Clustering Electricity Big Data for Consumption Modeling Using Comparative Strainer Method for High Accuracy Attainment and Dimensionality Reduction

Elnaz Azizi[1,*], Hamed Kharrati Shishavan[2], Behnam Mohammadi-Ivatloo[2] and Amin Mohammadpour Shotorbani[3]

[1] *Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran*
[2] *Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran*
[3] *The University of British Columbia, Canada*
**Corresponding author: e.azizi@modares.ac.ir*

Manuscript received 18 January, 2018; Revised 29 September, 2018; accepted 15 Febuary, 2019. Paper no. JEMT-1801-1058.

In smart grid, the relation between grid and customer is bidirectional. Therefore, analyzing load consumption patterns is essential for optimal and efficient operation and planning of smart grid in addition to precise load forecasting. However, emergence of the advanced metering infrastructure, which enables a two-way flow of data and power consumption between consumers and suppliers, has resulted in data explosion in smart grid applications. Because of the volume and velocity of data generation in recent years, traditional data analysis methods are inefficient. Therefore, new methods of analyzing such as "data mining", which segments data before analyzing and manipulating, are recommended. Clustering, as a well-known method in data mining, has extensively been employed in recent electricity industry. This article argues that even though clustering methods can be directly applied to raw data of electricity consumption, this approach is inefficient since it requires storage and processing of high-dimensional and high-volume data. Hence, it would be more beneficial to cluster consumption data in a space of reduced dimension. In this paper, the authors propose a new structure for dimension reduction to refine the electricity consumption data. This method aims to increase the accuracy and decrease the time of clustering. The results are compared with the famous method of dimension reduction, component analysis (PCA). The authors evaluate the proposed technique using datasets from Kaveh, an industrial area in Iran.

Keywords: *Smart grid, Advanced metering infrastructure, Data mining, Clustering, principal component analysis*.

http://dx.doi.org/10.22109/jemt.2019.115900.1058

Nomenclature

i, j, k	Index of cluster
m	Feature index for elements of x_i
$N_{Cluster}$	Total number of clusters
$N_{dataset}$	Number of members of the dataset
N_F	Number of features
c	Cluster's centroid
c_{c1},c_{c2}	Centroids of first and second clusters, respectively.
n_i, n_j, n_k	Number of members of the clusters i, j and k, respectively
r_a	A positive constant which defines the neighborhood of a data point.
r_b	A positive constant to seperate the cluster centers
a, b, λ	Coefficients
P_i	Probability of the occurrence of the cluster i
$Mean_{dataset}$	The mean of the dataset
D_{base}	Dispersion of base dataset
S_{factor}	Similarity factor
D_{x_i}	Density of each data point
D_{c1}	Density of the first cluster's centroid
x_i	Each observation in dataset
$x_i(m)$	The m^{th} feature of the observation x_i
d_{ik}, d_{jk}, d_{ij}	The pairwise distances between the clusters i and k, j and k, and i and j, respectively.

1. Introduction

As the use of electricity grows more and more in modern life, an increase in average use and peak demand is an unavoidable fact. It is expected that the world's power consumption will rise by 53% by 2035 [1]. Due to this fact, the interest in analyzing load consumption patterns is gaining more importance [2]. In recent years, researchers in the field of smart grid use different methods of data mining, such as clustering, for analyzing datasets about load consumption patterns [3]. Clustering is one of the famous methods in data mining. In recent years, there has been an increasing interest in using the clustering approaches in the engineering application especially in the field of electrical power system. Consumption pattern segmentation, load modeling and forecasting based on clustering are hot research topics in this field [4-6].

Through clustering the data measured by advanced metering infrastructure, we try to extract informative characteristics and the pattern of the consumption to attain a probabilistic load model [7].

