A method of mining electricity consumption behaviour based on CC-DWT

Fan Li1*, Yingjie Tian1, Yi Wu1, Yihan Liu1
1State Grid Shanghai Municipal Electric Power Company, Shanghai, 200000, China
*Corresponding author’s e-mail: li.fan@sh.sgcc.com.cn

Abstract. In this paper, a fusion clustering algorithm based on discrete wavelet transform called CC-DDWT is proposed. In this method, the preliminary feature construction of load curve is realized by multi-layer discrete wavelet transform, and the approximate component and detail component are formed, and then the clustering fusion is carried out. The experimental results show that the typical load patterns of a single user extracted by this method are more reliable than the traditional methods.

1. Introduction
Mining the power consumption behaviour characteristics of demand-side users is an important link in the intelligent process of power system[1]. The general mining method is to directly cluster load curves[2]. For example, McLoughlin et al. studied K-means, K-Medoid and SOM, and used Davies-Bouldin Index as an index to divide a single household into various groups according to the daily electricity patterns[3]. Li et al. proposed VDSL[4]. In addition, the load curve can be converted into frequency-domain data by signal processing method, so as to extract frequency-domain features for dimension reduction and clustering. For example, Sheikholeslami et al. proposed Wave Cluster for large spatial database and used the multi-influence characteristics of Wave transformation to identify clusters of arbitrary shapes with different levels of detail[5]. Saikia et al. took wave transform as a detection tool, decomposed the power quality disturbance represented by load curve, and further detected and classified it through fuzzy logic and neural network[6]. The experimental results showed that wave transform played a significant role in improving the accuracy of the model. In the current research, there is a lack of effective clustering methods for high-dimension data. The power load data has the characteristics of large volume, high dimension and time sequence, so it is difficult for traditional methods to solve the problem of high-dimension clustering. Therefore, this paper proposes a curve clustering algorithm based on discrete wavelet transform and evaluates its experimental performance from two perspectives of average performance and specific case analysis.

2. Load Curve Clustering Fusion Algorithm Based on Discrete Wavelet Transform

2.1. Structure of CC-DWT
CC-DWT includes two parts. The daily load curve is reduced and the data after dimension reduction is clustered and fused, so as to excavate the power consumption behaviour pattern. In the process of dimension reduction, multi-layer one-dimension discrete wavelet transform is used to extract the features of n-dimension daily load curve $X_{ld}$, and the approximate component $X_A$ and the detail component $X_D$ with the same lower dimension are obtained. Then, the contour information and the
detail information are processed by Z-score standardization simultaneously to obtain the normalized data $X_A'$ and $X_D'$. In the part of clustering, $X_A'$ and $X_D'$ are processed by DBSCAN clustering algorithm respectively to obtain the respective cluster groups $A$ and $D$. Finally, clustering fusion algorithm was used to optimize cluster groups $A$ and $D$ to get the final load mode cluster group $\text{Center}_{\text{id}}$. The specific CC-DWT algorithm is as follows:

Algorithm 1. CC-DWT.

Input: n-dimension power curve of a user, $X_{\text{id}}$
Output: typical power consumption curve of a user, $\text{Center}_{\text{id}}$

