

Analysis of energy consumption structure based on K-means clustering algorithm

Weizheng Kong^{1,a}, Yaohua Wang¹, Hongcai Dai¹, Liujun Zhao¹ and Chunming Wang¹

¹State Grid Energy Research Institute Co. Ltd, Binhe Road 18#, Future Science Park, Changping, Beijing 102209, China

Abstract. In order to solve the problem of huge and messy data in the process of analyzing energy consumption structure in different regions, an energy consumption structure analysis method based on K-means clustering algorithm is proposed, and the elbow method and contour coefficient method are used to analyze the data in Qinghai Province. The consumption structure was analyzed and the algorithm was verified. The results show that the algorithm can efficiently and quickly perform data mining and clustering based on local economic and environmental characteristics, which greatly improve the convenience of energy consumption structure analysis.

1 Introduction

At present, countries all over the world have higher and higher requirements for their domestic energy transition, and the global energy transition process is accelerating. Countries in the world are actively involved in promoting low-carbon energy under the common goal of the Paris Agreement. My country's 14th Five-Year Plan also proposes achieving the ambitious goal of achieving carbon peaks by 2030 and achieving carbon neutrality by 2060. The energy consumption structure of a region can reflect the local energy consumption habits of various cold nights, and can provide an important theoretical basis for the transformation of the energy consumption mode of the regional industrial structure and accelerate the process of low-carbon energy.

However, in actual operation, it is not possible to determine a unified index for evaluation due to the differences in the economic and natural environment of various regions. It is an extremely arduous task to integrate data to determine the index based on regional characteristics. The emergence of clustering analysis method has solved this problem well. The idea of clustering can be utilized for group a collection of physical or abstract objects into multiple classes composed of similar objects. Then using different analysis methods for different types of objects can save time greatly.

In recent years, domestic and foreign scholars have proposed many improved algorithms based on the K-means clustering algorithm and applied to various occasions, such as MinMax K-means algorithm^[1], Kmor algorithm^[2] and Seeded-Kmeans algorithm^[3]. Literature [4] proposed a staged clustering algorithm, but this algorithm has high time complexity. In the literature [5], the clustering method is applied to the power system data identification to eliminate the bad data.

This paper proposes a k-means clustering algorithm-based analytical method suitable for energy consumption structure in different regions. The K-means clustering algorithm combines local economic information and environmental information to quickly sort out several different objects, and select appropriate indicators to analyze the local energy consumption structure.

2 Introduction to K-means model

2.1 Introduction to K-means analysis method

K-Means clustering is a kind of fast cluster analysis. Fast cluster analysis is a gradual cluster analysis of large sample data specified by the user. It first classifies the data, and then gradually adjusts to obtain the final classification. K-Means clustering originated from a vector quantization method in signal processing, and now it is more prevalent in the field of data mining as a cluster analysis method. The purpose of K-Means clustering is to divide n points into k clusters, so that each point belongs to the cluster corresponding to its nearest mean (this is the cluster center), which is invoked as the cluster standard.

K-Means is a clustering algorithm that finds k clusters of a given data set. It is called K-Means because it can find k different clusters, and the center of each cluster uses the value contained in the cluster. The number k of clusters formed by the mean value calculation is specified by the user, and each cluster is described by its centroid (centroid), which is the center of all points in the cluster.

^a Corresponding author: kongweizheng@sgeri.sgcc.com.cn

2.2 K-means algorithm flow

The K-Means clustering algorithm assumes that the data sample set to be classified is $\{x_1, x_2, x_3, \dots, x_n\}$. For the number of clusters k , select k initial values as the cluster centers, and assign each sample to one of the k classes according to the minimum distance; then, adjust the centers of each class by recalculation the center of gravity of the class; The samples are allocated according to the minimum distance from the new center, and the cycle repeats until the cluster center is no longer adjusted (that is, the sum of the squares of the distance from each sample to the center of its category is the smallest).

The algorithm steps are as follows:

(1) Randomly select k sample features vectors as initial clustering centers: $Z_1^{(0)}, Z_2^{(0)}, Z_3^{(0)}, Z_4^{(0)}, \dots, Z_k^{(0)}$.

(2) Divide the samples in the feature vector set t to be allocated to one of the k classes one by one according to the principle of minimum distance, if

$$d_{ij}^n = \min_j [d_{ij}^n], i = 1, 2, 3, \dots, N, j = 1, 2, 3, \dots, k \quad (1)$$

Then judge $X_i \in W_l^{n+1}$.

