Classification Analysis Method for Residential Electricity Consumption Behavior Based on Recurrence Plot (RP) and Convolutional Auto-Encoder (CAE)

Zhiqing Sun1, Xinyu Deng2,3*, Xianghai Xu1, Weihong Hou1 and Shouxiang Wang3

1State Grid Hangzhou Power Supply Company, Hang Zhou 310009, China.
2Tianjin Xianghe Electric Co. Ltd., Tianjin 300000, China.
3School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
*Corresponding author’s e-mail: xy_deng@tju.edu.cn

Abstract. Load clustering is the foundation of big data mining in power distribution system. It is helpful for power companies to accurately grasp users' electricity consumption habits, improve power quality and develop demand response. To overcome the characteristic redundancy problem of the high-dimensional load data, the load clustering method based on RP and CAE is proposed. Firstly, the one-dimensional load curves are converted into two-dimensional recurrence plot to realize feature enhancement. Secondly, the advanced feature extraction capability of CAE is used to realize load feature extraction and dimension reduction. Finally, the spectrum clustering (SC) is used to analyze the user's electricity consumption patterns. The validity of proposed method is verified by Ireland Smart meter dataset.

1. Introduction
Classification analysis of residential electricity consumption behavior is mining the user's electricity consumption characteristics and obtaining typical load patterns from the massive load curve by clustering, which is the basis of implementing load management. The analysis results are helpful for power companies to perceive users' energy demand and carry out more efficient demand response management. Power companies can also innovate service patterns, such as establishing tiered electricity pricing policies to improve energy utilization efficiency, guiding residents to improve unreasonable energy usage habits, and conducting power quality optimization, to achieve a remote, friendly and interactive intelligent electricity distribution.

Lots of popular clustering algorithms have been successfully applied to load curve characteristics analysis, such as fuzzy C-means (FCM) clustering [1], density-based spatial clustering of applications with noise (DBSCAN) [2] and hierarchical clustering (HC) [3]. However, with the popularity of advanced metering infrastructure (AMI), traditional clustering analysis methods are confronted with challenges. The extensive deployment of smart meters brings massive data on household electricity consumption. The measurement cycle of smart meter reaches 1h, 30min, or even 15min, and the daily load curve contains 24, 48 or 96 points [4]. Such high-dimensional load data makes it possible to deeply explore users' electricity usage habits, but it also leads to the problem of feature redundancy, leading to an unsatisfactory clustering effect. Dimensionality reduction is an effective method to deal with high-
dimensional data analysis, which can be well applied to the load clustering based on smart meters. Commonly used dimensionality reduction algorithms are principal component analysis (PCA) [5][6], t-distributed stochastic neighbor embedding (t-SNE) [7] and wavelet transform [8]. There is also the segmented mean method [9], whose dimensionality reduction results are more interpretable. However, the direct dimensionality reduction method of the one-dimensional load curve may cause greater information loss and make the clustering effect worse.

To meet the challenge of high-dimensional load data, a dimensionality reduction method based on RP and CAE is proposed. RP converts the original one-dimensional load curve into two-dimensional image features to achieve feature enhancement. And CAE can perform non-linear feature extraction on image features. Finally, SC is used for cluster analysis.

The rest of this paper is organized as follows. Section II explains the proposed method. Section III presents the experimental settings and results. Finally, the paper is concluded in Section IV.

2. Methods

2.1. Recurrence plot

Recurrence plot is a feature transformation method for short time series and has been successfully applied in temperature change analysis, brainwave analysis and signal detection. Recurrence plot can encode one-dimensional features into two-dimensional image features, and the diagonally, horizontal and vertical lines of the recurrence plots can well reflect the chaos, stationarity and inherent similarity of the time series. Moreover, the recurrence plot has the advantages of low stability requirements and low sensitivity to noise. Therefore, this paper uses recurrence plot to encode the characteristics of daily load curve of residents to mine the implied characteristics of load curve and enhance the expression ability of effective features. For the given load sequence \( x(i) \), phase space reconstruction is carried out, and the formula is

\[
X(i) = \{x(i), x(i + \tau), \ldots, x(i + (m-1)\tau)\}
\]  

(1)

where \( \tau \) is delay time, and \( m \) represents the embedded dimension. Then the recurrence plot is formed using the following equation:

\[
E_{ij} = \|X_i - X_j\|
\]  

(2)

where \( \| \| \) denotes Euclidean distance. Take a 48-dimensional residential load curve as an example, and its daily load curve and the corresponding recurrence plot are shown in Figure 1.

![Figure 1. Resident load curve and recurrence plot.](image)

2.2. Convolutional auto-encoder

The CAE combines the AE with the convolutional neural network (CNN). The encoder and decoder are respectively composed of a CNN and a deconvolutional network. In the encoding process, the convolutional layer performs local feature perception on the original image features, then the pooling layer performs down-sampling on the features, and finally realizes feature extraction in the fully connected layer. The decoding process is opposite to the encoding process, but it does not require the
encoder and decoder to be completely symmetric. At the same time, multiple convolutional layers and pooling layers can be added in both encoder and decoder to carry out deeper feature extraction. The CAE architecture and parameter are shown in Figure 2. The final deconvolutional layer of the decoder is designed to transform the multi-layer 2d features into a single layer. The encoder and decoder use "Relu" as the activation function of the neural network and the output uses "Sigmoid" function. The initial learning rate of the network was set to 0.01. The optimizer is "Adam", and the loss function is mean square error (MSE).

