AN OUTLIER DETECTION METHOD OF LOW-VOLTAGE USERS BASED ON WEEKLY ELECTRICITY CONSUMPTION

To cite this article: Tang Xiao-feng *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **631** 042004

AN OUTLIER DETECTION METHOD OF LOW-VOLTAGE USERS BASED ON WEEKLY ELECTRICITY CONSUMPTION

TANG Xiao-feng [1] ,HUANG Rui [1] ,CHEN Qi[2], PENG Zhao-yi[2] , WANG Hao[2] , WANG Bi-heng[2]

[1.] POWER GRID JIANGSU ELECTRIC POWER CO.LTD YANGZHOU POWER SUPPLY BUREAU , Yangzhou 225000, China;

[2.]NARI TECHNOLOGY DEVELOPMENT CO.,LTD, Nanjing 210060,China

Abstract: A method of detecting abnormal electricity consumption behavior of low-voltage electricity consumption users based on weekly electricity consumption is proposed. Firstly, features of weekly electricity consumption curve data are extracted from the actual daily consumption data. Fuzzy C-means clustering algorithm is used to classify electricity consumption behavior. Typical weekly electricity consumption characteristic curves of low-voltage power users are obtained, and then low-voltage users with abnormal electricity consumption are marked based on distance. The effectiveness and feasibility of the algorithm are verified by the actual electricity consumption data of 15 stations in Yang-Zhou City.

1.Introduction

As the scale of the power grid continues to expand, the degree of informationization of power systems continues to increase, and the new generation of control, measurement, and data processing technologies such as the Internet of Things, cloud computing, and data mining continue to evolve. Big data applications have involved various business areas of power companies[1][2][3]. The research applies to the low-voltage user electricity behavior data mining algorithm and establishes the user electricity behavior abnormality monitoring and early warning system, which helps the power supply enterprise to make reasonable speculation and early warning to the user's abnormal power consumption behavior, and timely discover the user's electricity consumption process. The abnormal use of electricity, such as electricity theft, violation of regulations and over-capacity, is of great significance for grid security and reducing the cost of electricity auditing by power supply companies[4].

In recent years, relevant research has been carried out on the abnormal power consumption detection of users at home and abroad, and some progress has been made. At present, research on power anomalies is generally focused on industrial users, which is based on user's 15-minute load curve data and focus on clustering algorithms that include K-means clustering algorithm [5], fuzzy C-means clustering algorithm [6]][7], DBSCAN algorithm [8], etc. By clustering the 15-minute load data, a typical load curve is obtained, and then compared with the typical load characteristic curve of the corresponding industry, the users who may have abnormal power consumption are analyzed. With the application of deep learning and artificial neural networks in power systems, the research on abnormal power detection based on sparse coding model [9] and the research on user tampering behavior detection model based on BP neural network algorithm [10] [11] and so on. In addition, there are some abnormal power monitoring studies based on current, voltage, imbalance and local reference power data [12] [13].

Currently, the research on low-voltage users mainly focuses on the analysis of electricity characteristics [14] and the study of electricity consumption [15], and lacks research on the abnormality detection of power consumption for low-voltage users, because the number of low-voltage users is huge. For low-voltage users, the frequency of power collection is usually once a day. The data collection of voltage and current is not collected or passively collected in real time. Therefore, the effective data available for research is only the daily power. In view of this situation, the daily electricity basic data in the paper is applied to construct the weekly electricity consumption characteristic curve of low-voltage users, and the electricity consumption behavior is analyzed based on the fuzzy C-means clustering algorithm; the typical weekly electricity consumption curve of the three types of low-voltage power users is obtained. A distance-based method is then used to mark low-voltage users with abnormal power usage.

2.Classification of low-voltage users' electricity behavior based on weekly electricity consumption feature clustering

The classification of low-voltage users' electricity behavior is based on the user's electricity behavior habits, and the process of dividing the same power behavior into the same class. The specific process is shown in Figure 1.

