{x}, \mathbf{y})$ for every $\mathbf{y} \in N_n(\mathbf{x})$ according to (3); Calculate the local reachability density of \mathbf{x} according to (4); Calculate the LOF of \mathbf{x} ;
Step7:	Get the ranks of the samples in \mathcal{O} according to the descending order of their LOF first; Then, get the ranks of the samples not in \mathcal{O} according to the descending order of their LOF; The ranks of the samples in \mathcal{O} should be higher than that of the samples not in \mathcal{O} ;

By adding the clusters with a small number into \mathcal{O} , CLOF can detect the overlapping outliers effectively. Figure 4 shows an example of 2-dimensional outlier detection with CLOF. In Figure 4, the black points are outliers detected by CLOF while the hollow points are the non-outliers. And the red point is 5 overlapping outliers. From the distribution of LOFs in Figure 4, the points deviate from the normal majority more, the higher LOFs of these points are. The outliers detected by CLOF is accord with visual intuitive, and the overlapping outliers can also be detected, which proves the effectiveness of CLOF.

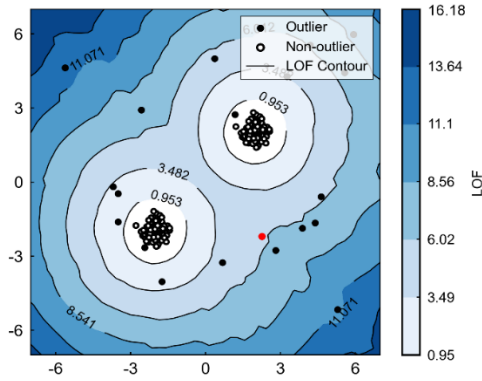


FIGURE 4. An example of 2-dimensional outlier detection by CLOF

C. DETECTION FRAMEWORK

Based on above methodology, we design corresponding detection framework for CLOF. The framework is composed of three modules: the preprocessing module, the detection module and the judgement module, as shown in Figure 5.

For an area that contains $|\mathcal{A}|$ consumers with their m -day load profiles, the preprocessing module firstly vectorize the load profiles each day to get the daily load vectors $\tilde{\mathbf{u}}_i = [\tilde{u}_{i,1}, \tilde{u}_{i,2}, \dots, \tilde{u}_{i,T}]^T$ of user i . For every daily load vector, the missing data are recovered as follows:

$$G(\tilde{u}_{i,t}) = \begin{cases} \text{mean}(\tilde{\mathbf{u}}_i) & \tilde{u}_{i,t} \in \text{NaN} \\ \tilde{u}_{i,t} & \text{otherwise} \end{cases} \quad (8)$$

where $\text{mean}(\tilde{\mathbf{u}}_i)$ is the average value of vector $\tilde{\mathbf{u}}_i$. In addition, there are some erroneous data in some conditions. Therefore, the preprocessing module also recover those data by the following equation according to “three-sigma rule of thumb”:

$$G(\tilde{u}_{i,t}) = \begin{cases} \frac{\tilde{u}_{i,t-1} + \tilde{u}_{i,t+1}}{2} & \text{if } \tilde{u}_{i,t} > 3\sigma(\tilde{\mathbf{u}}_i), \tilde{u}_{i,t-1}, \tilde{u}_{i,t+1} \neq \text{NaN} \\ \tilde{u}_{i,t} & \text{otherwise} \end{cases} \quad (9)$$

where, $\sigma(\tilde{\mathbf{u}}_i)$ is the standard deviation of vector $\tilde{\mathbf{u}}_i$. Next, every load vector is normalized by dividing it with its maximum.

Let us denote the normalized load vector of user i on d -th day as $\tilde{\mathbf{u}}_{i,d}^*$. For all the normalized vectors on d -th day, the detection module calculates $\text{LOF}_{i,d}$ utilizing CLOF. And it gives a rank list of the p -users on d -th day by the descending order of $\text{LOF}_{i,d}$. The rank of user i on d -th day is denoted as $\text{rank}_{i,d}$. After the detection module get all the rank list of m days, the judgement module calculates the average rank of user i according to (10).

$$\overline{\text{rank}}_i = \frac{1}{m} \sum_{j=1}^m \text{rank}_{i,d} \quad (10)$$

Finally, a user is considered committing electricity theft if his average rank is high.

