J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 8299--8312

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Multi cluster parallel spectral clustering algorithm for distribution area power network loss rate evaluation

Chang Liu[1,✉], Lin Xu[1], Qian Xie[1], Hua Zhang[1], Hua Yang[1], Shu Fang[2], Wei Wang[3], Shixuan Lv[3], Yinzhang Cheng[3], Guanliang Li[3]

[1] *Electric Power Research Institute of State Grid Sichuan Electric Power Company, Chengdu, Sichuan, 610041, China*

[2] *State Grid Sichuan Electric Power Company, Chengdu, Sichuan, 610041, China*

[3] *Electric Power Research Institute of State Grid Shanxi Electric Power Company, Taiyuan, Shanxi, 030002, China*

ABSTRACT

Rapid and accurate assessment of power network loss in the power system has become a key research topic for the vast and diverse dataset of power grid operation. This study integrates data mining techniques with typical scenario modeling concepts and innovatively designs a distribution area power network loss rate multi population parallel spectral clustering evaluation strategy that incorporates distribution characteristics. Firstly, clustering attributes are determined for power network loss evaluation, and a power network loss evaluation framework based on clustering algorithms is proposed. Based on power flow calculation, the distribution characteristics and indicator system of each node's output are analyzed; Secondly, in order to improve the clustering accuracy of power network loss evaluation, spectral clustering algorithm is introduced, and automatic algorithm design is carried out to address the issue of manually setting the initial number of clusters and cluster centers. Then, multi cluster partitioning and parallel computing methods are used to significantly improve the computational efficiency of spectral clustering algorithm; Finally, to verify the practicality of this method, a provincial power grid was selected as a case study. The results showed that this method not only has high accuracy in evaluating power network loss, but also has excellent computational efficiency, demonstrating good feasibility in practical engineering applications.

Keywords: multi cluster partitioning; Parallel computing; Spectral clustering; Assessment of power network loss rate; Data mining; Trend Calculation

1. Introduction

With the continuous advancement of the intelligentization process of the power system, the scale and diversity of power data have become increasingly prominent, leading to new manifestations of some

✉ Corresponding author.
E-mail address: 15927250759@163.com (C. Liu).
Received 05 March 2024; Revised 15 May 2024; Accepted 20 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-457
© 2025 The Author(s). Published by Combinatorial Press. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

traditional power problems. Exploring the deep integration path of data mining technology and traditional research paradigms in the power system has profound significance for accurately evaluating the state of the power system, formulating scientific decision-making plans, and efficiently allocating power resources [1-2].

The evaluation of power network loss, as a classic issue in the power system, traditionally relies on the power grid operation dataset during the evaluation period, and accurately depicts the distribution of power network loss through comprehensive power flow calculation [3]. However, when faced with massive power grid data, this method significantly affects evaluation efficiency due to the dramatic increase in computational complexity. In practical engineering operations, China often uses the power network loss conversion of typical days in four seasons to roughly estimate the annual power network loss. This method is simple but has limited accuracy. At the theoretical research level, probability evaluation methods are highly favored: Reference [4] uses linear fitting to correlate power network loss with node injection power, and approximately derives the probability distribution of power network loss during the evaluation period; Reference [5] uses Monte Carlo sampling to construct power grid operation scenarios, and then calculates the probability distribution of power network loss through power flow calculation. Although these methods have high accuracy, their computational efficiency is limited by the scale of the power grid and the amount of data. Additionally, linear fitting methods are not suitable for long-term evaluation as their linear error increases with the extension of evaluation time. In addition, reference [6] explored the power network loss method based on load measurement and theoretical calculation, and reference [7] proposed the theoretical power network loss calculation of low-voltage distribution area based on load electricity quantity and Newton Raphson method. Both are based on user energy meter data, highly dependent on network structure, line specifications, and other information. Regression analysis is also applied to the calculation of distribution area power network loss, but the establishment of regression equations is difficult and the calculation accuracy is insufficient. In recent years, the development of neural network theory has opened up a new path for calculating power network loss rate, which does not require the construction of mathematical models. With strong learning ability, generalization ability, and nonlinear processing ability, it can effectively fit the relationship between power network loss rate and feature parameters. However, current research mainly focuses on the calculation of power network loss in 10kV distribution networks, and there are still few applications for calculating power network loss rates in distribution areas [8-9].

This article focuses on the challenge of evaluating power network loss in large-scale datasets and proposes an innovative evaluation strategy that integrates hybrid clustering techniques. Cluster analysis, as a key branch of data mining, has shown extensive potential in the fields of new energy forecasting, load estimation, and electricity consumption pattern identification in the electricity industry. The core of this strategy is to first decompose the original clustering problem into multiple sub problems based on the differences in numerical types of clustering attributes; Subsequently, for each sub problem, carefully select suitable clustering algorithms for in-depth analysis; Finally, using clustering ensemble techniques, mixed clustering results are integrated. This strategy is particularly suitable for processing power grid datasets with large volumes and diverse data types. Its mixed clustering results can be extracted through random sampling techniques to obtain a small number of representative operating scenario sets, which can be used for power network loss assessment, significantly improving the efficiency of assessment in massive data environments. To verify the effectiveness and practicality of the strategy, this article conducted an in-depth analysis using an actual power grid case as the test object.

