Research on Power Behavior Analysis Based on Clustering

Xunjia Li\textsuperscript{a}, Tao Zhang\textsuperscript{b}

College of System Engineering, National University of Defense Technology, Changsha, China.

\textsuperscript{a}lixunjia0526@qq.com, \textsuperscript{b}zhangtao@nudt.edu.cn

Abstract. This research is based on data clustering to analyze power users’ electricity consumption behavior, and analyzes the load curve to obtain the users' electricity consumption characteristics and classify users according to the characteristics of electricity consumption behavior. It is of far-reaching significance to the power industry and socio-economic development. This paper introduces the principle and flow of the main clustering algorithm K-means, fuzzy clustering and neural network clustering algorithm. The data preprocessing method is given. The clustering algorithm and the optimal clustering number are determined by defining the clustering volatility and clustering accuracy index. Then the power consumption behavior of the user is analyzed through the load curves and characteristic parameters obtained through clustering.

Keywords: Clustering, data preprocessing, load curve, electricity behaviour.

1. Introduction

The research on user electricity behavior analysis based on clustering is the theoretical basis of load control, load forecasting, time-of-use price development and implementation, electricity anomaly detection and adjustment of power generation strategy. By studying the user's electricity behavior, understanding the load characteristics of electricity customers plays a crucial role in improving the entire electricity market and improving the power system. China's current classification of the power sector is not based on the user's power load curve, but by the user to report the type of economic sector by the professional sector verification. As a result, the load situation grasped by the power supply department does not match the real load information, which causes deviation from the actual load. As an unsupervised learning algorithm, cluster analysis calculates the similarity between each other based on the characteristics of load data, and then classifies the similarity by comparing the similarities of users' electrical behaviors. In recent years, this method has been widely used in the analysis of power user behavior.

Literature [1] adopts the method of probability and statistics to classify power users through the similarity between objects and the probability and statistics information as the clustering basis. Literature [2] considered the power load, economic and environmental factors synthetically, improved the BP neural network construction model by the genetic algorithm, and studied the power load clustering. In [3], a support vector machine (SVM) method was used to construct a nonlinear mapping and look for a hyperplane. Gaussian kernel procedures and specific algorithms were used to construct multiple clusters to classify users. Literature [4] uses wavelet transform method to solve the local
optimal problem that may exist in neural network, and then replenishes a multidimensional network structure in data space, and then returns the original feature space using wavelet transform, and performs cluster classification in this space. The literature [5] adopts the fuzzy clustering method to reflect the belonging clusters of the object by the degree of membership, which is an improvement on hard clustering. The literature [6] adopts a model-based clustering algorithm and selects the optimal mathematical model to fit the data. The selection is often determined by the characteristics of the data distribution. The literature [7] uses the neural network method to classify the end user's 3 types of feature load into 3 types of input vectors for visual clustering. Literature [8] uses multiple clustering methods to analyze, for example, using K-means for rough segmentation, and then use hierarchical clustering for the second fine division. Literature [9] uses spectral clustering for power load clustering, overcomes the deficiencies of common algorithms, does not need to determine the number of clusters and has no special requirements on the data. The algorithm is simple and requires no iteration and is also a practical algorithm.

To sum up, clustering method is a very important method in power user behavior analysis, but each method has its own limitations and does not apply to all data types and clustering requirements.

2. Clustering Algorithm and Data Preprocessing
The main clustering algorithms used in this paper include K-means, fuzzy clustering and neural network clustering algorithm. The algorithm principle and flow introduction of each algorithm are introduced below, and then the method of data preprocessing is introduced.

2.1. Clustering Algorithm

2.1.1. K-means clustering algorithm. By setting the parameter k, the algorithm divides the data set containing n objects into k clusters and makes the similarity within the cluster higher and the similarity between the clusters lower.

K-means clustering algorithm[10-11] is as follows:
Step1: Select k objects as the initial cluster center
Step2: Repeat
Step3: Calculate the average of the objects in the cluster and assign the input objects to the most similar cluster
Step4: Update the average of each cluster, that is, the average of all objects in the cluster
Step5: Until the clustering results no longer change

2.1.2. Fuzzy clustering algorithm. Fuzzy clustering takes the theory of fuzzy set theory as the research method. By introducing the concept of membership degree and calculating the membership degree of the object, the class of the object belongs is determined, and the non-peculiarity of the clustering result is avoided. The obvious difference of other methods is also the optimization and improvement of traditional hard clustering. The FCM algorithm[12-13] is a more classic algorithm in fuzzy clustering.