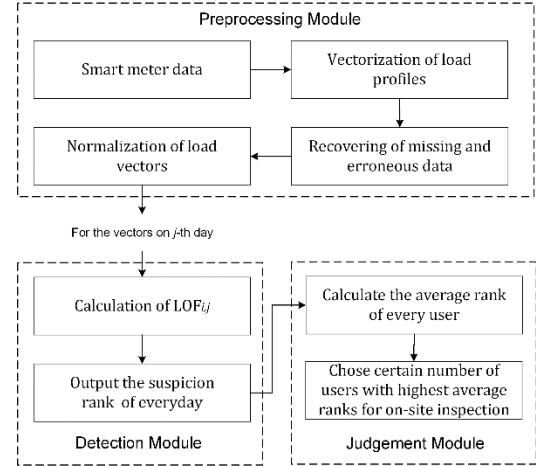


FIGURE 5. Framework of the CLOF detection method

IV. VALIDATION AND EVALUATION

A. DATASET

We use the realistic electricity consumption data released by SGCC as benign dataset. Because all the users involved in the dataset come from the areas whose line loss rates per month are lower than 3%, those data are considered ground truth. Table 2 presents detailed information about this dataset. Particularly, it contains the load profiles of 3000 single-phase (SP) users and 500 three-phase (TP) users within 285 days (from April 1, 2019 to December 31, 2019). Each load profile a day consists of 96 points with a time interval of quarter hour.

TABLE II
INFORMATION OF THE DATASET

Description	Information	
Types of users	SP	TP
Number of users	3000	500
Date	Apr. 1, 2019-Dec. 31, 2019	
Type of data	Each load profile a day consists of 96 points with a time internal of quarter hour	

We use the load profiles of all TP users in the dataset from August 1 to September 31, 2019 to conduct the experiments. The 500 TP users are randomly and evenly divided into several areas. For each area, several users are randomly chosen as electricity thieves. And certain types of attack functions are used to tamper with their load profiles. 40 of the 61 profiles of each fraudulent user are tampered with.

B. EVALUATION METRICS AND COMPARISON

To obtain comprehensive evaluation results, we use area under curve (AUC) [28] and mean average precision (MAP) [29] that are widely adopted classification evaluation criteria as performance metrics. The AUC is the area under the receiver operating characteristic (ROC) curve, which is the trace of true positive rate and false positive rate under different thresholds.

TABLE III
BEST EVALUATION RESULTS OF THE 5 METHODS WITH DIFFERENT ATTACK TYPES

Attack Type	AUC(%)					Attack Type	MAP@20(%)				
	PCC	MIC	CFSFDP	LOF	CLOF		PCC	MIC	CFSFDP	LOF	CLOF
Type1	82.73	80.11	59.37	63.08	62.77	Type1	84.52	83.33	26.19	29.58	27.75
Type2	61.70	69.85	63.00	71.06	69.89	Type2	54.29	67.66	49.37	57.81	58.96
Type3	59.62	67.93	71.25	72.03	72.20	Type3	49.53	61.38	54.69	63.27	66.16
Type4	51.08	62.64	81.75	80.20	79.86	Type4	35.51	57.24	59.88	65.68	62.24
Type5	50.20	61.11	83.48	85.50	85.11	Type5	28.66	53.17	63.29	68.58	67.71
Type6	41.25	49.13	89.70	92.61	90.08	Type6	20.05	34.25	69.91	73.31	71.92
Type7	34.51	39.75	54.34	60.18	91.84	Type7	12.25	18.27	28.14	33.23	73.11
MIX	63.23	67.97	71.25	73.44	81.50	MIX	53.16	60.15	69.34	70.10	73.35