1: $X_A$ and $X_D$ are obtained by using one-dimension discrete wavelet transform to reduce dimension of $\alpha$ layer on $X_{\text{id}}$;
2: $X_A'$ and $X_D'$ are obtained by Z-score normalization of $X_A$ and $X_D$;
3: for $\epsilon = 0.001, \epsilon \leq 1, \epsilon += 0.05$ do
4: $A_{\epsilon, \text{sample}_{\text{min}}} = \text{DBSCAN}(X_A', \epsilon, \text{sample}_{\text{min}})$;
5: $D_{\epsilon, \text{sample}_{\text{min}}} = \text{DBSCAN}(X_D', \epsilon, \text{sample}_{\text{min}})$;
6: $S_{A_{\epsilon, \text{sample}_{\text{min}}}}$ are obtained by calculating $A_{\epsilon, \text{sample}_{\text{min}}}$ and simplifying SSWC;
7: $S_{D_{\epsilon, \text{sample}_{\text{min}}}}$ are obtained by calculating $D_{\epsilon, \text{sample}_{\text{min}}}$ and simplifying SSWC;
8: end for
9: $\epsilon_{A, \text{sample}_{\text{min}}} = \arg \max_{\epsilon_{A, \text{sample}_{\text{min}}}} \left\{ S_{A_{\epsilon, \text{sample}_{\text{min}}}} \right\}$, $A = \text{DBSCAN}(X_A', \epsilon_A, \text{sample}_{\text{min}} A)$;
10: $\epsilon_{D, \text{sample}_{\text{min}}} = \arg \max_{\epsilon_{D, \text{sample}_{\text{min}}}} \left\{ S_{D_{\epsilon, \text{sample}_{\text{min}}}} \right\}$, $D = \text{DBSCAN}(X_D', \epsilon_D, \text{sample}_{\text{min}} D)$;
11: Execute clustering fusion Algorithm 2 for $A$ and $D$, and get $\text{Center}_{\text{id}}$;
12: return $\text{Center}_{\text{id}}$.

2.2. Dimension reduction of load curve

The high-dimension daily load curve is regarded as a discrete signal, and the infinite trigonometric function basis in Fourier transform is changed into a finite attenuated wavelet basis through wavelet transform, so as to achieve data dimension reduction.

$$WT(\alpha, \tau) = \frac{1}{\sqrt{\alpha}} \int^{-\infty}_{\infty} f(t) \ast \psi_{\frac{t-\tau}{\alpha}} dt$$

(1)

where the scale $\alpha$ represents the stretching degree of the wavelet function, and the shift $\tau$ represents shifting degree. 1-D Discrete Wavelet Transform is selected when the Wavelet Transform is selected. In dimension reduction, we choose the multi-level 1-D discrete wavelet transform to process the original data. After multi-layer 1-D DWT dimension reduction, the original data $X_{\text{id}}[n]$ becomes the approximate component $X_{\text{id},\text{AA}}$ and the detail component $X_{\text{id},\text{AD}}$ in the following form.

Then the set of n-dimension daily load curves $X_{\text{id}}$ produces two output sets called $X_{\text{id},\text{AA}} = \{X_{\text{id},1,\text{AA}}, X_{\text{id},2,\text{AA}}, \ldots, X_{\text{id},N,\text{AA}}\}$ and $X_{\text{id},\text{AD}} = \{X_{\text{id},1,\text{AD}}, X_{\text{id},2,\text{AD}}, \ldots, X_{\text{id},N,\text{AD}}\}$. In order to ignore the distance difference between curves, the approximate component set $X_{\text{id},\text{AA}}$ and the detail component set $X_{\text{id},\text{AD}}$ are normalized by Z-score.

2.3. Structure of CC-DWT

We need to carry out density clustering for the approximate component and the detail component obtained by dimension reduction and standardization simultaneously to reduce the information loss caused by feature extraction. DBSCAN is a typical density clustering algorithm. The algorithm defines the cluster as the maximum set of densely-connected samples and can extract cluster of any shape from noisy data. Since DBSCAN algorithm needs to input the neighbourhood parameter $\epsilon$ and $\text{sample}_{\text{min}}$, and the optimal parameter combination is unknown. Therefore, based on the idea of
network search, we performed DBSCAN clustering on the normalized approximate component and the detailed component within the value range of \( \varepsilon \in (0.001, 0.051, 0.101, \ldots, 1) \) and \( \text{sample}_{\text{min}} \in (2, 3, \ldots, 10) \). Then the optimal parameters of each cluster were determined by the simplified contour coefficient (SSWC). We selected the optimal parameter combination when SSWC is at its maximum. And after clustering of the standardized approximate component and detail component, we got \( A = \{A_1, A_2, \ldots, A_p\} \) and \( D = \{D_1, D_2, \ldots, D_q\} \).