FCM algorithm specific process is as follows:
Step1: Determine the number of clusters, initialize cluster centers and fuzzy indicators
Step2: Establish fuzzy similar matrix, initialize the membership matrix
Step3: Iterative calculation, until the objective function converges to a minimum value
Step4: Determine the class to which the data belong according to the membership matrix obtained from the iterative calculation

2.1.3. Neural network clustering algorithm. Neural network is a complex network composed of a large number of simple neurons interconnected with each other, which reflects to some extent many of the basic characteristics of the human brain. It is an extremely complex nonlinear dynamics system. Biological neurology related to the working principle of data processing information, and can learn to gain knowledge and improve the performance of the algorithm.
Self-organizing Maps (SOM) [14-16] is a classic neural network algorithm, which has the features of no supervision and high-dimensional visibility. The research results based on the theoretical method have been widely used in practice. The core idea of SOM is that the network inputs data objects at the input layer. The neurons in the competition layer compete for the response opportunity of input objects. Eventually, only unique neurons survive in the competition layer. We call the winning neurons, meanwhile, The weights associated with the neuron will change in a more beneficial and competitive direction. The "winner" is the classification of the input object.

SOM algorithm specific process is as follows:
Step1: Initialize the network, and assign a weight to each node in the input layer
Step2: Randomly select the input vector in the input sample to find the weight vector with the smallest distance from the input vector
Step3: Define the winning unit, adjust the weight to move it closer to the input vector in the area of the winning unit
Step4: Provide new samples for training
Step5: Shrink the radius of the field, reduce the learning rate, repeat, until less than the allowable value, the output of clustering results

2.2. Data Preprocessing
The power user load data used in this paper also has some missing values and outliers. For these data, it is impossible to directly perform cluster analysis. In order to get satisfactory and valuable conclusions, data preprocessing is often carried out in advance. This article focuses on the introduction and use of data Cleaning and data standardization are two methods.

2.2.1. Data Cleaning. There are some missing values and outliers in the data used in this paper, so we need to fill the missing data and smooth the noise of the abnormal data. For the missing data, the traditional method usually averages the data before and after the load curve where the missing load data exists to fill the average. For the abnormal value, by comparing the load level at the same time with the long and short days before and after the load, A number of similar "abnormal" values occur, and the "abnormal" value is a normal feature of the user's electrical behavior.

2.2.2. Data Standardization. Before clustering, we need to standardize the data (also known as normalization), that is, through the data processing, the data is limited within the set range, on the one hand to facilitate the latter part of the processing, on the other hand, the convergence of the algorithm accelerated. The common method of data standardization in this article is summation normalization:

Totalization of the summation: Calculate the sum of all the attribute values contained in each object, divide each attribute value by the sum of all the attribute values, let $x'_{i j}$ be the normalized data, and the calculation formula is:

$$x'_{i j} = \frac{x_{i j}}{\sum_{j=1}^{n} x_{i j}}$$ \hspace{1cm} (1)

Satisfy the following formula:

$$\sum_{j=1}^{n} x'_{i j} = 1 (i = 1, 2, \ldots, m)$$ \hspace{1cm} (2)

3. Clustering Based on Power User Load Curve
For the load curve clustering, the most important is to select the optimal algorithm for the experimental data and determine the optimal clustering number. We first understand the classification of China's
power industry, and then determine the optimal algorithm and clustering number according to the clustering volatility index and clustering accuracy index.

### 3.1. The Determination of the Optimal Algorithm

China’s power industry has a clear classification criteria, mainly divided into electricity prices based on classification and classification based on national economy, which can provide the basis for the determination of the number of clusters.

Based on the tariff classification, it is divided into seven categories, namely: residential electricity, non-residential lighting electricity, commercial electricity, agricultural electricity, agricultural irrigation electricity, large industrial electricity and non-industrial electricity; based on the national economy Industry classification is divided into 20 categories.