In addition to drawing the ROC curve, AUC can also be calculated as in (22) [14]:

$$AUC = \frac{\sum_{i \in \mathcal{C}} \text{rank}_i - 0.5|\mathcal{C}|(|\mathcal{C}|+1)}{|\mathcal{C}| \times |\mathcal{B}|} \quad (11)$$

where $|\mathcal{C}|$ is the number of fraudulent users, $|\mathcal{B}|$ is the number of benign users, and rank_i is the rank of user i in ascending order according to $\overline{\text{rank}_i}$. The value of AUC must be in (0, 1). And if it is closer to 1, the better result can be achieved.

MAP is usually used to estimate the quality of information retrieval. To calculate MAP, we first define P@S as

$$P@S = \frac{Y_s}{S} \quad (12)$$

where Y_s is the number of fraudulent users among the top S suspicious users; Given a certain number of R , MAP@ R can be calculated as in (24)

$$\text{MAP@R} = \frac{\sum_{i=1}^r P@S_i}{r} \quad (13)$$

where r is the number of fraudulent users among the top R suspicious users and S_i is the position of such i -th fraudulent user. It can be summarized that the value range of MAP@ R is [0, 1], which can measure how high-ranking the fraudulent users are in suspicious list. The closer the value of MAP@ R is to 1, the higher the fraudulent users rank. On the other hand, if there are no fraudulent users among top R suspicious users, the MAP@ R will be 0.

To demonstrate the effectiveness of the presented method, we use some other correlation sorting and unsupervised outlier detection methods for comparison.

1) PCC [30] A famous bivariate correlation measurement. Put the value of PCC as anomaly degree, the larger the PCC is, the more suspicious the user is.

2) Maximum information coefficient (MIC) [30]: A metric that can measure the degree of non-linear correlation between two vectors.

3) Clustering by fast search and find of density peaks (CFSFDP) [31]: A novel clustering algorithm based on density and distance. In [10], it was transformed into an outlier detection method to find fraudulent users.

4) Local outlier factor (LOF) [23]: a classic outlier detection method based on local density.

C. RESULTS

In this part, we divide the users into 10 areas, and randomly chose 6 users as electricity thieves. Thus, each area contains 50 users, and the ratio of electricity thieves is 12%. The test is repeated for 100 times by recombination of users and random selection of electricity thieves.

Table 3 gives the best values of AUC and MAP@20 of the 5 methods with the 8 attack types, in which type MIX indicates that the 6 electricity thieves randomly choose one of the seven attack functions. The best scores for each attack functions are bold. And Figure.6 shows the average AUC and MAP@20 of the 5 methods in 100 tests.

From Table 3 and Figure 6, the highest MAP@20 of outlier-based methods is only 0.323 in detecting type 1, whereas the lowest MAP@20 of correlation-based methods is 0.737. Meanwhile, the gap of AUCs of type1 between these two sorts of methods is also huge. This result demonstrate that the outlier-based methods perform poorly in detecting type1, while the correlation-based methods are far more capable of detecting this type. This is because the tampered load curves of type 1 are nearly identical with the ground truth ones after standardization, which makes these load curves still conform to the normal majority. On the other hand, when the load curves are modified to arbitrary shapes (e.g., type4, type5 and type6), the results are up-side-down. The outliers-based methods have quite high values of AUC in detecting type4, type5 and type6, especially LOF which is found to have the best performance in detecting type5 and type6. While correlation-based methods perform poorly in detecting these three attack types because the tampered load curves become quite random and the correlation no longer exists. For type2 and type3, there is not much difference of the performances of all 5 methods. However, for type7, all the methods are failed except CLOF. This is because type7 modified the load curves into straight curves which overlap together after normalization

The presented CLOF have taken the advantages from LOF and overcome its disadvantage in type 7. For type4, type5 and type6, for which LOF specializes in, the performance of our method is as good as LOF. For type 7 which LOF failed with, both the AUC and MAP@20 of CLOF are above 0.80. The results demonstrate that, CLOF maintain the excellent

performance of LOF in its specialized situations while achieving significant improvements in type 7, resulting in the best detection accuracy in type MIX. The AUC and MAP@20 of CLOF in detecting MIC increased by 12% on basis of those of LOF. It is worthwhile to mention that weight factors in type MIX alter the detection accuracy. Although we assume identical weights for the attack types, the CLOF method achieve improvements in accuracy for other nonextreme weight factors.