Recently, researchers in the field of smart grid have applied different methods of clustering on load consumption profiles, in different countries. The most known method of clustering, K-means, was applied on load datasets in different studies [8-10]. Researchers used K-means for clustering American residential load profiles in [9], and it was figured out that there are two main groups of users for each season. In addition, this article draws correlations to different profiles and found that some variables affect the shape of load profile significantly. As an instance, the hour of television watch per week and education levels have been drastically affected if someone works at home. In [10], K-means clustering method was applied on real power consumption and total harmonic distortion (THD) and the effectiveness of the combined method is verified by fuzzy logic technique. The authors in [8] have analyzed commercial and industrial load patterns. In this study, K-means was applied on data with different time resolutions, typical daily profiles (TDP) and typical weekly profiles (TWP). The authors in [11] have used dynamic clustering for verifying residential electricity users of Spain. They identify three main types of energy consumption users. Peak hours are different in these clusters. In [12], 1.2 million load patterns were verified. These load profiles are residential, industrial and commercial. Hierarchical clustering was used for analyzing this dataset. Due to the fact that, customers' behavior affects the consumption patterns, different studies has been done in the field of forecasting load pattern based on customers' behavior. The application of clustering in load forecasting was verified in [13-15]. For instance, in [10], k-means clustering was applied on data. After clustering, the similarities of customers' behavior of each cluster were studied. By considering these similarities, the load pattern of each group was forecasted. By combining of forecasts of each group, the entire load usage pattern was calculated. The load consumption patterns of a university campus were analyzed in [16]. K-means algorithm was applied on this data and the results were used for optimizing the usage of this area. In [17], it was shown that estimating load consumption based on K-means clustering has more accurate results. Due to the fact that K-means suffers from local optimal solutions, [18] presents a hierarchical K-means which overcame this problem and has more accurate results. Many articles used fuzzy C-means for analyzing the data and the results were compared with other methods of clustering. For instance, the authors in [19] analyzed the load profiles of seven areas of Italy by using fuzzy C-means (FCM) and linkage clustering methods. They applied these algorithms on raw data, normalized data, and on some features that were extracted from data. They aimed at segmenting data in three clusters. These clusters show working day's pattern, holiday's pattern and semi-holiday's pattern. Interestingly, they conclude that if they apply clustering in two steps and on features that were extracted from the main data, the load prediction will be more accurate. In [20], fuzzy clustering method has been applied on the data from smart grid to model the demand response. By considering dynamic prices, consumers reshape their consumption. It is showed that load shedding and valley filling can be resulted from demand response modeling. Electricity consumption patterns of China were analyzed in [21, 22]. FCM were applied on these data and the effectiveness of this algorithm was demonstrated based on different validity indexes. In [23], different methods of clustering including K-means, hierarchical, and the Gaussian mixture model-based clustering, were applied on load data to obtain optimal number of clusters. The result of this research has been used for defining better time of use pricing. Different methods of clustering such as K-means, K-medoid and self-organizing map have been applied on load data of Ireland in [5]. It was figured out the most proper algorithm for clustering these data is the self-organizing map method. The result of this clustering was used for extracting the relation between electricity pattern and customer manner. The authors in [24] have emphasized on the high voltage users' and industrial customers' load data. Different validity indices were calculated to find the most proper method of clustering and the optimal number of clusters.

Due to the velocity of production, volume and high dimensions of load datasets, many studies used different methods of dimension reduction to increase the speed of calculations [25-27]. For instance, in [25, 27], wavelet transform was applied on dataset before clustering in order to dimension reduction. Similarly, the authors in [28] have discussed different methods of clustering, and 18,200,000 load profiles were verified. Because of high dimension of this dataset, the principal component analysis was used for dimension reduction.

This work presented a novel method of dimension reduction, comparative strainer (CS) which has the following desirable characteristics in comparison with PCA: 1) CS omits the outliers of dataset, but PCA has a tendency to lose useful information, 2) CS reduces the time of analysis more than PCA. To prove these, the authors applied these methods on Kaveh industrial dataset. Kaveh industrial zone located at Saveh is the largest industrial zone in Iran. More than 500 factories have been located in this industrial zone. Due to the fact that more than 80% of electricity usage of Saveh is industrial, analyzing this data plays an important role in electricity management. The results of this analysis can be used for defining time-varying tariffs based on the peak hours of centroids of clusters. Researchers can take advantage of the results for proper planning of installing DGs and renewable sources.

