Electricity stealing time recognition method based on difference and K-means clustering

Yang Yining¹, Xue Yang¹, Cai Hui², Pan Boliang¹, Chen Jianyu², Li Jing²*, Guo Qian²

¹Zhejiang Huayun Information Technology Co., Ltd, Hangzhou, Zhejiang, 310000, China
²China University of metrology, Hangzhou, Zhejiang, 310018, China
*Corresponding author’s e-mail: ljagu@163.com

Abstract: The time of stealing electricity is an important element of stealing electricity. In order to solve the disadvantages of traditional manual determination of stealing time and make the verified stealing users pay electricity charges accurately, this paper proposes a new kind of method to determine the time of stealing electricity. In this method, the difference and clustering algorithms are combined through the electricity fluctuation rate. When the power fluctuation is small, the stealing time is determined through detecting an abnormal value with the difference distribution combined box diagram method; when the power fluctuation is large, the low power period of several consecutive days is found by clustering to determine the stealing time. The example demonstrates the high accuracy of this method and its value in practical applications.

1.Introduction
With the development of economy, people's living standards have been improved and their expenses on electricity consumption have increased accordingly. In order to decrease the electric cost, some corporations and individuals have the tendency to take risks to steal electricity in various ways. Electricity stealing not only makes power supply enterprises suffer from unnecessary economic losses, but also influences the order of supply and causes potential risks of the safety operation of grid.

Current researches on identification of electricity stealing mainly focus on the identification of electricity stealing behavior. Identification of stealing time is not being discussed in previous researches.

Literature [4] achieves an accurate recognition by combining clustering algorithm with outlier detection algorithm to analyse the abnormal electricity data. Literature [5] proposes a layered system of abnormal event detection, which leads to anomaly detection and real-time processing of trajectory data. Literature [6] proposes to find the optimal cluster of h-means clustering through particle swarm optimization and identify the abnormal data through fuzzy method combining clustering based on distance with outliers. Literature [7] suggests that when the fluctuation rate of electric quantity is large in a period time, it is considered that electric stealing happened during that time. Hence, the abnormal power and the corresponding time can be discovered according to this principle.

Therefore, this paper conducts a research aiming at the recognition of electricity stealing time. When an individual steals electricity on a certain day, electricity consumption will decrease sharply compared with the previous day; when an individual stops stealing, electricity consumption will rise
compared with the previous day. The difference algorithm can be used to find out the catastrophe point of these two cases. In addition, K-means algorithm can be adopted to recognize the abnormal electric quantity, and then to determine the time of stealing. Considering that different users may have distinct energy fluctuation and efficiency of calculation, this paper proposes a method to identify the time of electricity stealing combining with difference algorithm and K-means clustering to reduce the influence of fluctuation rate of electric quantity and to shorten the time of calculation.

2. Materials and Methods

2.1. Data Pre-Processing

The data collected in this paper is from electric energy information acquisition system. Due to the fact that electricity power gathering system is affected by external factors such as calculation failure, the collected data may be missing. In this case, it is necessary to pre-process the data to fill in missing values before being used. This paper fills missing values through linear interpolation. The specific steps are as follows:

1) Calculate the slope between known value before and after the missing data.

\[ k = \frac{(b_2 - b_1)}{(n + 1)} \]  

(1)

Where \(n\) is number of missing data, \(b_1\), \(b_2\) is the data before and after the missing data.

2) Calculate the corresponding missing data.

\[ a(i) = b_1 + k \times i \]  

(2)

Where \(i\) is the order number of missing data.

2.2. Algorithm Steps

This paper analyses data through difference algorithm and K-means clustering. The first step is to calculate the fluctuation rate of electric quantity. if the rate is small, difference algorithm combining with box-plot will be applied to find out the unusual point of electric quantity; if the rate is large, K-means clustering will be used to obtain stealing time.

The algorithm flow is in Fig.1, specific steps are as follows:

1) Process the electricity data to fill in missing data

2) Calculate the fluctuation rate of electric quantity of individuals, then choose the algorithm according to the fluctuation rate. After lots of experimental verifications, the threshold of rate is set to be 0.3.