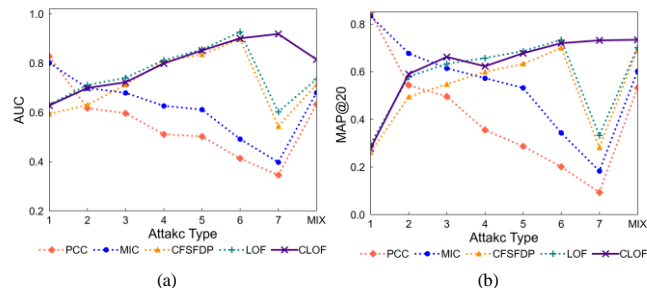


FIGURE 6. Evaluation results of the 5 methods with different attack types. (a) Average values of AUC in 100 tests. (b) Average values of MAP@20 in 100 tests.

Figure 7 (a) shows the standard deviations σ of AUC and MAP@20 in 100 experiments for the 5 methods when detecting type MIX. The σ_{AUC} of the 5 methods are all approximately 0.04, and CLOF has a minimum σ_{AUC} of 0.041. The $\sigma_{MAP@20}$ is distributed between 0.8 and 0.19. The $\sigma_{MAP@20}$ of CLOF is 0.084, and is lower than that of all other methods. The result shows that the CLOF method has a superior and stable performance when detecting type MIX.

Figure 7 (b) gives the average time consumption of the 5 methods for one detection of the whole 30500 load profiles. The test was done on AMD Ryzen 95900@4.7GHz desktop computer with 64GB RAM. Among the 5 methods, the CFSFDP is the most time consuming while PCC is the least. The time consumption of the CLOF is 4,76s, which still is the lowest in the three outlier-based methods.

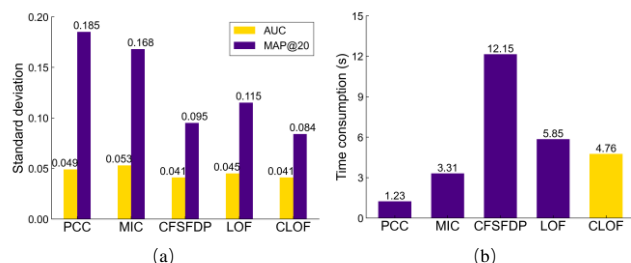


FIGURE 7. (a) The standard deviations of the evaluation results. (b) The time consumption of the 5 methods.

D. SENSITIVITY ANALYSIS

In this part, we attempt to explore the impact of number of electricity thieves on the accuracy of above 5 methods. We hold the number of users per area to 50 and change the number of electricity thieves from 2 to 16 (step size is 2). Figure 8

shows the AUC and MAP@20 of the 5 methods in this progress when detecting type MIX.

We can see from the AUC and MAP@20 values that PCC and MIC perform well under the conditions of fewer electricity thieves. However, with the number of electricity thieves increasing, the AUC and MAP@20 values of PCC and MIC drop rapidly. The three outlier-based methods all behave robustly against the number increasing of electricity thieves. Among them, CLOF maintains excellent performance for both AUC and MAP@20.

