Cascaded Clustering Analysis of Electricity Load Profile Based on Smart Metering Data

Mustafa Şen Yıldız[1], Kadir Doğanşahin[2], and Bedri Kekezoğlu[3]

[1]Kırklareli University, Kırklareli, Turkey
mustafasenyildiz@klu.edu.tr

[2]Artvin Çoruh University, Artvin, Turkey
dogansahin@artvin.edu.tr

[3]Yıldız Technical University, İstanbul, Turkey
bkekez@yildiz.edu.tr

Abstract

In the operation of deregulated power systems, consumption data is used effectively by system operators. Thanks to the developing measurement and communication technologies, measurement data with high temporal resolution can be obtained from many points within the power systems. Considering the number of consumers connected to power systems, the data in question is a big data. To deal with such a large amount data clustering analyzes are effectively used to identify consumers with similar behaviors in consumption data and to represent consumers with similar behaviors with a single load profile. Success of the clustering studies is related with the compatibility of the data to the selected algorithm and the appropriateness of the adopted approaches to the application of the algorithm to the data. In this study, a cascade clustering algorithm created with the k-medoids algorithm is proposed.

1. Introduction

In the transition from traditional to modern power systems, the most intense evolutions have been experienced at the distribution level. Along with the developing measurement and communication technologies, distribution networks have advanced by having higher observability and infrastructure suitable for the bidirectional data stream, which enables interaction with the consumer. Thus, passive distribution networks that only allow consumption activities in traditional power systems, have become an active system where users can participate in the operation.

Consumption behavior data is very important information for system operators in deregulated distribution systems that allow users connected to the system to participate in the operation of the system. This data is used in studies such as system planning, demand forecasting, demand management, and dynamic pricing. Considering the number of the customer and the diversity on customers' consumption behaviors, it is obvious that the data is huge in size and the content is not so uniform as it can be represented by a typical consumption profile. Rather than working with large volumes of data, it is more practical to characterize consumption behaviors and classify consumers according to their consumption characteristics. Computational efforts and time are saved in the studies realized by using representative load profiles obtained through these processes, which are performed effectively with the help of clustering algorithms.

There are many clustering algorithms in the literature [1]. These algorithms differ from each other in terms of data representation, similarity measurement, and clustering criterion. It should be noted that no one of these algorithms is always the best for clustering. The quality of the clustering analysis depends on achieving the minimum similarity between clusters and maximum similarity within clusters. It is highly related with the compatibility of the data to the algorithm and the appropriateness of the adopted approaches to the application of the algorithm to the data [2].

Commonly preferred clustering algorithms for load profile clustering studies can be given as K-means, Fuzzy c-means (FCM), Hierarchical clustering algorithms, and Self-Organization Mapping (SOM) [3]. Success of different clustering algorithms in clustering electricity consumption data have been surveyed by several studies, in a comparative context [4]–[6]. Considering criteria such as seasonality and consumer types, various approaches can be adopted, such as decomposing the data set before the clustering analysis and sending the obtained decomposed data sets separately to the clustering algorithm. A comprehensive review for the approaches adopted in load clustering studies is provided in [7].

In the electricity consumption data, it is possible to exist the consumers together with quite different consumption behaviors, especially in the structure of the deregulated power system. As a partitioning clustering methodology, k-medoids effectively dealt with the noise and outliers present in data; because it uses medoid for the partitioning of objects into clusters rather than centroid as in k-means [8]. Therefore, it is obvious that the k-medoids algorithm can be used effectively in load profile clustering.

In this study, a cascaded clustering method formed with the k-medoids algorithm has been proposed. In the first level of the proposed cascaded algorithm, the data of each consumer in the data set is clustered. Thus, instead of separating the data with criteria that cannot be defined precisely such as seasonality and consumer type, the decomposition is provided by considering the values in the data set directly. At the second level, the clusters obtained in the first level are assigned to common clusters with similar consumption behaviors of different consumers by re-clustering.

This paper is organized as follows; in the second section k-medoids algorithm is summarized and proposed cascaded algorithm is explained. In the third section, cases studies are presented and the results and findings are discussed. The last section is the conclusion remarks.