This work aims to propose a new method of clustering called "comparative strainer". In this method, the combination of dimension reduction and linkage-ward clustering is applied on dataset. This algorithm was applied on electricity dataset and for demonstrating the effectiveness of CS, the results are compared with PCA clustering. It is shown that the proposed method increases the accuracy and reduces the time of analysis more than PCA clustering. Figure 1 illustrates the general architecture of this research. The proposed comparative strainer clustering method is described in detail in Section 2.6.

2. Methodology

This section presents an overview of the geographical location of Kaveh, clustering and principle analysis component. Finally, comparative strainer clustering is introduced and the procedures of this algorithm are described.

2.1. The Geographical Location of Kaveh

Kaveh industrial city is located in Saveh country, Markazi province, Iran. Because of its appropriate location and being a neighbor to populated and industrial provinces (Tehran and Isfahan), this city is one of the most proper areas for building different companies and factories. Therefore, the electricity usage is high there and the energy management plays an important role there.

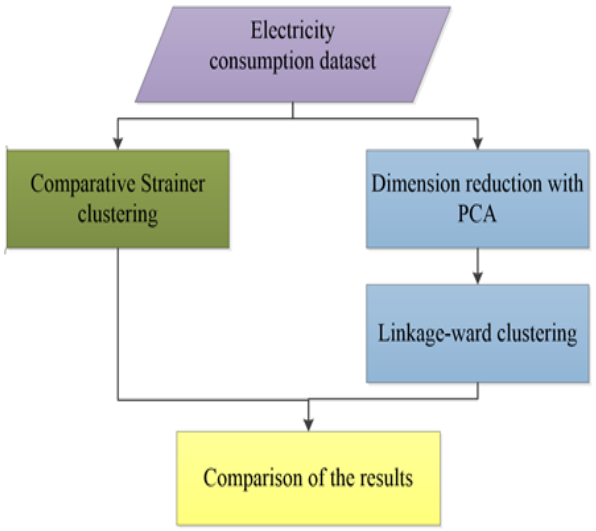

Fig. 1. Flowchart of the research

Fig. 2. Map of Kaveh in Markazi province of Iran

2.2. Data analysis

For the purpose of this work, the authors considered the actual industrial load data of industrial zone, Kaveh, in Iran during four years (2006-2009), which is measured each 60 minutes. Figure 3 displays load consumption pattern for 50 days of a year (2009). These days are chosen randomly from the main dataset.

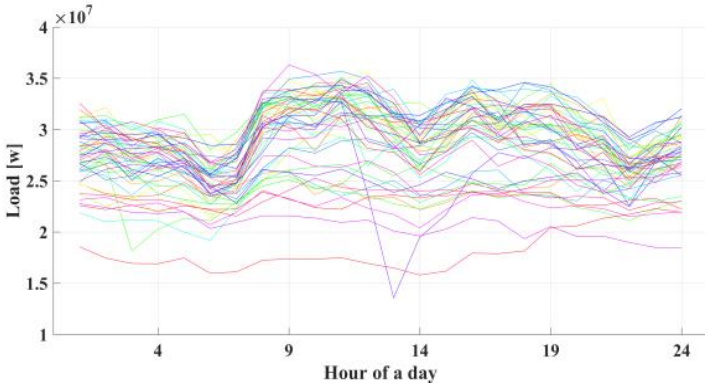

Fig. 3. Industrial load consumption pattern for 50 days of Kaveh, Iran

2.3. Clustering

The increasing amount of data being generated each year makes it critical to extract the useful information from that data. Because of the amount of data, traditional methods are not useful for analyzing the high-dimensional data. Therefore, new approaches of analyzing and verifying are recommended as data mining [29, 30], the process of extracting data set, analyzing it from many dimensions or perspectives, then producing a summary of the information in a useful form.

Clustering analysis is one of the most important and powerful branches of data mining and also a useful data analysis tool. Clustering methods segment the objects into optimally homogeneous groups, assuming measures of similarity among the objects. As it shown in Figure 4, clustering algorithms are broadly classified into overlapping and exclusive algorithms. Overlapping clustering algorithm assigns a data point to all clusters with different percentages. But, in exclusive clustering, each data just belongs to one cluster. Exclusive clustering can be further classified into hierarchical and partitioned clustering. Partitioned clustering directly divides data segments into a pre-determined number of clusters without building a hierarchical structure, whereas hierarchical clustering seeks to build a hierarchy of clusters with a sequence of nested partitions, either from singleton clusters to a cluster including all data segments or vice versa [31, 32].

