Typical Power Load Profiles Shape Clustering Analysis Based on Adaptive Piecewise Aggregate Approximation

Quan TANG¹, Quanming ZHANG¹, Yingying DENG¹, Mi ZHU¹, Yunche SU¹, Junyong LIU² and Xiaodi WANG²

¹Sichuan Electric Power Corporation Power Economic Research Institute, Chengdu 610000, China
²College of Electrical Engineering and Information Technology of SCU, Chengdu 610065, China

Corresponding author: 414363359@qq.com

Abstract. With the trend of increasing dimensions of collected power load data, the data dimension reduction and classification become essential pre-processing steps for data mining and further application. On the basis of adaptive piecewise aggregate approximation(APAA) and k-Shape algorithm, a novel method is proposed. In light of the fluctuation degree and shape characteristics of the original load profiles, the new load dataset with variable temple resolution replace the old one by APAA, and further k-Shape algorithm is adopted for lower dimension load profiles clustering. K-Shape algorithm cluster curves with distance SBD as similarity measurement and also a novel method to extract the representative centroids is mentioned. The experiment testifies that APAA-kShape algorithm has shorter calculating time, higher accuracy and represent better load patterns than other cluster algorithms.

1. Introduction

With the popularity of smart meters, huge amount of detailed end-user energy consumption data at different granularity are generated. Clustering as a useful tool to simplify the data, groups similar data sets into aggregated clusters by weighing the the similarity of load profiles[1]. Valuable information can be extracted from databases and pattern discovery for further application can be developed by clustering[2].

The shape of a curve is of great important, the most common clustering methodologies such as k-means, fuzzy C-means, self-organizing map, etc and their combination algorithm[3,4], most of which measure the similarity of profiles by traditional euclidean distance. [5] points out that all above methods have difficulty in distinguishing curves with close distance but huge shape discrepancy.

High temporal resolution is required to elaborately analyze daily load data, which is also accompanied by time-consuming computation. Dimension reduction methodologies are applied such as symbolic aggregate approximation, piecewise aggregate approximation, principal component analysis, sammon map and etc[6,7]. Most of them re-express the original data with the constant temporal resolution, which is unreasonable because constant temporal resolution ignores the fluctuation degree and will lead to loss of useful information.

In this paper, we propose a novel algorithm APAA-kShape. First, adaptive piecewise aggregate approximation(APAA) is applied to obtain the re-expressed data sets with variable temporal resolution, which retain the fluctuation features and the morphological characters of load profiles. Then, k-Shape
algorithm is used to cluster the low-dimension data sets. k-Shape algorithm belongs to the partitional clustering algorithm but has the capability of keeping the shape feature of time series data set. It puts forward a novel distance measurement SBD and a way to extract centroids on the basis of Steiner’s sequence. Further more, the experiment shows that our APAA-kShape algorithm can reconstruct original data effectively and our distance measurement is scale-, shift- and noise invariant. APAA-kShape algorithm can cluster load profiles with less computation time, higher accuracy and shows an outperforming than traditional time-series clustering method. The study proves the algorithm practical for high dimension power load profiles.

2. Adaptive Piecewise Aggregate Approximation

2.1. load profiles characteristics

2.1.1 Climbing accident. Within the certain period of time, a sharp change of a curve is called climbing accident, which can measure the fluctuation characteristic of that curve[8]. At the time window T, if the difference between the maximum and minimum value of a curve \( x_i \) exceeds a certain threshold, we define the variable \( \phi_i \):

\[
\phi_i = \begin{cases} 
1, & x_i \in \text{Climbing accident} \\
0, & x_i \notin \text{Climbing accident} 
\end{cases}
\]

For \( N \) load profiles, at the time window \( T \), we define the total number of climbing accidents as \( N_{\text{climb}} \).

2.1.2 Edge point extraction based on slope. Piecewise aggregation approximation(PAA) based on the change rate of slope is put forwards to detect the edge points of time series, which can retain the main shape characteristic of the curves and remove detail interference.