2. Materials and Methodology

The characteristics of the data sets and the variance level of the consumption profiles in the data are effective in the algorithm preference. Besides, the ease of application of the method to the data set is also important. In this study, the k-medoids clustering algorithm, which is a Partitional algorithm, has been preferred.

As in all partitioning clustering algorithms, the (k) value, which expresses the optimum number of clusters, must be determined in the k-medoids algorithm. Working with a high number of clusters increases the number of parameters and the computational effort, while working with a low number of clusters reduces the consistency [9]. Therefore, it is important to determine the optimum number of clusters for the success of clustering analysis. In this study, the elbow method, which is frequently preferred in cluster analysis, has been used. However, an analytical approach has been adopted rather than the graphical approach, which is widely used in the literature. The flow of the adopted elbow method approach has been given in Fig. 1.

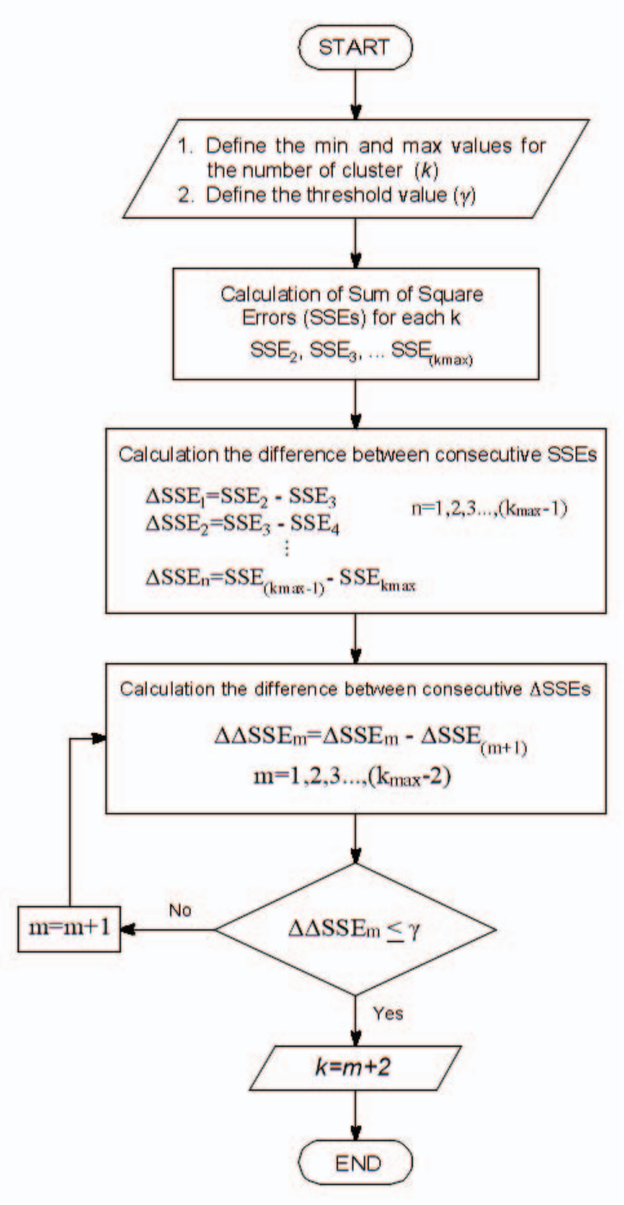
START
1. Define the min and max values for the number of cluster (k)
2. Define the threshold value (γ)
Calculation of Sum of Square Errors (SSEs) for each k
SSE_2, SSE_3, ... $SSE_{(kmax)}$
Calculation the difference between consecutive SSEs
$\Delta SSE_1 = SSE_2 - SSE_3$
$\Delta SSE_2 = SSE_3 - SSE_4$
⋮
$\Delta SSE_n = SSE_{(kmax-1)} - SSE_{kmax}$
$n=1,2,3...,(k_{max}-1)$
Calculation the difference between consecutive ΔSSEs
$\Delta\Delta SSE_m = \Delta SSE_m - \Delta SSE_{(m+1)}$
$m=1,2,3...,(k_{max}-2)$
$\Delta\Delta SSE_m \leq \gamma$
No
$m=m+1$
Yes
$k=m+2$
END

Fig. 1 Flow diagram of adopted elbow method approach

In the first step of applying the elbow method, the minimum and maximum values for the number of clusters are decided. In this study, these values are chosen as 2 and 40, respectively. For each k value, clustering is performed and the Sum of Squared Errors (SSE) is calculated. By calculating the difference between the SSEs for consecutive k values, the rate of SSEs change (ΔSSE) versus the increase on k values is obtained. The difference between successive DSSEs (ΔΔSSE) gives information about the linearity of this change. A threshold value close to zero is determined for DDSSE values, the point where linearization starts is determined and the number of clusters to be used is decided. In this study, the threshold value was determined as 0.3.