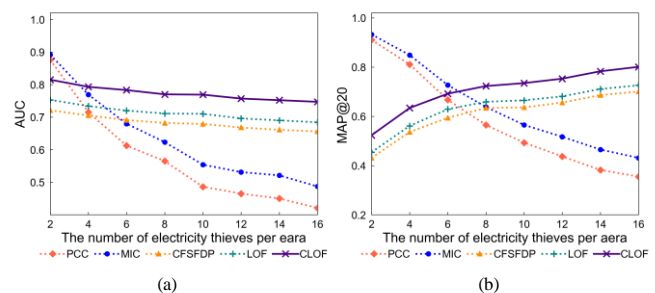


FIGURE 8. Performance of the 5 methods with different numbers of electricity thieves per area. (a) AUC values of the methods. (b) MAP@20 values of the methods.

V. CONCLUSION

In this study, we proposed a CLOF based method for electricity theft detection in AMI. By combining k -means and LOF together, this method utilizes LOF to calculated the anomaly degree of outlier candidates selected by k -means. And a detection framework for practical application is designed. Numerical experiments based on realistic dataset from SGCC with 7 attack types shows that, the proposed method exhibit excellent performance in all attack types except type 1. Thus, our method outperforms other approaches in detecting type MIX which is closer to the real scene. Considering the fact there is no one-fit-all solution to handle all sorts of attack types, the CLOF method is of high value in practical application.

However, there are also some limitations in the proposed method. First, the proposed method only analyzes electricity consumption data alone, which may contain limited information. In addition to meter reading data, the other information such as climatic factors (temperature), regional factors, and some electric factors (current and voltage) is worth being studied in the future. Second, our method dose not specialize in detecting linear FDI (type 1), which is adopted by most physical attacks. Therefore, it is worthwhile for us to investigate how to supplement the detection for linear FDI in next step.

REFERENCES

- [1] S.S.S.R. Depuru, L. Wang, and V. D Devabhaktuni, "Electricity theft: overview, issues, prevention and a smart meter based approach to control theft," *Energy Policy*, vol. 39, no. 5, pp. 1007-1015, 2001.
- [2] Northeast Group LLC, "Electricity theft and non-technical losses: global markets, solutions, and vendors," May, 2017. [Online] Available: <http://www.northeast-group.com/reports/Brochure-Electricity%20Theft%20&%20Non-Technical%20Losses%20-%20Northeast%20Group.pdf>.