The slope rate \( i g 1, i g 2 \) of one point \( x_i \) of the time series, can be calculated by comparing the point \( x_i \) with its adjacent points \( x_{i-1}, x_{i+1} \). If the change rate of slope exceeds the threshold \( d \), we define the point \( x_i \) as the edge point.

\[
\mu_i = \begin{cases} 
\|i g 1 - i g 2\| > d = 1 (x_{i-1}, x_{i+1} \text{ on the same side of } x_i) \\
\|i g 1 + i g 2\| > d = 1 (x_{i-1}, x_{i+1} \text{ on the different sides of } x_i) \\
0 
\end{cases}
\]

For \( N \) load curves, at the time window \( T \), we define the total number of the edge points as \( \epsilon \).

2.2. APAA detailed steps

Adaptive piecewise aggregate approximation(APAA) divides the time series into several segments of different length based on the curves’ characteristics without supervision. Take the normalized data set \( X = (x_1, \ldots, x_n) \) of length \( n \) and the threshold \( r \) and \( d \) as the input, here’s the detailed steps:
1) Divide \( n \) dimensions of data set \( X \) into \( n/4 \) segments with four data in each segment, calculating the number of \( \psi \) and \( \epsilon \) in each.
2) For all the segments, if \( \psi < \sigma_1 \) and \( \epsilon < \sigma_2 \), replace the four data in the segment with their average data; if \( \psi < \sigma_1 \) and \( \epsilon > \sigma_2 \), replace the four data with their maximum and minimum value.
3) Calculate the data dimensions
4) If not satisfied with the dimensions, enlarge the threshold and repeat step 2.
5) if satisfied with the dimensions, return data set \( \mathcal{X} \) with \( m \) dimensions and output the results.

The value of \( r, d, \sigma_1, \sigma_2 \) are chosen according to the practical situation.

3. k-shape algorithm
3.1. Time-series theoretical background

Time series data are sometimes distorted. Hence, a number of invariances (scaling invariance, shift invariance, noise invariance) have to be taken into account when it comes to compare time series data sets. An ideal distance measurement can achieve scaling-, shift- and noise invariance.

The two state-of-the-art distance measures for time-series comparison to determine their similarity and possibly capture more invariances are ED and DTW. ED is most widely used one, but is sensitive to the noise, scaling and shift distortions.

DTW offers a local optimal alignment by constructing an \( n \)-by-\( m \) matrix \( M \) to achieve distance measuring. A warping path is a set of matrix elements that defines a mapping between \( X = (x_1, \ldots, x_n) \) of length \( n \) and \( Y = (y_1, \ldots, y_m) \) of length of \( m \). Dynamic programming detect the best path to obtain the minimum distance DTW, which realizes the similarity comparison of two time series of different length.

\[
DTW(x, y) = \tau(m, n) \quad (3)
\]

\[
\tau(i, j) = ED(i, j) + \min \{\tau(i-1, j-1), \tau(i-1, j), \tau(i, j-1)\} \quad (4)
\]

However, the same weight is vested to the observed values, which may lead to misclassification. Also DTW is too time-consuming to cluster power load profiles with high dimensions.

3.2. Time series similarity

Each time series must be pre-processed before comparing the similarity.

Cross-correlation measure: Cross-correlation is a statistical measurement with which we keep \( \tilde{y} = (y_1, \ldots, y_n) \) static and slide \( \tilde{x} = (x_1, \ldots, x_n) \) to align the two sequences to achieve shift-invariance and calculate the similarity. Considering all possible shift \( s \), the cross-correlation sequence \( CC_w(\tilde{x}, \tilde{y}) \) with length \( 2m-1 \) is produced.

\[
CC_w(\tilde{x}, \tilde{y}) = R_{w-m}(\tilde{x}, \tilde{y}) = \begin{cases} 
\sum_{i=1}^{m} x_i + k \cdot y_i, & k \geq 0 \\
\sum_{i=1}^{m} x_i, & k < 0 
\end{cases} 
\approx \frac{1}{\sqrt{R_{0,w}(\tilde{x}, \tilde{x}) \cdot R_{0,w}(\tilde{y}, \tilde{y})}} \quad \text{(5)}
\]