2.1. k-medoids Clustering Analysis

K-Means is a basic but highly functional clustering algorithm. K-Medoids or Partitioning Around the Medoids (PAM) algorithm is an unsupervised partitional machine learning algorithm like K-means. The k-medoids algorithm is based on the medoids using the median data point, unlike the K-Means algorithm, which is based on the centroids taking the mean value in the selection of the cluster center. Medoid is the object in the cluster that has the minimum average distance from all other objects in the cluster. Thanks to this feature, it produces more accurate results against the data with high variance. It also provides a representation of clusters with real data [10].

$$J = \sum_{j=1}^{k} \sum_{i=1}^{N} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (1)$$

where, J is main function, $j = \{1,2, \dots, k\}$ is number of clusters, $i = \{1,2, \dots, n\}$ is number of data, c_j is jth medoid and x_i is ith data point [8].

2.2. Cascaded k-medoids Algorithm

In the first step of the algorithm, k medoids are selected among the objects. The distance of each object to the medoids is calculated and assigned to the cluster of the nearest medoid. In the second step, new medoids of the clusters are determined with the assigned objects. The distance of each object to the new medoids is calculated. Objects whose set needs to be changed according to the distance results are assigned to their new sets. The second step is iterated until no objects have changed sets. The flow of the k-medoids algorithm is given below.

- **Requires:** hourly load data for each user, k.
- **Choises:** Distance function, Replicates, Initial medoids selection method.
- **Steps**
  1. Choose random medoids as many as k,
  2. Associate all users to the closest medoids,
- **Iterations**
  3. Calculate new medoids,
  4. If it needs, swap user's cluster to closest medoids.

3. Case Studies

Load profile clustering is effectively used in studies such as system planning, demand forecasting, demand management and dynamic pricing. The success of these studies is directly related to the consistency of the representative load profiles (RLP) obtained by clustering algorithms. In addition to consistency, the number of clusters to be obtained after cluster analysis is also important. Having a large number of clusters may reduce the functionality of clustering. Achieving a higher representation with a smaller number of clusters indicates the effectiveness of clustering studies.

Table 1. Definition of the cases

Cases		**Adopted approach**	**Data size in clustering**
Case 1		A typical load profile has been created for each consumer by averaging the daily load profiles of each consumer throughout the year (*C1*)	342x24
Case 2		A daily average profile is created for each household by taking the average of the daily profiles of each household on weekdays throughout the year. ($C2_{Week}$)	342x24
		A daily average profile is created for each household by taking the average of the daily profiles of each household on weekends throughout the year. ($C2_{Weekend}$)	342x24
Case 3	Winter	for weekdays (as case 2) ($C3_{W.Week}$)	342x24
		for weekends (as case 2) ($C3_{W.Weekend}$)	342x24
	Spring	for weekdays (as case 2) ($C3_{S.Week}$)	342x24
		for weekends (as case 2) ($C3_{Spring.Weekend}$)	342x24
	Summer	for weekdays (as case 2) ($C3_{Summer.Week}$)	342x24
		for weekends (as case 2) ($C3_{Summer.Weekend}$)	342x24
Case 4	First Layer	The daily consumption of each consumer on all days of the year	342x365x24
		Number of RLP obtained in first layer	⬇ *k* ⬇
	Second Layer	RLPs obtained after clustering analysis performed in the first layer	(*k*)x24

In this section, the clustering of the data has been examined for four different cases. In three of these cases, the data has been sent into the clustering algorithm with different approaches in terms of data selection and aggregation. In the other case study, the data set has been directly subjected to the cascaded k-medoids clustering algorithm proposed in the study. Details about the studied cases are given in Table - 1. The results have been compared and the findings have been discussed at the end of this section.