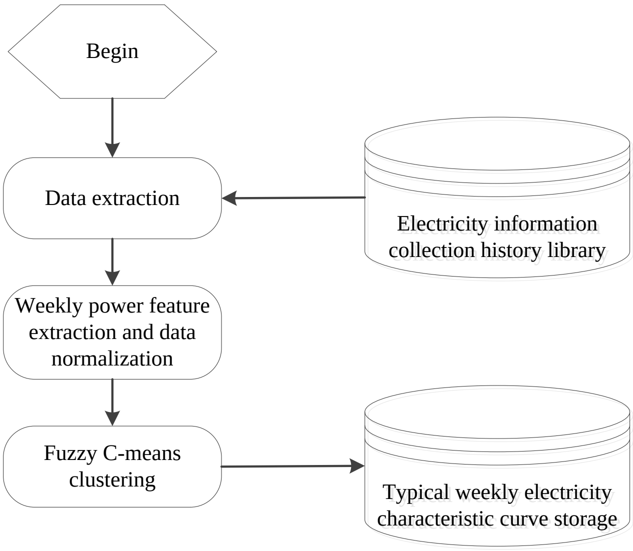

Begin
Data extraction
Electricity information collection history library
Weekly power feature extraction and data normalization
Fuzzy C-means clustering
Typical weekly electricity characteristic curve storage

Figure 1. Flow Chart of Classification of electricity consumption behavior

In order to calculate the accuracy of the results, the abnormal data is eliminated at the same time as the data extraction process, which includes continuous 0 data and missing data.

2.1 Weekly power feature extraction

Feature extraction is very important for data mining. Good features can improve the performance of models and algorithms. The main purpose of feature extraction is to reduce the feature quantity and dimension reduction, and make the model generalization ability stronger.

If the daily electricity consumption data is used directly as a feature, the dimension of the vector is very large, which is not conducive to the implementation of the algorithm. According to the characteristics of the low-voltage users, the weekly power consumption of the user is relatively stable. Therefore, the weekly electricity consumption is selected as the feature data, and the user's typical weekly electricity consumption characteristic curve is used as the cluster sample data.

The typical weekly electricity consumption characteristic curve of the user is calculated by the weighted average method, and the calculation method is as follows.

Assuming the sample set, the user's electricity consumption data for the i-week is $W_i = \{w_{i1}, w_{i2}, w_{i3}, w_{i4}, w_{i5}, w_{i6}, w_{i7}\}$, and a total of n weeks of history is extracted, then the weight coefficient of the *j-th* day of the *i-th* week is

$$r_{ij} = \frac{w_{ij}}{\sum_{i=1}^{n} w_{ij}} \quad j = 1,2 \ldots 7 \quad (1)$$

The weighted average power consumption of the user in j days is

$$W_{ij}^{'} = \sum_{i=1}^{n} (r_{ij} w_{ij}) \quad (2)$$

2.2 Data set normalization

In the power data, the user has a large difference in power consumption due to different types. If the power consumption is used as the clustering feature directly, users with large magnitude will have a great influence on the clustering result in the clustering process; therefore, the power consumption data needs to be normalized to limit the data to [0,1] within the scope. The standard deviation linearly transforms the original data and maps the result to the [0,1] interval. Let the data set X=$\{x_1,x_2,\cdots x_n\}$ that are a total of n data objects, $x_i = (x_{i1}, x_{i2}, \ldots x_{im})$that are m attributes, then normalized is as shown formula (3).

$$x_{ij}^{'} = \frac{x_{ij} - \min\{x_{ij}\}}{\max\{x_{ij}\} - \min\{x_{ij}\}} \quad (3)$$

Where $\max\{x_{ij}\}$ and $\min\{x_{ij}\}$ are the maximum and minimum values of the *j-th* attribute, respectively.

The algorithm of the standard deviation method is simple to implement, and the data is linearly transformed without changing the characteristic form of the data.

2.3 Fuzzy C-means clustering algorithm and clustering performance index

The fuzzy C-means clustering algorithm divides n vectors x_i (i=1,2,…, n) into c fuzzy groups, and obtains each cluster center to minimize the objective function.

The data set is to be clustered $X = \{x_1,x_2,\cdots x_n\}$, $x_k \in R^p$, each data has p features, and the data set has n elements, which are classified into c classes. The fuzzy C partition can be represented by a membership matrix $U = [u_{ij}]$, $(i = 1,2, \ldots c;\ j = 1,2 \ldots, n)$ of c rows and n columns, where u_{ij} indicates that the *j-th* item data belongs to the degree of membership of the *i-th* class, and the following conditions are satisfied.

$$\begin{gathered} \forall i, j, u_{ij} \in [0,1] \\ \forall j, \sum_{i=1}^{c} u_{ij} = 1 \qquad (4) \\ \forall i, \sum_{j=1}^{n} u_{ij} > 0 \end{gathered}$$

V is a collection of c cluster centers.

$$V = \{v_1,v_2,\cdots v_c\}, \ v_k \in R^p \quad (5)$$

The goal of the fuzzy C-means algorithm is to minimize the sum of the weighted distances of the points of the data set X to the respective cluster centers.

$$\begin{aligned} J_m(U,V) &= \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{\ m} d_{ij}^{\ 2} \\ &= \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{\ m} \left| x_j - v_i \right|^2 \end{aligned} \qquad (6)$$

Where $d_{ij}{}^2$ is the Euclidean distance between the element x_j and the center point v_i, and m is a fuzzy parameter, and the value is generally 2-7.