3) If the fluctuation rate of electric quantity is small, process data through difference at first, then use box-plot to detect the maximum and minimum abnormal data. The maximum data is the biggest data of data at the bottom edge; it is the point corresponding to the closing time of electricity theft. The minimum data is the smallest data of data at the top edge; it is the point corresponding to the start time of electricity theft; if the fluctuation rate of electric quantity is larger. When the smallest outlier appears later than the largest outlier, clustering will be used for identification.

4) If the fluctuation rate of electric quantity is large, cluster the electricity data and select the category with the lowest electricity consumption. In order to reduce the impact of low electric quantity due to holidays or power outages, etc., low electricity data points that exceed 7 consecutive days will be considered as power stealing time.
3. Algorithm Principle

3.1. The Fluctuation Rate of Electric Quantity
The fluctuation rate of electric quantity can be used to describe the fluctuation and discretization of electricity data. This helps the paper to choose a suitable algorithm. Adopting the coefficient of variation CV to describe the fluctuation of electricity consumption data, which is defined as,

\[ CV = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (d_i - \bar{d})^2}}{\bar{d}} \]  

(3)

\( d_i \) is daily power consumption of users, \( \bar{d} \) is means of daily power consumption. \( N \) is the number of days, when the \( CV \) is larger, the fluctuation degree is larger.

3.2. Difference Algorithm

3.2.1. The Principle of Difference
The first difference refers to the difference between two consecutive terms in the discrete function. For function \( y=f(x), \ x > 0, \) when \( x=0,1,2,\ldots, \) corresponding function values is \( f(0), f(1), f(2),\ldots, \)
denoted by \( y_1, y_2, y_3 \ldots \). When the independent variable changes for \( x-1 \) to \( x \), the variation of \( y = y(x) \) is:

\[
\Delta y_x = y(x) - y(x - 1) (x = 1, 2, \ldots)
\]

\( \Delta y_x \) is the first difference of the function at point \( x \), which always represents as:

\[
\Delta y_x = y_x - y_{x-1} (x = 1, 2, 3, \ldots)
\]

### 3.2.2. Box-Plot

After differential processing, it is necessary to find out the abnormal data, that is, the point of difference changes suddenly because of the mutation of electricity quantity. This paper adopts box-plot to verify the outliers.

Box-plot is a statistical graph used to display dispersion of a set of data. Box-plot uses five statistics in data to describe data, which is bottom edge, top edge, upper quartile, lower quartile and median, such as Fig. 2

Suppose a group of data has \( n \) data quantity, and sort them from small to large, then the lower quartile \( Q_1 \) is shown as:

\[
Q_1 = \frac{n+1}{4}
\]

The median \( Q_2 \) is shown as:

\[
Q_2 = \frac{n+1}{2}
\]

The upper quartile \( Q_3 \) is shown as:

\[
Q_3 = \frac{3(n+1)}{4}
\]

The top edge \( Q_4 \) is shown as:

\[
Q_4 = Q_3 + 1.5 \times (Q_3 - Q_1)
\]

The bottom edge

\[
Q_5 = Q_1 - 1.5 \times (Q_3 - Q_1)
\]

### 3.3. K-means Clustering Algorithm

The main idea of K-means clustering algorithm is to divide the data set into different clustering categories through iterative and make the criterion function of evaluating clustering performance optimized at last. Usually, the error square sum criterion is used to evaluate the clustering performance. For a given data set \( X \), it includes only descriptive attribute but not categorical attributes. Assuming that the data set \( X \) contains \( K \) cluster subsets \( X_1, X_2, \ldots, X_k \), the samples in each cluster subset are \( m_1, m_2, \ldots, m_k \), and the cluster centers of each cluster subset are \( m_1, m_2, \ldots, m_k \). The formula of the error square sum criterion is:

\[
E = \sum_{l=1}^{K} \sum_{p \in X_l} \|p - m_l\|^2
\]
\( m_i \) is clustering centers, \( p \) is sample values in a clustering.

K-means clustering algorithm inputs the number of clusters \( K \) and a database containing \( n \) objects, outputs \( K \) clusters to make the sum of squared errors is minimized. The algorithm steps are as the following:

1) Confirm the method to decide the distance, this paper uses Euclid distance, the formula of Euclid distance is as follows:

\[
D(g_1, g_2) = |g_1 - g_2|
\]

(12)

2) Randomly determine an initial cluster center for every cluster, so that there are \( K \) initial cluster centers, and then set the number of iterations.

