Journal of Computer and Communications, 2015, 3, 88-93
Published Online May 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.35011

Cluster Analysis of Electrical Behavior

Lin Liu

Lin Liu, School of Electrical and Electronic Engineering, North China Electric Power University, Beijing, China
Email: bihdliulin@163.com

Received February 2015

Abstract

In this paper, we apply clustering analysis of data mining into power system. We adapt K-means clustering algorithm to analyze customer load, analyzing similar behavior between customer of electricity, and we adapt principal component analysis to get the clustering result visible, Simulation and analysis using matlab, and this well verify cluster rationality. The conclusion of this paper can provide important basis to the peak for the power system, stable operation the power system security.

Keywords

K-Means Clustering Analysis, Principle Component Analysis, The Power System

1. Introduction

On the one hand, in the age of big data such a massive information, data affects our works and lives every second, data mining and clustering analysis is becoming more and more important, on the other hand, With the rapid development of our national economy, the power consumption is larger and larger. And our current power source is mainly rely on thermal power, in order to ensure the stable operation of power system, power dispatch and peak becomes more and more important. The clustering analysis to customer power load is a key link in power decision. Therefore, this paper will focus on the application of large data in power system. Clustering algorithm can be divided into different classification with different standards. Commonly used algorithms in clustering analysis include K-means clustering algorithm, agglomerative hierarchical clustering algorithm, SOM of neural network clustering algorithm, the FCM of fuzzy clustering algorithm, and so on [1]. By comparison, we discover that the K-MEANS program and the FCM program have good comprehensive performance, however the FCM program are too complex for us to use. The power system data is produced every second, so the K-MEAN program are outstanding for its highly efficiency. We select K-means clustering algorithm to analyze the customer power load. Thus may balance power load according to different classification. And this can provide different service to different kinds of customers. The characteristic of this article is: every detail is analyzed from rom the generation of customer power load to data clustering.

2. The Source Data of Power Load

The source of data used for clustering analysis in this paper comes from reference [2]. We sample the reference data, then interpolate, this makes data regeneration. We select 4 classifications of power load, each classification

How to cite this paper: Liu, L. (2015) Cluster Analysis of Electrical Behavior. *Journal of Computer and Communications*, **3**, 88-93. http://dx.doi.org/10.4236/jcc.2015.35011

respectively have 100 sets of data, a total of 400 sets of data. To analyze one day's power load, every 10 minutes for a sample, each data set contains 144 numerical (as is shown in **Figure 1)**.

3.1. The Algorithm and Process

Clustering is one of the important research topics in data mining, is the process of physical objects into multiple classes or clusters [3] [4]. The objects in the same cluster are as similar as possible, while objects in different clusters as different as possible. Clustering can handle different field types and discover clusters of arbitrary shape, it can process the abnormal data, Clustering is not sensitive to data order and less dependent of professional knowledge. K-means algorithm is one of the most classic clustering algorithms commonly used in the present, it has advantages in the following three aspects [5]:

It's quick and sample;

For large data sets with high efficiency and scalability;

It has nearly linear time complexity, and it is suitable for mining large data sets. K-Means clustering algorithm's time complexity is a function of *n*, *k*, and *t*. Where *n* stands for the number of objects in data sets, *t* stands for number of the iteration algorithm, *k* stands for the number of clusters.

So this paper uses the K-means clustering algorithm to analyze customer load, and design a flow chart as shown in **Figure 2**.

