Send Orders for Reprints to reprints@benthamscience.ae

The Open Cybernetics & Systemics Journal, 2015, *9,* 1099-1106

Research on the Influence of Several Factors on Clustering Results using Different Algorithms

Zhang Tiefeng* and Gu Mingdi

School of Electric and Electronic Engineering, North China Electric Power University, Baoding, Hebei, 071003, P.R. China

Abstract: Clustering analysis is an important technology in the field of pattern extraction and recognition. In order to find the influence of several factors on clustering results using different algorithms and support the decision for power load pattern extraction using clustering techniques, this paper develops the research on the influence of several factors on clustering results using different algorithms. In this paper, three data sets, five normalization methods and five known clustering algorithms including k-means, FCM, SOM, hierarchical clustering and spectral clustering are used, four experiments are designed and performed, they are the influence of the normalization methods on the clustering results, the dependence of clustering results on a data set, the algorithm stability and the sensitivity of clustering algorithm to the input order of the data. The results show that all the factors have obvious influence on the clustering results, and using maximum normalization and FCM algorithm in clustering procedure has the best performance for power load pattern extraction.

Keywords: Clustering analysis, FCM, normalization method, power load.

1. INTRODUCTION

Clustering analysis is one of the most important research topics in the field of pattern extraction and recognition [1]. At present, the research on the method of load pattern extraction has attracted much attention all over world, and various clustering algorithms are developed and applied to the power load pattern extraction [2]. In the power system, the load curve of power consumption data is collected by load measurement equipment. Load curve shows the electricity consumption behavior of all kinds of industry users, which has laid the foundation for mining useful information for electric power company through analysis and access the load pattern, and it is significant for power system dispatching and planning. The load pattern extraction and recognition of power users can provide theoretical basis for load forecasting [3], load control [4], abnormal electricity consumption detection [5-6] and designing electricity tariff offers, etc. The load pattern is extracted from the load curves by using the clustering algorithm, and the clustering center curve is the typical load profile. In the last few years, a lot of literatures mainly focus on the research of clustering algorithm. In [7], the k-means, hierarchical clustering, SOM and FCM, four basic ideas of the traditional clustering algorithm are introduced, and the Iris flower data set is used for the performance comparison. In [8], a new method for the clustering and pattern recognition of multivariate time series (CPT-M) based on multivariate statistics is presented. The algorithm comprises four steps that extract essential features of multivariate time series of residential users with emphasis on seasonal and temporal profile, among others. The method is successfully implemented and tested in the context of an energy efficiency program carried out by the Electric Company of Alagoas (Brazil). In [9], an improved global k-means clustering algorithm is proposed by presenting a novel method of generating the next optimal initial center with the enlightening of the idea of k-medoids clustering algorithm suggested by Parketal. In [10], a new weighted fuzzy C-Means (NW-FCM) algorithm is proposed to improve the performance of both FCM and FWCM models for high-dimensional multi-class pattern recognition problems. In [11], a FCM clustering algorithm based on attribute reduction is proposed, which has good performance in dealing with big data sets. In [12], several mature clustering algorithms are studied deeply and the advantages and disadvantages of these algorithms are summarized as well as the use scope.

However, clustering results are influenced by various factors in the steps of clustering procedure. Clustering algorithm is only one of the main factors, but few literatures mentioned other factors. In [13], when using FCM clustering algorithm to extract the typical load profile, the optimal cluster number is determined by DB index, and the accuracy of the method is proved. It is not enough to consider just that. In [14], the influence of using different data preprocessing methods on fuzzy c-means clustering results is analyzed. The results show that the sum normalization and maximum normalization are the best in the average accuracy. However, there may be a big difference in the results of each normalization method for different clustering algorithms. In [15], the influence of normalization methods on the clustering results is studied in clustering load curves, and the matching relations between normalization methods and clustering algorithms are obtained, but other factors are not mentioned. In

1874-110X/15

[16], the advantages and disadvantages of several typical clustering algorithms are discussed by comparing the clustering results on different data sets. The results show there is big difference in both sides of accuracy and efficiency when one algorithm with different datasets is analyzed by comparing with the same clustering of the data set under different algorithms. In [17], several distances are used in the k-means algorithm for clustering load curves and their influences on the clustering results are analyzed, the results indicate that the choice of distances is an important factor in power load pattern extraction using clustering techniques and those distances based Euclidean distance have relative higher accuracy than others, so we used Euclidean distance in this paper. Moreover, for some clustering algorithms, the input order of the data has a certain effect on the clustering results [18], which must be taken into account.