3.1. Data Structure and Data Preprocessing

The raw data used in the study is 4-year data (2014 - 2019) with 15 minutes resolution for 370 consumers from different classes (residential, commercial, industrial, etc.). One year of data is sufficient for the clustering analyzes planned to be carried out in the study. Therefore, the year with the least missing and outlier values has been detected and the remaining data belongs to three years have been removed from the raw data. Thereafter, consumers with missing and outlier values have been detected and cleaned from the data set. In order to reduce the data size, the data with 15 minutes resolution have been converted to hourly data, which is mostly preferred in load profile clustering studies.

After the data cleaning and reduction steps, a data set with a size of 342x365 has been obtained. The data set has been normalized in order to make the consumption behaviors in the data set independent of the size of the consumers' facilities. min-max normalization has been applied to the data set in order to make the consumption behaviors in the data set independent of the size of the consumers' facilities.

In Case 1, the data set has been sent to the clustering algorithm after estimating typical daily load profiles for each consumer. In Case 2, it has been decomposed as weekdays and weekends. In Case 3, the consumption data has been segmented by considering seasonal differences and six data set have been obtained by decomposing the data of each season as weekdays and weekends. Clustering analysis was performed 9 times, 1 in Case 1, 2 in Case 2 and 6 in Case 3. With the help of the elbow method, the optimum number of clusters for each clustering analysis is given in Fig. 2.

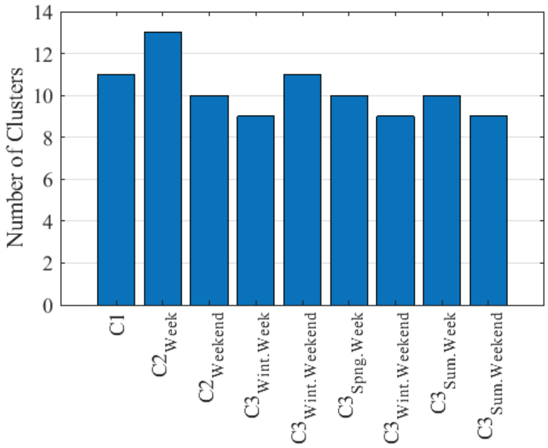

Fig. 2 Optimum number of clusters for each clustering analysis performed for Case 1, Case 2 and Case 3.

In Case 4, the proposed cascade clustering algorithm is directly applied to the data without the need for any prior decomposition or data processing. The decompositions performed by the practitioner in other cases are performed in the first level of the proposed cascaded algorithm without any need to external intervention. As a result of the clustering performed in the first level of the algorithm, the distribution of objects according to the number of clusters is given in Figure 3.

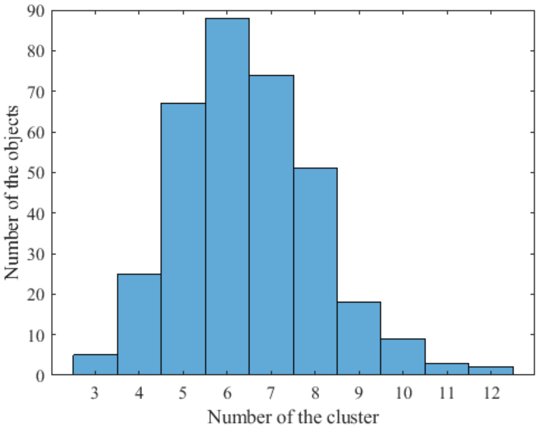

Fig. 3. Histogram of the distribution of objects according to the number of clusters after the first layer clustering

In the second step, the clusters obtained in the first step are subjected to the k-medoids clustering algorithm again and new cluster centers are determined for similar clusters belonging to different objects and these clusters are assigned to new cluster centers. After the first step, a total of 2213 clusters have been obtained. For the clustering analysis performed in the second step, the optimum number of clusters has been determined as 27 with the help of the elbow method.

Cluster centers produced as a result of clustering has been used as representative load profile (RLP) in order to compare the success of clustering analyzes for the studied cases. By using the representative profiles, the synthetic consumption data for a year for each users have been derived by considering the adopted situations of each case.