Where \( CC_w(\tilde{x}, \tilde{y}) \) is maximized is the place \( w \) that the optimal shift \( s \) can be calculated: \( s = w - m \). The cross-correlation coefficient is normalized to limit its range from -1 to 1 and 1 means highly correlated. And we derive the following distance measure SBD:

\[
SBD(\tilde{x}, \tilde{y}) = 1 - \max_w \frac{CC_w(\tilde{x}, \tilde{y})}{\sqrt{R_{0,w}(\tilde{x}, \tilde{x}) \cdot R_{0,w}(\tilde{y}, \tilde{y})}} \quad \text{(6)}
\]

Which takes value between 0 to 2, and 0 indicates perfect similarity for time-series sequences.

3.3. Time-Series Shape Extraction

Extracting meaningful centroids is a challenging task which critically depends on the choice of distance measurement and each cluster centroid represents the typical shape characteristics. Traditional k-means centroid is to compute the arithmetic mean of the corresponding coordinates of the sequences, but these centroids do not effectively represent the cluster shape characteristics.

The centroid computation can be an optimization problem where the objective is to find the minimum value of the sum of squared distances to all other sequences, which is Steiner’s sequence:

\[
c_j = \arg\min_{\mu_j} \sum_{x_i \in p_j} \text{dist}(\mu, x_i)^2, \quad \bar{w} \in R^m \quad \text{(7)}
\]

As cross-correlation captures the similarity, we express the computed sequence as the maximum value and use the previously computed centroid as reference to align all sequences.

\[
\hat{c}_k = \arg\max_{\tilde{c}_k} \sum_{x_i \in p_k} NCC_C(\tilde{x}, \tilde{c}_k)^2
\]

\[
= \arg\max_{\tilde{c}_k} \sum_{x_i \in p_k} (\tilde{x} \cdot \tilde{c}_k)^2 \hat{c}_k \quad \text{(8)}
\]
\[ \bar{c}_k^* = \underset{\bar{c}_k}{\operatorname{arg\, max}} \frac{\bar{c}_k^T \cdot Q^T \cdot S \cdot Q \cdot \bar{c}_k}{\bar{c}_k^T \cdot \bar{c}_k} \]  \tag{9} 

Where \( \bar{c}_k = \bar{c}_k \cdot Q \cdot Q = I - O / m \), \( I \) is the identity matrix and \( O \) is a matrix with all ones. The maximizer \( \bar{c}_k^* \), as the eigenvector that corresponds to the largest eigenvalue of the real symmetric matrix \( M \), is also our extracted centroids.

3.4. **k-Shape algorithm detailed steps**

Taking the load curve matrix \( X \) as input, the centroid of each cluster is extracted by cross-correlation measure and the data of each cluster is updated in each iteration.

1) Initial each cluster centroid \( \bar{c}_k \) and \( \bar{c}_k \) is 0 vector.

2) The SBD distance from each sample \( \bar{x}_i \) to all the centroids \( \bar{c}_k \) is calculated by equation 8. Assign \( \bar{x}_i \) into the cluster \( \bar{c}_k \) with smallest distance.

3) Calculate the centroid of each cluster and extract the shape characteristics.

4) Repeat the step 2 and the step 3. Stop iterating when the maximum number of iterations \( t \) is reached or the samples of each cluster do not change anymore;

5) Output cluster results.

4. **Experience and analysis**

4.1. **Simulation four types of typical daily load profiles**

Four types of typical daily load curves are simulated, a certain ratio of random noise \( \alpha \) is added and the curves are partially shifted. We simulate 200 load curves which are grouped into four typical clusters with 50 curves in each cluster.

200 simulating load profiles with 10\% noise in four clusters are shown in figure 1 and re-expressed low-dimension load curves are shown in figure 2, where the time window \( T \) is 1h, \( \tau \) is 5 and \( d \) is 2.