In conclusion, the clustering results are influenced by various factors in the steps of clustering procedure. Therefore, research on the influence of these factors on the clustering results is an essential task when we use clustering techniques for power load pattern extraction. So this paper developed the research on the influence of several factors on clustering results using different algorithms. In this paper, three data sets, five normalization methods and five known clustering algorithms including k-means, FCM, SOM, hierarchical clustering and spectral clustering are used, four experiments are designed and performed, they are the influence of the normalization methods on the clustering results, the dependence of clustering results on a data set, the algorithm stability and the sensitivity of the algorithm to the input order of the data.

This paper is organized as follows. Section 2 introduces the clustering procedure, describes the steps and presents the normalization methods. Section 3 first presents the data sets used in the paper, then introduces the design of experiment, finally gives the clustering results and analysis. Conclusions are presented in Section 4.

2. CLUSTER ANALYSIS THEORY AND METHODS

2.1. The Process of Clustering Analysis

The process of clustering data using clustering algorithms mainly includes the five steps as follows.

The first step: data selection. Get the data set and determine the data type. The data may be obtained from different heterogeneous data sources. Thus, the first step is to obtain data from various databases, files and non-electronic data sources. For example, a preliminary data selection of customers can be carried out by geographical region and voltage level (high, medium, and low). The daily chronological load curves for each individual customer are determined for each study period (month, season, and year).

The second step: data cleaning. Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

The third step: data preprocessing. Take load curves as an example, clustering load curves is based on the shape of a load curve but not by absolute MW values, so the data should be normalized, that is, scaled to a specific range. Normalization is particularly useful for clustering algorithms involving distance measurements. There are many methods for data normalization, such as min-max normalization, z-score normalization, and normalization by Mean-variance [15].

The fourth step: data clustering. Several clustering algorithms are used to cluster the normalized data. The choice of the clustering method in the step is very important to the whole process. The clustering methods used in this paper are: k-means, FCM, SOM [19], hierarchical clustering [20] and spectral clustering [21].

The fifth step: clustering analysis and evaluation. The clustering results obtained from the previous step are analyzed. For a given number of clusters, the composition and number of data in each cluster are analyzed. In power load pattern extraction, we can identify the TLPs of each customer by analyzing the distribution of load curves in the load patterns. Research in this area has proposed many different indicators [22, 23], such as the mean index adequacy (MIA), the clustering dispersion indicator (CDI), the similarity matrix indicator (SMI), the Davies-Bouldin indicator (DBI), the modified Dunn index, the scatter index (SI), and the mean square error [24]. Many studies [25] on clustering illustrate the applications and compare the results obtained by various unsupervised clustering algorithms based on these adequacy measures.

In this paper, the experiment mainly focused on the methods of normalization in the third step and other three factors influencing the performances of clustering algorithm in the fourth step, the distance measurement is based Euclidean distance.

The process of clustering analysis is shown in Fig. (**1**).

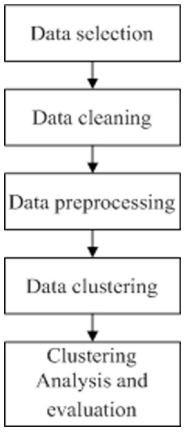


Fig. (1). The flow chart of clustering analysis.

2.2. Normalization Methods

Normalization is particularly useful for clustering algorithms involving neural networks, or distance measurements such as nearest-neighbor classification and clustering. Especially for distance-based methods, normalization helps prevent attributes with initially large ranges (*e.g.*, income) from

outweighing attributes with initially smaller ranges (*e.g.*, binary attributes) [18]. An attribute is normalized by scaling its values so that they fall within a small specified range, such as 0.0 to 1.0. Thus we must normalize the data set before clustering data to limit the sample data to a certain range, this is not only convenient for data processing, but also improve the convergence rate to shorten the running time of clustering. The normalization methods used in this paper are: the min-max, the mean-variance, the sum normalization, the z-score and the maximum normalization.

If a sample data set is *X*, the number is *n*, dimension is *p*.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{bmatrix}$$

where x_{ij} represents the j dimensional data of the i sample data.

The transformation is expressed as follows.