2. Principles of power network loss clustering evaluation

2.1. *Algorithm framework*

In response to the challenge of evaluating power network loss in large-scale datasets, this article focuses on data closely related to the distribution of power network loss from a data mining perspective. An innovative hybrid clustering strategy is used to quickly identify and partition typical power network loss feature patterns hidden in massive data. This strategy ensures that running data within the same cluster group exhibit similar power network loss distribution characteristics, while significant differences are observed between different cluster groups [10]. Subsequently, samples were extracted from each clustering group to construct a set of typical operating scenarios, and efficient approximation evaluation of power network loss was achieved using power flow calculation technology. The core of this strategy lies in accurately identifying clustering features that can map the relationship between power grid operation data and power network loss values, while selecting appropriate clustering algorithms to balance clustering accuracy and computational efficiency.

Figure 1 visually illustrates the implementation steps of this power network loss assessment method. Subsequent chapters will elaborate on the specific characteristics of power network loss assessment, including cluster feature analysis, deconstruction and solution strategies for clustering problems, as well as power network loss assessment methods based on clustering results.

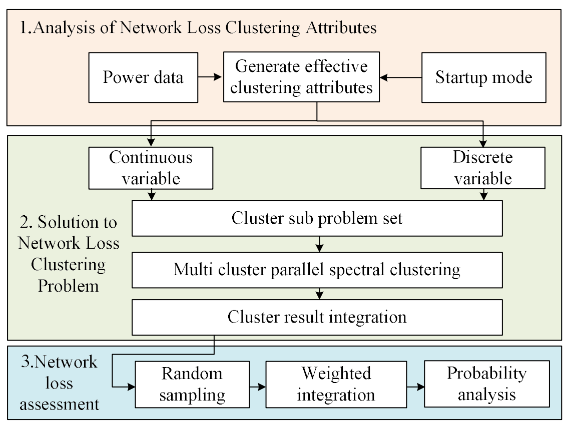

Fig. 1. Power network loss evaluation process based on hybrid clustering technology

2.2. *Attribute Analysis*

Based on the fundamental theory of power flow calculation [11-12], the core influencing factor of power network loss distribution is the power injection situation of each node. In the power grid operation dataset, the clustering elements closely related to power network loss include total load, transmission power of interconnection lines, and output power of generator nodes. However, if these power values are directly adopted as clustering elements, the increase in the number of grid nodes will lead to a significant leap in clustering dimensions, forming complex high-dimensional clustering challenges. This not only increases the computational burden, but may also dilute the core role of total load and tie line power in power network loss assessment, leading to a deviation in clustering effectiveness from the original intention.

To address the above challenges, this article proposes an innovative solution strategy. Considering that the sum of total load and connecting line power is approximately equivalent to the total power generation output, and the output power of generator nodes can be regarded as a comprehensive reflection of the distribution pattern of total power generation output and the output of each node, this paper decides to use the distribution pattern of output of each node as the clustering element, rather than directly using node output data. This distribution pattern is approximately characterized by the unit start stop status information of nodes. The specific implementation method is to convert the node output data into binary encoding form. The specific criteria for the conversion will be explained in detail later.

1) If $P_i \geq a P_i^{max}$ is satisfied, it can be inferred that node i is in a high output situation. Processing its power data yields 100 transcoded data.

2) If $P_i \geq P_{min}$ and $P_i < a P_i^{max}$ are satisfied, it can be inferred that node i is in a medium to low output state. Processing its power data yields transcoded data 010.

3) If $P_i < P_{min}$ is satisfied, it can be inferred that node i does not have any output situation and can be set to 0 to process its power data and obtain transcoding data 001.

Where, P_i is the power output of node i during normal operation; P_i^{max} is the maximum output power of node i during normal operation; a is a boundary coefficient used to distinguish the high and low output power of nodes. According to relevant standards, the value of the boundary coefficient is generally between 65% and 75% of the rated power. Here, a compromise is taken to set the value of a as 70%; P_{min} is the minimum output power of a node during normal operation, and P_{min} is set to 35% of the rated power according to relevant standards.

Based on specific conversion principles, this article converts the output power dataset of N nodes into a binary sequence consisting of 3N bits, which comprehensively reflects the start stop status of the entire power grid's units. It successfully achieves the transformation from N-dimensional clustering features to single dimensional clustering features, significantly simplifying the complexity of the clustering process. Based on the above considerations, the clustering features selected in the power network loss assessment in this article specifically include: total load, transmission power of interconnection lines, and the start stop status information of the entire power grid units (represented in 3N bit binary form).