2.4. Clustering Fusion

In Algorithm 2, the intersection is mainly used to determine the state of cluster groups in \( A \) and \( D \). If \( A = D \), then the final clustering effect is \( \text{Center} = A = D \); If \( A \neq D \), and \( A \cap D \neq \emptyset \), the same cluster groups in \( A \) and \( D \) are retained first, and then a new clustering is performed on the different cluster groups retained after the set intersection. However, since DBSCAN clustering will produce outliers, all possible cases where \( A \neq D \) and \( A \cap D = \emptyset \) cannot be reclustered for the remaining cluster groups after intersecting. In this case, our fusion clustering is meaningless. Thus, for the cluster groups in \( A \) and \( D \) are not exactly the same, but there is still an intersection, i.e. \( A_i \neq D_j \) and \( A_i \cap D_j = \emptyset \). In this case, we make the intersection \( A_i \cap D_j = C_m \). Then we arrange \( C_m \) from the largest to the smallest and determine whether it is counted as one of the final cluster groups according to the size ratio of \( C_m \). Finally, we determine whether the identified cluster group category is large enough to decide whether to recluster.

**Algorithm 2. Cluster fusion algorithm.**

Input: approximate component cluster group \( A \) and detail component cluster group \( D \)

Output: cluster group \( \text{Center}_{\text{id}} \) after fusion

1. \( p = \text{len}(C_A), q = \text{len}(C_D), C_{\text{sure}} = \emptyset, C_{\text{unsure}} = \emptyset; \)
2. \( \text{for} \ A_i \ in \ A (1 \leq i \leq p); \)
3. \( \text{for} \ D_j \ in \ D (1 \leq j \leq q); \)
4. \( \text{if} \ A_i = D_j \text{ then } C_{\text{sure}} = C_{\text{sure}} \cup D_j; \)
5. \( \text{if} \ A_i \neq D_j \ and \ A_i \cap D_j = \emptyset \ (i.e \ C_{ijm} = \emptyset) \)
6. \( \text{then } C_{\text{unsure}} = C_{\text{unsure}} \cup C_{ijm}; \)
7. \( \text{end for} \)
8. \( k = \text{max}(p, q) \)
9. \( \text{if } \text{len}(C_{\text{sure}}) = k \text{ then break } ; \)
10. \( \text{if } \text{len}(C_{\text{sure}}) < k ; \)
11. \( \text{sort } C_{\text{unsure}} \text{ by } \text{len}(C_m); \)
12. \( \text{for } C_m \ in \ C_{\text{unsure}} (1 \leq m \leq k – \text{len}(C_{\text{sure}})); \)
13. \( \frac{\text{len}(C_m)}{\text{sum}(\text{len}(C_n), nzm + 1)} > \frac{1}{k} \text{ then } C_{\text{sure}} = C_{\text{sure}} \cup C_m; \)
14. \( \text{end for} \)
15. \( \text{if } \text{len}(C_{\text{sure}}) < k \text{ then } \text{KMeans}(C_{\text{unsure}}, k – \text{len}(C_{\text{sure}})); \)
16. \( \text{Calculate the center of } C_{\text{sure}}. \)

3. Experiment

We used the 15-minute electricity load data of 536 users for 730 days from April 1, 2016 to March 31, 2018. Considering the diversity and reliability of evaluation indexes, we added four clustering performance evaluation indexes during the experiment: SWC、DB、Dunn and CH.
3.1. Performance comparison

The CC-DWT algorithm proposed in this paper will be compared with four control clustering methods. Table 1 summarizes all the comparison methods. The table shows the mean of the five performance metrics for 536 users and the percentage of performance improvement of the CC-DWT algorithm over the other optimal methods. It can be seen from Table 1 that the clustering performance of CC-DWT algorithm on data sets is superior to other comparison algorithms. Compared with the optimal results of other methods, the CC-DWT algorithm improves the clustering performance by 15.4%~199.4%. Among them, CC-DWT clustering algorithm actually fuses clustering on the basis of approximate classification and detailed classification clustering, and the performance of clustering results is improved compared with appro-signal and detail-signal, especially compared with appro&detail-signal, the same clustering effect is better. To some extent, it shows that the fusion clustering algorithm is very useful for the clustering results of approximate classification and detailed classification.