- [3] Q. Chen, K. Zheng, and C. Kang et al. "Detection methods of abnormal electricity consumption behaviours: review and prospect," *Automation of Electric Power Systems*, vol. 42, no. 17, pp. 189-199, 2018. (in Chinese)
- [4] Fujian Daily, "The first high-tech smart meter electricity theft case in China reported solved," 2013. [Online]. Available: <http://news.sina.com.cn/c/2013-12-08/081028915975.shtml>
- [5] A. A. Cárdenas, S. Amin, G. Schwartz, R. Dong and S. Sastry, "A game theory model for electricity theft detection and privacy-aware control in AMI systems," in *Proc. 50th Annu. Allerton Conf. Commun., Control, and Comput.* 2012, pp. 1830-1837.
- [6] S. Amin, G. A. Schwartz, A. A. Cardenas, and S. S. Sastry, "Game theoretic models of electricity theft detection in smart utility networks: Providing new capabilities with advanced metering infrastructure," *IEEE Control Syst.*, vol. 35, no. 1, pp. 66-81, Feb. 2015.
- [7] J. B. Leite, J. R. Sanches Mantovani, "Detecting and locating non-technical losses estimation in modern distribution system," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1023-1032, 2018.
- [8] L. M. R. Raggi, F. C. L. Trindade, V. C. Cunha, and W. Freitas, "Non-technical loss identification by using data analytics and customer smart meters," *IEEE Trans. Smart Grid*, vol. 35, no. 6, pp. 2700-2909, 2020.
- [9] S. Salinas, M. Li, and P. Li, "Privacy-preserving energy theft detection in smart grids: A P2P computing approach," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 257-267, Sep. 2013.
- [10] K. Zheng, Q. Chen, and Y. Wang et al., "A novel combined data-driven approach for electricity theft detection," *IEEE Trans. on Industrial Informatics*, vol. 15, no. 3, pp. 1809-1819, Mar. 2019.
- [11] M. De Nadai and M. van Someren, "Short-term anomaly detection in gas consumption through ARIMA and Artificial Neural Network forecast," in *Proc. IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS)*, Trento, Italy, 2015, pp. 250-255.
- [12] D. Mashima and A. A. Cárdenas, "Evaluating electricity theft detectors in smart grid networks," in *Research in Attacks, Intrusions, and Defenses*. Berlin, Germany: Springer-Verlag, 2012, pp. 210-229.
- [13] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216-226, Jan. 2016.
- [14] Z. Zheng, Y. Yang, X. Niu et al., "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Industrial Informatics*, vol. 14, no. 4, pp. 1606-1615, Apr. 2018.
- [15] Y. He, G. J. Mendis, J. Wei, "Real-time detection of false data injection attacks in smart grid: a deep learning-based intelligent mechanism," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2505-2516, 2017.
- [16] M. Ismail, M. F. Shaaban, M. Naidu, and E. Serpedin, "Deep learning detection of electricity theft cyber-attacks in renewable distributed generation," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3428-3437, 2020.
- [17] L. Tian and M. Xiang, "Abnormal power consumption analysis based on density-based spatial clustering of applications with noise in power systems," *Automation of Electric Power Systems*, vol. 41, no. 5, pp. 64-70, 2017. (in Chinese)
- [18] A. Takiddin, M. Ismail, U. Zafar, and E. Serpedin, "Robust electricity theft detection against data poisoning attacks in smart grids," *IEEE Trans. Smart Grid*, vol. 12, no. 3, pp. 2675-2684, 2021.
- [19] D. Birant, A. Kut, "Spatio-temporal outlier detection in large databases," *Journal of Computing and Information Technology*, vol. 14, no. 4, pp. 291-297, 2006.
- [20] E. W. S. Angelos, O. R. Saavedra, O. A. C. Cortes, 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.
- [21] C. C. O. Ramos, A. N. de Sousa, J. P. Papa, and A. X. Falcao, "A new approach for nontechnical losses detection based on optimum-path forest," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 181-189, Feb. 2011.
- [22] J. Yackle and B. Tang, "Detection of Electricity Theft in Customer Consumption Using Outlier Detection Algorithms," 2018 *1st International Conference on Data Intelligence and Security (ICDIS)*, 2018, pp. 135-140.
- [23] M. M. Breunig and H.-P. Kriegel et al., "LOF: Identifying density based local outliers," *ACM Sigmod Rec.*, vol. 29, no. 2, pp.93-104, 2000.
- [24] J. Tao, "Clustering-based and density outlier detection method," M.S. thesis, South China University of Technology, Guangzhou, China, 2014. (in Chinese)
- [25] A. Li, S. Teng, "Application of web mining technology to click fraud detection," *Computer Engineering and Design*, vol. 33, no. 3, pp. 954-962, Mar. 2012. (in Chinese)
- [26] Y. Teng, J. Wu, Z. Zhang, Z. Jiang, Q. Huang, "Online identification of measurement abnormality fault based on outlier detection for current transformer in high voltage shunt reactor," *Transactions of China electrotechnical society*, vol. 34, no. 11, pp. 2405-2414, Jun. 2019. (in Chinese)
- [27] J. Wang, B. Wang, M. Qu, and L. Zhang, "Hardware trojan detection based on improved Euclidean distance," *Computer Engineering*, vol. 43, no. 6, pp. 92-96, Jun. 2017. (in Chinese)
- [28] J. Davis and G. Mark, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233-240.
- [29] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2006, pp. 11-18.
- [30] D. N. Reshef et al., "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518-1524, 2011.
- [31] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492-1496, 2014.
- [32] W. Han and Y. Xiao, "Combating TNTL: Non-technical loss fraud targeting time-based pricing in smart grid," in *Proc. Int. Conf. Cloud Comput. Sec.*, 2016, pp. 48-57.