2.3. Index system

To optimize the effectiveness of clustering evaluation, it is necessary to reduce the dimensions of input variables. Based on the convenience of obtaining indicators, this article selects core electrical characteristic parameters closely related to the distribution area power grid architecture and load [13]. Specifically, the characteristics of the power grid architecture include the power supply radius and the total length of low-voltage lines, while the load characteristics cover the load rate and the proportion of residential electricity consumption (hereinafter referred to as residential electricity consumption ratio, to replace the expression of electricity nature and proportion).

1) Power supply radius X_1(m). X_1 is the total length of the load node located in the most remote location of the distribution area and the transformer to be studied and analyzed, which represents the rationality of the grid structure.

2) The total length of the low-voltage line is X_2(m). X_2 refers to the total length of low-voltage lines installed within the distribution area.

3) Load rate X_3(%). X_3 is the ratio of the total amount of electricity supplied in the power system to the capacity of the distribution transformer, which represents the average load within the distribution area.

4) Electricity usage ratio X_4(%). X_4 is the ratio of the total electricity usage to the total electricity supply in the power system, which characterizes the nature of electricity consumption within the distribution area.

The above four electrical characteristic indicators constitute the input variable system for distribution area analysis. Given the differences in the dimensions and ranges of values of each indicator, in order to ensure that the calculation process is not affected by this, it is necessary to standardize and preprocess the original data.

The specific process of standardizing the raw data given the number of feature parameters m and the total number of samples N is as follows [14]:

$$Z_{ij} = x_{ij} - \bar{x}_j / \sqrt{s_{ij}} \tag{1}$$

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} \tag{2}$$

$$s_{ij} = \frac{1}{N-1}\sum_{i=1}^{N} \left(x_{ij} - \bar{x}_j\right)^2 \tag{3}$$

Where, s_{ij} is the variance of the original data x_{ij}; Z_{ij} is the data obtained by standardizing the original data x_{ij}; $\bar{x}_j$ is the mean of the original data x_{ij}.

3. Multi cluster parallel spectral clustering algorithm

This section elaborates on the implementation details of the Multi cluster parallel spectral clustering algorithm, which mainly includes three components: preprocessing noise, sub cluster segmentation, and inter cluster similarity evaluation. The preprocessing noise stage aims to eliminate abnormal sample points that may mislead the results in advance. In the sub cluster segmentation process, the algorithm incorporates the MKN mechanism to optimize sub cluster generation, effectively avoiding the interference of neighboring density differences on segmentation performance. The evaluation of sub cluster similarity is the core of Multi cluster parallel spectral clustering algorithm, which directly affects the quality of the final clustering results. At the end of this section, we provided pseudocode for the Multi cluster parallel spectral clustering algorithm and conducted an in-depth analysis of its computational complexity.

3.1. Noise preprocessing

In multi cluster parallel spectral clustering, the local density of the original data samples is obtained based on k-nearest neighbors [15]:

$$\rho_i = \frac{k}{\sum_{j \in knn(i)} d(i,j)} \tag{4}$$

Where, $knn(i)$ is the k-nearest neighbor of the i-th original data sample.

Given that noise in the dataset may interfere with clustering performance, this article pre eliminates low-density data points based on equation (3) at the beginning of the clustering process. These points are identified and removed based on specific density thresholds.

$$\rho_q = \bar{\rho}_X - m \times \sigma_X \tag{5}$$

Where, the density mean and standard deviation of all points in the dataset are represented as $\bar{\rho}_X$ and σ_X, respectively. According to the "3σ principle" of Gaussian distribution, we have set the boundary range $\{1,2,3\}$ of noise coefficient to effectively distinguish low-density data points at different density levels. It should be emphasized that since noise points often have low density characteristics, high-density outliers were not taken into account when setting the density threshold in equation (5). Points below this density threshold will be considered noise and will not participate in the subsequent sub cluster partitioning process.

In response to the issue of the need to preset the initial number of clusters, this paper adopts the method of calculating the overall contour coefficient S_t of the clustering results to automatically determine the optimal k value. The contour coefficient, as an indicator for evaluating the effectiveness of clustering, the larger its value, the better the clustering effect. For sample point i, the specific calculation method for its contour coefficient is as follows [16]:

$$S(i) = \frac{p(i) - q(i)}{\max\{q(i), p(i)\}} \tag{6}$$

Where, $q(i)$ is the average distance between sample point i and the other sample points within its category; $p(i)$ is the average distance from sample point i to other sample points outside its category. The mean silhouette coefficient can be calculated as:

$$S_t = \frac{1}{N}\sum_{i=1}^{N} S(i) \tag{7}$$

This article innovatively proposes a selection strategy to address the challenge of selecting initial cluster centers. Firstly, a distribution area evaluation metric named P_E was constructed, with the specific definition as follows:

$$P_E = \sqrt{\sum_{j=1}^{m} w_j \left(Z_{ij} - Z_{j\min}\right)^2}, \quad i = 1,2,\cdots,N \tag{8}$$

Where, $Z_{j\min} = \min_{i\in[1,N]} Z_{ij}$, $j = 1,2,\cdots,m$; w_j has a value of 1, indicating the weight configuration of the jth indicator.

