Rural Developing Level Clustering Based on KMEANS From Electricity Perspective

Peng Li
Economic and Technical Research Institute
State Grid Henan Electric Power Company
Zhengzhou, China
1842029075@qq.com

Huixuan Li
Economic and Technical Research Institute
State Grid Henan Electric Power Company
Zhengzhou, China
2954767243@qq.com

Shiqian Wang
Economic and Technical Research Institute
State Grid Henan Electric Power Company
Zhengzhou, China
2284692441 @qq.com

Wenjing Zu
Economic and Technical Research Institute
State Grid Henan Electric Power Company
Zhengzhou, China
1679275226 @qq.com

Hongkai Zhang
Economic and Technical Research Institute
State Grid Henan Electric Power Company
Zhengzhou, China
hjkwork2020@163.com

Jianjun Wang*
School of Economics and Management
North China Electric Power University
Beijing, China
wangjianjunhd@126.com

2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS) | 978-1-6654-2498-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/TOCS53301.2021.9688988

***Abstract*—KMEANS cluster analysis is widely used in big data environments. Due to the electricity consumption is closely related to the industrial development and living conditions of rural industry production and residents, and it is forming the electricity big data environment, it has given a perspective to analyze the rural developing level from big data mining technology such as KMEANS clustering method. The study of rural developing is becoming an important issue at present. In this paper, we use KMEANS clustering the rural developing levels, and we identify 4 main factors from 14 factors at four aspects: prosperous industry, eco-friendly living, affluent living and agricultural development, and from the case study, 5 types of rural developing level are clustered by KMEANS technology. The case study is also proving the proposed method are effectiveness for rural developing level analysis in the current big data situation.**

Keywords—big data, KMEANS, clustering, rural developing

I. INTRODUCTION

The data mining method such as KMEANS clustering method can be used to identify the developing levels of the rural, and give the rural developing guideline and with the Internet technology and Management Information System with the rural developing, lots of data and messages have been recorded by computers and servers. Therefore, this paper selects the KMEANS clustering method to evaluate the level of rural development. As China's economy continues to grow rapidly and the level of urbanization continues to increase, Chinese rural are developing increasingly. As an essential element for people's lives and production, as well as the top priority of rural development and rural revitalization, the Central Work Conference will "do a good job in achieving carbon economic peaks and carbon neutrality". So how to make a decision of rural developing level is a hot issue at present [1-2].

As electricity is one of the main energy sources used in rural areas and is related to all aspects of rural production, the better life of rural residents, and rural production activities are relied on the electricity grid system. And during the long-term construction of the power system, it has accumulated many years of data, which is forming a big data space for power grids. In this case, we can use the big data technology such as clustering to make a scientific judgment and evaluation of the rural developing level to guide the rural revitalization development based on the clustering.

Some scholars are studied the rural developing problem with the big data environment. Li et al. studied the rural domestic wastewater treatment technology under the rural revitalization strategy to help promote the realization of rural ecological livability [3], Han et al. analyzed the current problems such as weak awareness of rural environmental management, and proposed relevant countermeasures that can help build a beautiful countryside [4], Lai et al. proposed the idea of building a rural revitalization institute, and analyzed the utility of building a rural revitalization institute from five aspects: industrial revitalization, talent revitalization, cultural revitalization, ecological revitalization and organizational revitalization [5], Yang proposes to realize the rural revitalization strategy based on agricultural economics to realize the prosperity of farmers' lives and the rapid development of rural agriculture as early as possible [6], Zhao et al. propose to promote the industrialization of agriculture from the perspective of energy consumption in agriculture, so as to achieve energy conservation and new type of agriculture [7], Fang proposes to promote the development of rural economy through the development of rural tourism culture and the formation of a tourism culture with Chinese characteristics [8], Zhao and others will introduce computerized big data into rural industrial development, promote rapid development of rural industries through big data, and further drive economic development forward [9]. It can be seen that prosperous industry, ecological livability, affluent living and agricultural development are favorable evaluation indicators for rural revitalization.