In the actual clustering process, for a fixed number of clusters, after multiple runs, we compare the clustering results and find that there are differences between multiple results because of the number of clusters and data actually set. So the result of the number of clusters is unstable and the difference is large. Therefore, we define the index of clustering volatility to scale the inconsistency of the algorithm results.

Assuming $T$ runs, $T \times k$ results are obtained for a fixed number of clusters $k$, and the $T \times k$ results are grouped into patterns. If the number of categories obtained by pattern grouping is equal to $k$, it means that the results of $T$ runs are the same. Otherwise, the results are different. By comparing the number of categories obtained by pattern grouping with the given $k$ value, we can see the difference in multiple clustering results. And the volatility of the situation. For this, we define cluster volatility index $CVI$:

$$CVI = \frac{\text{number of mode groupings for } T \text{-run results}}{\text{given number of clusters}} \quad (3)$$

The formula shows that the larger the index, the greater the difference between multiple results. By clustering the volatility index $B$, the optimal $k$ value can be chosen.

For a given number of clusters, taking into account the classification of the power industry in China, combined with the 7 categories based on electricity price and the 20 categories based on national economy, in order not to make the number of clusters too large or too small, we determine clustering from 5 The class performs 20 to each clustering algorithm, and calculates the volatility index under different cluster numbers. For each algorithm, run 6 times for each given class number and the results are given in the following table.

#### Table 1. 6 results for a given number of clusters

| Cluster Number | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  |
|----------------|-----|-----|-----|-----|-----|-----|-----|-----|
| K-means        | 5   | 8   | 8   | 10  | 10  | 12  | 15  | 15  |
| FCM            | 6   | 10  | 11  | 13  | 13  | 16  | 18  | 17  |
| SOM            | 6   | 8   | 9   | 9   | 12  | 14  | 16  | 18  |
| Cluster Number | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  |
| K-means        | 17  | 16  | 19  | 21  | 20  | 22  | 24  | 20  |
| FCM            | 18  | 22  | 23  | 23  | 24  | 25  | 27  | 28  |
| SOM            | 17  | 20  | 22  | 24  | 21  | 22  | 28  | 29  |
K-means algorithm's volatility index range: 1~1.36, average volatility index is 1.23; FCM algorithm's volatility index range: 1.2~1.67, average volatility index is 1.48; SOM algorithm's volatility index range: 1.125 ~ 1.5, the average volatility index of 1.36. It can be seen that for the three methods, K-means has the best stability, better convergence, SOM second, and FCM least stable.

3.2. Determination of the Optimal Number of Clusters

Based on the above performance of the three algorithms, the comprehensive volatility index $CVI_z$ (to a certain extent reflecting the volatility in the case of the number of clusters) is defined:

$$CVI_z = 0.5CVI_{K-means} + 0.3CVI_{SOM} + 0.2CVI_{FCM}$$ (4)

The order of priority for getting clustering number is: 5, 9, 17, 18, 7, 8, 14, 13, 10, 19, 12, 15, 20, 16, 6, 11.

In the results, there are only two kinds of relationship between the object and the object, namely: within the class and between classes. When the objects in the same class are more compact, the objects between the classes are more sparse, indicating that the distinction between different classes is more significant and the better the effect is. For poorly-performing clustering results, class and class intervals are fuzzy, boundaries are not clear, and there is overlap between classes and classes. In order to describe this clustering effect, we define the index of clustering accuracy ($Pr$):

Before defining this metric, we give three definitions: the sum of the distances in the same class ($Within_S$) and the sum of the distances between the different classes ($Between_S$) and the total sum of the distances ($Total_S$).

First of all, the relationship between the three is:

$$Total_S=Within_S+Between_S$$ (5)

Assume that the data set contains $N$ objects and the number of clusters is $K$. For the $k$-th class ($k = 1, 2, 3, \ldots, K$), there are $N_k$ objects.