(1) Min-max normalization. The method is a linear transformation of the sample data. Suppose the k sample data is $x_k = (x_{k1}, x_{k2}, \cdots, x_{kp})$, the map f is shown as (1).

$$x_{kp} \to f(x_{kp}) = \frac{x_{kp} - x_{k\min}}{x_{k\max} - x_{k\min}} \tag{1}$$

where $x_{k\max} = \max(x_{k1}, x_{k2}, ..., x_{kp})$, $x_{k\min} = \min(x_{k1}, x_{k2}, ..., x_{kp})$, $k = 1,2,\cdots,n$. After min-max normalization, the value of original sample data set will be scaled to [0,1].

(2) Mean-variance normalization. Suppose the k sample data is $x_k = (x_{k1}, x_{k2}, ..., x_{kp})$, the map f is shown as (2).

$$x_{kp} \to f(x_{kp}) = \frac{x_{kp} - x_{kmean}}{x_{k\,\mathrm{var}}} \tag{2}$$

where

$x_{kmean} = mean(x_{k1}, x_{k2}, ..., x_{kp})$, $x_{k\,\mathrm{var}} = \mathrm{var}(x_{k1}, x_{k2}, ..., x_{kp})$, $k = 1,2,\cdots,n$. After mean-variance normalization, the value of original sample data set will be scaled to [-1,1].

(3) Sum normalization. It is to use various attribute values of each sample divided by the sum of all the data of the original data set. The equation of sum normalization is shown as (3).

$$x'_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \tag{3}$$

$$\sum_{i=1}^{m} x'_{ij} = 1 \tag{4}$$

where x_{ij} is the original data, x'_{ij} is the data after sum normalization, $i = 1,2,\cdots,n$, $j = 1,2,\cdots,p$.

(4) Z-score normalization. In this method, the attribute values of the sample data set are normalized based on the mean and standard deviation of x_j. The equation of z-score normalization is shown as (5).

$$x'_{ij} = \frac{x_{ij} - \overline{x}_j}{S_j} \tag{5}$$

where x_{ij} is the original data, x'_{ij} is the data after z-score normalization, S_j is the standard deviation of the sample data. $i = 1,2,\cdots,n$, $j = 1,2,\cdots,p$. After z-score normalization, the new value satisfies that the average value is 0 and the standard deviation value is 1.

$$\overline{x}'_j = \frac{1}{m}\sum_{i=1}^{m} x'_{ij} = 0 \tag{6}$$

$$S'_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m} (x'_{ij} - \overline{x}'_{ij})^2} = 1 \tag{7}$$

(5) Maximum normalization is to use each value of sample data set divided by the maximum of the sample data set. The equation of maximum normalization is shown as (8).

$$x'_{ij} = \frac{x_{ij}}{\max_i \{x_{ij}\}} \tag{8}$$

where $i = 1,2,\cdots,n$, $j = 1,2,\cdots,p$. For the new value, the maximum value is 1, other is less than 1.

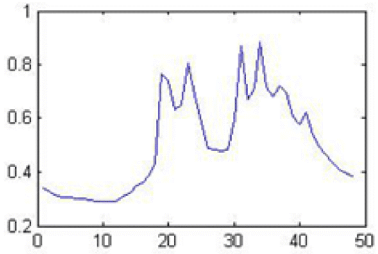
3. EXPERIMENT AND RESULT ANALYSIS

3.1. Data Set

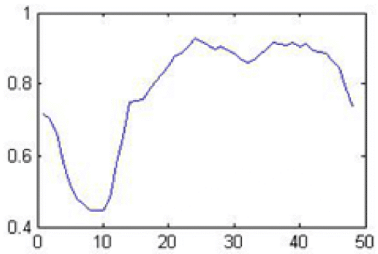
There are three different data sets used in this experiment. The data set 1 is IRIS data set from UCI database. This data set contains 150 samples, which is about three plants of Setosa, Versicolor and Virginica, and each sample contains four attributes of the sepal length, sepal width, petal length, and petal width (Unit: cm).

Data set 2 contains 147 samples, which is the actual power consumption from five kind of industry users, these daily load curves are obtained by an automatic meter reading system with time periods in steps of 30 min. Five typical load profiles generated by these load curves represent five load patterns, and it is identified that each load data are strictly belong to a certain industry, so the data set can be used as standard data set. The typical profiles can be obtained by computing centroid of each cluster with maximum normalization, they are shown in Fig. (**2**) (vertical axis: load, unit: kW; horizontal axis: time, unit: 30-min).

