An Improved Cuckoo Search Clustering Method for Line Loss Data of Transformer District with DGs

Lingang Yu¹, Dongwen Wu¹, Zhiqiang Hu², Aiqing Yu³*, Liang Zhu¹, and Jun Wang¹

¹State Grid Jiangxi Electric Power Research Institute, 330096, Nanchang, China
²Jiangxi Electric Power Co, Ltd, State Gride, 330077, Nanchang, China
³College of Power Engineering, Shanghai University of Electric Power, Shanghai 200090, China

* Corresponding author: yuaiqing@shiep.edu.cn

Abstract: For the low-voltage transformer district with distributed generations (DGs), the traditional theoretical calculation method of line loss is not applicable. This paper presents a novel clustering method for line loss data of transformer district with DGs, which combined an improved Cuckoo Search algorithm and K-Means clustering algorithm. Firstly, the influence factors of line loss are screened based on the maximum information coefficient, and the line loss index system is established. Secondly, an improved cuckoo search clustering algorithm is proposed to cluster the sample data set to reduce the dependence on the initial clustering center. Finally, the simulation results of 410 samples from a certain area with photovoltaic power supply show the accuracy and effectiveness of the proposed method. The simulation results show that the proposed method is accurate and effective.

1. Introduction
Reducing the line loss in the low voltage platform area of distribution network is conducive to energy conservation and emission reduction. It is of great significance to promote the low-carbon development of China's energy enterprises and achieve the double carbon goal. In view of the difficulty in calculating the theoretical line loss value of fine transformer district and the access of distributed power generation, the line loss calculation based on big data mining and intelligent algorithm is the main research direction at present. In reference [1-2], K-means clustering algorithm is used to cluster the line loss set in the transformer district, and the cluster data samples are modeled by BP, RBF neural network and other methods. However, the results of K-means algorithm depend on the initial clustering center, and the improper setting of BP network structure parameters is easy to lead to local convergence. Literature [3-4] combines the cuckoo search algorithm with k-means algorithm to optimize the selection of initial clustering center, so as to greatly improve its clustering effect and accelerate its convergence. Literature [5-7] proposes expert sample database, feature selection method and deep learning to calculate line loss. However, the above literature is highly subjective in the selection of line loss data set in the transformer district, and does not consider the selection basis of electrical characteristic indexes containing distributed power generation in the current transformer district, and the amount of calculation is large. Therefore, based on the correlation analysis method of maximum information coefficient (MIC), this paper analyzes and screens the main factors affecting the line loss in the low-voltage transformer district of distributed generation access, and proposes an improved cuckoo search based K-means clustering.
algorithm (ICS kmeans), which defines the weighted Euclidean distance. The arctangent function is used to define the adaptive step size and the adaptive nest elimination probability, which ensures the diversity among populations and improves the clustering accuracy. Finally, the effectiveness of the proposed algorithm is verified by using actual line loss data in a certain region.

2. MIC-Based Selection of Impact Factors for Line Loss in Active Low-Voltage Transformer District

2.1 Correlation Analysis Based on the Maximum Information Coefficient

Experts and scholars such as Reshef first proposed the concept of maximum information coefficient [8] in 2011, which is based on the theory of normalized mutual information and information entropy. MIC can measure any kind of correlation between different random variables. In the process of analyzing the correlation based on big data, it is found that the correlation between the line loss influencing factors is relatively complex and nonlinear, and conventional methods cannot describe it. Therefore, the correlation analysis based on MIC in the article selects the influence factors of the line loss in the active low-voltage transformer district.

