Clustering analysis method of power grid company based on K-means

Wenting Wang¹, Qiang Ma², Yong Liu², Ning Yao³, Jing Liu³, Zhaoxuan Wang³, HongLin Li⁴*
¹STATE GRID SHANDONG ELECTRIC POWER RESEARCH INSTITUTE, Jinan, Shandong Province, 250000, China
²STATE GRID SHANDONG ELECTRIC POWER COMPANY, Jinan, Shandong Province, 250000, China
³STATE GRID ZIBO POWER SUPPLY COMPANY, Zibo, Shandong Province, 255000, China
⁴School of Computer Science and Technology, Harbin Institute of Technology at Weihai, Weihai, Shandong Province, 264209, China
*Corresponding author’s e-mail: 2191110608@stu.hit.edu.cn

Abstract. With the development of network and the expansion of power grid business, the power grid traffic becomes more and more complex and huge. It is more and more difficult to identify the power grid business through the traffic, and the identification accuracy also needs to be improved. In order to solve this problem, this paper proposes a clustering scheme based on K-means to classify the power grid business through the grid traffic, using elkan K-means algorithm improves the efficiency of the algorithm and finds the most suitable K value for power grid business data.

1. Introduction
In recent years, cluster analysis is booming, and it plays a very important role in all walks of life. However, cluster analysis is not widely used in power grid business, and it is still in the initial stage.

Nowadays, the concept of smart power grid wins great support among the people. Today global industrial becoming more and more smarter and informative, cluster analysis is also an indispensable part of the power grid business. With the development of society, the power grid business is expanding. China’s power grid is the largest power grid in the world, and the power grid business is also the most complex in the world. Therefore, while the power grid business is expanding, the amount of data is also increasing. The demand of smart grid is informatization and intellectualization. It is urgent to apply cluster analysis to smart grid.

Smart grid can be roughly divided into four areas: advanced smart grid measurement system, advanced distribution operation, advanced transmission operation and advanced asset management. Business analysis is an indispensable part of advanced asset management. Only when you know your own business well, can you effectively evaluate your own assets. For cluster analysis, it is more convenient for the grid system to evaluate its own assets. If we know the business clustering, we can merge the similar businesses according to the clustering results, which can improve the work efficiency.
At present, the classification of power grid business still relies on manual classification and the data is scattered and messy. It is difficult to achieve efficient classification, which deviates from the current development trend of smart grid. There is an urgent need for new and efficient business classification methods to satisfy the requirements of smart grid.

With the expansion of power grid business, the complexity of power grid internal business has increased geometrically. It is inevitable that there will be overlapping or similar situations in the business with the previous classification methods. In the cluster analysis, we can use the results of the analysis to integrate, redivide the business of the power grid and simplify the departments, which can not only improve the work efficiency, but also get rid of many problems Repetitive work. So people have more time for innovation experiments[1-6].

In view of the above problems, this paper proposes an improved k-means algorithm to solve the problem of power grid business classification, and uses Calinski-Harabasz Index algorithm to evaluate the effect of clustering analysis.

2. Correlation Technique

Academics have proposed many different methods for cluster analysis of flow, such as traditional cluster analysis based on port classification, cluster analysis based on payload characteristics classification, and classification using machine learning. Two other kinds of network technology have shown more and more drawbacks with the development of network technology, such as port classification can not recognize dynamic ports...Machine learning method can overcome these shortcomings very well, has a high recognition accuracy, and data constraints are small. Relatively speaking, it is a more effective method to classify the traffic of power network, which has a large amount of data. K-means algorithm is a common and effective algorithm in cluster analysis, and can often be seen in traffic analysis[7].

2.1. Principle of K-means algorithm

K-means algorithm is an iterative clustering algorithm using Euclidean distance.

The principle of K-means algorithm is that given the classification of n objects, each object has x features. Then n objects are selected as cluster centers, and the distance between each object and the cluster center is calculated. The objects with the smallest distance are assigned to the cluster centers. After each sample is assigned, the cluster centers are recalculated according to the existing objects in the cluster. This process repeats until a termination condition is met. The termination conditions in this paper are that no objects are reassigned to different clusters, no cluster centers change, and the sum of squared errors is locally minimum [8-11].
K-means algorithm is one of the most widely used and mature clustering analysis algorithms nowadays. It has a high frequency in industry and information industry. It also has strong scalability and robustness when dealing with large datasets such as power grid.