Clustering allows researchers to compress a big amount of data into single data, corresponding to the representative of that cluster. This causes data dimension reduction, which raises the speed of analyzing the data.

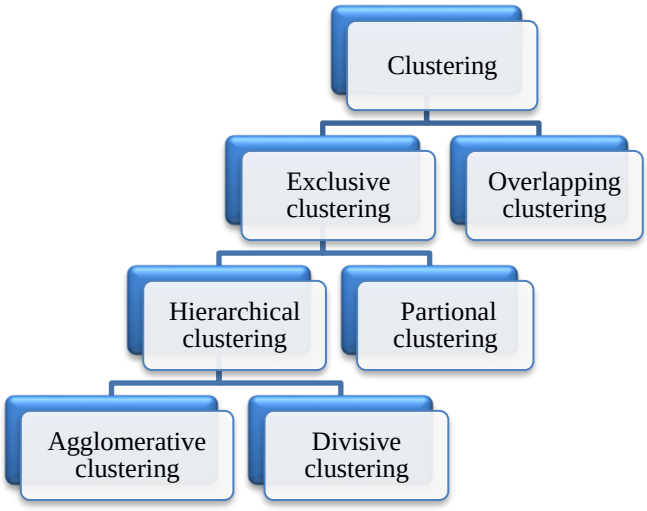

Fig. 4. Classification of clustering methods

There are many ways of clustering N data points (in a 24-dimensional space) in to K classes. We applied the linkage-ward (LW) clustering, which is one of the most common algorithms for partitioning [33]. LW is a hierarchical clustering method. This algorithm begins with considering each data as a cluster, and then clusters are merged together iteratively until only K clusters are remained. The merging policy analyzes the dissimilarity between existing clusters, and chooses the two clusters to be merged which guarantee the minimum increase of a particular objective function of interest. Overall, with hierarchical clustering, the sum of squares starts out at zero, because every point is in its own cluster, and then grows as we merge clusters. Ward's method keeps this growth as small as possible. In Ward's method the objective function, merging cost, is the

sum of squares from the points to the centroids, which is shown in the following equation:

$$d_{i+j,k} = ad_{ik} + ad_{jk} + bd_{ij} + \lambda\left|d_{ik} - d_{jk}\right| \quad \textbf{(1)}$$

where the coefficients a, b, and c are defined as:

$$a = \frac{n_i + n_k}{n_i + n_j + n_k}, b = -\frac{n_k}{n_i + n_j + n_k}, \lambda = 0$$

Clusters with minimum distance, based on the equation (1), merged together in each updating step. Figure 5 depicts the flowchart of this method.

The mean of the members of each cluster is chosen as centroid of each cluster using (2).

$$c_i = \frac{1}{n_i}\sum_{i=1}^{n_i} x_i \quad \textbf{(2)}$$

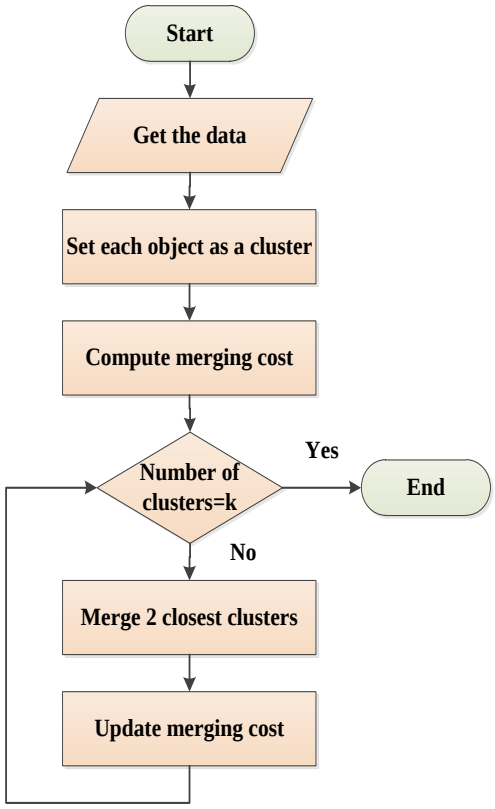

Fig. 5. Linkage-ward flowchart

2.4. Optimal number of clusters

For choosing optimal number of clusters, we used subtractive clustering. Subtractive clustering is a method to find the optimal number of data points to define a cluster centroid based on the density of surrounding data points. In the first step of this method, each data point is a candidate for a cluster center, then the density of each of them is calculated based on the following equation:

$$D_{x_i} = \sum_{j=1}^{n} \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{\left(r_a/2\right)^2}\right) \quad \textbf{(3)}$$

A data point will have a high-density value if it has many neighboring data points. After the density measure of each data point has been calculated, the data point with the highest density measure is selected as the first cluster center. Let Xc1 be the point selected and Dc1 its density measure. Next, the density measure for each data point xi is revised by following equation:

$$D_{x_i} = D_{x_i} - D_{c1}\exp\left(\frac{\left\|x_i - c_{c1}\right\|^2}{\left(r_b/2\right)^2}\right) \quad \textbf{(4)}$$

After the density calculation for each data point is revised, the nest cluster center cc2 is selected, and all of the density calculations for data points are revised again. This process is repeated until a sufficient number of cluster centers are generated.

2.4. Probability

The probability of occurrence of each cluster can be calculated by the following equation:

$$P_i = \frac{n_i}{\sum_{j=1}^{N} n_j} \quad \textbf{(5)}$$

2.5. Principal component analysis

When large multivariate datasets are analyzed, it is often desirable to reduce their dimensionality. One of the most important methods of dimension reduction is the principle component analysis (PCA). This method involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called as the “principal component”. In other words, PCA aims at reducing the diversion of a d-dimensional dataset by projecting it into a k-dimensional subspace (where k<d), in order to increase the computational efficiency while retaining most of the information. The PCA method transforms the data to a new coordinate system such that the first principal component accounts for as much of the variability in the data as possible [34, 35].

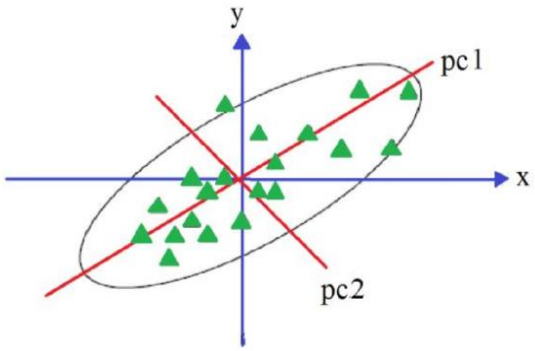

Fig. 6. First and second principle components of data points

2.6. Comparative Strainer method

The first point of this method is reducing the dimension of dataset before clustering and the second point is how to omit outliers from datasets. In comparative strainer (CS) method, the dataset is divided into smaller datasets, and then clustering is applied on one of these subsets and centroid of each cluster is measured. Now the strainer should identify the outliers or update centroids by comparing members of other subsets and centroids. The proposed method is summarized into 7 steps as:

Step 1: Separation of data

We separate each year’s data.

Step 2: Dispersion of data

For choosing the base dataset, we calculate the dispersion of four datasets. This calculation is based on the maximum distance between load patterns in each dataset which is measured based on Euclidean norm (6). The dataset with the maximum dispersion is chosen as the base dataset.

$$dist = \sqrt{\sum_{\substack{i=1 \\ j=i+1}}^{N_{dataset}} \sum_{m=1}^{Num_F} (x_i(m) - x_j(m))^2} \quad \textbf{(6)}$$

Step 3: Mean of datasets

Here the mean of each dataset is calculated.

$$Mean_{dataset} = \frac{\sum_{i=1}^{N_{dataset}} x_i}{N_{dataset}} \quad \textbf{(7)}$$

Step 4: Ratio

The ratio between each dataset's mean and the base dataset' mean is calculated in this step. Then each dataset is multiplied by the ratio.

Step 5: Cluster

The clustering method is applied on the base dataset and representatives are calculated.