| Table 1. Comparison of user typical load characteristic methods. |
|---------------------------------------------------------------|
| description | dimension | data type | SWWC | SWC | DB | CH | Dunn |
| CC-DWT | Experimentai method | 2 | Frequency domain | 0.54 | 0.47 | 1.17 | 36.88 | 0.33 |
| original | Standardize the original daily load curve | 96 | Time domain | 0.18 | 0.09 | 1.71 | 31.96 | 0.17 |
| appro-signal | Multi-level 1-D DWT approximation component standardization | 2 | Frequency domain | 0.11 | -0.12 | 2.42 | 21.69 | 0.15 |
| detail-signal | Multi-level 1-D DWT detail component standardization | 2 | Frequency domain | 0.02 | 0.19 | 4.08 | 7.18 | 0.24 |
| appro&detail-signal | Multi-level 1-D DWT approximation and detail component | 4 | Frequency domain | 0.13 | 0.07 | 2.25 | 3.18 | 0.17 |

3.2. Case analysis

Figure 1(a) shows 730 original daily load curves of User A in two years. Figure 1(b-f) presents the clustering results of these daily load curves by five methods. We can see that the user has two main power consumption modes. We guess that the user should have a normal job and may often leave no one living in the home, and the electricity consumption is stable and less. In addition, Figure 1(b, e) further extracts the power consumption characteristics of the user. For example, the user also uses a lot of electricity in the daytime. We can speculate that the user does not need to go to work because of holidays and other reasons. In Fig. 1(b), there are two curves, one showing an obvious concave shape, and the other showing a steady rise. However, it can be clearly seen that the electricity consumption increases greatly after 8 o’clock at night. It can be found that these two power consumption modes should be the power consumption modes when users are at work. Moreover, all power consumption fluctuations are more significant in summer and winter.
4. Conclusion
In this paper, CC-DWT based on DWT is proposed. Based on the historical load curve of a single user, the algorithm decomposes its dimensionality reduction into approximate component and detail component by multi-layer discrete wavelet transform, and performs DBSCAN clustering for the two components respectively. Then, the clustering results are fused to obtain the final typical power consumption characteristics of a single user. This method not only reduces the information loss in data dimensionality reduction, but also solves the problem that clustering clusters have no intersection in the fusion process due to DBSCAN clustering outliers. CC-DWT was tested on real data sets of power users. Experimental results show that the proposed method has the best clustering performance compared with the other four methods. The clustering effect of this method and the practical significance of the typical electricity consumption characteristics of the user are illustrated by the case analysis of random users.

Acknowledgments
This paper is supported by Construction of power big data information platform for supply and demand side intelligent management (No. 2019-GYHLW-01031) and Research on electricity data value mining and business model exploration demonstration (No. 52094019007B).

References
[1] Tu, C., He, X., Shuai, Z., et al. (2017) Big data issues in smart grid–A review. Renewable and Sustainable Energy Reviews, 79: 1099-1107.
[2] Chen, T., Qian, K., Mutanen, A., et al. (2017) Classification of electricity customer groups towards individualized price scheme design. In: 2017 North American Power Symposium (NAPS), pp.1-4.
[3] Mcloughlin, F., Duffy, A., Conlon M.A. (2015) Clustering approach to domestic electricity load profile characterisation using smart metering data. Applied Energy, 141: 190-199.
[4] Li, H.B., Zhang, Z., Wang, X.Y., et al. (2019) Electricity consumption behaviour analysis based on time sequence clustering. IOP Publishing, 1168(3): 032011.
[5] Sheikholeslami, G., Chatterjee, S., Zhang, A.D. (2000) A wavelet-based clustering approach for spatial data in very large databases. The VLDB Journal, 8(3-4): 289-304.
[6] Saikia, L.C., Borah, S.M., Pait, S. (2010) Detection and classification of power quality disturbances using wavelet transform, fuzzy logic and neural network. In: Annual IEEE India Conference (INDICON), pp.1-5.