International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 6, December 2018, pp. 4829~4835
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i6.pp4829-4835

Advanced SOM & K Mean Method for Load Curve Clustering

Phan Thi Thanh Binh[1], Trong Nghia Le[2], Nui Pham Xuan[3]
[1,3]Department of Electrical and Electronics Engineering, HCMC University of Technology, Vietnam
[2]Department of Electrical and Electronics Engineering, HCMC University of Technology and Education, Vietnam

Article Info

Article history:

Received Feb 1, 2018
Revised Jun 30, 2018
Accepted Jul 22, 2018

Keyword:

Cluster analysis
K-mean
PSO
SOM
Subtractive clustering

ABSTRACT

From the load curve classification for one customer, the main features such as the seasonal factors, the weekday factors influencing on the electricity consumption may be extracted. By this way some utilities can make decision on the tariff by seasons or by day in week. The popular clustering techniques are the SOM & K-mean or Fuzzy K-mean. SOM &Kmean is a prominent approach for clustering with a two-level approach: first, the data set will be clustered using the SOM and in the second level, the SOM will be clustered by K-mean. In the first level, two training algorithms were examined: sequential and batch training. For the second level, the K-mean has the results that are strongly depended on the initial values of the centers. To overcome this, this paper used the subtractive clustering approach proposed by Chiu in 1994 to determine the centers. Because the effective radius in Chiu's method has some influence on the number of centers, the paper applied the PSO technique to find the optimum radius. To valid the proposed approach, the test on well-known data samples is carried out. The applications for daily load curves of one Southern utility are presented.

Corresponding Author:

Phan Thi Thanh Binh,
Department of Electrical and Electronics Engineering,
HCMC University of Technology,
268 Ly Thuong Kiet steeet, 10 district, Ho chi minh city, Vietnam.
Email: pttbinh@hcmut.edu.vn

1. INTRODUCTION

The load curve classification has one important meaning: the utility can draw the own feature for each group in one class of consumer [1]-[2]. Here the main features such as the seasonal factors, the week day factors, influencing on the electricity consumption may be extracted. By this way some utilities can make decision on the tariff by seasons or by day in a week. Some utilities will have the different prices on electricity for winter, summer. Others will take the prices for working days in difference with those for the weekend with the very clear purpose: to shift loads from working days to the weekend. Many utilities design their demand response policy for each customer group having the same form of load curves [3].

Load curve classification is the clustering with the large number of input data. The daily load curve for years or months must be considered. From the point of data mining, the way of clustering big data is necessary to extracting useful information. Many authors concentrated on data clustering basing on the K-mean algorithm because it is rather easy to implement and apply even on large data set. Jung, *et al* used K-means algorithms combining with principal component analysis to analyze and classify user data efficiently [4]. But as mentioned in [5]-[7], K-means has the results that strongly depended on the initial values of the centers, so this will influence on the clustering results. To over come this drawback, Bedboudi, *et al* used the combining K-mean and genetic algorithm, meanwhile Sahu, *et al* used the Adaptive K-mean [5], [6]. Chiu presented the subtractive method to remove the influence of center initialization [7]. For load curve clustering, many works are based on dimensionality reduction in order to simplify the models or reduce the computation time such as [8]-[10]. Here the feature selection or construction is the main key for clustering.

For example, [10] proposed three ways to construct the features, exceptionally are suitable for smart metering: conditional filters on time-resolution based features, calibration and normalization, and using profile errors.

Other works continue to use the advantages of K-mean algorithm and combine with the dimensionality reduction algorithm for load curve clustering. The popular clustering techniques, based on this combining, are the SOM & K-means. With the large number of input data, SOM & K-means is a prominent approach for clustering. In [11] this technique is with two-level approach: first, the data set will be clustered using the SOM by sequential training algorithm. The result here is a set of prototype vectors. In the second level, the SOM will be clustered by K-mean. But this method contains the weak points of K-mean so does not have the high accuracy. Besides, the sequential training algorithm for SOM is time consumption.

To take the full advantage of SOM & K-mean with the big data, to over come its drawback, this paper will use the subtractive clustering method for the second level. However, choosing the effective radius is one key question of clustering procedure. We proposed applying the PSO technique to find the optimum radius in order to improve the accuracy. The paper also used another training way in SOM- the batch training algorithm to enhance the calculating time. To validate the proposed method, the Fuzzy K-mean algorithm will be also applied to give the comparison.