At present, few researches are focus on the rural developing levels in the big data environment, this paper uses clustering method to analyze the rural developing status to divide the rural developing levels, it can give a better judgement of the rural developing. Due to the electricity is one of the most widely used energy sources in rural areas, and rural electricity consumption is closely related to farmers' life, agricultural production and industrial development. Therefore, we are clustering the rural developing level from the electricity perspective.

II. Collection data for Rural developing leveld

From the rural developing strategy of Chinese government, the rural developing level clustering data should collect from four aspects: industrial prosperity, affluent living, agricultural development and ecological livability were selected to construct a comprehensive evaluation index system for rural revitalization, the comprehensive evaluation index system is shown in Table I.

(1)Industrial prosperity. The category of industrial prosperity mainly includes four aspects: the proportion of electricity consumption in rural industries, the electricity consumption in rural manufacturing industries, the electricity consumption in rural tourism industries and the electricity consumption in other non-energy-consuming industries in rural areas. In the process of evaluating the revitalization of the countryside, the proportion of electricity consumption in the industry can fully illustrate the input of the countryside in industrial development, and the electricity consumption of each industry is also an important indicator of the development of the countryside industry. Among these four indicators, the electricity consumption ratio of rural industries can most intuitively reflect the prosperity of rural industries, and the electricity consumption of manufacturing industries is as important as that of tourism, followed by that of other high-energy-consuming industries.

(2) Ecological livability. The eco-livability category takes into account the percentage of clean electricity consumption, vegetation coverage, and the coverage of water and electricity grid facilities. The percentage of clean electricity consumption refers to the proportion of electricity generated by clean energy to the total electricity consumption of residents; the vegetation coverage refers to the proportion of green plants planted in the countryside to the overall countryside residential area; and the coverage of water and grid facilities refers to the convenience of water and electricity in the countryside.

(3)Living affluence. The category of living affluence requires consideration of per capita electricity consumption, social electricity intensity, the Engel's coefficient of residents, and the number of appliances owned by residents per 100 households. When the per capita electricity consumption of rural residents increases, it proves that the living standard of residents has improved to a certain extent, which can reflect the living condition of residents more intuitively. The per capita consumption of electricity/per capita disposable income of residents reflects the proportion of residents' income invested in the use of electrical appliances, and reflects the affluence of residents' life.

(4)Agricultural development. In rural development, agriculture is a very important industry, and agricultural development is also related to whether rural revitalization can be realized. The scope of agricultural development mainly includes four aspects, namely, electricity consumption for cultivation, electricity consumption for breeding, and electricity consumption per mu of field. The electricity consumption per mu of field is the ratio of the electricity consumption of rural agricultural development to the number of cultivated lands, which reflects the advanced degree of rural agricultural development to a certain exten 1.9 cm

TABLE I. COMPREHENSIVE EVALUATION INDICATOR SYSTEM OF RURAL REVITALIZATION

Target layer	Guideline layer	Indicator layer	Unit	Type
Rural revitalization A	Industrial prosperity B_1	Percentage of electricity consumption in rural industries C_1	%	Positive
		Percentage of electricity consumption in rural manufacturing C_2	%	Positive
		Percentage of electricity consumption in rural tourism C_3	%	Positive
		Percentage of electricity consumption of other non-high-energy-consuming industries C_4	%	Positive
	Ecological livability B_2	Percentage of clean electricity consumption C_5	%	Positive
	Living affluence B_3	Electricity consumption per capita C_8	KWh	Positive
		Social electricity consumption intensity C_9	KWh/yuan	Positive
		Resident Engel Coefficient C_{10}	%	Negative
		Resident ownership of electrical appliances per 100 households C_{11}	Set	Positive
	Agricultural development B_4	Electricity consumption for planting C_{12}	KWh	Positive
		Electricity consumption for farming C_{13}	KWh	Positive
		Electricity consumption per acre of field C_{14}	KWh/a	Positive

III. Kmeans clustering model of rural developing level

As a large number of sample data has been formed for power grid technological transformation projects over the years, a big data environment has been formed. So lots of the data mining techniques of statistical data analysis are widely used in many fields, including machine learning, data mining, pattern recognition, image analysis, and biological information. Clustering is the static classification of similar objects into different groups or more subsets (subset), so that the member objects in the same subset have similar attributes, commonly included in the coordinate system Shorter spatial distance, etc.

