Cluster analysis of time series K-Means line loss based on DWT distance

Liming Song, Zhimin Chen, XinXin Meng, Shuai Kang

1State Grid Qinhuangdao Electric Power Supply Company, Beijing, China
2State Grid Jibei Electric Power Company Limited Smart Distribution Network Center, Beijing, China
*aemail: dmslm@163.com
b*email: 1225942208@qq.com
cemail: 18716010619@163.com
demail: 15650756765@163.com

Abstract. This paper constructs an indicator system composed of inherent attributes and time characteristics of the line based on the line loss, and proposes a K-Means line loss cluster analysis model based on this indicator system. The line is classified according to the clustering results. The result is 314.51 on the CH index (Calinski Harabasz Index), 0.19 on the Silhouette Coefficient (Silhouette Coefficient), and a running time of 0.508 s. Compared with the traditional algorithm, it is greatly improved. The field of line loss analysis has guiding significance.

1. Introduction
After the electric energy is transformed into high-voltage current through the transformer, it is transmitted to the power supply area and unit through the distribution network, and then the voltage is reduced by the substation for users to use. In this process, electric energy loss occurs anytime and anywhere. The process of power loss is called line loss. This is the result of the inherent attributes of the line and cannot be avoided. However, the power supply unit can improve the management capabilities and methods to seek to minimize the line loss during power supply to increase corporate revenue. This paper introduces K-means line loss clustering in line loss management to help power supply units actively implement loss reduction management and increase energy conservation and efficiency.

With the development of "Internet + Smart Energy", "Energy Internet", and other energy development models, China's electricity consumption information collection system has entered a new stage of 2.0 construction. Collection and full cost control" provides data support for the smart operation and management of the smart grid [1]. Many scholars in the field of online loss analysis have participated in the research and published relevant results. The literature [2] based on the knowledge map of the station area line loss anomaly analysis proposed to use the knowledge map to determine the cause of the station area line loss abnormality, According to the power data in the grid information system, the line loss anomaly knowledge map is constructed, analyzes the cause characteristics and builds judgment rules, and builds the judgment method of the abnormal cause of the line loss in the district through the deduction grid algorithm; Literature [3] Distribution network line loss management based on big data value mining in the analysis method, it is proposed to
preprocess and store the line loss big data through data cleaning, use the K-means clustering algorithm to analyze the line loss characteristics, and mine the value information. The literature [4] low-voltage station line based on data mining technology in the research of damage detection, it is proposed to mine the useful information behind the smart meter data to realize the detection of users' abnormal electricity consumption behavior, prevent electricity theft and leakage, and reduce the line loss. The weighted LOF algorithm is used to perform weighted outlier analysis on massive user data, which effectively completes the location of abnormal power users. Literature [5] In the research of multi-system fusion line loss analysis based on k-means clustering, for the problem of multiple sources of line loss data in the power grid, based on SG-CIM and marketing, PMS, GIS multi-system fusion data, a method based on Multi-system fusion line loss analysis method of k-means clustering: The above literature mostly uses data mining analysis, k-means and other algorithms to analyze the line loss data, and the effect needs to be improved. At the same time, these documents ignore a problem, which is the line loss rate. It is a time-related data, and the characteristics of its time series should be fully considered. Therefore, this article will extract one month's line loss data, combined with the current advanced deep learning algorithm, to analyze the line loss data and give relevant guidance opinion.

The line loss rate directly reflects the power loss of the line in the distribution network. Therefore, this index is an important evaluation index to measure the line loss. Its level also directly reflects the health of the power supply unit. It is occupied in research and production. It is an important position, so this article will use this indicator as an evaluation indicator to use the time series K-means line loss clustering analysis method to analyze the line loss situation in detail, and give corresponding countermeasures.