3.2. The Steps of K-Means Algorithm

Randomly select *k* points as the initial clustering center $\mu_1, \mu_2, \ldots, \mu_k$ in the data sets $\{x_i\}_{i=1}^{N}$ and N is the number of samples

On the *i* of sample points x_i in the data sets μ_i, calculate Euclidean distance between it and the clustering center, and get its category label

$$\mu_j(i) \leftarrow \arg\min_i \left\| x_i - \mu_j \right\|^2 \quad i = 1,\ldots,N;\ j = 1,\ldots,k \tag{1}$$

Recalculate the k cluster centers, according to type (2)

$$\mu_j = \frac{1}{N_j} \sum_{x_i \in \mu_j} x_i, j = 1,\ldots,k \tag{2}$$

In the formula, N_j is the number of objects in clusters μ_j

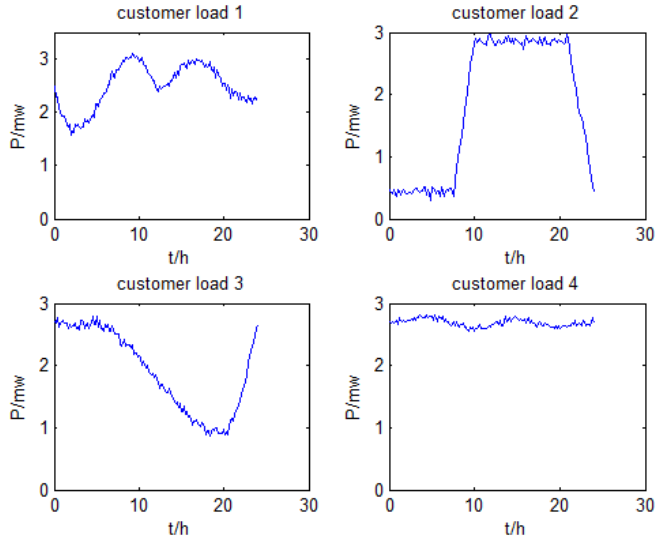

Figure 1. Source data: customer load.

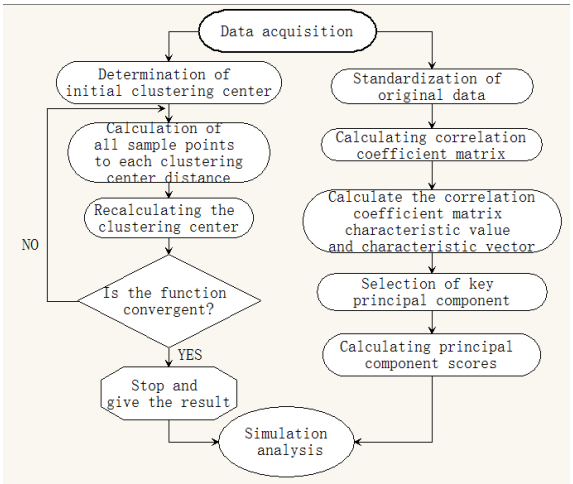

Figure 2. Flow chart.

Repeat step 2) and step 3), until it reaches the convergence criterion function. The evaluation of convergence is based on the square error criterion, as shown in Formula (3).

$$E = \sum_{i=1}^{k} \sum_{\mu_i} \left| x - m_i \right|^2 \tag{3}$$

In the formula, E is the sum of square error of all the objects in the database; x is a point in space; m_i is the average value of the cluster u_i. This objective function makes the generated clusters as compact as possible and independent.

Using the above K-means algorithm, cluster analysis was performed on the data obtained, thus draw customer load can be divided into 4 categories obviously, and the clustering center is shown in **Figure 5**.

4. Visualization of Clustering Results

Because of the use of the large amount of data, selected 144 sampling moments every day, we can't express the clustering results directly. In order to get the clustering results visual, we use the method of principal component analysis (PCA) to study the clustering results.

PCA is a mathematical method of dimensionality reduction. It can take many variables with certain correlation into a set of new independent variables [6]. Use as few variables as possible to express as much information as possible, this is one of the basic principles of PCA. By clustering analysis, 144 dimensional data is mapped to a 3 dimensional space, then analysis. That is, select 3 principal components.

The following is the introduction about the calculation steps of PCA standardization processing of the original data.

