A Two-stage Pattern Recognition Method for Electric Customer Classification in Smart Grid

Bo Peng, *Student Member, IEEE* ,Can Wan, *Member, IEEE,* Shufeng Dong, *Member, IEEE,* Jin Lin, *Member, IEEE,* Yonghua Song, *Fellow, IEEE,* Yi Zhang, Jun Xiong

***Abstract*—Identifying the consumption patterns of electric customers and grouping them to classes according to their load characteristics can be very meaningful for power supply and demand side management in smart grid. Previously, tariff structures are mainly based on the type of activity. However, the type of activity and electrical behavior of the customer have poor relationship. Using clustering techniques to classify customer according to load curves is more meaningful. This paper proposes a two-stage clustering algorithm combining supervised learning methods to classify electric customer. Firstly, clustering results are obtained based unsupervised learning method. Clustering method and number to get the result of first-stage are selected via the clustering evaluation index. Secondly, customers are reclassified using supervised learning algorithm. Different supervised learning algorithms for second-step reclassification are compared in the case studies. Case studies show that second-step reclassification can make up for the weakness of first-step clustering in load shape similarity.**

***Index Terms*—Load clustering; data mining for power system; load shape; supervised learning algorithm**

I. Introduction

With the development of smart grid, more and more smart meters are installed into distribution networks[1]. Consumption behaviors of customers are known through load curves data collected from smart meters. The demand side data is important for the operation and control of the distribution network [2],[3]. Clustering technology is very useful for data mining in smart gird. In competitive electricity market with severe uncertainties[4]-[6], performing effective customer classification according to customers' electrical behavior is important for setting up new tariff offers [7]. In the past, tariff structures are mainly based on the type of economic activities. However, the types of economic activities have poor relationship with the electrical behavior of the customer [7]. Using clustering techniques to classify customers according to load curves is more meaningful. Besides, conducting load pattern analysis by clustering load curves is also important for load forecasting [8]-[10], making marked strategies [11], and demand side management [12], [13].

Various methods for clustering load curves have been used in recent years. such as K-means[14],[15], fuzzy c-means(FCM)[16],[17], hierarchical methods[18],[19], self-organizing map (SOM)[20], support vector machine (SVM)[21],[22], subspace projection Method [23]. Most of these methods are based on load curves of the full dimension. The advantage of clustering based on the Euclidean distance of the full dimension load curve is to consider the value of the full time period of the load curve, with the most information. However, the above methods also have shortcoming that they cannot fully guarantee the similarity of time series shapes as the Euclidean distance only represents the similarity of geometric mean distance [23]. In addition, the above method is easily affected by noise and peak value. Some other indices have been extracted to represent the load shape information, such as the rate of load and peak valley ratio, peak load rate, valley load rate [18],[24]. The same type of users should have similar load shape indexes. But the load shape indices are the reduction of the dimensionality of the original curve. Clustering the load curve directly according to the load shape indices has much information loss [18].

In order to improve the deficiency of the whole dimension load curve clustering, this paper proposes a two-stage clustering algorithm combining the whole dimension load curve and load shape indices. Firstly, clustering results are obtained using classical clustering methods. Clustering method and number are selected via the clustering evaluation index. Many indices are proposed to evaluate the clustering result in the previous paper, such as mean index adequacy (MIA), clustering dispersion indicator (CDI), similarity matrix indicator (SMI), *Davies–Bouldin index* (DBI), within cluster sum of squares to between cluster variation (WCBCR) and so on [7],[18],[19],[24]. In the paper , MIA is used to evaluate the clustering performance and clustering number is selected by finding the ''knee'' of the MIA curve [14], [15]. Secondly, customers are reclassified using supervised learning algorithm based on load shape indices. In the study, daily load data of 10kV distribution network in a Chinese Southeast coastal city are utilized for load clustering. Different supervised learning

This work was partially supported by National High-Technology Research and Development Program (863 Program) of China (2014AA051901), and China Postdoctoral Science Foundation (2015M580097).

B. Peng, S. Dong and Y. Song are with College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 21410170@zju.edu.cn, yhsongcn@zju.edu.cn).