The work is organized as the following: some mathematics definition such as SOM, K-mean, Fuzzy K-mean, PSO will be mentioned in Section 2; the proposed algorithm (denoted as Advanced SOM & K means) will be presented in Section 3 with some tests on the famous data set; finally, one case study will be presented in Section 4, comparing the results of different algorithms such as SOM & K-means, Fuzzy K-mean.

2. SOME MATHEMATIC DEFINITIONS

2.1. SOM

The SOM consists of a regular, usually two-dimensional 2D grid of map units. Data points lying near each other in the input space are mapped onto nearby map units. The SOM can be interpreted as a topology preserving mapping from input space onto the 2-D grid of map units.

In our work, the two algorithms for training of the maps were carried out: sequential training algorithm and batch training. The neuron whose weight vector is closest to the input vector is called the best-matching unit (BMU) denoted by *c*. In the batch training algorithm, instead of using a single data vector at a time, the whole data set is presented to the map before any adjustments are made (hence the name "batch"). In each training step, the data set is partitioned according to the Voronoi regions of the map weight vectors, i.e. each data vector belongs to the data set of the closest map unit. After this, the new weight vectors are calculated as follows:

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{ic}(t) x_j}{\sum_{j=1}^{n} h_{ic}(t)} \quad (1)$$

where: ***t*** denotes time; xj is an input vector; $\boldsymbol{h_{ci}(t)}$ the neighborhood Kernel around the winner unit; $c = \arg\min_k \left\{ \left\| x_j - m_k \right\| \right\}$ is the index of the BMU of data sample, with m_k is synaptic weight vector k .

2.2. The K-mean algorithm

The K-mean-algorithm is a well-known algorithm in clustering field. For each cluster number K, the procedure follows a simple way to classify a given data set and looks like that:

$$F = \sum_{i=1}^{k} \sum_{j=1}^{n} \left\| x_j - z_i \right\|^2 \rightarrow \min \quad (2)$$

where, $\|.\|$ is the Euclidean distance between x_j and z_i.; z_i- is the center of the i^{th} cluster; *k*- is the number of clusters centers; *n*-number of data. The Davies-Bouldin (DB) index is applied for hard clustering [6]. The optimal number of clusters corresponds to the minimum value of DB index.

2.3. The subtractive method

Consider a collection of *n* data points $\{x_1, x_2 ... x_n\}$ in an M dimensional space. If each data point is considering as a possible cluster center, then the potential of data point x_i will be:

$$P_i = 1\sum_{k=1}^{n} e^{-\alpha\|x_k - x_i\|} \quad (3)$$

with $\alpha = 4/r_a^2$. The constant r_a is effectively the radius defining a neighborhood. The data point with the highest potential is selected as the first cluster center. Let x_1^* be the location of the first cluster center and P_1^* be its potential value. The potential of each data point x_i is revised by the formula:

$$P_i \Leftarrow P_i - P_1^* e^{-\beta\|x_k - x_1^*\|} \quad (4)$$

with $\beta = 4/r_b^2$, where r_b is the effective radius and be equal to 1.25 r_a. The data point with the highest remaining potential is selected as the second cluster center. The process is then continued further until the remaining potential of all data points falls below some fraction of the potential of the first cluster center P_1^*.

2.4. The PSO [12]

PSO was based on the phenomenon of collective intelligence inspired by the social behavior of bird flocking or fish schooling. The fitness function is evaluated for each particle in the swarm and is compared to the fitness of the best previous position for that particle $pbest_t$ and to the fitness of the global best particle among all particles in the swarm *gbest*. After finding the two best values, the i^{th} particles evolve by updating their velocities and positions according to the following equations:

$$V_i^{k+1} = wV_i^k + c_1 rand_1 * (pbest_i - s_i^k) + c_2 rand_2 * (gbest_i - s_i^k) \quad (5)$$

$$s_i^{k+1} = s_i^k + V_i^{k+1} \quad (6)$$

where: s^k -current searching point; s^{k+1} -modified searching point; v^{k-} -current velocity; v^{k+-} -modified velocity; *rand1 and rand2*- the random values in (0,1) following a normal distribution; *c1* and *c2* are constants called acceleration coefficients; w-some weighted coefficient. The values of *c1* and *c2* control the weight balance of *pbest* and *gbest* in deciding the particle's next movement.