Let J_m find the partial derivatives of u_{ij} and v_j, and the necessary condition for minimizing the objective function is

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}{}^m x_j}{\sum_{j=1}^{n} u_{ij}{}^m} \quad (7)$$

and

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{m-1}}} \quad (8)$$

The algorithm steps of fuzzy C-means clustering are as follows.

1) Initializing the membership matrix U to satisfy the condition of equation (4).

2) Calculating *c* cluster centers v_i, *i=1, 2,...,c* by using equation(7)

3) Calculating the target value function according to equation (6). If the value function is less than a certain threshold or the amount of change is less than a certain threshold, the algorithm stops.

4) Otherwise, recalculating the membership matrix U by using equation (8) and return to step 2).

The fuzzy C-means algorithm is simple in design and easy to implement by computer. The disadvantage is that the value of the initialization membership matrix is good or bad, and it is easy to fall into the local saddle point.

The selection of the cluster number and the initial membership matrix in the fuzzy C-means algorithm has a great influence on the clustering result. The value of the general cluster number c is between 2 and$\sqrt{n}$. The number of clusters is reversely reversed by using the silhouette coefficient index to measure the performance of the cluster.

The calculation steps of the contour coefficient index are as follows.

1) Calculating the average distance $\mathrm{a_i}$ of the sample x_i to other samples of the same type. The smaller the $\mathrm{a_i}$, the more accurate the sample classification.

2) Calculating the average distance b_{ij} of all samples of the sample x_i to other classes, and defining $b_i = min(b_{i1}, b_{i2}, \dots, b_{in})$, the larger the b_i, the less the sample *i* belongs to other clusters.

3) According to the in-cluster dissimilarity a_i of the sample *i* and the dissimilarity degree b_i between the clusters, the silhouette coefficient of the sample *i* is defined as

$$\mathrm{S_i} = \frac{\mathrm{b_i - a_i}}{\mathrm{max(a_i, b_i)}} \quad (9)$$

The closer $\mathrm{S_i}$ is to 1, the more reasonable the sample *i* clustering is; the $\mathrm{S_i}$ is close to -1, indicating that the sample *i* should be classified to another cluster; if $\mathrm{S_i}$ is approximately 0, the sample *i* is on the boundary of the two clusters.

According to the experimental data of CH index and silhouette coefficient index, the clustering effect is best when the number of clusters selected in the project is 3.

3.Power anomaly detection

The specific steps of the power anomaly detection process are as follows.

1) The daily electricity consumption of the user is obtained from the collection of electricity consumption information, and the weekly electricity consumption curve data in the user detection period are gained by the weighted average and normalization algorithm;

2) The Euclidean distance between the user's weekly electricity consumption curve and the three kinds of typical weekly electricity consumption characteristic curve is calculated; the Euclidean distance is defined as follows.

$$d = \sqrt{\sum_{i=1}^{7}(x_i - l_i)^2} \quad (10)$$

Where X={x_1， x_2 …, x_7} and L={l_1， l_2 …, l_7} are the user's actual weekly load curve and the weekly load characteristic curve of the cluster center, respectively.

3) If the Euclidean distance of the user data to be detected is less than the threshold of the normal power consumption, the user is considered to belong to the normal user of the class, and the abnormal power consumption detection process is terminated. If the Euclidean distance between the user and the three cluster centers exceeds the normal power consumption threshold, the user is marked as an abnormal power user, and the user data is recorded in the abnormal user data table.

4.Experiment analysis

The experimental data of this project is derived from the actual data of users under the low-voltage station area of Yangzhou Power Supply Company in 2018. The structure is shown in Table 1.

Table 1. Raw Data Structure

Code	Name	Type	Is it empty
Cons_no	user ID	NUMBER16）	No
YEAR_MON	Record Time	DATE	No
KWH_1	No. 1 electricity consumption	NUMBER(11,4)	
…	…	…	…
KWH_31	No. 31 electricity consumption	NUMBER(11,4)	

After data extraction and weighted averaging, the weekly power consumption weighted average characteristic curve of user number 9103603119 is shown in Figure 2.