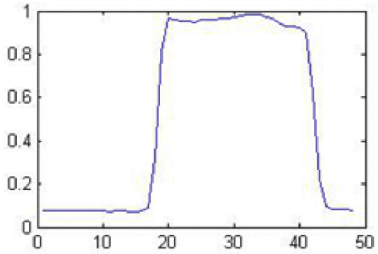
Data set 3 contains 138 samples, which is the actual power consumption from five kind of industry users, these daily load curves are obtained by an automatic meter reading system with time periods in steps of 30 min. Five typical



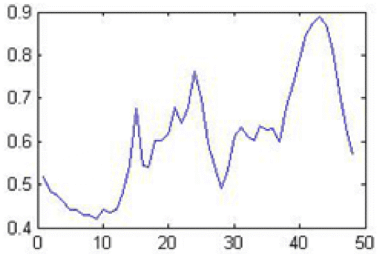
(2-1) Electronics



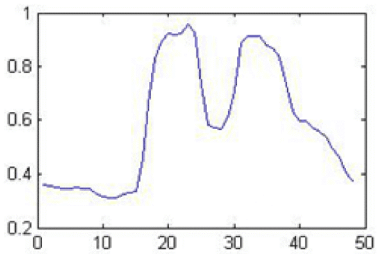
(2-2) Food



(2-3) Business



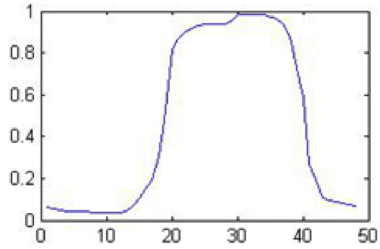
(2-4) Smelting



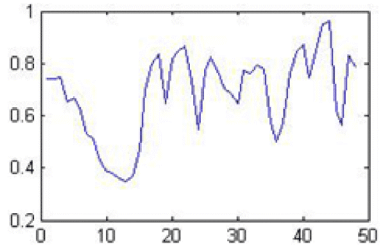
(2-5) Textile

Fig. (2). Typical load profiles of 5 industries.

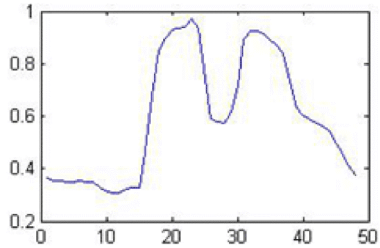
load profiles generated by these load curves represent five load patterns, and it is identified that each load data are strictly belong to a certain industry, so the data set can be used as standard data set. The typical profiles can be obtained by computing the centroid of each cluster with maximum normalization, they are shown in Fig. (**3**) (vertical axis: load, unit: kW; horizontal axis: time, unit: 30-min).



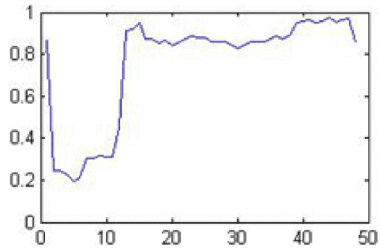
(3-1) Business



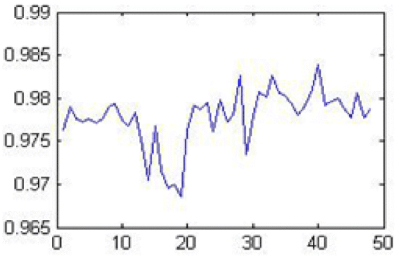
(3-2) Machining



(3-3) Textile



(3-4) Utilities



(3-5) Chemical

Fig. (3). Typical load profiles of 5 industries.

3.2. Experimental Design and Analysis of Results

(1) The influence of the normalization methods on the clustering results

Because of the different nature of the electric power users, the power consumption data values may vary greatly, and the order of magnitude may be different, thus making the clustering results unreliable. The relevant literature [18] also points out that the clustering time and clustering accu-

Table 1. Clustering results of k-means using five normalization methods.

Normalization Method	The Number of Clustering Error	The CPU Time(s)	Accuracy (%)
Maximum	13	0.26s	90%
Min-max	21	0.16s	85%
Z-score	30	0.23s	80%
Sum	17	0.18s	87%
Mean-variance	39	0.24s	67%

Table 2. Clustering results of FCM using five normalization methods.