In business, clustering can help market analysts find different customer groups from the basic customer base, and use purchasing patterns to characterize different customer groups. In biology, clustering can be used to derive the classification of plants and animals, classify genes, and gain an understanding of the inherent structure of the population. Clustering can also play a role in the determination of similar

areas in the Earth observation database, the grouping of car insurance policy holders, and the grouping of houses in a city according to the type, value and geographic location of the house. Clustering can also be used to classify documents on the Web to discover information. And so on, clustering has a wide range of practical applications.

The common clustering method is KMEANS[10-12], The KMEANS clustering algorithm originated from a vector quantization method in signal processing, and is now more popular in the field of data mining as a clustering analysis method. The purpose of *k*-average clustering is to divide n points into *k* clusters, so that each point belongs to the cluster corresponding to the nearest mean, also called cluster center, and use it as the cluster standard. There is no relationship between KMEANS clustering and k-nearest neighbors.

KMEANS is a clustering algorithm that finds K clusters in a given data set. The reason why it is called K-means is because it can find K different clusters, and the center of each cluster uses the value contained in the cluster The average value is calculated. The number of clusters K is specified by the user, and each cluster is described by its centroid, which is the center of all points in the cluster. The biggest difference between clustering and classification algorithms is that the target category of classification is known, while the target category of clustering is unknown.

The basic idea of KMEANS is very simple. For a given sample set, divide the sample set into k-clusters according to the distance between the samples. Make the sample points in the clusters as close together as possible, and make the distance between the clusters as large as possible.

In the k-means algorithm, because of manually selecting the value of *k* and initializing the random centroid, the result will not be exactly the same every time, and because of the manual selection of the value of *k*, we need to know whether the value of *k* we selected is reasonable and whether the clustering effect is good, but in the reality big data environment, it is that our data will not only have two features, generally there are more than a dozen features, and observing more than a dozen dimensions of space is good for us. Therefore, we need Equation 1 to calculate Sum of Squared Error (SSE) to help us judge the performance of clustering.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{1}$$

In which, E is the sum of the square errors,SSE, p is a point in the space, m_i is the average value of the cluster C_i.

The smaller the SSE value, the closer the data points are to their centroids, and the better the clustering effect. Because the error is squared, points that are far from the center are more important. One way to definitely reduce the SSE value is to increase the number of clusters, but this violates the goal of clustering. The goal of clustering is to improve the quality of clusters while keeping the number of clusters unchanged.

The process of the k-means algorithm is as follows: first select k-objects from n data objects as the initial cluster centers, and for the remaining objects, assign them to the cluster centers according to their distance to these cluster centers. The cluster that is most similar, which is represented by the cluster center, then calculate the cluster center C_i of each new cluster obtained, and repeat this process until the standard measurement function starts Convergence.

From the electricity perspective, the data environment provide a lot of data, and based on the algorithm of K-means as follows, we can get the rural developing levels.

```
while(t)
    for(int i=0;i <n;i++)
        for(int j=0;j <k;j++)
            Calculate the distance from point i to class j
    for(int i=0;i <k;i++)
        Find all data points that belong to your category
        Modify center point of these data points
End
```

Before we clustering the rural developing data, we compare each pair factors of the factors. The top six factors of rural developing clustering pair plots are shown in Fig 1. From the figure, we can clearly see that the C3,C4,C5,C6 are the better factors of the clustering than C1, and the algorithm are passed C2 factors due to lots of lack values. Similarly, when we finished all of the factors in Table I. We decided to use C3, C4, C5, C6 for clustering. And the results are shown in Fig 2.