Step 6: *Sfactor*

For comparing the similarity of data points with centroid of each cluster, *Sfactor* is calculated based on equation (8).

$$Sfactor = \frac{D_{base}}{N_{Cluster}} \quad \textbf{(8)}$$

Step 7: Compare and strain

Compare the members of each dataset with the centroid of each cluster based on Euclidean norm. This norm measures the distance between each representative and load patterns in other datasets.

If *dist* is lower than *Sfactor*, it means that the load pattern is correlated with a centroid and the centroid is updated. Else, it will be saved in a matrix named as *Pmatrix*. This matrix is used to save outliers.

3. Results

The LW clustering method is applied on the studied data in MATLAB. Based on the subtractive clustering method, ra varying from 0.8 to 1, the proper number of clusters for this dataset is 3, which can be considered as low, median and high electricity users. The mean of the members of each cluster is chosen as representative for each one. We consider three types of industrial users as low, median and high electricity users.

In first step, we apply LW clustering on the raw data, then after using PCA to have dimension reduction, LW method is applied. Finally, we use our new method (i.e. CS algorithm) for clustering this dataset based on the following steps.

Step 1: Separation of data

We have the industrial consumption pattern of four years. Therefore, we will have 4 datasets.

(data2006, data2007, data2008, data2009)

Step 2: Dispersion of data

As it is illustrated in Table 1, the dataset of 2009 has the maximum dispersion. This dataset is chosen as the base dataset.

Step 3: Mean of datasets

Table 2 displays the mean of each dataset.

Step 4: Ratio

Table 3 illustrates the ratio between datasets.

Step 5: Cluster

The LW method is applied on the base dataset and representatives are calculated.

Step 6: *Sfactor*

We have 3 clusters here. Therefore, n is 3 in *Sfactor*.

Step 7: Compare and strain

By comparing other data points with the centroids, the outliers removed and the centroids are updated with homologous ones.

Figure 7 depicts the representatives of each cluster in different methods of clustering as:

I- Raw clustering (original)

II- PCA clustering, and

III- CS clustering

Table 4 displays the sum of error between the original representatives (from raw clustering) and the representatives in other methods base on the Euclidean norm. The probability of occurrence of each cluster is displayed in table 5 and the time of clustering of three methods is showed in table 6.

It is figured out that about 0.1% of dataset is outlier, and the CS method omits these outliers.

Table 1. Dispersion of each dataset

	2006	2007	2008	2009
Dispersion (MW)	73.59	106.24	109.93	119.48

Table 2. Mean of load consumption in each year (MW)

	2006	2007	2008	2009
1	22.759	25.321	26.082	25.372
2	23.161	25.769	26.543	25.820
3	22.227	24.730	25.473	24.780
4	21.977	24.452	25.187	24.501
5	21.316	23.716	24.429	23.764
6	20.091	22.354	23.026	22.399
7	21.645	24.082	24.806	24.130
8	24.095	26.808	27.614	26.862
9	24.978	27.791	28.626	27.846
10	24.919	27.725	28.558	27.780
11	25.367	28.223	29.072	28.280
12	24.805	27.598	28.427	27.653
13	23.659	26.323	27.115	26.376
14	22.045	24.527	25.265	24.577
15	23.556	26.208	26.996	26.261
16	24.350	27.091	27.906	27.146
17	23.705	26.374	27.167	26.427
18	23.938	26.634	27.434	26.687
19	23.446	26.086	26.870	26.138
20	22.472	25.002	25.754	25.005

21	21.192	23.579	24.287	23.626
22	20.223	22.500	23.176	22.545
23	21.361	23.766	24.481	23.814
24	21.931	24.400	25.133	24.449

Table 3. Ratio between the mean of each dataset

Mean_2009/mean_2008	1.028
Mean_2009/mean_2007	0.998
Mean_2009/mean_2006	0.897

Table 4. Sum of error between centroids of each clustering and original centroids

	PCA	CS
Error (w)	1.1093e+08	3.6777e+07

Table 5. Probability of occurrence

	PCA	CS
Error (w)	1.1093e+08	3.6777e+07

Table 6. Time computation comparison

	Cluster 1	Cluster 2	Cluster 3
Raw clustering	0.4994	0.4364	0.0641
PCA and clustering	0.6873	0.2500	0.0627
CS and clustering	0.5900	0.3501	0.0556