Normalization Method	The Number of Clustering Error	The CPU Time(s)	Accuracy (%)
Maximum	0	0.39s	100%
Min-max	0	0.38s	100%
Z-score	35	0.39s	77%
Sum	0	0.40s	100%
Mean-variance	33	0.34s	75%

Table 3. Clustering results of SOM using five normalization methods.

Normalization Method	The Number of Clustering Error	The CPU Time(s)	Accuracy (%)
Maximum	31	14.20s	80%
Min-max	31	14.26s	80%
Z-score	4	14.49s	97%
Sum	31	14.66s	80%
Mean-variance	31	14.39s	80%

racy of the data before and after normalization have different effects. Accordingly, in this paper we use the min-max, the mean-variance, the sum normalization, the z-score and the maximum normalization to normalize the data, and use five known clustering algorithms including k-means, FCM, SOM, hierarchical clustering and spectral clustering to process the data set 3 for clustering performance analysis and comparison. The experiment compared the clustering performance mainly from the following three indicators: (1) the number of clustering error: The total number of the whole data set of clustering error. (2)The CPU time(Unit: s). (3) Accuracy: $avg = \frac{1}{k}\sum_{i=1}^{k}\frac{m_i}{n_i}$, where k is the clustering number of data sets, n_i is the number of samples in cluster i, m_i is the correct sample number of cluster in cluster i. Because of the instability in the process of clustering algorithm, when running several times the results under the given number is often different. So the data in the table are the average result after running 5 times, similarity measure used is Euclidean distance.

From the above experimental results, we can see the accuracy of k-means clustering algorithm is higher when using maximum normalization and sum normalization. The accuracy of FCM clustering algorithm can achieve 100% when using maximum normalization, the min-max normalization and the sum normalization. For SOM, when using z-score normalization, the accuracy is 97%, but the CPU time is too long to handle big data sets. The hierarchical clustering algorithm has the best performance when using maximum normalization and sum normalization, the accuracy is 100%. The accuracy of spectral clustering is higher when using maximum normalization and sum normalization.

(2) Dependence of clustering results on a data set

Because different data sets have different features, clustering algorithms may have different performance when dealing with different data sets. Therefore, we performed clustering experiment with different data sets for these clustering algorithms. Performance analysis results are shown in Table **6**. (The maximum normalization and Euclidean distance are used).

Table 4. Clustering results of hierarchical clustering using five normalization methods.

Normalization Method	The number of Clustering Error	The CPU Time(s)	Accuracy (%)
Maximum	0	0.05s	100%
Min-max	116	0.03s	17%
Z-score	34	0.09s	78%
Sum	0	0.11s	100%
Mean-variance	116	0.02s	17%

Table 5. Clustering results of spectral clustering using five normalization methods.

Normalization Method	The number of Clustering Error	The CPU Time(s)	Accuracy (%)
Maximum	11	0.39s	92%
Min-max	83	0.15s	43%
Z-score	24	0.19s	82%
Sum	17	0.04s	85%
Mean-variance	103	0.08s	24%

Table 6. Data dependence analysis.

Clustering Algorithm	Data Set 1 IRIS		Data Set 2 Power Load Data 1		Data Set 3 Power Load Data 2	
	The Number of clustering Error	Accuracy (%)	The Number of Clustering Error	Accuracy (%)	The Number of Clustering Error	Accuracy (%)
k-means	4	97%	28	81%	13	90%
FCM	31	70%	0	100%	0	100%
SOM	14	91%	31	80%	31	80%
Hierarchical clustering	51	66%	60	57%	0	100%
Spectral clustering	9	94%	32	79%	11	92%

From Table **6**, it can be seen that the clustering performance of hierarchical clustering is worse for IRIS data set, and data set 2. Although the FCM clustering algorithm has some dependence on the data sets, the accuracy of dealing with the power load data can reach 100%. There are no obvious conclusions for other algorithms.