MIC uses mutual information and meshing methods to calculate. Let $X$ and $Y$ be the random variables in the data set $D$, where $X = \{x_i\}, Y = \{y_i\}, i = 1, 2, 3, \ldots, n$ where, $n$ represents the number of samples. Divide the data set $D$ into a grid of $G = (a, b)$. Under different divisions, the maximum mutual information is taken and further normalized to obtain the maximum information coefficient, whose definition is shown in equation (1).

$$
\text{MIC}(X,Y) = \max_{G} \sum_{x \in X} \sum_{y \in Y} \log \frac{p(x,y)}{p(x)p(y)}
$$

Where $p(x,y)$ is the joint probability density of $X$ and $Y$, and $p(x)$ and $p(y)$ are the marginal probability densities of $X$ and $Y$ respectively.

It can be seen from equation (1) that the greater the MIC value, the greater the correlation of the variables, and vice versa.

2.2 Selection of Impact Factors for Line Loss in Active Low Voltage Transformer District

In this paper, the following 13 characteristic indexes are collected as the basic indexes to construct the characteristic system of the line loss index in the active transformer district by combining the traditional calculation formula of the line loss theory, the distributed power model and the data of the electricity information acquisition system.

1. Grid type indicators: line length, power supply radius;
2. Capacity indicators: total number of users, number of distributed power users;
3. Power indicators: total meter forward and reverse active power, power generation of distributed power users, power supply and power sales;
4. Operational indicators: head voltage, ending voltage, power factor of the total meter, three-phase unbalance degree.

The MIC described in Section 2.1 is used to carry out correlation analysis of influence factors and dimension reduction screening for the actual line loss data of several low-voltage transformer district with PV(Photovoltaic) in a certain region. The comparison results of the MIC and Spearman rank correlation coefficients are shown in Figure 1. (Spearman rank correlation coefficient method is a commonly used method to analyze the linear correlation between two variables). It can be seen from the figure that the indicators within the shadow threshold ($\pm 0.1$) show that there is no linear correlation between the factors and the line loss, and 9 indicators will be eliminated. In fact, there is a nonlinear correlation between the line loss and the impact factor. Therefore, MIC can be used to analyze the nonlinear correlation between the impact factor and the line loss rate and determine the importance of the impact factor.
The MIC is further used to discriminate the non-linear redundancy characteristics among impact factors. It can be seen from the thermal map of the impact factor that No.7 factor is redundant with No.12 and No.13 factor. No.12 and No.13 factor are redundant. With reference to Figure 1, No.12 and No.13 factor are eliminated, and a characteristic system of line loss indicators in the active low-voltage transformer district composed of 11 impact factors has been established to pave the way for the subsequent refined line loss calculations.

3. An Improved Cuckoo Search Clustering Method for Line Loss Data of Transformer District with DGs

3.1 Clustering process of line loss data in active low voltage transformer district
The dimensions of electrical characteristic parameters of line loss data set are different, and zero-mean normalization is adopted to normalize the data.
Where, $M$ is the median; $\sigma$ is the absolute standard deviation.

Based on the normalized data set, the definition of MIC in Section 2.1 is firstly used to analyze the correlation between line loss influence factors and line loss for screening, so as to establish the characteristic system of line loss indicators in the active transformer district. Then, according to the data set of the index characteristic system, different types of sub-data sets are obtained by ICS-Kmeans clustering. The process is shown in Figure 3.

3.2 An Improved Cuckoo Clustering Algorithm

The normalized platform samples are clustered by data driven, which provides the basis for realizing the fine line loss calculation. Aiming at the problem that the clustering effect depends on the initial clustering center, the article proposes an improved cuckoo search clustering algorithm. The main improvements are:

(1) Initial population generation: each bird's nest represents a solution, that is, a set $C, C \in R^{k \times P}$ of $k$ cluster centers, and $P$ is the number of impact factors. Each cluster center of each population was randomly generated in the sample data set, and it is used as the initial nest to reduce the sensitivity of K-means to single-group initial clustering and ensure diversity..