In the formula, d_{ij}^n represents the distance between the center $z_j^{(n)}$ of X_i and W_l^n , and the superscript represents the number of iterations, so a new cluster $W_l^{n+1}, l = 1, 2, 3, \dots, k$ is generated.

The various center distances after reclassification are calculated as:

$$z_j^{(n+1)} = \frac{1}{n_j^{n+1}} \sum_{x_i \in W_j^{n+1}} x_i, j = 1, 2, 3, \dots, k \quad (2)$$

In the formula, n_j^{n+1} is the number of samples contained in the W_j^{n+1} category.

If $z_j^{(n+1)}$ is approximately equal to $z_j^{(n)} (j = 1, 2, \dots, k)$, it ends; otherwise, go to step $k = k + 1$ through c to continue the cycle.

3 K-means clustering method of determining k value

When using the k-means clustering method, it is often necessary to determine the value of the cluster number k first. Due to the lack of precise control over a large amount of data, it is impossible to prepare to determine the number of clusters k . The choice of k value often needs to be determined according to the characteristics of the data sample. With the help of the elbow method and the contour coefficient method, the value of k can often be judged more accurately, which increases the accuracy of the cluster analysis results and enhances the credibility of the cluster analysis results. The specific introduction of the two algorithms is as follows:

3.1 Elbow Method

The core indicator of the elbow method is SSE (sum of the squared errors), and the formula is calculated as:

$$SSE = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2 \quad (3)$$

Among them, C_i is the i -th cluster, p is the sample point in C_i , m_i is the centroid of C_i (the mean value of all samples in C_i), and SSE is the clustering error of all samples, representing the quality of the clustering effect.

The core idea of the elbow method is: as the number of clusters k increases, the sample division will be more refined, and the degree of aggregation of each cluster will gradually increase, so the error square and SSE will naturally gradually become smaller. Moreover, when k is less than the true number of clusters, since the increase of k will greatly increase the degree of aggregation of each cluster, the SSE will decrease greatly, and when k reaches the true number of clusters, the result of increasing k is The degree of aggregation returns will quickly decrease, so the decline of SSE will decrease sharply, and then it will level off as the value of k continues to increase. That is to say, the relationship between SSE and k is in the shape of an elbow, and this elbow The k value corresponding to the part is the true number of clusters of the data, which is why this method is called the elbow method.

3.2 Contour coefficient method

The core indicator of this method is Silhouette Coefficient. The contour coefficient of a sample point X_i is defined as follows:

$$S = \frac{b - a}{\max(a, b)} \quad (4)$$

Among them, a is the average distance between X_i and other samples in the same cluster, called the cohesion degree, and b is the average distance between X_i and all samples in the nearest cluster, called the separation degree. The definition of the nearest cluster is:

$$C_j = \arg \min_{C_k} \frac{1}{n} \sum_{p \in C_k} |p - X_i|^2 \quad (5)$$

Where p is a sample in a certain cluster C_k . In fact, to put it simply, the average distance of all samples from X_i to a certain cluster is used as a measure of the distance from the point to the cluster, and the cluster closest to X_i is selected as the closest cluster.

After calculating contour coefficients of all samples, the average contour coefficient is obtained by averaging. The value range of the average contour coefficient is $[-1, 1]$, and the closer the distance between the samples in the cluster, the farther the sample distance between the clusters, the larger the average contour coefficient, the better the clustering effect. Then, the k with the largest average profile coefficient is the optimal number of clusters.

In the process of selecting the value of k , it is often necessary to use a combination of two methods. In the case of large data samples, the determination of the k

value is crucial to the accuracy of the K-Means method. Therefore, it is necessary to use different k value determination methods to comprehensively consider the value of k to reduce the clustering deviation caused by the value of k.

4 Analysis of energy clustering results based on K-means algorithm

The application of K-Means clustering algorithm requires the help of SPSS software and the use of python algorithm to determine the value of k. Before using this method to analyze the energy consumption data, the contour coefficient algorithm and the elbow algorithm in python are used to determine the value of k.

Use K-Means clustering algorithm to analyze energy consumption big data to obtain key information such as cluster center data, iteration history, and cluster members. According to this method, it is possible to have a more intuitive understanding of the energy consumption structure. Using the province's electricity energy

consumption data in 2018 in Qinghai Province, select twelve major electricity energy consumption industries, and perform k-means clustering algorithm analysis based on monthly energy consumption data. This method can effectively cluster the months with the same energy consumption structure in different months based on the electricity energy consumption data, and can intuitively understand the energy consumption composition of different consumption time periods, which is of great significance to the transmission and deployment of electricity energy, and is helpful for combination Optimize the composition of energy consumption and increase the comprehensive utilization rate of energy.