The original 96-dimension data is reduced to 63-dimension by APAA, which can clearly represent the morphological characteristics of each cluster(bi-peak, tri-peak, peak-period, peak avoidance). We compare APAA with the constant temporal resolution PAA and measure the cluster results by CP value. The average CP value of each re-expressed cluster by our method is 1.891 which is significantly smaller than the CP value 2.548 of PAA, slightly larger than the CP value 1.797 of the original data, which proves the effectiveness of our method, which validated our method.

![Figure 1](image1.png)  
**Figure 1.** Simulation of 200 typical load curves with 10\% noise

![Figure 2](image2.png)  
**Figure 2.** Low-dimension simulating load curves with APAA method (\( \alpha=10\% \))
The APAA-kShape detailed steps are shown in figure 5, and the clustering results of the k-means and the APAA-kShape algorithms are shown in figure 3 and figure 4 respectively. k-means fails to distinguish the curves in the cluster 3 and cluster 4 and shows some deviation with its accuracy 86%, while the cluster accuracy rate of APAA-kShape reaches 100%.

With the certain noise ratio, each algorithm has been run 10 times, and the accuracy results are shown in figure 6. k-means clustering accuracy rate is inconsistent in 10 times clustering, which is noise sensitive and poorly stable., while APAA-kshape average cluster accuracy rate is much higher. The experiment confirms APAA-kshape with noise invariance and good robustness.

![Figure 3. Clustering results of k-means method](image)

![Figure 4. Cluster results of APAA-kShape method](image)

![Figure 5. APAA-kshape processing](image)

![Figure 6. Comparison of cluster accuracy with different noise ratio](image)
4.2. Measured data
Taking the daily load curves of 1600 users on a working day in April, 2015 in a Chinese city as the research object, the energy consumption data is collected at every 15 min interval so that 96 data points are obtained from one user in that day. All the data has gone through data cleaning process to identify the outlines and impute the missing data. After data cleaning, 1589 practical daily load curves are collected. At the stage of APAA, we set time window $T$ as 1h, $r$ as 20, $d$ as 5.

Ding, Wang[9] have made a uniform comparison experiment on the existing methods for measuring the similarity of time series, and pointed out that the DTW distance measurement may be the best method at present. The APAA-kShape and the DTW are adopted to cluster 1589 daily load curves respectively, the optimal cluster number of APAA-kShape is 8 based on elbow method[10], as well as DTW. The clustering results are shown in figure 7 and figure 8.

These 8 clusters of APAA-kShape can be classified as the following 4 type: single-peak, double-peak, multi-peak and avoidance peak. The typical daily load curves extracted by our algorithm is reasonable and each one has obvious shape feature. Each cluster are different in peak value, peak valley time and basic load of all the clusters. The extracted typical load curves can represent the typical user behavior patterns and load characteristics. The optimal cluster number of DTW is also 8, but the shape characteristic of the curves in each cluster is vague. Though the avoidance-peak type users(the cluster 2 and the cluster 3) in figure 8 can be distinguished accurately, for the single-peak, double-peak, multi-peak type users, DTW performs worse than APAA-kShape. Choosing 100 curves from 1589 daily load curves, the performance of DTW, k-Shape, APAA-kShape are comprehensive compared by cluster validity index and running time, the results are shown in the Table 1. Among these three algorithms, the APAA-kShape shows great advantage.