C. Wan is with Department of Electrical Engineering, Tsinghua University, Beijing 10084, China, and also with Department of Electrical Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: can.wan.hk@ieee.org).

J. Lin is with Department of Electrical Engineering, Tsinghua University, Beijing 10084, China.

Y. Zhang is with State Grid Fujian Electric Power Research Institute, Fuzhou 350007, China.

J. Xiong is with State Grid Xiamen Electric Power Supply Company, Xiamen 361000, China.

978-1-5090-4075-9/16/$31.00 ©2016 IEEE

algorithms including SVM, random forest and K nearest neighbors (KNN) are utilized to improve the performance of load shape indices and the loss of the original curve information. Case studies demonstrate that SVM has better performance at the second stage. The computation efficiency of the reclassification method is also investigated at the second stage. In general, the proposed method can improve the full dimensions load curve clustering in load shape similarity.

This remainder of the paper is organized as follows: The mathematical backgrounds are introduced in Section II .The basic idea of the algorithm is presented in Section III. In Section IV, case studies are conducted and analyzed. Finally, the conclusion is given in Section V.

II. Mathematical Background

A. Clustering Method

1) K-means Clustering Algorithm

K-means algorithm is a partition based clustering algorithm [25]. Firstly, *k* points are chosen as the clustering centers; then samples close to the center point are classified to it. The iteration are repeated until the objective function converges or to the maximum number of iterations.

The objective function of *K*-means algorithm is defined as，

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d^2(x_j, c_i) \tag{1}$$

where $d^2(x_j, c_i)$ is the Euclidean distance from the sample point to the cluster center; k is the number of clusters; $n_i$is the number of samples in cluster i; c_i is the cluster center of the cluster i.

2) Fuzzy c-Mean Algorithm (FCM)

The fuzzy *c*-means is similar to *K*-means algorithm, and the final clustering results are obtained by iteration [26]. Different from *K*-means algorithm, each of the samples belong to different cluster centers through the membership degree, rather than only the slave of only one center. The objective function of Fuzzy *c*-mean algorithm is defined as,

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} u_{ij}^m d^2(x_j, c_i) \tag{2}$$

where u_{ij} is the membership degree of x_j for cluster i, m is the fuzzy index, u_{ij} satisfies the following constraint,

$$\sum_{i=1}^{k} u_{ij} = 1 \tag{3}$$

3) Ward Algorithm

The Ward algorithm is a hierarchical clustering method [27]. First each sample is regarded as a separate class, and then samples are combined step by step. Each merge chooses two clusters that make the minimum increasing of the sum of squares of deviations.

B. Cluster Evaluation Index.

Commonly used load clustering evaluation indicators are MIA, clustering dispersion indicator (CDI), SMI, DBI, WCBCR and so on [18],[19],[24], this paper uses the MIA index to evaluate the performance of clustering. In this paper, the distance between two vectors is defined as,

$$d(l^{(i)}, l^{(j)}) = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (l_t^{(i)} - l_t^{(j)})^2} \tag{4}$$

where $l^{(i)}$ is the feature vector of the ith user, $l_t^{(i)}$ is the tth element of $l^{(i)}$. T is the number of time periods. The MIA index is defined as,

$$MIA = \sqrt{\frac{1}{K} \sum_{j=1}^{K} (\sum_{l \in \Omega_j} d^2(\omega_j, l) / N_j)} \tag{5}$$

where Ω_j is the set of all load of class j, ω_j is the average of the vector of class j, N_j is the total number of class j. Smaller MIA value shows better clustering effect.

C. Classical Supervised Learning Algorithm

1) Classification and Regression Tree (CART)

CART is a kind of decision tree algorithm with two fork tree [28]. Each non leaf node divides the current sample into two subsets according to some attribute. Gini index is used to choose the attribute, defined as,

$$Gini(\mathrm{D}) = 1 - \sum_{j=1}^{m} p_j^2 \tag{6}$$

where m is the number of clusters for the data set **D**, and p_j is the percentage of cluster j. The messier the classification is, the greater the Gini index is. For each node, the origin set **D** is divided into two sets $\mathbf{D_1}$ and $\mathbf{D_2}$ according to each attribute of the sample. Then the change of the Gini index can be obtained and expressed as,

$$Gini(\mathbf{D}) - p*Gini(\mathbf{D1}) - q*Gini(\mathbf{D2}) \tag{7}$$

where p is percentage that **D1** account for **D**, q is the percentage that **D2** account for **D**. The purpose of the attempt to divide the node using each attribute is to find a partition, which is the biggest of Gini index change. Then the attribute value is used as the optimal branch.