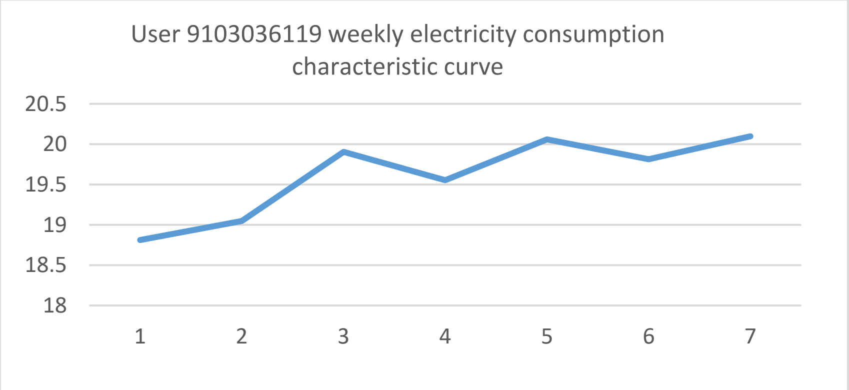


Figure 2 .Characteristic curve of weekly electricity consumption

FIG. 3 is a normalized graph of the data of FIG. 2, and it can be seen that the normalized shape of the curve remains substantially unchanged after normalization.

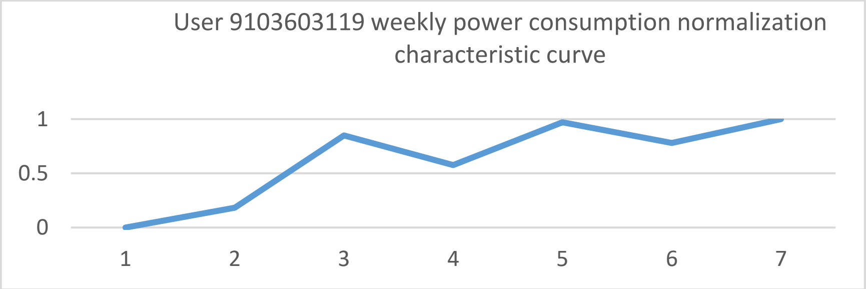


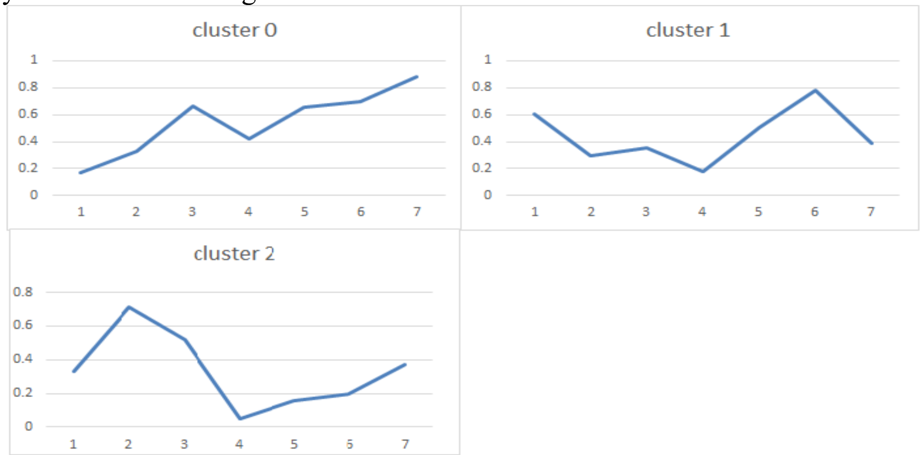
Figure 3. Normalized Characteristic curve of weekly electricity consumption

Table 2 is the contour index of the clustering results when the number of clusters is 2, 3, and 4 respectively. When the number of clusters is 3, the silhouette coefficient index is the largest, which achieves the best clustering effect.

Table2. The performances of clustering indexes

Number of categories	2	3	4
silhouette coefficient value	0.268	0.324	0.291

FIG. 4 is a weekly power consumption characteristic curve of three types of cluster centers after fuzzy C-means clustering.


cluster 0

cluster 1

cluster 2

Figure 4. Characteristic curve of weekly electricity consumption for 3 Clustering Centers

Table 3 shows the Euclidean distances between the three types of normal users and one tampering sample in the electricity anomaly test with the three cluster centers during the power stealing cycle.