![Figure 2. Convolutional auto-encoder.](image)

Figure 3 (a) is the original recurrence plot corresponding to a daily load curve, and Figure 3 (b) is the recurrence plot reconstructed by eigenvector. The texture and color of the two are highly similar, which verifies the advanced feature extraction capability of CAE. For convenience, the feature extraction method proposed is denoted as "RP-CAE".

![Figure 3. Effect of feature reconstruction.](image)

2.3. Spectrum clustering
SC is a novel clustering method based on graph theory, which takes all samples as points in the space, calculates the similarity weights of two points according to certain rules to form edges, and then cuts the graph to achieve clustering. SC transforms the clustering problem into the optimal programming problem of the graph and solves the global optimal solution, which can effectively divide the clustering structure of arbitrary shape. The algorithm flow of SC is as follows:
Algorithm Spectrum clustering

**Input:** The eigenvector of RP-CAE and clustering number \( k \).

**Output:** Clusters \( A_1, A_2, \ldots, A_k \).

1. Calculate the adjacent matrix \( W \) and degree matrix \( D \).
2. Calculate the Laplace matrix \( L \).
3. Calculate the eigenvalues of \( L \), take the minimum of the first \( k \) eigenvalues and calculate their eigenvectors \( u_1, u_2, \ldots, u_k \).
4. Construct the eigenvector matrix \( U = \{u_1, u_2, \ldots, u_k\} \).
5. \( K \)-means is used to cluster the row vectors of matrix \( U \) to form clusters \( C_1, C_2, \ldots, C_k \).
6. Output the clustering results \( A_1, A_2, \ldots, A_k \), where \( A_j = \{j \mid y_j \in C_i\} \).

2.4. Metrics

Load clustering is an unsupervised clustering problem. Good clustering results should have high in-class similarity and low inter-class similarity. Therefore, the Davies-Bouldin Index (DBI) and Dunn Index (DI) are adopted as the evaluation index of clustering effect. The calculation formula is

\[
DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j} \left( \frac{C_i - C_j}{d_{min}(C_i, C_j)} \right)
\]

\[
DI = \min_{i \in 1 \ldots K} \left( \min_{i \neq j} \left( \frac{d_{min}(C_i, C_j)}{\max diam(C_i)} \right) \right)
\]

where \( K \) denotes clustering number, \( C_i \) is the average of the inter-sample distances in class \( i \) and \( c_i \) indicates the clustering centroid. \( d_{min}(C_i, C_j) \) represents the minimum distance between the sample of class \( i \) and \( j \), and \( diam(C_i) \) is the maximum distance between samples in class \( i \). It means that the clustering is successful when the DBI is big or DI is small.

3. Results

The experiments are conducted on a desktop PC with Ubuntu 18.04 operating system, Ryzen3700x CPU, and RTX2070 GPU. The deep learning algorithm is implemented using the open-source platform TensorFlow 1.15.0.

3.1. Dataset

The dataset used in this paper was Ireland smart meter dataset provided by the Commission for Energy Regulation (CER) [10]. This dataset contains the load data of 6059 residential consumers over 536 days (7/21/2011 - 1/6/2013) at an interval of 30 minutes. The dimension of original daily load data is 48, and the dimension after feature extraction is 15.

Electricity consumption behavior of residents has strong randomness and volatility, in only one-day load curve cannot accurately reflect the typical pattern, while the scope for too long-time habit might change. Therefore, this section takes the users’ average monthly load curve as their typical electricity consumption pattern, and the specific time range for August 1, 2012 - August 31, 2012. Different load sizes of users may lead to wrong clustering results, so the max-min normalization is adopted to scale the daily load data to \([0,1]\). In the original dataset, there were 6,059 residential users. After removing 8 invalid users whose power was always 0, the dimension of load matrix to be clustered was 6,051 \( \times 48 \). The optimal clustering number was determined as 10 by Gap statistic [11].

3.2. Comparison of dimensional reduction effects

This section compares the feature extraction method proposed with several dimensional reduction algorithms, namely PCA, Kernel PCA, undercomplete auto-encoder (UAE) and long short-term
memory auto-encoder (LSTM-AE). The load data was reduced from 48 to 15 by those dimension reduction algorithms, and the SC results were shown in Figure 4 (a). Compared with clustering by original load data, the DBI and DI of the dimensionality reduction strategy are smaller and larger, indicating that the dimensionality reduction strategy is effective. Among several dimensionality reduction algorithms, the method based on auto-encoder is superior to traditional PCA and Kernel-PCA. And then, benefiting from the feature enhancement of RP and feature extraction of CAE, the effect of RP-CAE is better than that of LSTM-AE and UAE. The results verify the correctness of the RP-CAE applied to load feature extraction.