2) Random Forest

Random forest algorithm is an ensemble learning algorithm [29] . It is the combination of many CARTs. The training samples for each tree are independent of each other. For each tree, it uses a training set that is sampled from the total training set with replacement. For each tree, about 37% of the original samples are not used. Training features for each tree are selected from the origin features randomly for a certain proportion. The final result of the random forest is,

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \tag{8}$$

where, $H(x)$ represents the combination model; $h_i()$ is a single decision tree model; Y is the output variable; I is the indicator function, which equals to 1 when $h_i(x) = Y$ and to 0 otherwise.

3) Support Vector Machine

Support vector machine is a supervised learning model that analyzes data used for classification and regression analysis [30]. The learning strategy of support vector machine is to find the optimal hyper plane to maximize the classification interval. Consider a linear classifier function,

$$\mathrm{f}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b \tag{9}$$

where $\boldsymbol{w}$ and b are parameters; $\boldsymbol{x}$ is the input feature vector.

The optimization process is formulated as,

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{10}$$

$$\text{s.t. } y_i f(x_i) \geq 1-\xi_i, \xi_i \geq 0, i = 1,...,n \tag{11}$$

Where C is used to control the trade-off between the model complexity (first term) and empirical risk (second term). The slack variables ξ_i is the misclassification penalties for the samples of the training set that violate the constraints. For the low dimensional non separable problem, the SVM maps the sample space x to the high dimension space by the nonlinear transformation $\varphi(x)$ to make it linearly separable in the high dimension space. So it introduces the kernel function to solve the nonlinear classification problem.

4) K Nearest Neighbors

KNN is a simple and effective supervised learning algorithm [31]. It first calculates the similarity between the training samples and the sample to be predicted. In this paper, Euclidean distance is used as the measure of similarity. The Euclidean distance of vector $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as,

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{t=1}^{N}(x_t - y_t)^2} \tag{12}.$$

Then find the most similar K samples according to the distance and use the mode of the K samples to be the predicting class.

III. Two-stage Clustering Algorithm

A. Load Shape Indices

The advantage of clustering based on the Euclidean distance of the full dimension load curve is to consider the values of the load curve under the full time period, with the most information. However, clustering based on load curves of the full dimension curves cannot fully guarantee the similarity of time series shapes, as the Euclidean distance only represents the similarity of geometric mean distance. Actually, not every point of the load curves is important in practice. Load shape indices can be derived from the load curve data to represent the load shape information and the key information that people care, such as peak load value, valley time value and load level at different system load pressure time: peak load time, valley load time, flat load time. Using the comprehensive load shape indices in clustering would be meaningful to obtain the key information of load curves. In this paper, five load shape indices are defined.

Load factor can represent the peak load information and the whole load curve shape, which is defined as,

$$I_{\mathrm{LF}} = \frac{P_{av}}{P_{\max}} \tag{13}$$

where P_{av} is the average load of all the day, $P_{\max}$ is the maximum load of all the day.

Peak valley factor that can denote variation interval of the load and is defined by,

$$I_{\mathrm{PV}} = \frac{P_{\max} - P_{\min}}{P_{\max}} \tag{14}$$

where $P_{\min}$ is the minimum load of all the day.

Peak load factor represents the average load level at system peak time, and can be defined as,

$$I_{\mathrm{P}} = \frac{P_{av.peak}}{P_{av}} \tag{15}$$

where $P_{av.peak}$ is the average load during system's peak load time. In the study, 9:00-17:00 is the system's peak load time.

Flat load factor expresses the average load level at system flat time, represented by,

$$I_{\mathrm{F}} = \frac{P_{av.flat}}{P_{av}} \tag{16}$$

where $P_{av.flat}$ is the average load during system's flat load time. In the study, 6:00-9:00 and 17:00-21:00 are the system's flat load times.