$Total_S$: The sum of the distances of all objects in the entire collection.

$$Total_S = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \left| x_i - x_j \right|$$ (6)

$Within_S$: The sum of the distances of all data objects in each of the $K$ classes, and sums the $K$ classes.

$$Within_S = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \left| x_i - x_j \right|$$ (7)
**Between**: $K$ Arbitrary arbitrary extraction of two classes, calculating all objects in a class with all objects in another class and the distance, and then add the distance of all combinations.

$$\text{Between} = \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j| - \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} |x_i - x_j|$$ (8)

For a fixed data set, $Total$ are fixed. Here we define the index of clustering accuracy ($Pr$), defined as follows:

$$Pr = \frac{\text{Between} - \text{Total}}{\text{Total}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j| - \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} |x_i - x_j|}{\sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|}$$ (9)

From the formula, we find that for fixed $Total$, when $Within$ is smaller, $Between$ will be bigger, and the clustering precision will be bigger. In the result, the separation between clusters is obvious and the objects in clusters are compact and dense.

We test the number of clusters from 2 to 50 in order to obtain the curve of clustering accuracy:

![Figure 2. The influence of clustering number on clustering accuracy](image)

As the number of clusters increased from 2 to 5, the clustering accuracy increased rapidly from 52.5% to 76.3%. As the number of clusters continued to increase, the growth rate of the curve slowed down and reached 85.2% when it reached 12. After that, with the increase of the number of clusters, the curve grows very slowly and gradually converges smoothly. Finally, when the number reaches 50, the clustering accuracy stabilizes at 90.1%.

Thus, as the number of clusters increases, the distance between the clusters decreases rapidly at the beginning, and the interval between the clusters becomes obvious, and the boundary division effect is remarkable. With the further increase in the number of clusters, the clustering accuracy has slowed down, indicating that the convergence speed of each cluster becomes slower and the shape of clusters begins to be fixed. When the number of clusters increases again, it shows that some clusters begin to split into smaller clusters. At that time, the formation of new cluster does not significantly help the growth of clustering accuracy. At this time, the clustering accuracy grows extremely slowly, and slowly approaching a certain value, it is basically meaningless to increase the number of clusters. Thus, for the actual data set, we must not only ensure the accuracy of clustering, but also does not make the computational complexity too high, we should select the appropriate number of clusters, so that the clustering accuracy is maintained at a high level, and the computational complexity Degree is within acceptable range. For this dataset, clustering is selected if the number of clustering is about 10, so that the above two requirements can be satisfied at the same time and the satisfactory result can be obtained.

Combining the two indicators, we finally determined that 9 is the optimal value.
3.3. Load Curve Characteristic Parameter

The load curve describes the changes in user power consumption over a period of time, where the horizontal axis represents time and is usually divided by a fixed time interval. The vertical axis represents the number of power loads. The load during the day is collected and displayed on the map, called the daily load curve.

For the daily load curve, the characteristic parameters are as follows:
- Daily maximum load \( P_{\text{max}} \)
- Daily minimum load \( P_{\text{min}} \)

Daily average load \( P_{\text{ave}} \):
\[
P_{\text{ave}} = \frac{\int_0^{24} P \, dt}{24}
\]  
(10)

Daily load factor \( R_1 \): for the load curve, there are maximum and minimum values, and the daily load factor is to shift load, calculate the ratio of \( P_{\text{ave}} \) and \( P_{\text{max}} \). The calculation formula is as follows:
\[
R_1 = \frac{P_{\text{ave}}}{P_{\text{max}}}
\]  
(11)

Daily peak-to-valley difference rate \( R_2 \): characteristic parameters describing the difference between \( P_{\text{max}} \) and \( P_{\text{min}} \). The formula is as follows:
\[
R_2 = \frac{P_{\text{max}} - P_{\text{min}}}{P_{\text{max}}}
\]  
(12)

Day maximum load utilization time \( R_3 \): the time of total load of a day consumed at the maximum daily load rate, the formula is as follows:
\[
R_3 = \frac{\int_0^{24} P \, dt}{P_{\text{max}}}
\]  
(13)

Among the six characteristic parameters given above, the first three parameters characterize the most basic characteristics of the single load curve, the latter three curves fully describe the common characteristics of the integrated daily load curves for different power industries, with clear industry differentiation.

4. Clustering Results Analysis

The load data used in this paper is the actual electricity consumption data of 6435 electric users in the recent two years from January 1, 2009. These data are collected through the smart meter reading system. The daily collection interval is 30 minutes and the daily total of 48 data points the value at each data point is the amount of electricity used for that time period, and the unit of data collected is kilowatt-hours.