(3) Algorithm stability analysis

Most clustering algorithms cluster the data set under the given clustering number, and there are often different results after running several times. So the stability analysis of clustering algorithm is very important. Here we use stability index to measure the inconsistency of the performance of the clustering algorithm when dealing with the data sets with different feature components. The stability index [1] is expressed as follows:

The stability index=the number of groups of the load curves by their load patterns of T-run/the given number of clusters

If the number of groups is equal to the number of clusters, that is, the stability index is 1, and the algorithm is stable. For the stability index, the smaller the stability index, the better. In this experiment, T is 5, so the data in the below table is the stability index of 5-run. When the data set 1 is analyzed, the given clustering number is 3, the power load data (including 5 load patterns) are analyzed and the given clustering number is 5. (The following algorithms are normalized by the maximum normalization and the similarity measure using Euclidean distance)

From Table **7**, it can be seen that k-means algorithm is stable in processing IRIS data set, but it is unstable in processing power load data set. The FCM algorithm is unstable in processing IRIS data set, while the clustering results are stable when processing two power load data sets. Although hierarchical clustering is stable in dealing with three data sets, but the accuracy of clustering is too low.

(4) The sensitivity analysis of the algorithm to the input order of the data.

Table 7. Stability analysis of algorithms.

Stability Index	Data Set 1 IRIS	Data Set 2 Power Load Data 1	Data Set 3 Power Load Data 2
k-means	1	2.2	1.2
FCM	1.7	1	1
SOM	1	0.8	0.8
Hierarchical clustering	1	1	1
Spectral clustering	3	1.2	1.4

For any clustering algorithm, if the stability index is 1, we can perform the experiment about the sensitivity analysis of the algorithm to the input order of the data by adjusting the order of data set input for several times to analyze its sensitivity to the input order of the data.

From Table **7**, it can be seen that k-means algorithm is stable when processing the data set 1. So we changed the order of the data set 1 5 times and clustered it 5 times respectively (the maximum normalization used). The results show that the input order of the data has a certain degree of effect on the stability and accuracy of the clustering algorithm, the stability index becomes 1.25, and the average accuracy of clustering becomes 95%. In this paper, the spectral clustering algorithm uses the classical clustering algorithm to clustering the feature vector, and the traditional algorithm is k-means. From table 7, it can be seen that the results of spectral clustering are unstable and sensitive to the input order of the data. At the same time, it can be seen that FCM algorithm is stable when processing power load data. So we changed the order of the power load data set (5 times) and clustered it 5 times respectively (the maximum normalization used), the results show that the FCM algorithm is still stable. The stability index is 1 and the average accuracy of clustering is 100%.

CONCLUSION

In this paper, we study the influence of the normalization methods on the clustering results, the dependence of clustering results on a data set, the algorithm stability and the sensitivity of the algorithm to the input order of the data. In our experiments, we use three data sets, five known algorithms and five normalization methods, and analyze the influence from above four aspects.

Some conclusions are as follows:

The results show that the factors have obvious influence on the clustering results, and each factor has different influence on the clustering results using different algorithms, more influence factors should be considered in the future study.

For load pattern extraction, using the maximum normalization in data preprocessing step and using the FCM algorithm in data clustering step, the clustering results are stable and the accuracy is up to 100%. It also tells us a "good" combination of factors may exist for our clustering analysis.

The influence of the factors on the clustering results can't be ignored, especially for data set, when we use clustering techniques to extract patterns from special data set for application purposes, we need do some new experiments to find the influence of the factors on the clustering results, then make a further decision.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was financially supported by the Natural Science Foundation of Hebei Province (No. F2015502047).

REFERENCES

[1] T.F. Zhang, G.Q. Zhang, J. Lu, X.P. Feng, and W.C. Yang, "A New Index and Classification Approach for Load Pattern Analysis of Large Electricity Customers," *IEEE Transactions. Power Systems*, vol. 27, no. 1, pp. 153-160, Feb. 2012.

[2] F. McLoughlin, A. Duffy and M. Conlon, "A clustering approach to domestic electricity load profile characterization using smart metering data," *Applied Energy*, vol. 141, pp. 190-199, 2015.

[3] H. L. Willis, A. E. Schauer, J. E.D. Northcote-Green, and T. D.Vismor, "Forecasting distribution system loads using curve shape clustering," *IEEE Transactions. Power Application Systems*, vol. PAS-102, no. 4, pp. 893-901, 1983.

[4] J. Nazarko, A. Jurczuk, and W. Zalewski, "ARIMA models in load modeling with clustering approach," in *Proc. IEEE Russia Power Tech*, St. Petersburg, Russia, Jun. 27-30, 2005, pp. 1-6.

[5] A. H. Nizar, Z. Y. Dong, and Y.Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Transactions. Power Systems*, vol. 23, no. 3, pp. 946-955, Aug. 2008.

[6] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions. Power Systems*, vol. 21, no. 2, pp. 933-940, May 2006.