2.5. Fuzzy K-means (FKM) [13]

FKM is one clustering method with high flexibility having the following objective function:

$$F = \sum_{i=1}^{K}\sum_{j=1}^{n} w_{ij}^{\alpha} d^2(z_i, x_j) \to \min \quad (7)$$

where α is a weighting exponent; w_{ij} is the value of membership function and $d(z_i, x_j)$ is the Euclidean distance between x_i and the center z_i of *i* cluster. For determining the final number of clusters, there are many criteria are applied. This paper used the methods in [14] based on the principles Bellmand – Zadeh.

3. THE PROPOSED ALGORITHM

The proposed algorithm (denoted as Advanced SOM & K means) will be shown in Figure 1. The batch training approach is used and the training time will be enhanced. Here the Subtractive clustering is applied to find out the initial centers for K-means. Traditionally, the radius r_a in (3) has the values from 0.15 to 0.8. Our examining shows that the smaller the r_a is, the large the number of clusters will be received. So, the optimum radius is the one that will lead to the smallest value of DB index. To find out the suitable radius r_a, this paper applied the PSO algorithm.

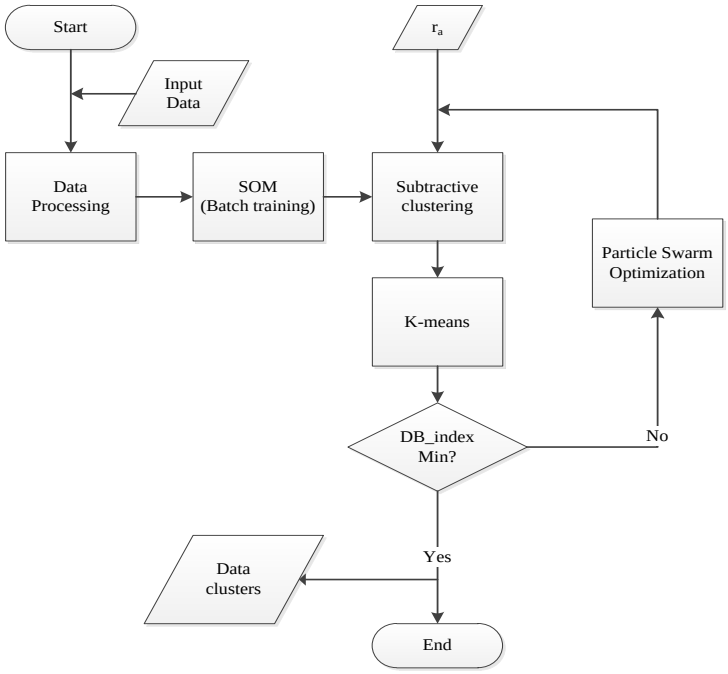

Figure 1. The proposed algorithm

4. EXPERIMENTAL STUDIES

4.1. Testing on the well-known data samples:

Three real and famous data sets (Iris, WBCD, Wine) are taken These data sets are used in many works for testing the clustering technique. The Iris Plants Database [15] contains 150 samples (4 attributes in each sample) and was clustered into 3 classes: Iris Setosa; Iris Versicolour; Iris Virginica (50 samples for each class). The Wisconsin Breast Cancer Database [16] was built from the University of Hospitals. It contains 683 test (10 attributes in each test) and was clustered into 2 classes: benign (65.5%) and malignant (35.5%). The last one [17] is the data obtained from a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituent found in each of the three types of wines. Three algorithms: SOM & K-mean, FKM, and Advanced SOM & K-mean are applied and the results are given in Table 1. From Table 1, the conclusion is that Advanced SOM & K-mean has the best result.

Table 1. Testing results on well-known data samples

Data sample	Number of the correct cluster	Algorithms		
		SOM & K-means	FKM	Advanced SOM & K-means
Iris	3	2	2	3
WBCD	2	3	2	2
Wine	3	2	7	3

4.2. Application for load curve clustering

4.2.1. The input data

The 365 daily load curves of one utility in the South of Vietnam are the input data. Each load curve is regarded as the vector of 24 attributes (24 hours). The Euclidean distances between two load curves *j* and *k* will be defined as:

$$d_{jk} = \sqrt{\sum_{i=1}^{24} (x_{ij} - x_{ik})^2} \quad (8)$$

where, x_{ij}-load at *i*-hour of *j*-load curve.

4.2.2. Extract the information

From the clustering process, by looking into each cluster, the main factors characterized each cluster may be extracted. For example if the load curves in one cluster are belonged to the rainy season, while in other cluster-the dry season, then it can say that there is a necessity to form a seasonal tariff. And if there are the different clusters by weekend and working day, the weekend day tariff must be formed.