![Figure 7. Load curves clustering results of APAA-kShape method](image1)

![Figure 8. Load curves clustering results of DTW method](image2)

| Cluster number | APAA-kShape | k-Shape | DTW |
|---------------|-------------|---------|-----|
| Time          | SIL         | DBI     | CH  |
| 3             | 18.41       | 0.312   | 2.050 | 154.2 |
| 4             | 15.35       | 0.341   | 1.318 | 143.0 |
| Time          | SIL         | DBI     | CH  |
| 3             | 24.47       | 0.455   | 1.501 | 116.1 |
| 4             | 29.89       | 0.399   | 1.784 | 144.6 |
| Time          | SIL         | DBI     | CH  |
| 3             | 36.70       | 0.194   | 2.518 | 170.8 |
| 4             | 59.39       | 0.073   | 2.186 | 116.1 |

Table 1. Clustering property comparison among APAA-kShape, k-Shape and DTW method
5. Conclusion
With the challenging of increasing dimensions of power load data and the time-series distortions, in this study, first, APAA with variable temporal resolution is adopted based on the curve fluctuation degree and curve shape characteristics. Then, the re-expressed low-dimensional data will be clustered by k-Shape. Our method propose a distance measurement to weigh the similarity of the curves, which is scaling-, shift- and noise invariance. And also a way to extract meaningful centroids is proposed. The experiment testify that:
1) APAA with variable temporal resolution can reduce the data dimensions effectively without affecting the cluster results. It is not only able to improve the calculation speed of the algorithm, but also retain the main shape characteristic.
2) The shape extraction method mentioned in this study can better reflect the tendency of sharp increase or drop of the aggregated curves than the traditional arithmetic average method.
3) The algorithm performs much better than DTW on the calculation time and the cluster validity index.

Acknowledgments
The authors would like to appreciate the support by State Grid Corporation Scientific Research Project(SGSCJY00JHJS201700009) and National Natural Science Foundation of China(51437003).

References
[1] Zhao Wenqing, Gong Yaqiang.Load curve clustering based on Kernel K-means. Electric Power Automation Equipment, 2016, 36(6) p203.
[2] Yi Wang, Qixin Chen, Chongqing Kang, et al. Load profiling and its application to demand response: a review. Tsinghua Science and Technology, 2015, 20(2): 117-129.
[3] Wang Chun, Gao Yuanhui. Dynamic Reconfiguration of Distribution Network Based on Optimal Fuzzy C-means Clustering and Improved Chemical Reaction Optimization. Proceedings of the CSEE, 2014, 34(10): 1682-1691.
[4] Chen Bocheng, Liang Bing, Zhou Yuebo, et al. An application of SOM neural network in customer classification. Systems Engineering. Theory & Practice, 2004, 3(8): 8-14(in Chinese).
[5] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: Experimental comparison of representations and distance measures. PVLDB, 2008, 1(2):1542–1552.
[6] Keogh E, Chu S, Hart D, et al. An Online Algorithm for Segmenting Time Series. Proc of the IEEE International Conference on Data Mining, San Jose,USA,2001: 289—296
[7] Keogh E, Chakrabarti K, Pazzani M, et al. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Journal of Knowledge and Information Systems, 2001, 3(3): 263—286
[8] Ouyang Tinghui, Zha Xiaoming, Qin Liang, et al. Modeling on selection of the time window for ramp events Prediction. Proceedings of the CSEE, 2015, 35(13): 3204-3210.
[9] Wang X, Mueen A, Ding H, et al. Experimental comparison of representation methods and distance measures for time series data. Journal of Data Mining and Knowledge Discovery, 2013, 26: 275-309.
[10] R.L.Thorndike, Who belongs in the family? Psychometrika, 1953, 18(4) 267–276.

|  | 10.74 | 0.347 | 1.669 | 134.2 | 31.21 | 0.375 | 1.927 | 110.2 | 49.21 | 0.01 | 2.873 | 69.68 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 24.48 | 0.531 | 1.267 | 165 | 40.30 | 0.563 | 1.051 | 199.4 | 76.31 | 0.128 | 3.276 | 81.62 |
| 3 | 46.16 | 0.215 | 1.647 | 125.3 | 53.12 | 0.453 | 1.547 | 90.19 | 91.80 | 0.198 | 2.028 | 87.12 |
| 4 | 29.01 | 0.16 | 2.245 | 19.97 | 60.26 | 0.405 | 1.203 | 82.28 | 103.8 | 0.041 | 2.461 | 63.82 |