Figure 4 (b) shows the clustering effect of RP-CAE under different dimension reduction number. When the dimension is 15, the clustering effect reaches the best. And if the dimension is too high or too low, the effect becomes worse. The reason is that when the dimension is too high, there is still feature redundancy in the data, and when the dimension is too low, it is difficult to fully represent information of the load curve. Therefore, the number of dimensions should be slightly higher than the number of clusters.

3.3. Comparison of clustering methods
This section compares several different clustering algorithms, namely SC, FCM [1], gaussian mixture model (GMM) clustering and hierarchical agglomerative clustering (HAC). The experiment tested the clustering effect under the original dimension and RP-CAE dimension reduction strategy respectively, and the results were shown in Figure 5 (a) and (b). SC is superior to the other three methods in terms of indicators, and all clustering algorithms are significantly improved after the application of RP-CAE. The results further verify the effectiveness and correctness of the proposed method.
3.4. Classification results of users' electricity consumption behavior

The proposed method is used to divide the residential electricity consumption behavior into 10 categories, and the cluster centroid is taken as the typical residential electricity consumption pattern, as shown in Figure 6. There are unimodal, bimodal, and stationary patterns in residential electricity consumption behavior. Pattern 1, 4, 5, 6, 8 and 10 are unimodal, and their load curves will reach the peak at a single time point or a small period. For example, the peak values of pattern 1, 4 and 8 are concentrated at 22:00, 17:30 and 7:00 respectively, while the peak of pattern 5, 6 and 10 are at 18:30-21:00, 21:30-00:00 and 9:00-13:00 respectively. Pattern 3 is bimodal, with two electricity consumption peaks around 10:00 am and 21:30. Pattern 2, 7 and 9 are stationary which has three stages: rising, stable and falling. The stationary period is 7:30-21:30, 9:30-16:30 and 13:00-21:30 respectively. It is worth noting that the stationary period may refer to the absolute stability of the user's load curve, such as pattern 7, or it may refer to the user's total electricity consumption remaining stable over a while, such as pattern 2 and 9, where the former is more regular and stable than the latter.

4. Conclusion

In the face of high-dimensional and massive resident load data, the key to load clustering lies in how to extract features from the original data. To reduce the loss of information and improve the effect of feature extraction, this paper proposes a residential electricity consumption behavior analysis method based on RP-CAE and SC. On the premise of data pre-processing and determining the optimal cluster number, the electricity consumption behavior of Irish residents was divided into 10 patterns. It is verified that RP-CAE has the better ability to extract load curve features than PCA and LSTM-AE by DBI and DI indexes, and can achieve a good clustering effect by combining with SC. The influence of the dimension on clustering effect is also discussed. The results show that the optimal dimension is slightly higher than the number of clusters.
Acknowledgments
This work was supported by the Science and Technology Project of SGCC [grant numbers SGZJHZ00HLJS2000871].

References
[1] Qi Y, Luo B, Wang X, Wu L. (2018) Load pattern recognition method based on fuzzy clustering and decision tree. In: 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2). Beijing. pp. 1–5.
[2] Li K, Ge X, Lu X, Wang F, Mi Z. (2019) Meta-Heuristic Optimization Based Two-stage Residential Load Pattern Clustering Approach Considering Intra-cluster Compactness and Inter-cluster Separation. In: 2019 IEEE Industry Applications Society Annual Meeting. Baltimore. pp. 1–8.
[3] Wang X, Chen L, Tao W, Principle AB. (2018) Research on Load Classification Based on User's Typical Daily Load Curve. In: 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2). Beijing. pp. 1-4.
[4] Flath C, Nicolay D, Conte T, Van Dinther C, Filipova-Neumann L. (2012) Cluster analysis of smart metering data: An implementation in practice. Business & Information Systems Engineering. 4: 31-39.
[5] Wen L, Zhou K, Yang S. (2019) A shape-based clustering method for pattern recognition of residential electricity consumption. Journal of Cleaner Production. 212: 475-488.
[6] Zhao Q, Li H, Wang X, Pu T, Wang J. (2019) Analysis of users’ electricity consumption behavior based on ensemble clustering. Glob Energy Interconnect. 2: 479–88.
[7] Liu LP, Wang Q, Dong MN, Zhang ZY, Li Y, Wang ZZ, Wang SQ. (2020) Application of K-Means+ Algorithm Based on t-SNE Dimension Reduction in Transformer District Clustering. In: 2020 Asia Energy and Electrical Engineering Symposium (AEEES). Chengdu. pp. 74-78.
[8] Jiang Z, Lin R, Yang F, Wu B. (2018) A fused load curve clustering algorithm based on wavelet transform. IEEE Transactions on Industrial Informatics. 14: 1856-1865.
[9] Zhang T, Gu M, Zhang G, Lu J. (2016) A fast load pattern extraction approach based on dimension reduction and sampling. In: 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Vancouver. pp. 1253-1258.
[10] Irish Social Science Data Archive. (2012). Commission for Energy Regulation (CER) Smart Metering Project. http://www.ucd.ie/issda/data/commissionforenergyregulationcer/
[11] Tibshirani R, Walther G, Hastie T. (2001) Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B. 63: 411–423.