Valley load factor represents the average load level at system valley time, which is defined by,

$$I_{\mathrm{V}} = \frac{P_{av.valley}}{P_{av}} \tag{17}$$

where $P_{av.valley}$ is the average load during system's flat load time. In the study, 0:00-6:00 and 21:00-24:00 are the system's valley load times.

B. Algorithm Introduction

The flow chart of the proposed two-stage algorithm is given in Fig. 1. Firstly, clustering results are obtained using clustering methods with different clustering nembers. Clustering method and number to get the result of first-stage are selected via the clustering evaluation index. Secondly, customers are reclassified using supervised learning algorithms based on load shape indices.

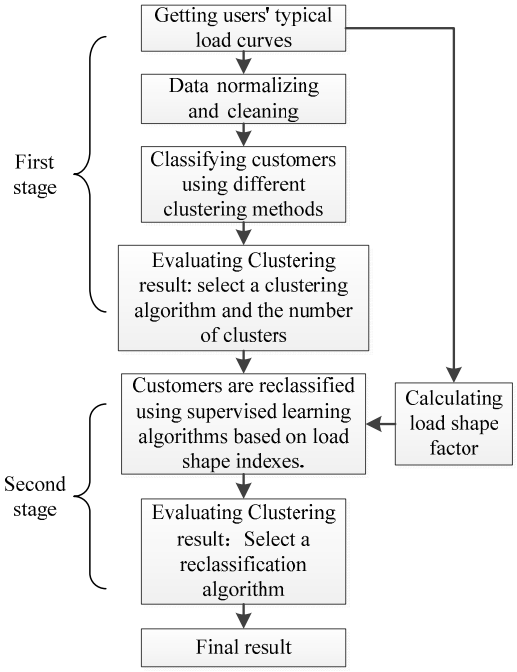

Fig.1 The flow chart of the proposed two-stage clustering algorithm for electric load.

1) First Stage

Typical load curve should be selected before clustering, which can be derived from the average value of the daily load curves for a period of time, or load curve at a typical day, such as, peak load day. The power consumption magnitudes of customers have great differences. Clustering customers should be on basis of the load shapes, not the absolute value. So load curves should be normalized before using clustering methods. Then different clustering algorithms with different clustering numbers are employed to cluster the full dimensions of the load curve. The clustering algorithm and the number of clusters used to obtain the final result are selected through the load clustering evaluation indices.

2) Second Stage

The basic idea of the reclassification of load is that if a given sample has low similarity with the samples of the same cluster, but has low similarity with the samples of a different cluster, then this sample should be reclassified to the cluster with high similarity. The reclassification algorithm is expressed as the following four steps,

1. Select the *i*th sample (x_i, y_i) ,where x_i is the input feature vector (load shape indices), y_i is the class the sample belongs to at the first stage.
2. Use all the other loads without the *i*th sample as the training sample $\{(x_k, y_k)\}_{k=1,k\neq i}^{n}$ to train a classifier.
3. Reclassify the *i*th load is to the new class using the classifier in step 2.
4. Repeat the reclassification until all the load are reclassified

IV. Case Study

A. Description of Experiments Data

In the study, daily load data of 10kV distribution network in a Chinese Southeast coastal city are utilized. Electric daily loads with 96 points of 2782 customers in a summer day are collected as the origin typical load curves. It should be emphasized that the dataset is much larger than previous literature and contains various kinds of customers including commercial, municipal, governmental, industrial, residential, institutional customers and even the street lamp.

B. Numerical Results and Analysis

1) First Stage

In this paper, we first use the full dimensional load curve clustering method to obtain preliminary results. Load curve data with missing value are eliminated and, 2672 load curves are remained. Each load curve is normalized using the peak load as the reference. Then *K*-means, FCM and Ward method are used to cluster the load curves with the cluster number from 4 to 25. MIA indicators of the clustering results of the three algorithms are shown in Fig. 2. As the clustering number grows to 8, the decreasing of the MIA becomes slow. So the *K*-means clustering algorithm with clustering number 8 is the knee of all the clustering results. In this paper, we choose the first stage result using *K*-means clustering algorithm with clustering number 8. The boxplots of the clustering results are shown in Fig. 3.The *x* axis denotes the number of the time period of the daily load curves. The *y* axis denotes the normalized load. It is the same with other load figures.