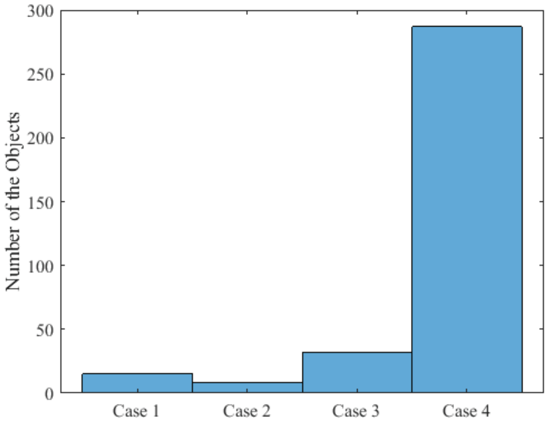

Fig. 4. Distribution of objects according to the cases in which they receive the smallest RMS value

After the clustering analysis for Case 1, 11 RLP have been obtained. Consumption data of each consumer for a year has been created by repeating the representative load profile to which the consumer has been assigned 365 times. To move from normalized values to actual values, each consumer has been multiplied by its maximum value.

The same approach has been applied for each case and one-year synthetic consumption data has been derived for all consumers in all cases. For each case, the RMSE values of all users between synthetic data and real data have been calculated. The RMSE values obtained for each user's four cases have been compared and it has been examined in which case the lowest RMSE value has been provided. Obtained results have been illustrated in Fig. 4.

As can be seen from the figure, the clustering analysis performed with the proposed clustering algorithm has a much higher success than those performed for the other three cases. Out of 342 consumers, 287 had the lowest RMSE value in Case 4.

4. Conclusions

Consumption behavior of consumers in power systems is a characteristic feature. The consumption behavior of the same consumer during the year may differ according to the days of the week and the seasons. Various approach may be adopted to increase the success of the clustering study by decomposing the data to identify the intervals when similar consumption behaviors are exhibited and by applying the cluster analysis separately to the decomposed data. Considering the number and diversity of consumers in the system, it is very difficult to implement this approach correctly. Consumption behavior of a consumer in the summer months may be similar to the behavior of another consumer in the winter months. On the other hand, it is not possible to determine the seasonal intervals accurately.

These uncertainties can be avoided with the cascaded clustering algorithm proposed within the scope of the study. In the first level of the proposed algorithm, representative consumption profiles of each user are determined, and in the second level, these consumption behaviors are clustered again. Thus, similar consumption behaviors of different consumers at different time intervals can be assigned to the same cluster center.

5. Acknowledgement

This study has been performed in the scope of the doctoral thesis study titled A New Approach to Demand Response in Smart Distribution Systems, which is in progress within the Graduate School of Science and Engineering of Yıldız Technical University.

6. References

[1] L. Rokach, “A survey of Clustering Algorithms,” in *Data Mining and Knowledge Discovery Handbook*, Boston, MA: Springer US, 2009, pp. 269–298.

[2] R. Xu and D. WunschII, “Survey of Clustering Algorithms,” *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005.

[3] K. Zhou, S. Yang, and C. Shen, “A review of electric load classification in smart grid environment,” *Renew. Sustain. Energy Rev.*, vol. 24, pp. 103–110, Aug. 2013.

[4] G. Chicco, R. Napoli, and F. Piglione, “Comparisons Among Clustering Techniques for Electricity Customer

Classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006.

[5] F. Rodrigues, J. Duarte, V. Figueiredo, Z. Vale, and M. Cordeiro, "A Comparative Analysis of Clustering Algorithms Applied to Load Profiling," in *Machine Learning and Data Mining in Pattern Recognition*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 73–85.

[6] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, Jun. 2012.

[7] S. Yilmaz, J. Chambers, and M. K. Patel, "Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management," *Energy*, vol. 180, pp. 665–677, Aug. 2019.

[8] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," *Renew. Sustain. Energy Rev.*, vol. 120, p. 109628, Mar. 2020.

[9] I. Khan, Z. Luo, J. Z. Huang, and W. Shahzad, "Variable Weighting in Fuzzy k-Means Clustering to Determine the Number of Clusters," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 9, pp. 1838–1853, Sep. 2020.

[10] L. Kaufman and J. P. Rousseeuw, "Clustering by Means of Medoids," in *Proceedings of the Statistical Data Analysis Based on the L1 Norm Conference*, 1987, pp. 405–416.