Construct the minimum electrical characteristic parameter vector for distribution area samples, which is composed of the minimum values of power supply radius parameters, total length of low-voltage lines, electrical load conditions, and residential electricity consumption ratio. In addition, P_E is defined as the Euclidean distance between the electrical characteristic parameter vector of a certain sample and the minimum electrical characteristic parameter vector mentioned above.

3.2. *Subcluster division*

After removing low-density data points, the sub cluster is composed of all data points that share the same identification point, which is initialized according to the following definition:

Definition 1: Representing point indicators. Its initial state is set to an empty set, $Rep(i) = 0$. If $j \in knn(i) \wedge i \in knn(j)$ and $\rho_j > \rho_i$, then $Rep(i) = \mathrm{argmin}_j\, d(i,j)$ can be obtained.

According to the local density characteristics determined by formula (3) and the representative point principle described in Definition 1, each sample belongs to the subgroup to which its representative point belongs. If condition $Rep(i) = 0$ is met, select that point as the core and construct a new subpopulation. It is worth noting that the uniqueness of Definition 1 lies in the fact that its representative point must be located within the k-nearest neighbor range of that point, which satisfies the condition $j \in knn(i) \wedge i \in knn(j)$. This feature enables the Multi cluster parallel spectral clustering algorithm to demonstrate superiority in handling adjacent subpopulations with uneven density distribution. If Definition 1 involves the condition $i \in knn(j)$, then the sample and its representative point form a k-nearest neighbor relationship, otherwise, only the k-nearest neighbor relationship is maintained between the two. Based on this, we label the presence or absence of the $i \in knn(j)$ attribute in Definition 1 with mutual k-nearest neighbor relationships and k-nearest neighbor relationships, respectively.

3.3. *Subcluster similarity*

After completing the subpopulation partitioning, the Multi cluster parallel spectral clustering algorithm needs to evaluate the similarity between any two subpopulations in order to guide the subsequent merging steps. The newly introduced similarity evaluation index focuses on the relative ratio of density between subgroups rather than absolute values. Figure 2 shows examples of subpopulation partitioning with varying densities on the Pathbased dataset [17], where marker 01 is located in the intersection area of low-density subpopulations, while marker 02 is located at the edge of subpopulations with different densities. According to equation (2), under the premise of controlling for other variables being consistent, the higher the density at the intersection of subgroups, the shorter the distance between two subgroups, and the higher the similarity. Therefore, adjacent subgroups in region 02 are more likely to be merged. Based on this, this article compared the density of the intersection area of subpopulations with adjacent areas to determine the relative ratio of density similarity rather than absolute values. The subsequent definition elaborates on the specific calculation method of similarity between subgroups.

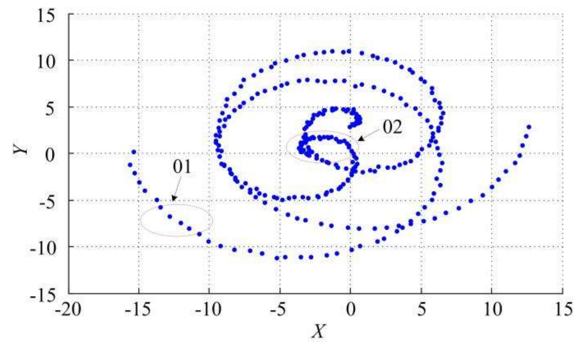

Fig. 2. Similarity of Intersection between Subclusters of Different Density

Definition 2: Intersection factor indicator. Let CE_i be the extended set of spectral clustering algorithm sub cluster C_i, which is the union of sample points in spectral clustering sub cluster C_i. The intersection factor of spectral clustering sub clusters C_i and C_j can be calculated as follows [18]:

$$perc_{ij} = sign_{ij} \times sign_{ji} \times \frac{|CE_{ij}|}{|C_i| + |C_j|} \tag{9}$$

Where, CE_{ij} is the intersection of the spectral clustering sub cluster extension set CE_i and CE_j. $sign_{ij} = sign\left(|CE_i \cap C_j|\right)$ and $sign(\cdot)$ are sign functions. If $x > 0$ is satisfied, then $sign(x) = 1$ can be obtained; When $x = 0$, $sign(\cdot) = 0$.

Definition 3: Connectivity metric. Based on the mutual distance between sample points C_i and C_j in spectral clustering, their connectivity can be defined as:

$$con_{ij} = mean(\{d(p,q)\}) \tag{10}$$

Where, $p \in CE_i \cap C_j, q \in CE_j \cap C_i$。

Definition 4: Density distance similarity index.

$$pavg_{ij} = \frac{\min\left(\bar{r}_{c_i}, \bar{r}_{c_c}, \bar{r}_{CE_g}\right)}{\max\left(\bar{r}_{c_i}, \bar{r}_{c_j}, \bar{r}_{CE_g}\right)} \tag{11}$$

Where, $\bar{r}_s$ is the mean of all density values between data points in the sample dataset s.