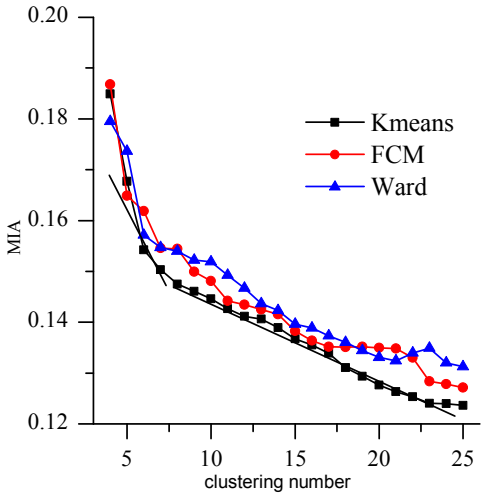

Fig. 2 Relationship between MIA index and number of clusters.

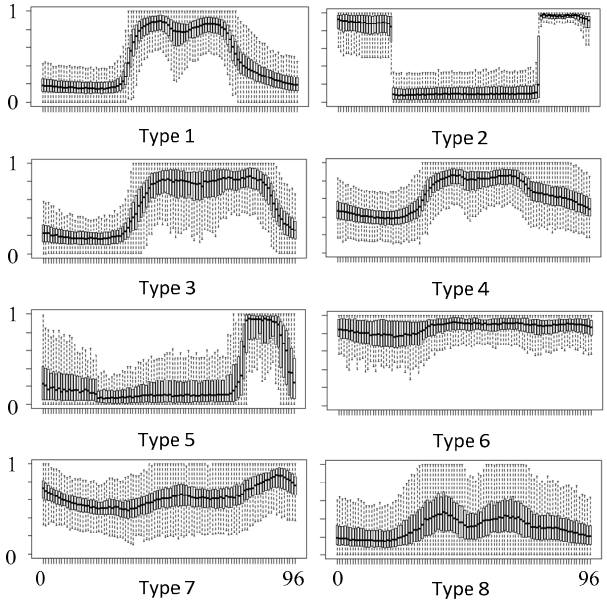

Fig.3 Boxplots of first-stage clustering.

2) Second Stage

In this paper, the decision tree, random forest, SVM and KNN algorithm are used to carry out the second stage classification. For each sample, the left samples are used to train the classifier. Given the sample number N, to complete the reclassification of all the samples, the training progress of the classifier should be implemented for N times. But for random forest algorithm, due to the repeatable sampling, for each tree, about 37% of origin sample are not used. In the study, all the samples are used to train a random forest and each sample is predicted using the trees whose training set do not contain this sample. It is the same meaning as the proposed method. Only single training process of random forest is needed to complete the second stage reclassification. KNN is an inert learning method, which does not need pre

training. SVM is chosen as the second stage reclassification algorithm due to its reliable performance in classification. 241 loads are reclassified at the second stage. Figs. 4 and 5 show the load curves of type 1 reclassified to type 3 and type 4 (type 1 only reclassified to these two types). Load in Fig. 3 has evening load, so it is more reasonable to be classified to type 3. Load in Fig. 4 has night load, so it is more reasonable to be classified to type 4. To consider load shape indices, these customers can be reclassified more reasonable than the origin clustering results.