4.1 Analysis of K-Means clustering algorithm results

Using SPSS data analysis software, perform K-Means clustering on the sample data, and obtain partial clustering results as shown in Table 1.

Table 1. Table of initial cluster centers

classification	Clustering				
	1	2	3	4	5
Citizens	19540.2612	15765.1372	17704.9511	15744.5534	18493.8792
Rural residents	7856.9186	7975.9624	8384.6482	7860.0289	8243.8910
agriculture	526.9819	1138.9364	389.5265	538.3904	783.3294
Animal husbandry	180.6198	134.0580	145.3942	131.3386	168.4473
Agriculture, forestry, animal husbandry and fishery and its supporting activities	557.2403	1513.6104	327.9276	1614.8847	1144.7412
manufacturing	505472.8350	478717.4735	457499.8698	495447.1705	486905.2059
Housing construction industry	1487.3614	1031.1193	854.6268	1056.4404	1646.7153
Building decoration, decoration and other construction industry	2193.4861	1546.5381	812.6569	1635.2354	2113.9365
Railway transportation industry	5545.4228	6158.8040	4820.4378	5691.6141	7333.8820
Road transport industry	1132.7645	844.3621	1028.3811	864.9606	1099.1732
Telecommunications, radio and television and satellite transmission services	1690.1314	1604.6013	1490.5628	1636.0317	1858.2070
Internet and related services	1103.8075	1734.9519	1002.8348	1864.9360	1517.8005

This table shows that when using the K-Means clustering algorithm, the algorithm randomly selects the sample center according to the characteristics of the data sample. After the initial sample center selection, the distance between each data point in the sample data and the respective sample data center needs to be calculated. Re-adjust the data sample clustering center, and the data change process of the sample center is shown in Table 2.

Table 2. Cluster center iteration table

Iteration	Changes in cluster centers				
	1	2	3	4	5
1	.000	1391.701	4651.718	.000	1668.159
2	.000	.000	.000	.000	.000

Since there is no change or only a small change in the cluster centers, convergence is achieved. The maximum absolute coordinate change of any center is .000. The current iteration is 2. The minimum distance between the initial centers is 8779.441.

Table 2 shows the number of iterations for the cluster centers when using the K-Means clustering algorithm. From the table, it can be seen that when the iteration proceeds to the second time, the change value of the entire cluster center is 0, which means that the cluster center is not Change again, that is, the best cluster center of the current data sample, that is, the distance between each data sample and its cluster center is the closest, and the final cluster center is shown in Table 3.

Table 3. The final cluster center table

classification	Clustering				
	1	2	3	4	5
Citizens	19540.2612	16903.0926	17773.8961	15744.5534	17154.1332
Rural residents	7856.9186	8036.3692	8177.5601	7860.0289	8107.3488
agriculture	526.9819	599.3412	526.3204	538.3904	936.4461
Animal husbandry	180.6198	143.4220	158.5538	131.3386	154.9133
Agriculture, forestry, animal husbandry and fishery and its supporting activities	557.2403	1333.2732	928.5450	1614.8847	1112.2287
manufacturing	505472.8350	478781.2066	461978.8552	495447.1705	486250.8035
Housing construction industry	1487.3614	1081.8613	1318.7179	1056.4404	1405.6895
Building decoration, decoration and other construction industry	2193.4861	1612.1178	1590.7546	1635.2354	1863.6595
Railway transportation industry	5545.4228	5684.4424	5356.6938	5691.6141	6738.3374
Road transport industry	1132.7645	931.3973	1051.2329	864.9606	969.6592
Telecommunications, radio and television and satellite transmission services	1690.1314	1595.3811	1587.1369	1636.0317	1703.7784
Internet and related services	1103.8075	1468.7739	1180.8035	1864.9360	1502.4303

It can show in table 3 that the final cluster number is 5, which is initially determined by the two k-value determination algorithms. There are data points with similar structure around each cluster center, so that different types of energy consumption composition types can be seen. It can be clearly seen from the table that the energy consumption of the manufacturing industry in cluster No. 1 is very huge, while the energy consumption of manufacturing industry in the No. 3 cluster center is relatively small compared to other cluster centers. At the same time, there is a big difference in energy consumption between the No. 1 and No. 5 cluster centers. The energy consumption of animal husbandry, agriculture, forestry, animal husbandry and fishery and its auxiliary activities differs greatly in energy consumption between No. 1 and No. 4 cluster centers. The electric energy consumption of urban residents differs greatly between cluster centers No. 1 and No. 4, while the difference in electric energy consumption of rural residents among the cluster centers is not large. The