Definition 5: Similarity index of density standard deviation.

$$std_{ij} = \frac{\min\left(s_{c_i}, s_{c_j}, s_{CE_j}\right)}{\max\left(s_{c_i}, s_{c_j}, s_{CE_g}\right) + s_{c_g}} \tag{12}$$

Where, s_s is the standard deviation of all density values between data points in the sample dataset s, and set C_{ij} is the union of spectral clustering sub clusters C_i and C_j.

Then, the formula for calculating the spacing between spectral clustering sub clusters can be obtained as follows:

$$newd_{ij} = perc_{ij} \times con_{ij} \times \left(1 - pavg_{ij}^2\right) \times \left(1 - std_{ij}\right) \tag{13}$$

When calculating the distance between sub clusters in equation (13), only scenes where adjacent sub clusters overlap are marked as scene $perc_{ij} > 0$. For sub clusters that are not directly connected, their distances are not calculated directly, but are derived using the Floyd algorithm based on the distance information of adjacent clusters. Set the geodesic distance between sub clusters obtained through equation (14) as metric $newd_{ij}$, and then use Gaussian kernel function to derive the similarity SIM_{ij} between clusters. The detailed mathematical expression can be found in equation (15).

$$\mathrm{SIM}_{ij} = \exp\left(-\left(\frac{newgd_{j}}{\sigma}\right)^2\right) \tag{14}$$

$$\sigma = \mathrm{mean}\left(\left\{\mathrm{newd}_{ij} \mid \mathrm{perc}_{ij} > 0\right\}\right) \tag{15}$$

3.4. *Algorithmic program*

Algorithm1:Multi cluster parallel spectral clustering algorithm。

Input: Sample set $X \in \mathbb{R}^{n \times m}$, number of sub clusters $nclust$, nearest neighbor parameter k of clustering algorithm, noise factor μ;

Output: Cluster output $\{S_i\}, i = 1, \ldots, nclust$;

Step 1: Normalize sample X by performing the following operations on the v-dimensional features of the data sample: $x_{.v} = (x_{.v} - \min(x_{.v}))/(\max(x_{.v}) - \min(x_{.v}))$;

Step 2: Determine the initial number of clusters and initial cluster centers using equations (6-8);

Step 3: Using equation (4), the local density index of the data sample can be obtained;

Step 4: Using equation (5) combined with the set coefficient μ, the density threshold ρ_q for data sample sub cluster partitioning can be obtained, and the sample points $\{i \mid \rho_i < \rho_q, i = 1, \ldots, n\}$ located in low-density areas can be removed;

Step 5: For the remaining sample points with high-density indicators after removing the low-density sample points mentioned above, the updated local density of the samples can be obtained using equation (3);

Step 6: According to Definition 1, the representative point $Rep(i)$ of the input sample data can be obtained, and based on this, the sub clusters of the sample can be calculated;

Step 7: Using equation (14), the similarity index SIM_{ij} between the updated sample sub clusters can be obtained;

Step 8: Use the similarity matrix obtained above to update the cluster labels and obtain $label(C_i)$;

Step 9: Subclusters with the same label can be merged due to their high similarity, i.e. $x_p \in C_i$, $x_q \in C_j$. If $label(C_i) = label(C_j)$ is satisfied, $label(x_p) = label(x_q)$ can be obtained;

Step 10: Return the clustering output $\{S_i\}$.

3.5. *Computational complexity*

Under the framework of Algorithm 1, spectral clustering technology relies on the k-means method to partition the reduced dimensional data and obtain clustering results. Due to the randomness of this process, multiple executions may result in inconsistent clustering results. To address this issue, we adopted a fixed random seed strategy in the experimental design to ensure that the k-means algorithm produces the same output results every time it is executed, while keeping other variables constant.

In the framework of Algorithm 1, the first three steps constitute the preprocessing stage, which aims to standardize data and eliminate low-density noise. Its computational complexity is mainly affected by nearest neighbor search and is marked as complexity $O(n\log(n))$. Although the operation of step 4 is the same as that of step 2, it applies to the denoised dataset, so the complexity is slightly reduced, approximately to complexity $O(n\log(n))$ (due to the small proportion of noise). Steps 5 and 6 focus on calculating the similarity between sub clusters: the complexity of step 6 depends on the number of sub clusters $nsc(<<n)$, denoted as complexity $O(nsc^3) + O(nsc^2)$; Step 5 is responsible for finding representative points, with a complexity of $O(nk)$. The core of steps 7 and 8 is to perform spectral clustering, with an overall complexity of $O(nsc^3)$. In summary, the overall time complexity of Algorithm 1 is $O(2n\log(n)) + O(nsc^3) + O(nsc^2) + O(nk) + O(nsc^3) \sim O(2n\log(n))$, $nsc, k << n$. This indicates that the Multi cluster parallel spectral clustering algorithm ensures efficient computational performance with its low complexity.