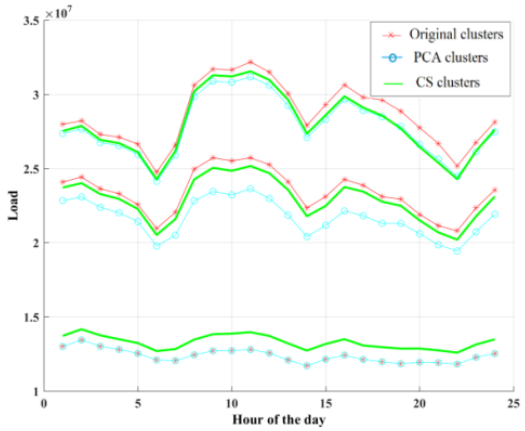

Fig. 7. Centroid of clusters in each method of clustering

4. Conclusions

In this study, the linkage-ward method as one of the famous methods of hierarchical clustering is applied on the industrial power consumption pattern of Kaveh industrial area in Iran, which is measured during four years. We have three kinds of clustering in this study; I- raw clustering (i.e. original clustering of the entire dataset), II- PCA clustering (i.e. applying the PCA technique for dimension reduction prior to clustering), and III- the proposed comparative strainer (CS) clustering (i.e. applying the clustering on a small part of dataset, and then straining the outliers through comparing). The results of these methods are compared with each other. The aim of this comparison is to provide evidence that the proposed method of dimension reduction not only reduces the time of analyzing but also increases the accuracy of clustering. Therefore, this algorithm can be effective for applying on big dataset, before clustering. Due to the fact that by installing advanced metering infrastructure, we face big data in the field of smart grid, CS can be considered as a powerful method in this field. Since Kaveh will face with the power supply problem, power suppliers can extract informative material about the actual consumption from the result of this paper and define proper tariffs.

References

[1] Birol, F.. "World energy outlook." *Paris: International Energy Agency* 23, no. 4 (2008): 329.
[2] Broeer, T., J. Fuller, F. Tuffner, D. Chassin, and N. Djilali. "Modeling framework and validation of a smart grid and demand response system for wind power integration." *Applied Energy* 113 (2014): 199-207.
[3] Stimmel, C. L. *Big data analytics strategies for the smart grid*. Auerbach Publications, 2016.
[4] Henao, N., K. Agbossou, S. Kelouwani, Y. Dubé, and M. Fournier. "Approach in nonintrusive type I load monitoring using subtractive clustering." *IEEE Transactions on Smart Grid* 8, no. 2 (2017): 812-821.
[5] McLoughlin, F., A. Duffy, and M. Conlon. "A clustering approach to domestic electricity load profile characterisation using smart metering data." *Applied energy*141 (2015): 190-199.
[6] Quilumba, F. L., W. Lee, H. Huang, D. Y. Wang, and R. L. Szabados. "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities." *IEEE Transactions on Smart Grid* 6, no. 2 (2015): 911-918.
[7] Chicco, G.. "Overview and performance assessment of the clustering methods for electrical load pattern grouping." *Energy* 42, no. 1 (2012): 68-80.
[8] Frost, A. E., M. Azaza, H. Li, and F. Wallin. "Patterns and temporal resolution in commercial and industrial typical load profiles." *Energy Procedia* 105 (2017): 2684-2689.
[9] Rhodes, J. D., W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber. "Clustering analysis of residential electricity demand profiles." *Applied Energy* 135 (2014): 461-471.
[10] Selvam, M. M., R. Gnanadass, and N. P. Padhy. "Fuzzy based clustering of smart meter data using real power and THD patterns." *Energy Procedia* 117 (2017): 401-408.
[11] Benítez, I., A. Quijano, J. Díez, and I. Delgado. "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers." *International Journal of Electrical Power & Energy Systems* 55 (2014): 437-448.
[12] Zhang, P., X. Wu, X. Wang, and S. Bi. "Short-term load forecasting based on big data technologies." *CSEE Journal of Power and Energy Systems* 1, no. 3 (2015): 59-67.
[13] Quilumba, F. L., W. Lee, H. Huang, D. Y. Wang, and R. L. Szabados. "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities." *IEEE Transactions on Smart Grid* 6, no. 2 (2015): 911-918.
[14] Diao, L., Y. Sun, Z. Chen, and J. Chen. "Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation." *Energy and Buildings*147 (2017): 47-66.
[15] Panapakidis, I. P. "Clustering based day-ahead and hour-ahead bus load forecasting models." *International Journal of Electrical Power & Energy Systems* 80 (2016): 171-178.
[16] Pérez-Chacón, R., J. Luna-Romera, A. Troncoso, F. Martínez-Álvarez, and J. Riquelme. "Big data analytics for discovering electricity consumption patterns in smart cities." *Energies* 11, no. 3 (2018): 683.
[17] Al-Wakeel, A., J. Wu, and N. Jenkins. "K-means based load estimation of domestic smart meter measurements." *Applied energy* 194 (2017): 333-342.
[18] Xu, T., H. Chiang, G. Liu, and C. Tan. "Hierarchical K-means method for clustering large-scale advanced metering infrastructure