4.1. Clustering Based on Daily Load Curve

After the previous work done in the above, K-means method is used to cluster the average daily load curve of 6435 users within two years, and the data are processed by the summation standardization method, the number of clusters is set to 9, as shown in Fig.3 and Fig.4, clustering results and the typical representative curve for each class are plotted.
Figure 3. The clustering result of clustering number 9

Figure 4. The center representative curve for each class

For the clustering results of users and typical representative curve of these 9 classes, analyze the electricity user's electricity consumption behavior by calculating curve characteristic parameters.

The range of characteristic parameter values for the nine types of curves is as follows:

Table 2. 9 types of curve characteristic parameter range

| Class | R₁   | R₂       | R₃       |
|-------|------|----------|----------|
| 1     | 0.205-0.545 | 0.728-0.975 | 5.041-12.987 |
| 2     | 0.182-0.722 | 0.562-0.971 | 5.689-16.784 |
| 3     | 0.151-0.720 | 0.453-0.966 | 5.035-16.470 |
| 4     | 0.220-0.675 | 0.656-0.997 | 6.245-15.912 |
| 5     | 0.257-0.887 | 0.056-0.892 | 7.081-22.323 |
| 6     | 0.250-0.669 | 0.628-0.964 | 6.257-15.928 |
| 7     | 0.231-0.479 | 0.810-0.997 | 5.631-11.400 |
| 8     | 0.158-0.680 | 0.644-0.969 | 4.684-16.098 |
| 9     | 0.183-0.613 | 0.695-0.970 | 5.118-14.826 |
It can be seen from the above table that the range of variation of the three characteristic parameters of each type of curve is relatively large, mainly due to the edge data of each cluster in the clustering results, and these data have large deviations from the center curve of the class. The difference between the characteristic parameters and the load center representative curve is also large.

The characteristic parameter values for the 9 typical representative curves are as follows:

Table 3. 9 typical representative curve characteristic parameters

| Class | \( R_1 \) | \( R_2 \) | \( R_3 \) |
|-------|-----------|-----------|-----------|
| 1     | 0.452     | 0.845     | 10.868    |
| 2     | 0.648     | 0.785     | 15.567    |
| 3     | 0.534     | 0.739     | 12.831    |
| 4     | 0.476     | 0.784     | 11.427    |
| 5     | 0.759     | 0.478     | 18.238    |
| 6     | 0.556     | 0.768     | 13.360    |
| 7     | 0.389     | 0.923     | 9.343     |
| 8     | 0.621     | 0.774     | 14.910    |
| 9     | 0.488     | 0.839     | 11.729    |

The table reflects the most representative of these types of users of electricity consumption characteristics of the parameters, such as the load curve of the fifth category of the lowest rate of day-to-peak valley, while the largest daily load maximum utilization time; the seventh category and the fifth category are just the opposite.

4.2. Analysis of Impact of Various User Loads on Total Load

In order to determine the impact of these 9 types of users on the total load, the total load curve of all 6435 power users is extracted, that is, all load curves are grouped into a class. The actual calculation finds that the curve is the same with the curve obtained by calculating the average of all user loads.

The total load curve and characteristic parameters are as follows:

![Figure 5. Total load curve](image)

Table 4. Total load curve characteristic parameters

|               | \( R_1 \) | \( R_2 \) | \( R_3 \) |
|---------------|-----------|-----------|-----------|
| Total load curve | 0.643     | 0.728     | 15.429    |

In order to determine the degree of similarity between the various load curves and the total load curve, a distance measurement method is used: the distance \( d \) between the two curves is defined to describe the similarity (the smaller the \( d \), the greater the similarity), \( i \) represents 48 time scales, \( P_{total}, P_{sub} \) respectively represent the load values of various types of load curves and total load curves at that time. In the actual calculation, data normalization makes the data value smaller, and in order to distinguish
the distance calculation value clearly, 48 data points are all enlarged by 100 times and then calculated. The formula is as follows:

\[ d = \sqrt{\frac{1}{48} \sum_{i=1}^{48} (p_{total,i} - p_{sub,i})^2} \]  

(14)

According to the calculation formula, the following results are obtained:

Table 5. Distance metric similarity

| Class | Similarity |
|-------|------------|
| 1     | 18.9732    |
| 2     | 4.3441     |
| 3     | 43.8421    |
| 4     | 164.3584   |
| 5     | 7.3978     |
| 6     | 11.4416    |
| 7     | 164.8413   |
| 8     | 8.8509     |
| 9     | 5.8169     |

According to the distance measurement method, it can be seen that the highest similarity to the total load curve is Class 2, 9, 5 and 8, and the lowest similarity is Class 4 and 7.