(2) Fitness calculation: According to the correlation analysis of line loss influencing factors based on MIC in Section 2.1, the weighting coefficient $w_p$ of the $p$-th impact factor is defined as the ratio of the MIC value of this factor to the sum MIC of all factors, as shown in equation (3). The K-Means algorithm based on weighted Euclidean distance is presented to perform clustering calculation, as shown in equation (4). SSE (the sum of error squares) is selected as the criterion function, and it is defined as a fitness function to evaluate each bird's nest, as shown in equation (5).

$$w_p = \frac{MIC_p}{\sum_{i=1}^{P} MIC_i}$$

$$\text{SSE} = \sum_{i=1}^{N} \sum_{j=1}^{K} d_{ij}^2$$

$$w_p = \frac{MIC_p}{\sum_{i=1}^{P} MIC_i}$$

(3)
dist(d_i, d_j) = \sqrt{\sum_{p=1}^{P} w_p (d_{ip} - d_{jp})^2} 

\text{(4)}

SSE = \sum_{i=1}^{k} \sum_{d \in E_i} \text{dist}(e_i, d)^2 

\text{(5)}

Where \( E_i \) is the \( i \)-th cluster; \( e_i \) is the cluster center of the cluster; \( d \) is the sample data in the cluster.

(3) Adaptive levy flight update: Perform adaptive levy flight on the cluster center represented by each bird's nest, and update its position according to equation (6)

\[
\begin{align*}
C_{t+1}^{p_0} &= C_{t}^{p_0} + \alpha \left( C_{t}^{p_0} - C_{t}^{\text{best}} \right) \odot L(\lambda) \\
pop_i &= 1, 2, \ldots, \text{popsize}
\end{align*}
\]

\text{(6)}

\[
\alpha = \frac{(\alpha_{\text{min}} + \alpha_{\text{max}}) \cdot \arctan(t - \frac{\text{maxiter}}{2})}{\pi}
\]

\text{(7)}

\[
L(\lambda) \sim u = t^{-\lambda} (1 < \lambda \leq 3)
\]

\text{(8)}

Where \( C_{t}^{p_0} \) is the combination of the cluster centers of the \( t \)-th generation of the \( \text{pop}_i \)th bird's nest; \( \odot \) is the point-to-point multiplication; \( C_{t}^{\text{best}} \) represents the optimal solution of the \( t \)-th generation; \( \alpha \) is the variable used to control the adaptive step size, expressed by equation (7). Its shape is an arctangent function, and different step lengths are used to search at different stages, which can improve the search accuracy and prevent trapping into a local optimal solution. The step length changes range is \((\alpha_{\text{min}}, \alpha_{\text{max}})\), and the maximum number of iterations is \( \text{maxiter} \); \( L(\lambda) \) is a random search vector, it obeys the Levy distribution of the parameters \((1 < \lambda \leq 3)\), as shown in equation (8).

(4) Adaptive bird’s nest elimination probability \( p_a \) : its form is also an arctangent function, which increases the elimination probability in the later stage of the algorithm and ensures the diversity of the population. As shown in equation (9), the bird’s nest elimination probability range is \((P_{a_{\text{min}}}, P_{a_{\text{max}}})\).

\[
\begin{align*}
P_a &= \frac{(P_{a_{\text{min}}}, P_{a_{\text{max}}}) \cdot \arctan(t - \frac{\text{maxiter}}{2})}{\pi} \\
&= \frac{P_{a_{\text{min}}} + P_{a_{\text{max}}}}{2}
\end{align*}
\]

\text{(9)}

The flow of the ICS-Kmeans algorithm is shown in Figure 4.
4. Case analysis

In this paper, sample data of 410 typical load days containing photovoltaic low-voltage stations in a certain area are used to cluster the line loss data and compare with other algorithms to verify and analyze the scientific validity of the proposed method from multiple angles.

4.1 Clustering algorithm performance analysis

The performance of the proposed clustering algorithm is analyzed below, setting the parameters maxiter=100, the number of bird nests popsize=20, the range of discovery probability (0.25, 0.55), and the range of step change (0.0001, 1).