2. background
Grid line loss analysis is an important technical indicator of the power system. It is also a comprehensive technical and economic indicator that measures the management level and management capability of power enterprises, such as power generation, transmission, transformation, distribution, and electricity consumption. Each link may cause line loss. At the same time, the line loss in the distribution network is messy, the number of users is large, and the business system is complex, which causes the line loss to remain high. Previously, the main method of line loss analysis still remained in manual experience, lacking an effective method of thinking. Therefore, this article will introduce the time series K-means, and add the inherent attributes and time characteristics of the line to perform cluster analysis on the line loss.

3. Basic knowledge
K-means K-Means algorithm is a common clustering algorithm, which divides the data set into K clusters, and each cluster is represented by the average of all samples in the cluster, and the average is called the centroid.

The algorithm flow is as follows:
1. Completely randomly select K points from the sample as the initial centroid.
2. Calculate the distance from each sample to each centroid, and divide the samples into clusters corresponding to the nearest centroid
3. Calculate the mean of all samples in each cluster, and use the mean to update the centroid of the cluster
4. Repeat step 2 And 3, until the position change of the centroid is less than the specified threshold (0.0001) or reaches the maximum number of iterations.

In the K-means clustering process, in order to introduce the time series characteristics of the line, the time series line loss clustering process in this article DWT (Dynamic Time Warping) is used to measure the distance between samples, and the LB_keogh algorithm is used to accelerate the clustering process. The main idea of the LB_keogh optimization algorithm is to first calculate the LB (lower bounding) to deal with the sequence that cannot be the most matching sequence. Sequence, the
calculation formula is as follows:

\[ \text{LB}_\text{keogh}(Q, C) = \sum_{i=1}^{n} \begin{cases} (c_i - u_i)^2, & q_i > u_i \\ (c_i - l_i)^2, & q_i < l_i \\ 0, & \text{other} \end{cases} \]

\( u_i \) and \( l_i \) are the upper and lower envelopes of \( Q \), \( u_i = \max(q_{i-r}:q_{i+r}) \), \( l_i = \min(q_{i-r}:q_{i+r}) \), s.t: \( j - r \leq i \leq j + r \), \( r \) is a sliding window distance.

4. Model building
As there are many sources of line loss data in the power grid, such as low efficiency of investigation and large amount of data, this paper uses time characteristics plus line inherent attributes as the extraction characteristics of line data to establish a new index system to further refine the line data. The first is data acquisition. With the help of smart meters, manual meter reading, etc., the power consumption data of the line and the inherent attributes of the line are collected, and then the collected data is preprocessed to eliminate the repeated useless data, and the missing data is lighter. Fill in the data with the average value, then extract the time characteristics of the line, and then splice the time characteristics of the line data with the inherent attributes of the line, standardize the spliced data, pass it into the model, and finally output the clustering results and analyze.

4.1. Data collection
First, collect the data required for the research, including line name, line number, voltage level, date, power supply, power sales, line loss rate, total number of special transformers, total number of station transformers, weather conditions, etc., according to the collection. The data extracts time characteristics. Here, the data of February is used as the time characteristic index to establish an index system. The power consumption data of the line includes two major categories: file data and meter power data. The file data is in the marketing business system, and the meter electricity data is in the electricity consumption information collection system. External factors such as weather, economy, and electricity prices that affect changes in electricity consumption data can be collected through the social data management system.

Categorize the data source so that data loss or malicious tampering can be traced back to the specific location in time. Use A, B, and C to represent archive data, meter data and external factor data (weather status, economic conditions, and electricity price policy), respectively. The number of sampling lines is \( n \), and the data format is:
4. Data preprocessing

4.2. Data cleaning.
Due to the actual system operation, it is impossible to work in the same way as all expected situations in the production process of the enterprise, and abnormal data will inevitably be generated at this time. Therefore, in order to make the experimental results true and reliable, the method of discarding is used for outliers in this article; for the missing values in the data set, the average value of the same line in the relative time range is used for filling.