By analyzing the impact of various types of user load on the total load, the user group with the highest similarity to the total load can be obtained, and the influence of such users on the entire power grid is also the largest. When the load of the power grid is too large, load control can be performed on the power user groups that have the greatest impact on the total load through user-demand management to ensure the safe and reliable operation of the power grid.

5. Conclusion

This paper introduces the principle and flow of the main clustering algorithm K-means, fuzzy clustering and neural network clustering algorithm, gives the method of data preprocessing, through the definition of clustering volatility and clustering accuracy index to determine the clustering algorithm and the optimal number of clusters, and then through the cluster of load curves and characteristic parameters of the users’ electricity behavior analysis, and by distance measurement method to determine the various types of load curve and the total load curve similarities, thus analyzing all types of user load impact on total load. Due to space constraints, the weekly load curve, monthly load curve clustering results not further shown in this article.

The clustering algorithm used in the next step will be improved, for example, the best K-means algorithm in the experiment, the improved K-means ++, ISODATA and Kernel K-means, improved to improve accuracy.

References

[1] Li Guohe, Liu Wensheng. Curve Segmentation Based on Local Clustering of Probability and Statistics [J]. Computer Engineering and Applications, 2006, 13: 162-165.

[2] Li Lingchun, Tian Li. Prediction of short-term power load based on genetic algorithm and BP neural network [J]. Journal of Anhui University of Technology and Science, 2009, 24 (3): 57-60.

[3] G. Chicco and I. S. Ilie. "Support Vector Clustering of Electrical Load Pattern Data," IEEE Transactions on Power Systems, 2009, 24: 1619-1628.

[4] Zhao Xiaoli. Systematic Marginal Price Forecasting Based on Wavelet Least Square Support Vector Machine Model [J]. Journal of Shanghai University of Electric Power, 2009, 25(3):284-287.
[5] R. F. Chang and C. N. Lu. "Load profiling and its applications in power market," presented at Power Engineering Society General Meeting, IEEE, 2003.

[6] Shi, J. Q. and Wang, B. Prediction and Clustering with Mixtures of Gaussian Process Functional Regression Models, Statistics and Computing, 2008, 18 (3): 267-283.

[7] Li Zhiyong, Wu Jingying, Wu Weilin et al. Power user load curve clustering based on self-organizing map neural network[J]. Power System Automation, 2008, 32(15): 66-70, 78.

[8] Duan Qing, Zhao Jianguo, Luo Xi. Multi-clustering analysis of daily load curve based on similar shape and its application[J]. Journal of Electrical Applications, 2008, 27(20): 53-56.

[9] Li Tianyun, Li Xiang, Liu Huijun, Wang Hongtao. Store load classification based on spectral clustering [J]. Jilin Electric Power, 2008, 30 (5): 4-8.

[10] J. A. Hartigan, M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm[J]. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979, 28(1):100-108.

[11] Wagstaff K, Cardie C, Rogers S. Constrained K-means Clustering with Background Knowledge[C]// Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2001:577-584.

[12] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2):191-203.

[13] Hathaway R J, Bezdek J C. Fuzzy c-means clustering of incomplete data[J]. Systems Man & Cybernetics Part B Cybernetics IEEE Transactions on, 2001, 31(5):735-744.

[14] Kohonen T. The self-organizing map[J]. IEEE Proc Icnn, 1990, 1(1–3):1-6.

[15] Ressom H, Wang D P. Adaptive double self-organizing maps for clustering gene expression profiles[J]. Neural Networks the Official Journal of the International Neural Network Society, 2003, 16(5):633-640.

[16] Taşdemir K, Mérényi E. Exploiting data topology in visualization and clustering of self-organizing maps[J]. IEEE Transactions on Neural Networks, 2009, 20(4):549-562.