Table 3. Euclidean Distance from Cluster Center

User	Cluster center 1	Cluster center 2	Cluster center 3
Normal user of cluster 1	0.340	1.400	0.934
Normal user of cluster 2	1.518	0.235	0.766
ormal user of cluster 3	0.963	1.236	0.283
Suspicious tamper samples	0.904	1.806	1.888

When the threshold value is 0.6, the suspicious sample households can be correctly removed 100%. According to the low-voltage station area of Yangzhou Power Supply Company in 2018, the actual data of the electricity users includes 3 tamper samples. The test results show that the screening rate of suspicious sample households reaches 100%, and the false positive rate is zero.

5.Conclusions

In the paper, a low-voltage user power anomaly detection method based on the low-voltage user's weekly power consumption characteristic curve is proposed. The clustering center is obtained by clustering the normalized low-voltage user's weekly power consumption characteristic curve; then, the Euclidean distance from the clustering center is used to determine whether the user's power consumption behavior is abnormal. The experimental results show that the clustering method and abnormality determination based on the weekly power characteristic curve proposed in the paper are feasible and effective for low-voltage users. In the future, if the conditions permit, the abnormality of the tag will be used to periodically check, such as once every hour, continuous monitoring for 7 days, etc., the voltage, current, unbalance rate and other electrical energy measurement data from the relevant system to further discriminate the tampering index.

References

[1]ZHANG Pei. Application Status and Prospect of Big Data in Electric Power[J]. Power and Electrical Engineers. 2014,12:27-30

[2]WANG Xiang-hong, CHEN Xiao-yi. Current Situation and Development Trend Study of Cloud-computing And Power Big Data Technology[J]。SHAANXI Electric Power, 2017, Vol45（6）:70-72

[3]YAN Yingjie, SHENG Gehao, CHEN Yufen, et al. An Method for Anomaly Detection of State Information of Power Equipment Based on Big Data Analysis[J]. Proceedings of the CSEE Vol35（1）：52-59

[4]ZHUANG Chi-jie，ZHANG Bin，HU Jun, et al. Anomaly Detection for Power Consumption Patterns Based on Unsupervised Learning[J]. Proceedings of the CSEE。2016 Vol36（2）：379-387

[5]CHEN C, Cook D J. Energy Outlier Detection in Smart Enviroments[J].Artificial Intelligence and Smarter Living,2011,11: 07.

[6]Angelos E W S,Saavedra O R,Cortés O A C,el al:Detection and identification of abnormalities in customer comsumption in power distribution systems[J]. IEEE Transaction on Power Delivery,2011,26(4):2346-2442

[7]WANG Peng-wu,SUN Zhi-jie,Fu Jun。Application of Fuzzy C-means Clustering Algorithm in inspection of Abnormal Electricity[J]. North China Electric Power 2016，4：14-18

[8]WANG Gui-lan，ZHOU Guo-liang，ZHAO hong-shan。Fast Clustering and Anomaly Detection Technique for Large-sacle Power Data Stream[J]. Power Automation System 2016，Vol（4）：14-18。

[9]ZHOU Li，ZHAO Lu-jun，GAO wei-guo。Application of Sparse Coding in Detection for Abnormal Electricity Consumption Behaviors [J]. Power System Technology。2015，39（11）：3182-3188

[10]WU Jian, LIN Guo-qiang, WANG Xiao-hui, et al. Research on Electricity Theft Behavior Detection Model Based on Neural Network Algorithm[J]。Electric Power ICT，2017 15（12）：36-40

[11]CAO Zheng,YANG Jing-fei,LIU Xiao-na.Study and application of preventing system from stealing power based on BP neural network[J].Interal Journal Hydroelectric Energy,2011,29(9):199-202

[12]SHANG Qiu-feng，LV Peng-peng，LI Meng-chan, et al。The Abnormal Electricity Monitoring Method Based on Local Power Reference[J]. Electrical Measurement & Instrumentation 2014 Vol51（8）：1-5

[13]ZENG hu。Clustering Analysis of Electric Energy Measurement Data and Research on Detection of Electricity Theft[D]. Kunming university of science and technology，2017.

[14]LEI Jingsheng，YU Xiucheng。Fuzzy C-means Clustering-based Algorithm for the Analysis of Regional Electric Power Characteristics[J]。Journal of Shanghai University of Electric Power. Vol33(2)：196-200

[15]ZHANG Su-xiang, LIU Jian-ming, ZHAO Bing-zhen, et al. Cloud computing-based analysis on residential electricity consumption behavior [J]. Power System Technology. 2013,37（6）：1542-1546