4.2.2 Time characteristic extraction.
In order to effectively classify the line loss status, it is necessary to sort out and extract the characteristic indicators that can reflect the line loss status, based on the original data information, line loss rate information, and time change information contained in the collected line loss records. Where the line loss rate information can be calculated from the daily power supply and electricity sales data. Taking the line as the research object, the 28-day data reflecting the line loss characteristics are extracted as the data time characteristic analysis index.

4.2.3 Data standardization.
Due to the inconsistency of dimensions among various features, direct input to the model will lead to deviations in the model, and even large numbers eat decimals. The direct impact is that features with large dimensions occupy greater weight, and features with larger dimensions. Small features have little or no impact. At the same time, the dimensions are inconsistent, which will increase the running speed of the model and reduce the efficiency of the model. Therefore, in view of the above points, in order to accelerate the training of the model and make the model more objectively reflect the data, it should be normalized or standardized. This article uses normalized processing, even if the result of the line data is in the interval \([0,1]\). The formula is as follows:

\[
X_w = \frac{X_i - X_{\text{imin}}}{X_{\text{imax}} - X_{\text{imin}}} \tag{5}
\]

In the formula, \(X_i\) is the ith line data, \(X_w\) represents the normalized ith line data, and \(X_{\text{imax}}\) and \(X_{\text{imin}}\) respectively represent the maximum and minimum values of the line data under the characteristic sequence.

4.2.4 The inherent attributes and time characteristics of splicing lines.
Integrate the preprocessed data together and classify them, and the extracted time characteristic variables are denoted as \(B'\). The data set \(F\) that constitutes a single line is as follows:

\[
F = \left( X_j; B'_j \right), F \in R^{n \times (a+b+c+b')} \tag{6}
\]

4.3. Training model
Input the preprocessed data into the model, and the experimental results are shown in the figure below.
Figure 2. Line loss rate centering and clustering results
5. Discussion Analysis
It can be seen from the above that the time series K-Means line loss cluster analysis model based on DWT distance proposed in this paper can be better applied to line cluster analysis. Its CH (Calinski Harabasz Index) index is 314.51, and the contour coefficient SC (Silhouette Coefficient) is 0.19, and the running time is 0.508s. Generally speaking, the time series K-Means line loss cluster analysis model based on DWT distance proposed in this paper can better deal with the line loss cluster analysis, and it has great development prospects in future applications.

6. Conclusion
The time series K-Means line loss clustering analysis method based on DWT distance proposed in this paper shows great application value in the field of line loss analysis. In the test data, the time it takes is significantly shorter than that of the traditional algorithm, which can improve the clustering density very well, that is, to minimize the distance between classes, but the distinction between classes is slightly blurred. This is not only related to the model, but also related to the model. The data space is related. Since the test data is the actual data in the industrial production process, the reason needs to be studied in depth. However, in general, the model still shows a huge room for development. The promotion of this model will not only help power companies to accurately grasp the line loss situation within their jurisdiction in time, but also can efficiently achieve scientific loss reduction and reduce line loss.

References
[1] GONG, G.J., LU, Z.M., LU, J.(2009). Research on Electricity Consumption Behavior of Users Based on Deep Learning. Proceedings of the CSEE.29(00):1-8.
[2] GAO, Z.P., ZHAO, Y., ZHANG, T.T., etc.(2021).Analysis of Abnormal Line Loss in Transformer Area Based on Knowledge Graph. Electric drive.51(17):69-80.
[3] QIU, Y.I., ZHANG, C., ZHANG, X.Y., etc.(2021). Analysis of Line Loss Management Method of Distribution Network Based on Big Data Value Mining. Inner Mongolia electric power.39(4):65-77.
[4] CHENG, H., WANG, J.F., HU, C.P., etc.(2021). Research on Line Loss Diagnosis of Low Voltage Station Based on Data Mining Technology. Computing Technology and Automation. 40(2):61-65.