[7] X.P. Feng and T.F. Zhang, "Comparison of four clustering methods," *Microcomputer and Application*, vol. 29, no. 16, pp. 1-3, 2010.

[8] A.M.S. Ferreira, C.H. De Oliveira Fontes, C.A.M.T. Cavalcante and J.E.S. Marambio, "Pattern recognition as a tool to support decision making in the management of the electric sector. Part II: A new method based on clustering of multivariate time series," *International Journal of Electrical Power and Energy Systems*, vol. 67, pp. 613-626, May. 2015.

[9] J.Y. Xie, S. Jiang, C.X. Wang, Y. Zhang and W.X. Xie, "An improved global K-means clustering algorithm," *Journal of Shaanxi Normal University (Natural Science Edition)*, vol. 38, no. 2, pp. 18-22, 2010.

[10] C.-C. Hung, S. Kulkarni and B.-C. Kuo, "A New Weighted Fuzzy C-Means Clustering Algorithm for Remotely Sensed Image Classi-

fication," *IEEE Journal. Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 543-553, June. 2011.

[11] X.Y. Lu, X.Y. Li and H.J. Shi, "FCM clustering aglrotithm based on attribute reduction," *Computer Engineering and Design,* vol. 31, no. 18, pp. 4062-4064+4127, 2010.

[12] Y. Qin, H. Wang and Q.H. Zhou, "The Research of Clustering Algorithm in Data Mining," *Network security technology and Application*, no. 1, pp. 65-66, 2014.

[13] H. Le Capitaine and C. Frelicot, "A Cluster-Validity Index Combining an Overlap Measure and a Separation Measure Based on Fuzzy-Aggregation Operators," *IEEE Transactions. Fuzzy Systems*, vol. 19, no. 3, pp. 580-588, June. 2011.

[14] L.Q. Liu, Q.L. Ding and T.F. Zhang, "Data Pretreatment Method of Fuzzy C-Means Clustering Effects," *Electric Power Science and Engineering*, vol. 27, no. 8, pp. 24-27+46, Aug. 2011.

[15] T.F. Zhang, M.D. Gu, F. Lv and R. Gu."The Influence on Clustering Results of Electricity Load Curves Using Different Clustering Algorithms with Different Data Normalization methods," *The 11th International FLINS Conference on Decision Making and Soft Computing (FLINS2014), João Pessoa (Paraíba), Brazil*, August 17-20, 2014.

[16] J.G. Sun, J. Liu and L.Y. Zhao, "Clustering Algorithms Research," *Journal of Software*, vol. 19, no. 1, pp. 48-61, Jan. 2008.

[17] T.F. Zhang, F. Lv and R. Gu, "The influence on clustering results of electricity load curves using different distances," *2013 the 3rd International Conference on Frontiers of Manufacturing Science and Measuring Technology, Kun Ming*, 2013.

[18] J.W. Han and K. Micheline, Data Mining Concepts and Technique , 2nd ed. *Elsevier (Singapore) Ltd*, 2006 [E book].

[19] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*. vol. 43, no. 1, pp. 59-69, 1982.

[20] R.J. Sanchez-Garcia, M. Fennelly, S. Norris, N. Wright, G. Niblo, J. Brodzki and J.W. Bialek, "Hierarchical Spectral Clustering of Power Grids," *IEEE Transactions. Power Systems*, vol. 29, no. 5, pp. 2229-2237, Sept. 2014.

[21] E. Elhamifar and R. Vidal, "Sparse Subspace Clustering: Algorithm, Theory, and Applications," *IEEE Transactions. Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765-2781, Nov. 2013.

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

[23] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Load pattern-based classification of electricity customers," *IEEE Transactions. Power Systems*, vol. 19, no. 2, pp. 1232-1239, May 2004.

[24] E. Carpaneto, G. Chicco, R. Napoli, and M. Scutariu, "Electricity customer classification using frequency-domain load pattern data," *International Journal of Electrical Power & Energy Systems*, vol. 28, no. 1, pp. 13-20, 2006.

[25] G. J. Tsekouras, P. B. Kotoulas, C. D. Tsirekis, E. N. Dialynas, and N. D. Hatziargyriou, "A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers," *Electric Power Systems Research*, vol. 78, no. 9, pp. 1494-1510, 2008.

Received: June 10, 2015 Revised: July 29, 2015 Accepted: August 15, 2015