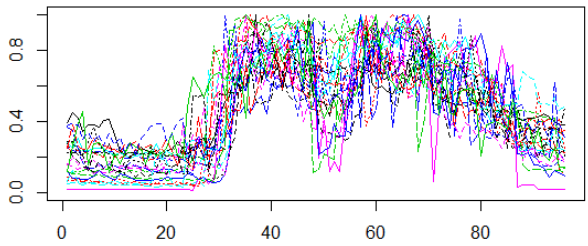

Fig.4 Load curves of type 1 reclassified to type 3

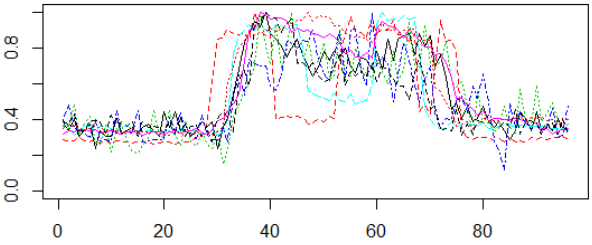

Fig.5 Load curves of type 1 reclassified to type 4

C. Comparison of Cluster Evaluation Indices

The MIA value, defined in (5), of the load shape indices is denoted as MIA1.The MIA value of the load curves is denoted as MIA2. MIA indices of load clustering only based on load curves and different supervised learning algorithms for the second stage load curve clustering are shown in Table I. It can be seen from Table I that the classic *K*-means method has the highest MIA1 with 0.07707 and the lowest MIA2 with 0.1507. After the second stage, all the performance of load shape indices are better than the origin method, but the performance of the load curves are slightly worse. Considering both the load shape index and the whole dimension load curve information, the final clustering results can be evaluated by MIA1*MIA2. The product is not as sensitive to the order of magnitude difference of different variables as summing. It can be seen from the Table I that the utilization of SVM for the two classification performs the best with the lowest MIA1*MIA2 of 0.01120, followed by KNN and random forest. It can be concluded that the proposed two-stage algorithm can improve the classification of load shape similarity in the clustering progress.

TABLE I

COMPARISON OF MIA INDEXES

Algorithm	MIA1	MIA2	MIA1*MIA2
K-means	0.07707	0.1507	0.01161
Random Forest	0.07488	0.1518	0.01137
SVM	0.07345	0.1525	0.01120
kNN	0.07364	0.1541	0.01135
CART	0.07430	0.1552	0.01153

D. Comparison of Computing Time

The computing times of different supervised learning algorithms under different sample sizes are compared in Table II. The computation time of SVM algorithm is the longest at each sample size, 1310 times longer than random forest and 19 times longer than KNN when sample size is 2500. Random forest algorithm uses out of bag data and it requires only one training time. Random forests have obvious advantages in the computational efficiency, about 85% shorter than the second shortest algorithm KNN.

TABLE II

COMPUTING TIME FOR DIFFERENT SUPERVISED LEARNING ALGORITHMS

Reclassification method	Sample size				
	500	1000	1500	2000	2500
Random Forest	0.29	0.55	0.91	1.34	1.62
SVM	147.87	400.23	766.30	1341.53	2122.91
KNN	1.06	3.28	5.45	7.96	11.52
CART	4.13	16.07	35.47	60.81	106.26

V. CONCLUSION

This paper presents a novel two-stage clustering method for classification of electric load curve. Firstly, clustering methods are based on the Euclidean distance of the load curve to get the initial results, and then supervised learning algorithm are used to reclassify the load based on comprehensive load shape factors defined in this paper. Different supervised learning algorithms are compared in improving the performance of load shape indices and the loss of the original curve information. Case studies demonstrate that SVM has better performance at the second stage. The computation efficiency of the reclassification method is also investigated at the second stage. SVM consumes the longest time, and random forest is the fastest. In general, the proposed method can improve the full dimensions load curve clustering in load shape similarity. Load can be reclassified more reasonable than the classic methods.

REFERENCES

[1] Q. Li, Z. Xu, and L. Yang, "Recent advancements on the development of microgrids," *J. Mod. Power Syst. Clean Energy*, vol. 2, no. 3, pp. 206-211, Sep. 2014.

[2] I. D. d. C. Mendaza, I. G. Szczesny, J. R. Pillai, and B. Bak-Jensen, "Flexible demand control to enhance the dynamic operation of low voltage networks," *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 705-715, Mar. 2015.

[3] Y. Liang, L. He, X. Cao, and Z. J. Shen, "Stochastic control for smart grid users with flexible demand," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2296-2308, Dec. 2013.

[4] C. Wan, Z. Xu, Y. L. Wang, Z.Y. Dong, and K.P. Wong, "A hybrid approach for probabilistic forecasting of electricity price," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 463-470, Jan. 2014.