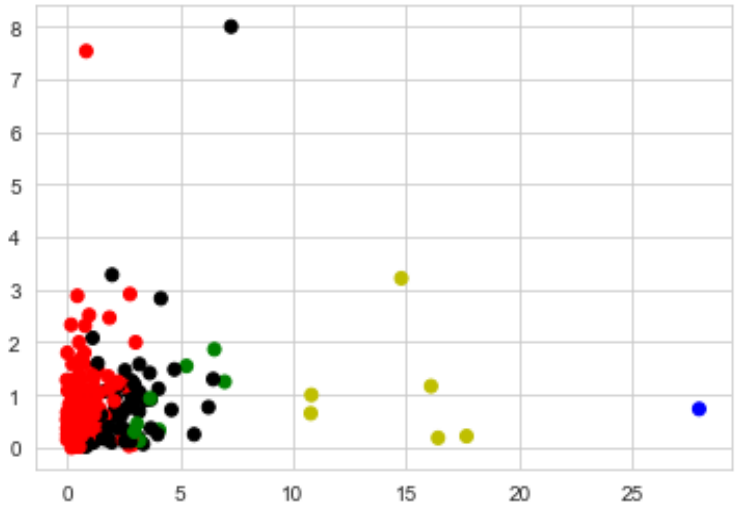

Fig. 1. The results of clustering based on kmeans

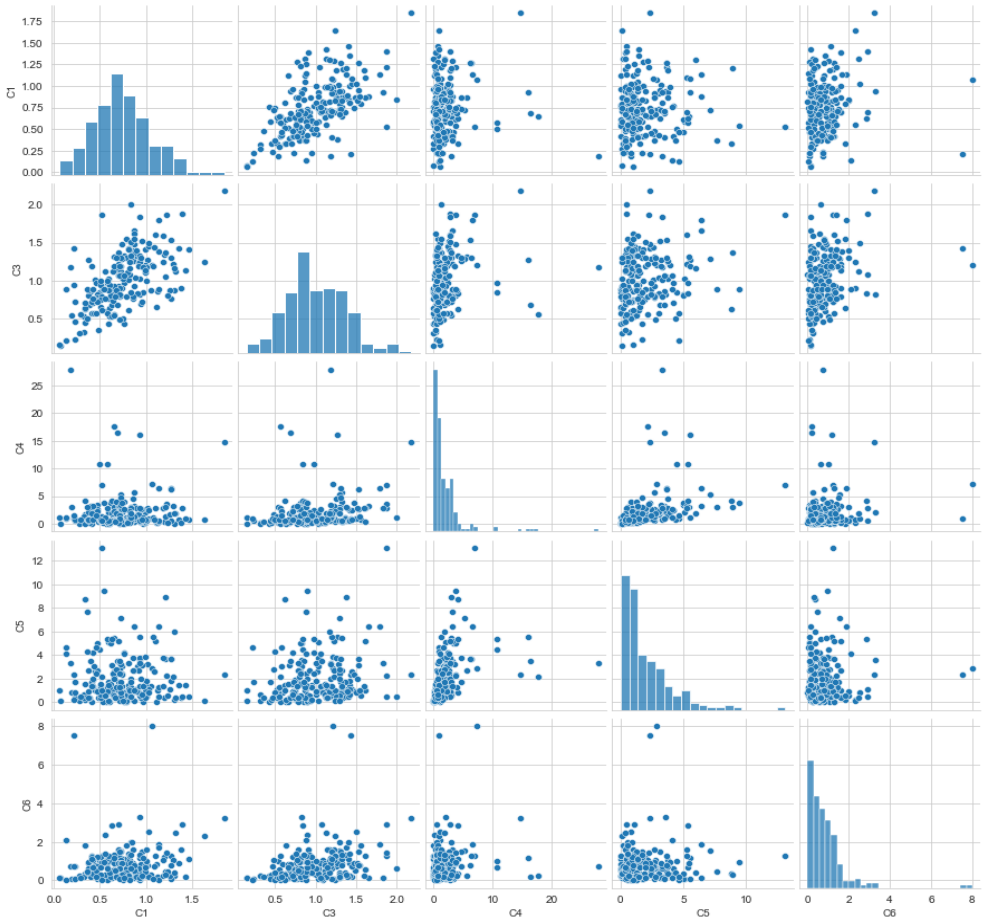

Fig. 2. The top five factors of rural developing clustering pair plot

From the Fig 2. we can see that the rural developing are dividing into 5 types, the best type is only one sample in the top 200 data, and there are 6 rural divided into better type, most of the data are divided in starting level and the second level, it is very similar as the actual situation.

IV. Conclusion

In this paper, we use KMEANS clustering model to analyze the rural developing levels of China from the electricity perspective. With the KMEANS clustering results, it can be clearly seen that the clustering can divide the rural developing level into 5 types, and it will clearly give some ideal developing rural developed samples of the other lower developing countries. And the main research contents of this paper are as follows.