In the execution process of multi cluster parallel spectral clustering, the main consumption requirements are focused on preserving k-nearest neighbor information and sub cluster similarity matrices, which determines the spatial complexity of multi cluster parallel spectral clustering to be $O(nk)+O(nsc^2)$.

4. Experimental analysis

4.1. Experimental setup

To verify the effectiveness of the aforementioned method, we selected a low-voltage distribution area of a provincial power grid as an example and collected a total of 8748 hours of operational data throughout the year to evaluate network losses. The power grid structure includes 421 nodes, 568 network branches, 42 power generation nodes, 93 load nodes, and 6 connection channels. We comprehensively evaluated the computational performance by analyzing key indicators such as the accuracy, convergence speed, and stability of the power network loss rate calculation.

Firstly, 590 samples were selected, each containing 5 independent variables ($x_1(m)$: power supply radius; $x_2(m)$: total length of low-voltage lines; $x_3(\%)$: load rate; $x_4(\%)$: proportion of residential electricity consumption; $d(\%)$: power network loss rate) and 1 dependent variable. Use optimized clustering algorithm to process these 590 standardized samples. Calculate the performance indicator P_E for standardized data and arrange the samples in ascending order of P_E values. The results are shown in Table 1, with P_E values ranging from 0.608 to 11.479.

Table 1. Initial Cluster Center Evaluation Index (PE)

Distribution area number	Value
1	0.608
2	0.673
3	0.765
...	...
589	10.037
590	11.479

We gradually increased the initial number of clusters k, trying from 3 to 8 one by one, and calculated the total silhouette coefficient S_t of the clustering results at each k value. The results are summarized in Table 2. Through comparative analysis, it was found that when k is set to 7, the total contour coefficient reaches its maximum value, indicating that the clustering effect is optimal at this time. Based on this, this article determines the initial number of clusters to be 7.

Table 2. Clustering silhouette coefficient test

k	silhouette coefficient	k	silhouette coefficient
3	0.3965	6	0.4087
4	0.4192	7	0.5196
5	0.3841	8	0.4468

According to P_E sorting, divide the samples into six groups, and select the central sample as the initial clustering core for each group. Please refer to Table 3 for details. Next, perform k-means clustering analysis. The number of samples included in each clustering group is as follows: 151 in the first group, 302 in the second group, 7 in the third group, 37 in the fourth group, 85 in the fifth group, and 8 in the sixth group, totaling 590.

Table 3. Initial Cluster Centers

No.	Cluster center (initial)				PE value
	x1/m	x2/m	x3/%	x4/%	
1	125	861	87.15	6.38	2.25
2	208	1449	82.46	8.56	1.96
3	197	1873	91.64	7.47	3.11
4	382	1736	91.83	14.61	3.76
5	266	1058	96.85	6.29	3.81
6	590	4219	79.34	10.45	4.69

4.2. Error result analysis

Based on the six cluster analyses mentioned above, this study decided to extract 10 data points from each cluster, totaling 60 samples, to form a representative set of operating scenarios for evaluating network loss. To verify the accuracy of the proposed power network loss assessment technique, we calculated the overall power network loss rate of the system and the loss of overloaded lines based on 8748 hours of annual power flow data, which served as benchmark reference values. Meanwhile, we adopt the following two methods for comparative analysis [19-20]:

1) Seasonal representative day method: Following engineering practice conventions, this study selects January 15th, April 15th, July 15th, and October 15th as representative days of the four seasons each year. Through trend analysis of 96 time points on each of these four dates, the power network loss data of each is obtained to summarize the characteristics of power network loss throughout the year.

2) Linear approximation method: Firstly, we calculate the average power flow state of the power grid; Next, using sensitivity analysis techniques, determine the linear relationship coefficient between node injection power and network loss; Finally, based on these coefficients, approximate the linearized power network loss of the power grid under different operating states, and complete the analysis of the total power network loss consumption.

Table 4 shows the comparison of the average power network loss rate of the power grid under different methods. The results show that the average power network loss rate obtained by clustering method is closest to the benchmark value, with an error rate of only 0.11%.

Table 4. Mean Power Network Loss Rate

Method	Relative error/%	Mean power network loss rate/%
Benchmark prediction model	0.00	1.163
Proposed method	0.11	1.175
Seasonal representative day method	4.86	1.107
Linear approximation method	22.69	0.914

In Figure 3, we present the cumulative probability distribution curves of the three different methods for evaluating power network loss rate, for a more in-depth comparative analysis.