[5] C. Wan, M. Niu, Y. Song, and Z. Xu, "Pareto Optimal Prediction Intervals of Electricity Price," *IEEE Trans. Power Syst.*, vol. PP, no.99, pp.1-3, 2016.

[6] C. Wan, Z. Xu, P. Pinson, Z.Y. Dong, and K.P. Wong, "Probabilistic forecasting of wind power generation using extreme learning machine," *IEEE Trans. Power Syst.*,vol.29, no.3, pp.1033-1044, May 2014.

[7] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 381–387, Feb. 2003.

[8] F. L. Quilumba, W.-J. 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 Trans. Smart Grid, vol. 6, no. 2, pp. 911–918, Mar. 2015.*

[9] M. Espinoza, C. Joye, R. Belmans, and B. DeMoor, "Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series," *IEEE Trans. Power Syst.*,vol. 20, no. 3, pp. 1622–1630, Aug. 2005.

[10] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 411-419, Jan. 2014.

[11] R. F. Chang and C. N. Lu, "Load profiling and its applications in power market," in Proc. *IEEE Power Eng. Soc. General Meeting*, Jul. 13–17, 2003, vol. 2.

[12] S. Valero, M. Ortiz, C. Senabre, C. Alvarez, F. J. G. Franco, and A. Gabaldon, "Methods for customer and demand response policies selection in new electricity markets," *IET Gen., Transm., Distrib.*, vol. 1,no. 1, pp. 104–110, 2007.

[13] H. T. Roh and J. W. Lee, "Residential demand response scheduling with multiclass appliances in the smart grid," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 94-104, Jan. 2016.

[14] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers,"*IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1120–1128, Aug.2007.

[15] N. M. Kohan, M. P. Moghaddam, S. M. Bidaki, and G. R. Yousefi, "Comparison of modified k-means and hierarchical algorithms in customers load curves clustering for designing suitable tariffs in electricity market," in *Proc. 43rd Int. Universities Power Engineering Conf.*, Padova, Italy, Sep. 1–4, 2008, pp. 1–5.

[16] Z. Zakaria, K. L. Lo, and M. H. Sohod, "Application of fuzzy clustering to determine electricity consumers' load profiles," in *Proc. IEEE Int. Power and Energy Conf.*, Putra Jaya, Malaysia, Nov. 28–29, 2006, pp.99–103.

[17] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Jarventausta, " Enhanced load profiling for residential customers," *IEEE Trans. Power Del.*, vol. 29, no. 1, pp. 88–96, Feb. 2014.

[18] G. Chicco, R. Napoli, F. Piglione, M. Scutariu, P. Postolache, and C. Toader, "Emergent electricity customer classification," *Proc. Inst.Elect. Eng., Gen., Transm., Distrib.*, vol. 152, no. 2, pp. 164–172, Mar.2005.

[19] 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.

[20] S. V. Verdu, M. O. Garcia, C. Senabre, A. G. Marin, and F. J. G. Franco, "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *IEEE Trans. Power Syst.*, vol. 21, no. 4, pp. 1672–1682, Nov. 2006.

[21] G. Chicco and I. S. Ilie, "Support vector clustering of electrical load pattern data," *IEEE Trans. Power Syst.*, vol. 24, no. 3, pp. 1619–1628,Aug. 2009.

[22] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp.1162–1171, Apr. 2010.

[23] M Piao，HS Shon，JY Lee，KH Ryu, "Subspace Projection Method Based Clustering Analysis in Load Profiling," *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 2628-2635, Nov. 2014.

[24] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 596–602, May 2005.

[25] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. Roy. Statist. Soc. Series C (Appl. Statist.)*, vol. 28, no. 1, pp. 100–108, 1979.

[26] J. C. Bezdec, *Pattern Recognition With Fuzzy Objective Function Algorithms*.New York: Plenum, 1981.

[27] Ward, J.H.: 'Hierarchical grouping to optimise an objective function', *J. Am. Stat. Assoc*., 1963, 58, pp. 236–244.

[28] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees.* Belmont, California: Wadsworth, 1984.

[29] L. Breiman, "Random forests," *Mach. Learn*., vol. 45, no. 1, pp. 5–32, 2001

[30] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support VectorMachines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[31] T.M. Cover, and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transaction on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.