3) Assign the samples in the sample set to the nearest cluster according to the minimum distance principle.

4) Use the mean value of samples in clustering as the new clustering center

5) Repeat step (3) and step (4) \( n \) times.

6) Cluster end, get \( K \) clusters.

4. Result and discussion

Use the algorithm in this paper to calculate the daily electricity consumption data of 8 users who have been verified for electricity theft in a certain area of Zhejiang Province from April 1, 2017 to September 30, 2017. The discriminant results of stealing time are shown in Table.1

Table.1 Discrimination of Power Stealing Time of Special Transformer Users in 2017

| Household number | fluctuation rate | result       | Actual power theft time | True/False |
|------------------|------------------|--------------|--------------------------|------------|
| 1                | 0.214            | 6-17—6-28   | 6-18—6-29                | True       |
| 2                | 0.594            | 6-29—7-27   | 6-30—7-31                | True       |
| 3                | 0.413            | 4-30—5-12   | 9-1—9-12                 | False      |
| 4                | 0.331            | 9-8—9-20    | 9-6—9-21                 | True       |
| 5                | 0.134            | 6-11—6-9    | 6-10—6-22                | False      |
| 6                | 0.380            | 7-24—7-31   | 7-24—7-31                | True       |
| 7                | 0.196            | 7-22—7-29   | 7-22—7-30                | True       |
| 8                | 0.309            | 9-12—9-20   | 9-11—9-22                | True       |

After a lot of experimental calculations, the threshold of CV is determined to be 0.3. It can be discovered from Table.1 that the method of this paper can correctly determine the power-stealing time of 6 users except User 3 and User 5. The time error is within one or two days.

Besides, this paper compares the running time of user 1 and user 2 whose volatility is around the threshold with the running time of only using the clustering algorithm. This is shown in Table.2. From Table.1 and Table.2, it can be observed that the power fluctuation rate of user 1 is 0.214. Using the differential algorithm, the time of calculation is 0.52s, Using the clustering algorithm, the time of calculation is 1.83s, which is 3 times that of the differential algorithm. User’s power fluctuation rate is 0.594, using the clustering algorithm, the operation time is 1.81s. Therefore, for users whose fluctuation rate is less than 0.3, directly using the difference method for identification can save certain computing resources.

Table.2 Comparison of Operation time between Difference Method and Clustering Method

| Household number | Time of calculation(s) | Time of calculation of clustering(s) |
|------------------|------------------------|-------------------------------------|
| 1                | 0.52                   | 1.83                                |
| 2                | 1.81                   | 1.81                                |

The following cases choose users 1 to 3 as examples to explain the results of the algorithm. Fig.3, 4, and 5 show the results of the power theft time of user 1, user 2, and user 3 respectively.
Fig. 3 Identification Result of Power Stealing Time of User 1

Fig.3(a) is the daily power consumption curve of user 1. The fluctuation rate of user 1’s power is 0.214, which is less than 0.3, using the difference algorithm. From the difference result in Fig.3(b), it is obvious that the user’s power consumption is abnormal. With the help of the box-plot method, the minimum and maximum abnormal values of the user are -0.25 and 0.17 respectively, as shown in Fig.3(c). The minimum abnormal value and maximum abnormal value of the user correspond to June 17 and June 28 in Fig.3(b) respectively. This is consistent with the actual electricity theft time.
Fig. 4 Identification Result of Power Stealing Time of User 2

Fig.4(a) is the power consumption curve of user 2. The fluctuation rate of the user is 0.594, using the clustering algorithm. Fig.4(b) is obtained by clustering the user’s power data. "•" is the lowest class which occurs from June 29 to July 27, that is, the minimum power consumption exceeds 7 consecutive days. It is believed that this time period is the power stealing time, which is basically consistent with the actual power stealing time from June 30 to July 31.