energy consumption of the housing construction industry is relatively larger in the 2nd and 5th buildings. Building decoration, decoration, and other construction industries have large differences in energy consumption between No. 1 and No. 4 cluster centers. The power and energy consumption of road transportation, telecommunications, radio and television, and satellite transmission services are relatively stable between No. 1 to No. 5 cluster centers, and the power and energy consumption does not fluctuate much, while the railway transportation industry is in No. 3 and No. 5 clusters. There is a great difference in the power and energy consumption between centers, and the Internet and related services have a big difference in the power and energy consumption between the No. 1 and No. 4 cluster centers. By comparing the distances of different energy cluster centers, the differences between different cluster centers can be seen intuitively. The differences between the five cluster centers are shown in Table 4.

Table 4. Distance table of cluster centers

number	Clustering				
	1	2	3	4	5
1		26846.661	43537.777	10826.597	19427.901
2	26846.661		16838.288	16714.517	7570.915
3	43537.777	16838.288		33548.580	24327.621
4	10826.597	16714.517	33548.580		9404.768
5	19427.901	7570.915	24327.621	9404.768	

Different clustering centers have different characteristics of energy consumption structure, and the uniqueness of their characteristics helps the energy dispatching department to formulate reasonable and effective power and energy allocation plans. It can be

seen from Table 5 that the distance between the cluster center No. 1 and the cluster center No. 3 is very large, which means that the cluster points around the two types of cluster centers are significantly different in the power and energy consumption structure. At the same time, the

distance between cluster center No. 3 and cluster center No. 4 is also very large, which represents a significant difference in the power energy structure between the two types of cluster centers. Through research on the power energy consumption structure of each cluster center, the industry composition of the power energy consumption of each cluster center can be analyzed.

5 Conclusion

Using this method to analyze energy consumption data of Qinghai Province is mainly based on the different types of energy consumption as variables, and the energy consumption date as the basis for the compilation of the case. This method can be utilized to analyze the energy consumption structure of Qinghai Province, which has important implications for energy allocation and dispatch. The cluster centers are defined according to different types of energy consumption users, so each type of cluster is significantly different in the composition of energy consumption. Therefore, specific analysis can be carried out for different clusters, and the energy consumption structure of different clusters can be analyzed, so as to determine feasible energy allocation schemes on different dates and peak shift allocation to ensure the continuous and stable operation of Qinghai Province's energy system.

Electricity-based clean energy is the main energy consumption species in the energy consumption system of Qinghai Province. Combining the K-Means clustering algorithm can analyze the law of Qinghai Province's energy consumption structure, and in-depth study of the energy consumption structure of Qinghai Province. At the same time, according to the law of energy consumption time, energy allocation is carried out in an orderly manner to ensure the smooth and orderly operation of the power energy system in Qinghai Province. At the same time, this method can provide a combined optimization structure, formulate specific power distribution plans for different power usage dates, and improve the comprehensive utilization of energy.

6 Summary and outlook

K-Means clustering algorithm can analyze the energy consumption characteristics of various industries and analyze the trend of energy consumption structure based on energy consumption data, which is helpful to deepen the transformation of energy consumption structure, promote the development of clean energy, and ensure the stable operation of the energy system. However, because the algorithm needs to continuously adjust the classification and adjustment of the booring objects, and constantly calculate the new cluster center point after adjustment, when the amount of data is too large, the time cost of the algorithm is very large, and the edge distributed computing method can be considered in the future. Combined with the K-means algorithm, it is used in the analysis of energy consumption structure.

ACKNOWLEDGMENTS

This research was financially supported by the SGCC Technology Project- The research on integrated simulation method and practical technology of think tank research platform.

References

1. G Tzortzis, A Likas,. The MinMax Kmeans clustering algorithm[J].Pattern Recognition, 2014, 47 (7) :2505-2516.
2. G Gan, K P Ng. K-means clustering with outlier removal[J].Pattern Recognition Letters , 2017 (90) :8-14.
3. Y L Luo, Q Y Yu, et al. Outlier-eliminated k-means clustering algorithm based on differential privacy preservation[J]. Applied Intelligence the International Journal of Artificial Intelligence Neural Networks & Complex Problem Solving Technologies, 2016.
4. R Samrin, D Vasumathi. Hybrid Weighted K-Means Clustering and Artificial Neural Network for an Anomaly-Based Network Intrusion Detection System[J]. Journal of Intelligent Systems, 2016, 27.
5. J L Meng, D C Liu. A new method for identifying bad data in power system based on Spark and cluster analysis[J]. Power System Protection and Control, 2016,44(03):85-91.