Assuming the sample observation data matrix is

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Then the original data were standardized according to the following methods

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{\sqrt{\mathrm{Var}(x_j)}} (i = 1, 2, \cdots, n; j = 1, 2, \cdots, p)$$

where, $\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$

$$\mathrm{Var}(x_j)=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)\quad (j=1,2,\cdots,p)$$

Calculation of sample correlation coefficient matrix

For the sake of convenience, assuming that the original data standardization is still denoted by X , the correlation coefficient matrix after data standardization is

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

where

$$r_{ij}=\mathrm{cov}(x_i,x_j)=\frac{\sum_{k=1}^{n}(x_i-\overline{x}_i)(x_j-\overline{x}_j)}{n-1}\qquad (n>1)$$

The calculation of characteristic value and corresponding characteristic vector of relation coefficient matrix R.

Characteristic value: $\lambda_1,\lambda_2,\cdots,\lambda_p$

Characteristic vector:

$$a_i=(a_{i1},a_{i2},\cdots,a_{ip}),i=1,2,\cdots,\ p$$

Select the important principal components, and give the expression

Through principal component analysis, we can get *p* principal components, but because the variance of each principal component is decreasing, the quantity of information is declining. In practical analysis, according to the principal component contribution to select the first *k* principal components, usually the accumulative contribution rate can reach more than 85%, in order to ensure the integrated variables carry most information of original variables.

where

$$\text{contribution rate}\ =\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

The calculation of principal component scores

According to the original data standardization, the principal component values are calculated from the expression for each sample, you can get all new sample data in each principal component, that is, the principal component scores. The specific form is as follows:

$$\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1k} \\ F_{21} & F_{22} & \cdots & F_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{nk} \end{bmatrix}$$

where

$$F_{ij}=a_{j1}x_{i1}+a_{j2}x_{i2}+\cdots+a_{jp}x_{ip}$$
$$i=1,2,\cdots,n;\ j=1,2,\cdots,k$$

Since we finished the clustering analysis, principal component analysis, then make the scatter of clustering results, by MATLAB. As shown in **Figure 5**.

5. Simulation Analysis

In order to verify the rationality of K-means algorithm used in this paper, this paper uses MATLAB simulation

analysis to explain.

Figure 1 is the customer load source data before clustering analysis, **Figure 3** is each clustering center after clustering, two picture comparison, all customer load 1 were clustered into the D category, the customer load 2 were clustered into the B category, the customer load 3 were clustered into A category, the customer load 4 were clustered into the C category, which can prove that our clustering method is reasonable.

Figure 4 depicts all points to the clustering center distance. Where

Subgraph A conveys the distance from all 4 kinds of customer load to A clustering center, we can see that all points of customer load 3 is nearest to A clustering center

Subgraph B conveys the distance from all 4 kinds of customer load to B clustering center, we can see that all points of customer load 2 is nearest to B clustering center

Subgraph C conveys the distance from all 4 kinds of customer load to C clustering center, we can see that all points of customer load 4 is nearest to C clustering center

Subgraph D conveys the distance from all 4 kinds of customer load to D clustering center, we can see that all points of customer load 1 is nearest to D clustering center

This is consistent with the definition of the clustering center, the relationship is also compatible with **Figure 1** and **Figure 3** exhibited by.

In the earlier analysis based K-means clustering and principal component analysis, draw the visualization map according to clustering analysis results, as **Figure 5**, from the figure, we can also directly obtained that the customer load can be divided into 4 categories, these four categories corresponding to four types of customer load types in the source data, further proved the correctness of this clustering analysis.