4.2.3. Implementation

As implementation, here the daily load curves of one utility in the year of 2012 were used. The tariff is TOU (time of use) and is the same for all day in week. All three algorithms have the same number of cluster (2 clusters) called holiday cluster and normal day cluster. There is no show of the rainy and dry season clusters. All of Sunday and public holidays are belonged to the holiday cluster. This result is consistent because the HochiMinh city is with the tropical climate, and on the other hand, there are many industrial parks and invested abroad enterprises so that the difference in load by seasons is not clearly. The Holiday cluster contained all of Sunday and public holidays according to Vietnam's Labor Code. So that there are 63 days in standard holiday cluster can see in Figure 2. It emphasizes the necessity to form the different prices on electricity for working days and Holidays. But the result shows that there are more than 63 days in the holidays cluster. There are some Saturdays and working days falling into the holiday cluster can see in Table 2.

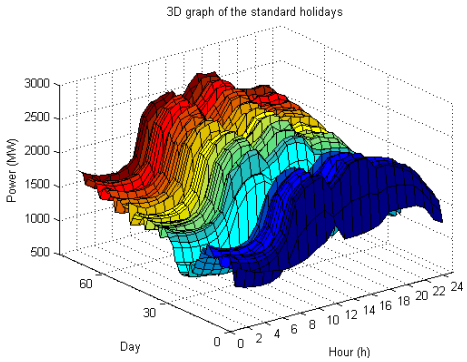

Figure 2. The load curves of 63 standard public holidays

There are differences in the results of 3 algorithms can see in Table 2. To consider the result accuracy of three algorithms, the distance of each different day to center of the standard holidays (63 days) and the normal days will be calculated can see in Table 3.

The results of FKM and Advanced SOM & K-means are coincided except for 4 days (Saturdays: 4-Feb., 11-Feb., 18-Feb, 6-Oct.). According to FKM, theses days belonged to the holiday cluster. But from Table 3, theses Saturdays have the distance to the center of the standard normal day cluster smaller than of the standard holiday cluster. It means that these 4 Saturdays must belong to the normal day cluster. And that means FKM is less accurate than Advanced SOM & K-means.

The results of SOM & K-mean and Advanced SOM & K-means are coincided except for one day (Tuesday: 31-Jan). This Tuesday has the distance to the center of the standard normal day cluster larger than the standard holiday cluster and must be belonged to the standard holiday cluster. So, the Advanced SOM & K-means algorithm gets the better accuracy than SOM & K-means.

Table 2. Number of weekdays in Holiday cluster for 3 algorithms

Weekday	Algorithms		
	SOM & K-means	FKM	Advanced SOM & K-means
Monday	7	7	7
Tuesday	3	4	4
Wednesday	3	3	3
Thursday	2	2	2
Friday	2	2	2
Saturday	4	8	4
Sunday	53	53	53

Table 3. Distance of all different days to standard holiday cluster's center (SHCC) and normal day cluster's center (SNDCC)

Day	Avg. dist. to the SHCC	Avg. dist. to the SNDCC
31-Jan-12	1030.48	1988.70
4-Feb-12	1742.42	1052.60
11-Feb-12	1753.69	977.21
18-Feb-12	1742.35	1014.14
6-Oct-12	1820.48	1046.44

This emphasizes the fact that Advanced SOM & K-means algorithm overcome the weak point of those algorithm based on the K-mean, and the choosing of optimal radius in Subtractive method enhances the accuracy.

4.2.4. Compare in time calculation domain

Changing SOM training by the batch training algorithm greatly reduces training time. Besides, applying the Subtractive clustering algorithm to get initial center in K-means can lead to quite fast solution the performance tests were made in a computer with 4 GBs of memory and 2.4 GHz Intel Core i3 CPU and have the following results: SOM & Kmeans-1599(s); Advanced SOM & K-means-62(s); FKM - 488 (s).

5. CONCLUSION

The data analysis presented in this work has been tested and validated using real data of one utility and the well-known data samples. Among three algorithms examined in this paper, the proposed Advanced SOM & K-means has the better result and smallest time for calculating. This algorithm overcomes some disadvantages of traditional SOM & K-means, FKM. In the results, the daily consumption behavior of a real utility has been analyzed by clustering and it shows that it is necessary to make different electricity prices for working days and for weekends. This algorithm can also be used for clustering different groups of customers-the basic for applying different tariff for different customer classes. For the future works, the study of possibility to apply this algorithm for detecting time zones of Time-of-Use tariff will be carried out.