2.2. Calinski-Harabasz index cluster evaluation method
Because of the confidentiality of power network business, it is necessary to evaluate the algorithm without knowing the actual classification of the business. We can only judge the clustering effect by the degree of aggregation within a cluster and the degree of dispersion between clusters. Common algorithms are the outline factor algorithm and the Calinski-Harabasz Index algorithm. The Calinski-Harabasz Index algorithm is several hundred times faster than the outline factor and more intuitive. So we chose the Calinski-Harabasz Index algorithm\[12-13\].

The evaluation criteria of this algorithm are that the larger the covariance between the cluster and the other cluster, the better the covariance of the cluster itself. Calinski-Harabasz Index fraction value $S(k)$ can be used to calculate:

$$S(k) = \frac{\text{tr}(B_k)m-k}{\text{tr}(W_k)k-1}$$

(1)

$M$ is the number of training set samples, $K$ is the number of categories, $B_k$ is the covariance between clusters, $W_k$ is the covariance of the cluster itself, and $\text{tr}()$ is the trace of the matrix\[14-16\].

3. Classification of network flow services based on K-means
For the classification of power grid business, the first step is to extract sample characteristics, extract what we need. Then the weighted features be calculated, using Elkan K-means algorithm to calculate
the distance between sample points and centroid. And then use Calinski-Harabasz Index algorithm to evaluate the classification effect of this k value after training, so as to select the most appropriate K value.

![Algorithm flow chart](image)

Based on K-means clustering analysis, an important point is the selection of dimension (characteristics). There are many attributes in traffic. Moore[17] mentioned 248 attributes to describe the complete flow. In k-means, the more dimensions, the longer training time will be required, but the classification may be more accurate. Considering influence factor such as: the time cost, training difficulty, operability accuracy and so on. We decide to use these following attributes:

| Feature name | Feature description                       |
|--------------|------------------------------------------|
| Time         | Duration of flow                         |
| Port         | Source port and destination port         |
| Address      | Source address and destination address    |
| Protocol     | Protocol used                             |

Most of the selected attributes are fixed, not presented numerically, but in a fixed form of attributes, such as ports, etc. For these attributes, considering that the information is sent continuously and the fixed attributes are constant over a period of time, so the fixed attributes are weighted by the duration of time.

In traditional K-means, each iteration calculates the distance to the center of mass for all sample points. In Elkan K-means, the calculation of unnecessary distances is reduced, and two rules are obtained by using the mathematical basis that the sum of two sides is larger than the third side:

The first rule is for one sample point X and two centroids j1, j2. If we have calculated the distance D (j1, j2) between these two centroids, if we find that 2D (x, j1) = D (j1, j2) is calculated, then we know that D (x, j1) = D (x, j2); the second rule is for one sample point X and two centroids j1, j2. We can get D(x, j2) > max{0, D(x, j1)D(j1, j2)}. Considering the real-time nature of power network services, the Elkan k-means algorithm can effectively reduce the operation time[18-20].

In the K-means algorithm, the most important thing is to select the k-value, which should be larger when the network business is complex.
4. Experiment and analysis

4.1. Experimental data
The experimental data used in this paper is provided by Zibo power grid company of China. The data is used in the form of flow, and the flow of power grid in a period of time is used as the original. Due to the confidentiality of the data, the power grid business is not easy to open to the public, so the Calinski-Harabasz Index method is used to judge the quality of clustering classification.

4.2 Experimental results and the analysis
For the weight experiment, the k value data of 2-20 is intercepted and presented. It can be clearly seen from the figure that after having the weight, compared with the case without the weight, the Calinski-Harabasz index algorithm score is significantly improved, and the clustering effect is better.

![Figure 3: Weight before and after comparison](image)

For the selection of k value in k-means algorithm, the experimental results show that the k-value is always on the rise before 505, and the k-value is mainly on the decline after 505. It can be seen that the k value which is most suitable for power grid business data is 505, and the Calinski-Harabasz index algorithm has the highest score in this k value.

The K value of power grid business clustering classification is larger. Because there are many and complicated power grid businesses, there are many classifications. From Figure 4, it can be seen that after 320, the score of Calinski-Harabasz index algorithm rises exponentially. Through experiments, the score reaches the maximum at 505, and after 505, the score drops exponentially.
After using Elkan k-means algorithm to replace the traditional K-means algorithm, the data processing speed has been significantly improved. Because the data processing speed is related to the sample capacity, the larger the sample capacity is, the more complex the selection and calculation of cluster center in the algorithm will be. It is not a linear relationship. Through experiments, Elkan k-means algorithm is proved to be effective. The data processing speed of Elkan K-means algorithm and K-means algorithm is shown in the following figure (the unit of time is seconds, and the unit of data scale is G):