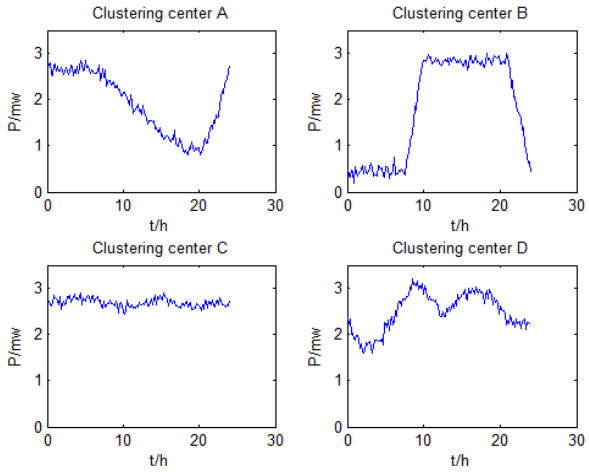

Figure 3. Clustering center.

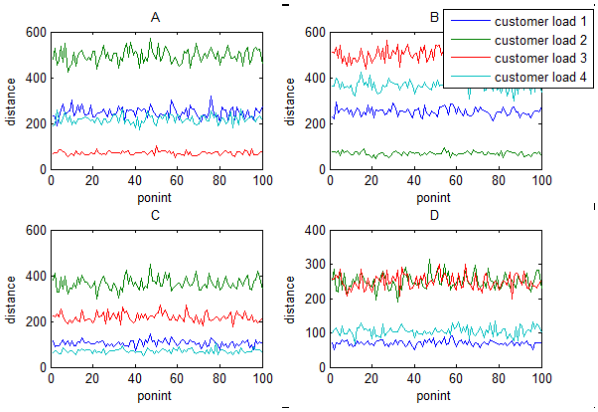

Figure 4. The distance between each point and each clustering center.

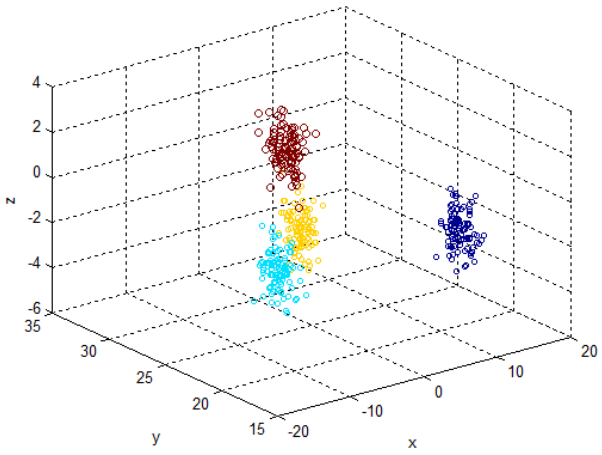

Figure 5. The scatter of the clustering results.

6. Conclusion

In this paper, the K-means clustering method in data mining was used in power system on the clustering analysis power load of customer, and the method of principal component analysis was used on the clustering results visualization, fully prove the rationality and correctness of the clustering. To provide an important basis for the power system decision, and ensure the stable operation of power system.

References

[1] Feng, X.P. and Zhang, T.F. (2010) Comparison of Four Clustering Methods. *Microcomputer and Application*, **16**.

[2] Zhang, M.M., Chen, J.Q., Wang, K., Peng, B. and Wu, H. (2014) The Multi Time Scale Coordinated Orderly Power Use Centralized Decision Method. *Automation of Electric Power Systems*, **38**, 70-77.

[3] Xu, J., Huang, Y.L. and Li, F. (2004) Research on Comparing the Sequential Learning with Batch Learning for K-Means. *Computer Science*, **31**, 156-158.

[4] Keogh, E. and Pazzani, M. (1998) An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback. *Proceedings of the* 3*rd International Conference of Knowledge Discovery and Data Mining*, The Association for the Advancement of Artificial Intelligence, New York, 239-241.

[5] Zhang, S.X., Liu, J.M., Zhao, B.Z. and Cao, J.P. (2013) Analysis of Cloud Computing of Residential Consumption Behavior Model Based on. *Power System Technology*, **37**, 1542-1546.

[6] Zhuo, J.W., Li, B.W., Wei, Y.S. and Qin, J. (2014) Application of MTLAB in Mathematical Modeling. The Beihang University Press, Beijing, 39-41.