Fig. 5 Identification Result of Power Stealing Time of User 2

Table.1 shows that the power fluctuation rate of user 3 is 0.413, using the clustering algorithm, the identification result is from April 30th to May 12th. The actual stealing time is from September 1st to September 12th. It can be seen from Fig.5(a) that the user has a low battery during the two periods.
from April 30th to May 12th and from September 1st to September 12th. The clustering puts two period time into the lowest class. Moreover, the number of their consecutive days exceeded 7 days, as shown in Figure 5(b), but the clustering algorithm finally considers from April 30 to May 12 as the time of power stealing, which resulted in a wrong identification.

5. Conclusion
This paper proposes a way to determine electricity stealing time. The method analyses the data through comparing difference algorithm and clustering algorithm. It chooses suitable algorithms according to different fluctuation rate of electricity. The experiment results demonstrate that, this method not only has a high accuracy, but also saves the time of calculation by solving the traditional shortcomings of manually determining the time of stealing electricity. Furthermore, the proposed method has considerable value in practical applications.

However, the algorithm currently still has its shortcomings. For example, when users have low power consumption for several times, the use of this algorithm may identify wrongly; when the user steal electricity intermittently or steal again after being caught, this method may not be able to make accurate determination. In the future, other methods could be introduced to improve the algorithm in this paper so as to improve the accuracy of identification.

Acknowledgments
This research was supported by the basic public welfare research program of Zhejiang Province (LGG20E070003)

References
[1] Wang Qingning, Zhang Donghui, sun Xiangde, et al. Research and application of anti-stealing power system based on GA-BP neural network [J]. Electrical measurement and instrumentation, 2018,55 (11): 35-40
[2] Dou Jian, Liu Xuan, Lu Jizhe, et al. Research on anti-stealing method based on big data of electricity consumption information collection [J]. Electrical measurement and instrumentation, 2018,55 (21): 43-49
[3] Huang Rongguo, Lu Chunguang, Yao Li, Xu Tao, Zhou you. Design and implementation of anti-stealing power simulation test platform [J]. Electrical measurement and instrumentation,2018,55 (21): 43-49
[4] Li Ning, Yin Xiaoming, Ding Xuefeng, et al. A power stealing identification method based on clustering and outlier detection algorithm [J]. Electrical measurement and instrumentation, 2018,55 (21): 19-24
[5] Lee S C, Nevatia R. Hierarchical abnormal event detection by real time and semi-real time multi-tasking video surveillance system[J]. Sung Chun Lee, Ram Nevatia. Machine Vision and Applications. 2014, 25(1): 1253-1269.
[6] Karami A, Guerrero-Zapata M. A fuzzy anomaly detection system based on hybrid PSO-Kmeans algorithm in content-centric networks[J]. Neurocomputing, 2015, 149: 1253-1269.
[7] Li Ning, Wang Weifeng, Cai Hui, et al. Research on identification method of stealing electricity based on outlier mining [J]. Journal of China University of metrology, 2018, 29 (03): 331-337
[8] Li Menghui, Sang Xiaotian, Tian Zhenchuan, et al. Application of box diagram in the detection of abnormal value of cotton heterozygosity in the Yangtze River Basin [J]. Hubei agricultural science, 2016, 55 (11): 2895-2898 + 2954
[9] Xie Zhiwei, Wen Ruigang, Meng Anbo, et al. Data processing and prediction of construction personnel based on box diagram and isolated forest [J]. Journal of engineering management, 2018, 32 (05): 92-96
[10] Wang Dan, Huang Xiaoli. Noninvasive load identification based on improved clustering and DTW [J]. Information and communication, 2018 (07): 27-31
[11] Teng Minggui, Wang Rujing, Ma Xianzhang. Spatial clustering based on the least square fitting error criterion [J]. Minicomputer system, 2005 (07): 1221-1224

[12] Zhou Jiangli, Tao Yizheng, Ji bin, Xie Yonghui. Optimization method of initial clustering center based on minimizing the sum of squares of errors [J]. Computer engineering and applications, 2018, 54 (15): 48-52

[13] Hu Wenyan, Chen Hongkun, Yang Xiaonan, Xie Cheng. Urban distribution network cluster planning method based on improved k-means clustering [J]. Electrical measurement and instrumentation, 2016, 53 (21): 10-15