It can be seen that compared with the traditional K-means algorithm, the Elkan K-means algorithm has faster processing speed. This is because the missing values of power grid data are less, the Elkan k-means algorithm plays a greater role and improves the efficiency.
5. CONCLUSION
This paper proposes a clustering analysis method of power grid business based on K-means. Starting from the flow data of Zibo power grid, the characteristics suitable for analysis are extracted through its analysis. Through weighted calculation, Elkan k-means algorithm is used to improve the efficiency of training, and Calinski-Harabasz index algorithm is used to find the most suitable K value for power grid business classification. Through experiments, the following conclusions are drawn:

1. The effect of weighted data processing is significantly better than that of untreated data.
2. Compared with the ordinary K-means algorithm, the efficiency of Elkan k-means algorithm is greatly improved.
3. In K-means algorithm, the most suitable K value is 505.

Acknowledgments
The work of this paper is funded by the Science and Technology Project of STATE GRID SHANDONG ELECTRIC POWER COMPANY (No. 5206031900F2).

REFERENCES
[1] R. Molina, Jesús, M. Núñez.(2014)Business Models in the Smart Grid: Challenges, Opportunities and Proposals for Prosumer Profitability. Energies, 7(9):6142-6171
[2] C. Ivanov, T. Saxton, J. Waight. (2016) Prescription for Interoperability: Power System Challenges and Requirements for Interoperable Solutions. Power and Energy Magazine, IEEE, 14(1): 30-39
[3] M. Zeng, Y.Q. Yang, L.H. Wang, J.H. Sun. (2016) The power industry reform in China 2015: Policies, evaluations and solutions. Renewable and Sustainable Energy Reviews, 57,: 94-110
[4] S.F. Zhang, Y.Q. Jiao, W.J. Chen. (2017) Demand-side management (DSM) in the context of China's on-going power sector reform. Energy Policy, 100:1-8.
[5] Y.N. Wu, J. Chen, L.R. Liu. (2011) Construction of China's smart grid information system analysis. Renewable and Sustainable Energy Reviews, 15(9):4236-4241.
[6] P. Shah; K. Gehring. (2016) Smart Solutions to Power the 21st Century: Managing Assets Today for a Better Grid Tomorrow. Power and Energy Magazine, IEEE, 14(2): 64-68.
[7] D. J. Arndt and A. N. Zincir-Heywood. (2011) A Comparison of three machine learning techniques for encrypted network traffic analysis. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications Paris, France. pp. 107-114
[8] S. Abudalfia; M Mikki. (2013) K-means algorithm with a novel distance measure. Turkish Journal of Electrical Engineering & Computer Sciences, 21(6): 1665-1684.
[9] A. Likasa, N. Vlassisb, Jakob, J. Verbeekb . (2003) The global k-means clustering algorithm. Pattern Recognition, 36(2):451-461
[10] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu. (2002) An efficient k-means clustering algorithm: analysis and implementation.IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7):881-892
[11] D. Steinley. (2006) K-means clustering: A half-century synthesis. Br J Math Stat Psychol, 59:1-34.
[12] K. Kryszczuk , P. Hurley. (2010) Estimation of the Number of Clusters Using Multiple Clustering Validity Indices. In: El Gayar N., Kittler J., Roli F. (Eds) Lecture Notes in Computer Science. Multiple Classifier Systems. Berlin. vol 5997.
[13] X. Wang, Y.S. Xu. (2019) An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. Materials Science and Engineering Conference Series:2019 .vol 569(5).
[14] Q. Zhaoa, P. Fräntib. (2014) A sum-of-squares based index for cluster validity. Data & Knowledge Engineering,92:77-89.
[15] R. C. D. Amorima, C. Hennig. (2015) Recovering the number of clusters in data sets with noise features using feature rescaling factors. Information Sciences,,324:126-145
[16] Moore D, Crogan M. (2005) Discriminators for use in flow-based classification. Technical Report, Intel Research, Cambridge.

[17] A. W. Moore. (2005) Discrete content-based classification — a data set. Technical report, Intel Research, Cambridge.

[18] W. Li and A. W. Moore. (2007) A Machine Learning Approach for Efficient Traffic Classification. In: 2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, Istanbul. pp. 310-317

[19] C. Elka. (2003) Using the Triangle Inequality to Accelerate K-Means. Proceedings of the Twentieth International Conference on International Conference on Machine Learning: 147–153

[20] Greg, J. Hamerly, C. P. Elkan. (2002) Alternatives to the k-means algorithm that find better clusterings. Proceedings of the eleventh international conference on Information and knowledge management: 600–607