ACKNOWLEDGEMENTS

The authors would like to thank the HCMC University of Technology and HCMC University of Technology and Education for their supports.

REFERENCES

[1] G. Chicco, *et al.*, "Customer characterization options for improving the tariff offer," *IEEE Trans. Power Syst*, vol/issue: 18(1), pp 381-387, 2003.
[2] D. Gerbec, *et al.*, "Determination and allocation of typical load profiles to the eligible customers," *in Proc IEEE Bologna Power Tech*, Bologna Italy, 2003.
[3] S. Valero, *et al.*, "Methods for customer and demand response policies selection in new electricity markets," *IET Gener. Transm. Distrib*., vol/issue: 1(1), pp. 104-110, 2007.
[4] S. H. Jung, *et al.*, "Prediction Data Processing Scheme using an Artificial Neural Network and Data Clustering for Big Data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol/issue: 6(1), pp. 330-336, 2016.
[5] A. Bedboudi, *et al.*, "An Heterogeneous Population-Based Genetic Algorithm for Data Clustering," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol/issue: 5(3), pp. 275-284, 2017.
[6] M. Sahu, *et al.*, "Parametric Comparison of K-means and Adaptive K-means Clustering Performance on Different Images," *International Journal of Electrical and Computer Engineering (IJECE)*, vol/issue: 7(2), pp. 810-817, 2017.
[7] S. L. Chiu, "Fuzzy model identification based on cluster estimation," *Journal of Intelligent and Fuzzy Systems*, vol/issue: 2(3), 1994.
[8] N. Jin, *et al.*, "Subgroup discovery in smart electricity meter data," *Industrial Informatics, IEEE Transactions on*, vol/issue: 10(2), pp. 1327-1336, 2014.
[9] I. Dent, *et al.*, "Variability of behaviour in electricity load profile clustering; who does things at the same time each day," *in Advances in Data Mining. Applications and Theoretical Aspects, ser. Lecture Notes in Computer Science*, P. Perner, Ed. Springer International Publishing, vol. 8557, pp. 70-84, 2014.
[10] R. Al-Otaibi, *et al.*, "Feature Construction and Calibration for Clustering Daily Load Curves from Smart Meter Data," *Industrial Informatics, IEEE Transactions on*, vol/issue: 12(2), pp. 645-654, 2016.

[11] S. V. Verdú, *et al.*, "Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the use of Self-Organizing Maps," *IEEE Transactions on power systems*, vol/issue: 21(4), 2006.
[12] M. El-Tarabily, *et al.*, "A PSO – Based on Subtractive Data Clustering Algorithm," *International Journal of Research in Computer Science*, vol/issue: 3(2), pp. 1-9, 2013.
[13] N. R. Pal and J. C. Bezdek, "On Cluster Validity for the Fuzzy c-means model," *IEEE Trans, Fuzzy syst*., vol/issue: 3(3), pp. 370-379, 1995.
[14] P. T. T. Binh, *et al*., "Determination of Representative Load Curve based on Fuzzy K-Means," *Proc. PEOCO*, 2010.
[15] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annual Eugenics*, vol. 7, Part II, pp. 179-188, 1936.
[16] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," *SIAM News*, vol/issue: 23(5), pp. 1-18, 1990.
[17] Forina, *et al*., "An Extendible Package for Data Exploration," *Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies*, Via Brigata Salerno, 16147 Genoa, Italy.

BIOGRAPHIES OF AUTHORS

Phan Thi Thanh Binh received Ph.D. degree in electrical engineering from Kiev Polytechnique University, Ukraine in 1995. Currently, she is a Assos. professor and lecturer in the Faculty Electrical and Electronics Engineering, HCMUT. Her main areas of research interests are power systems stability, power systems operation and control, load forecasting, data mining.

Trong Nghia Le received his M.Sc. degree in electrical engineering from Ho Chi Minh City University of Technology and Education (HCMUTE), Vietnam, in 2012. Currently, he is a lecturer in the Faculty Electrical and Electronics Engineering, HCMUTE. His main areas of research interests are load shedding in power systems, power systems stability, load forecasting and distribution network.

Nui Pham Xuan received his M.Sc. degree in electrical engineering from Ho Chi Minh City University of Technology, Vietnam, in 2013. Currently, he works at Quality Assurance and Testing Center 3 (QUATEST 3). His main area of research interests is data mining.