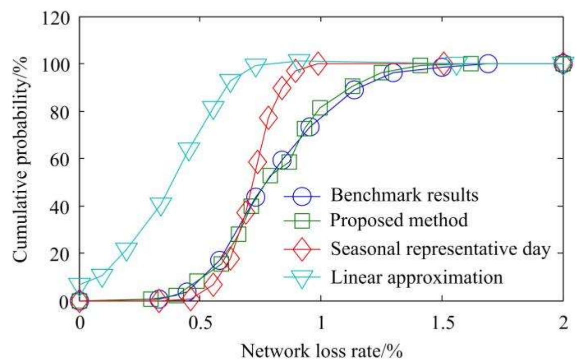

Fig. 3. Cumulative probability curve of power network loss rate

Table 5 lists the accuracy metrics of three methods in evaluating the probability distribution of power network loss rate, including correlation factors with benchmark results and root mean square error (RMSE), to measure the performance of each method.

Table 5. Evaluation Results of Power Network Loss Rate

Method	RMSE/%	Correlation factor
Proposed method	1.42	0.9994
Seasonal representative day method	12.69	0.9852
Linear approximation method	32.85	0.8967

According to the results in Figure 3, the cumulative probability curve of power network loss rate obtained by clustering method is highly consistent with the benchmark curve; Table 4 shows that the correlation coefficient between the power network loss rate distribution of clustering method and the benchmark distribution is as high as 0.9994, and the root mean square error (RMSE) of cumulative probability is as low as 1.42%. Given that the typical operating mode set is constructed based on random sampling and may affect the accuracy of power network loss assessment, this study conducted 55 repeated power network loss rate analysis experiments. The results showed that the average annual power network loss rate of the power grid in the 55 experiments reached 1.1473%, with a deviation of only 0.05% from the benchmark results. The RMSE of the probability distribution of power network loss rate was only 2.74%, which was better than the typical daily method and linearization method. These experimental results strongly demonstrate that the power network loss evaluation method proposed in this paper has high stability and accuracy, and is less affected by random sampling.

4.3. Analysis of Power Network Loss Results

For a heavy-duty transmission line in the power grid, this paper conducted a power network loss assessment and compared the annual average power network loss results of different methods, as shown in Table 6. Analysis shows that the average power network loss obtained by clustering method is closest to the benchmark value, with a relative error evaluation index of only 1.05%, which is lower than other algorithms.

Table 6. Results of Average Power Network Loss

Method	Relative error/%	Mean power network loss/MW
Benchmark prediction model	0.00	1.9527
Proposed method	1.05	1.9537
Seasonal representative day method	25.69	1.4698
Linear approximation method	25.78	1.4852

Figure 4 shows the cumulative probability curves of three power network loss evaluation methods, with power network loss ranging from 0 to 12MW. Table 7 provides a detailed list of various indicators used to evaluate the accuracy of the probability distribution of power network loss. These two sets of data together reveal the excellent accuracy of clustering methods in evaluating power network loss. In order to gain a deeper understanding of the potential impact of random sampling on the accuracy of power network loss assessment, this study carefully designed 55 independent power network loss analysis experiments. Experimental data shows that the average power network loss of the overloaded line accurately reaches 1.9537MW, with a slight deviation from the benchmark value of only 1.05%. The mean root mean square error (RMSE) of the power network loss probability distribution is also controlled at an extremely low 3.47%. This series of experimental results fully confirms that clustering methods also have excellent stability and evaluation accuracy in the field of power network loss analysis.

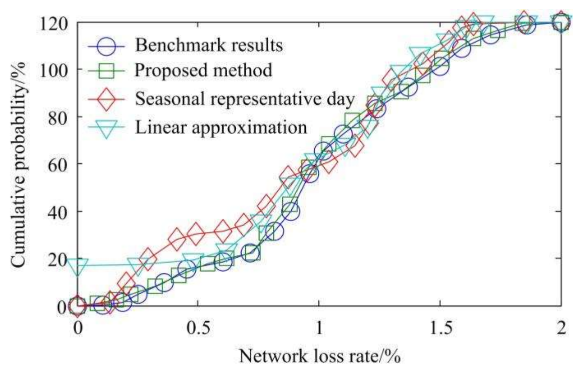

Fig. 4. Cumulative probability curve of power network loss

Table 7. Evaluation Indicators for Power Network Loss

Method	Mean square error/%	Correlation factor
Proposed method	1.43	0.9997
Seasonal representative day method	8.26	0.9876
Linear approximation method	8.17	0.9893

Previous studies have shown that the power network loss assessment method introduced in this article is easy to operate. By using cluster sampling techniques to obtain 60 representative operating mode datasets, it can efficiently replace 8748 hours of operating data throughout the year for power network loss assessment, demonstrating excellent computational speed and assessment accuracy.