data." *IEEE Transactions on Power Delivery* 32, no. 2 (2017): 609-616.
[19] Semeraro, L., W. Crisostomi, A. Franco, A. Landi, M. Raugi, M. Tucci, and G. Giunta. "Electrical load clustering: The Italian case." In *IEEE PES Innovative Smart Grid Technologies, Europe*, pp. 1-6. IEEE, 2014.
[20] Pereira, R., A. Fagundes, R. Melicio, V. M. F. Mendes, J. Figueiredo, and J. C. Quadrado. "Fuzzy subtractive clustering technique applied to demand response in a smart grid scope." *Procedia Technology* 17 (2014): 478-486.
[21] Zhou, K., C. Yang, and J. Shen. "Discovering residential electricity consumption patterns through smart-meter data mining: A case study from China." *Utilities Policy*44 (2017): 73-84.
[22] Zhou, K., S. Yang, and Z. Shao. "Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study." *Journal of cleaner production* 141 (2017): 900-908.
[23] Crow, M. L. "Clustering-based methodology for optimal residential time of use design structure." In *2014 North American Power Symposium (NAPS)*, pp. 1-6. IEEE, 2014.
[24] Panapakidis, I., M. Alexiadis, and G. Papagiannis. "Evaluation of the performance of clustering algorithms for a high voltage industrial consumer." *Engineering Applications of Artificial Intelligence* 38 (2015): 1-13.
[25] Jiang, Z., R. Lin, F. Yang, and B. Wu. "A fused load curve clustering algorithm based on wavelet transform." *IEEE Transactions on Industrial Informatics* 14, no. 5 (2018): 1856-1865.
[26] Lin, S., F. Li, E. Tian, Y. Fu, and D. Li. "Clustering load profiles for demand response applications." *IEEE Transactions on Smart Grid* (2017).
[27] Mets, K., F. Depuydt, and C. Develder. "Two-stage load pattern clustering using fast wavelet transformation." *IEEE Transactions on Smart Grid* 7, no. 5 (2016): 2250-2259.
[28] Chelmis, C., J. Kolte, and V. K. Prasanna. "Big data analytics for demand response: Clustering over space and time." In *2015 IEEE International Conference on Big Data (Big Data)*, pp. 2223-2232. IEEE, 2015.
[29] Felici, G.. *Mathematical methods for knowledge discovery and data mining*. IGI Global, 2007.
[30] Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.
[31] He, Q.. "A review of clustering algorithms as applied in IR." *Graduate School of Library and İnformation Science University of İllinois at Urbana-Compaign* 6 (1999).
[32] Jain, A. K., M. N. Murty, and P. J. Flynn. "Data clustering: a review." *ACM computing surveys (CSUR)* 31, no. 3 (1999): 264-323.
[33] Webb, A. R. *Statistical pattern recognition*. John Wiley & Sons, 2003.
[34] Shlens, J.. "A tutorial on principal component analysis." *arXiv preprint arXiv:1404.1100* (2014).
[35] Abdi, H., and L. J. Williams. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics* 2, no. 4 (2010): 433-459.