(1) From the electricity big data environment, we selected 14 factors from four aspects: prosperous industry, eco-friendly living, affluent living and agricultural development, and the factors are under different guideline levels. And we have collected the related data to make the case study. At first, we pair compare the factors from 14 factors, and finally find out 4 factors are suitable for the KMEANS clustering. From the results, it has found out that the clustering centers are 5 are most suitable for rural developing level analysis, and it has proved the proposed method are effectiveness.

(2) Construct a comprehensive index system for rural revitalization and consider the influencing factors affecting rural revitalization comprehensively. In this paper, the indicators of rural revitalization considered are mainly four aspects: prosperous industry, eco-friendly living, affluent living and agricultural development, and the indicators under different guideline levels are different, among which the category of prosperous industry mainly includes four aspects: the proportion of electricity consumption in rural industry, electricity consumption in rural manufacturing industry, electricity consumption in rural tourism industry and electricity consumption in other non-energy-consuming rural industries.

References

[1] Z. R. Zhao. “Background of Rural Revitalization Strategy Study on the Way Path of Rural Wind Civilization-- Taking Huangdian Town, Lanxi City, Zhejiang Province as an Example”, International Journal of Social Science and Education Research, 2021, vol. 4, no. 8, 2021.

[2] H. Li and C. X. Liao. “Study on the Development Path of High Efficiency Water Saving Agriculture in Chengdu Plain under the Background of Rural Revitalization Strategy”, IOP Conference Series: Earth and Environmental Science, vol. 768, no. 1, pp. 012045, 2021.

[3] X. Y. Li, “Rural Domestic Sewage Treatment Technology Application in Conghua District of Guangzhou under the Rural Revitalization Strategy”, IOP Conference Series: Earth and Environmental Science, vol. 621, no. 1, pp. 012097, 2021.

[4] X. J. Han. “Study on the dilemma and Countermeasures of rural environmental pollution control under the background of Rural Revitalization Strategy”, E3S Web of Conferences, vol. 237, pp. 01030, 2021.

[5] Y. F. Lai et al. “Building Rural Revitalization College and Implementing Rural Revitalization Strategy”, Management Science and Research, vol. 10, no. 0, pp. 37-41, 2021.

[6] H. Yang. “Study on Rural Revitalization Strategy Based on Agricultural Economics”, Learning & Education, vol. 9, no. 5, pp. 18, 2020.

[7] L. H. Zhao. “Research on the Development Guarantee of the Diversified Supply of Agricultural Energy in Changchun City Based

on the Rural Revitalization Strategy", International Journal of Frontiers in Sociology, vol. 2, no. 9, pp. 223-233, 2020.

[8] M. Fang. "Explore the Rural Revitalization Strategy Under the Revitalization of Tourism Culture to Revitalize the Rural Path", International Journal of Education and Teaching Research, vol. 1, no. 4, 2020.

[9] L. Y. Zhao et al. "Research on the Implementation Path of Rural Revitalization Strategy Based on Computer Big Data and Industrial Revitalization", Journal of Physics: Conference Series, vol. 1648, no. 2, pp. 022165, 2020.

[10] Retno Supriyanti, Ahmad Rifai, Yogi Ramadhani, Wahyu Siswandari. "Characteristics Identification of Myeloblast Cell Using K-Means Clustering for Uncontrolled Images", International Journal of Machine Learning and Computing, vol. 9, no. 3, pp. 351-356, 2019.

[11] Rajan Gupta, Sunil Kumar Muttoo, Saibal K. Pal. "Meta-Heuristic Algorithms to Improve Fuzzy C-Means and K-Means Clustering for Location Allocation of Telecenters Under E-Governance in Developing Nations", International Journal of Fuzzy Logic and Intelligent Systems, vol. 19, no. 4, pp. 290-298, 2019.

[12] Hichem Omrani, Benoit Parmentier, Marco Helbich, Bryan Pijanowskif. "The land transformation model-cluster framework: Applying k-means and the Spark computing environment for large scale land change analytics", *Environmental Modelling & Software*, vol. 111, pp. 182-191, 2019.