5. Conclusion

This study focuses on the challenges of power network loss assessment in large-scale datasets and innovatively proposes a power network loss assessment strategy that integrates hybrid clustering techniques. This strategy focuses on the dataset of power grid operation modes, taking into account the uniqueness of power data and using a unique hybrid clustering algorithm for in-depth analysis. Based on the clustering results, a typical set of operation modes is extracted to support power network loss assessment. To verify the practicality of this method, this study introduced an actual power grid case, and the results showed that compared with traditional methods, the hybrid clustering analysis driven power network loss evaluation not only significantly improved accuracy, but also had a simple implementation process and excellent computational efficiency. In addition, the hybrid clustering technique has shown great potential for expansion, and its application exploration in power data intensive fields such as new energy grid integration planning and grid economic operation evaluation will become one of the key directions of this study in the future.

References

[1] Liu Y, Ćetenović D, Li H, et al. An optimized multi-objective reactive power dispatch strategy based on improved genetic algorithm for wind power integrated systems[J]. International Journal of Electrical Power & Energy Systems, 2022, 136: 107764.

[2] Gu J, Cao X Y, Fu Y, et al. Experimental measurement-device-independent type quantum key distribution with flawed and correlated sources[J]. Science Bulletin, 2022, 67(21): 2167-2175.

[3] Yang C, Yao W, Wang Y, et al. Resilient event-triggered load frequency control for multi-area power system with wind power integrated considering packet losses[J]. IEEE Access, 2021, 9: 78784-78798.

[4] Anteneh D, Khan B, Mahela O P, et al. Distribution network reliability enhancement and power loss reduction by optimal network reconfiguration[J]. Computers & Electrical Engineering, 2021, 96: 107518.

[5] Yang L, Wang R, Zhou Y, et al. An Analytical Framework for Disruption of Licklider Transmission Protocol in Mars Communications[J]. IEEE Transactions on Vehicular Technology, 2022, 71(5): 5430-5444.

[6] Long H, Chen C, Gu W, et al. A data-driven combined algorithm for abnormal power loss detection in the distribution network[J]. IEEE Access, 2020, 8: 24675-24686.

[7] Kano N, Tian Z, Chinomi N, et al. Comparison of renewable integration schemes for AC railway power supply system[J]. IET Electrical Systems in Transportation, 2022, 12(3): 209-222.

[8] Lee H Y, Lee J, Kim N W, et al. Comparative analysis of photovoltaic performance metrics for reliable performance loss rate[J]. IET Renewable Power Generation, 2023, 17(4): 1008-1019.

[9] Bento M E C. Physics-guided neural network for load margin assessment of power systems[J]. IEEE Transactions on Power Systems, 2023, 39(1): 564-575.

[10] Kunya A B, Abubakar A S, Yusuf S S. Review of economic dispatch in multi-area power system: State-of-the-art and future prospective[J]. Electric Power Systems Research, 2023, 217: 109089.

[11] Bretas A S, Rossoni A, Trevizan R D, et al. Distribution networks nontechnical power loss estimation: A hybrid data-driven physics model-based framework[J]. Electric Power Systems Research, 2020, 186: 106397.

[12] Wang H, Huang Z, Zeng X, et al. Enhanced anticarbonization and electrical performance of epoxy resin via densified spherical boron nitride networks[J]. ACS Applied Electronic Materials, 2023, 5(7): 3726-3732.

[13] Rawa M, Alkhalaf S, Mihet-Popa L, et al. An Efficient Scheme for Determining the Power Loss in Wind-PV Based on Deep Learning[J]. IEEE Access, 2020, 9: 9481-9492.

[14] Revanesh M, Gundal S S, Arunkumar J R, et al. Artificial neural networks-based improved Levenberg–Marquardt neural network for energy efficiency and anomaly detection in WSN[J]. Wireless Networks, 2024, 30(6): 5613-5628.

[15] Chen Z, Amani A M, Yu X, et al. Control and optimisation of power grids using smart meter data: A review[J]. Sensors, 2023, 23(4): 2118-2129.

[16] Fernando X, Lăzăroiu G. Spectrum sensing, clustering algorithms, and energy-harvesting technology for cognitive-radio-based internet-of-things networks[J]. Sensors, 2023, 23(18): 7792-7806.

[17] Mahdavi M, Javadi M S, Catalão J P S. Integrated generation-transmission expansion planning considering power system reliability and optimal maintenance activities[J]. International Journal of Electrical Power & Energy Systems, 2023, 145: 108688.

[18] Naderi E, Mirzaei L, Trimble J P, et al. Multi-objective optimal power flow incorporating flexible alternating current transmission systems: application of a wavelet-oriented evolutionary algorithm[J]. Electric Power Components and Systems, 2024, 52(5): 766-795.

[19] Nazir Z, Bollen M. Operational risk assessment of transmission Systems: A review[J]. International Journal of Electrical Power & Energy Systems, 2024, 159: 109995.

[20] Anteneh D, Khan B, Mahela O P, et al. Distribution network reliability enhancement and power loss reduction by optimal network reconfiguration[J]. Computers & Electrical Engineering, 2021, 96: 107518.