From a statistical point of view, select different $k$ values, run 50 times respectively, and use ICS-Kmeans based on traditional Euclidean distance criterion function and classic K-Means algorithm and basic cuckoo K-Means clustering algorithm (CS-Kmeans) for clustering. The comparison results are shown in Table 1. It can be seen from Table 1 that when the $k$ value is small, the average value of the ICS-Kmeans algorithm is better than or equal to the other two algorithms. As the value increases, the average difference gradually increases, which shows that the ICS-Kmeans algorithm has improved the problem that the results of the K-Means algorithm are too dependent on the initial cluster centers. The comparison result of the optimal value shows that the ICS-Kmeans algorithm is better at solving local optimal problems.

| $k$ | average criterion function | optimal criterion function |
|-----|-----------------------------|-----------------------------|
|     | K-Means | CS-Kmeans | ICS-Kmeans | K-Means | CS-Kmeans | ICS-Kmeans |
| 2   | 1956.8  | 1956.7   | 1956.7   | 1956.7  | 1956.7   | 1956.7   |
| 3   | 1618.3  | 1513.4   | 1513.4   | 1513.4  | 1513.4   | 1513.4   |
| 4   | 1413.9  | 1297.7   | 1297.3   | 1299.2  | 1297.2   | 1297.2   |
| 5   | 1178.9  | 1097.7   | 1097.7   | 1128.3  | 1097.7   | 1097.7   |
| 6   | 1007.8  | 942.5    | 941.8    | 950.7   | 941.5    | 941.4    |
| 7   | 897.6   | 818.9    | 818.7    | 829.8   | 818.6    | 818.6    |
| 8   | 801.1   | 755.3    | 754.4    | 762.1   | 743.2    | 742.3    |
The discrete point detection graph is used to analyze the influence of the weighted Euclidean distance proposed in the article on the clustering effect. Under the same value selection (for example $k=3$), set the detection threshold to 15 and compare the discrete points of the weighted Euclidean distance and the traditional Euclidean distance, as shown in Figure 5 (a) and (b). The method proposed in this article combines the data-related characteristics, and the clustering effect is better.

![Figure 5. Sample discrete point detection diagram when $k=3$](image)

4.2 Determination of clustering parameters

In this paper, the clustering parameter values are determined based on the best initial classification map of the elbow rule. Take the average of the fitness function under different values to draw the best initial classification map, as shown in Figure 6. In the graph, it can be seen that as the value of the initial classification becomes larger, the average value drops rapidly. After the inflection point shown in the figure, the descending speed begins to slow down, and the inflection point is the best initial classification. For this practical example, $k=3$ is the best initial classification.

![Figure 6. The best initial classification diagram](image)

When $k=3$, the fitness of the ICS-Kmeans algorithm increases as the evolutionary algebra increases and its changing trend is shown in Figure 8. It can be seen from Figure 7 that within the first 5 generations of the evolution, the algorithm has obtained the clustering optimization results, and the solution efficiency is relatively high. The results of the clustering algorithm output, the percentage of samples of type 1 is 3%, the percentage of samples of type 2 is 17%, and the percentage of samples of type 3 is 80%. The clustering results are visualized by t-SNE (t-distributed stochastic neighbour embedding), and the final display results are shown in Figure 8.
5. Conclusion

In this paper, for the analysis of line loss in the active low-voltage transformer district, MIC is adopted to screen the influence factors of line loss, and the characteristic system of line loss index in the active transformer district is constructed. A clustering algorithm based on improved cuckoo search is proposed for line loss analysis in active low-voltage transformer district. The diversity of the population is improved by improving the adaptive bird nest selection probability, and the adaptive step size is proposed to enhance the ability of the algorithm to jump out of the local optimal solution. The clustering algorithm reduces the complexity of line loss data analysis, provides technical support for the subsequent line loss analysis of low-pressure platform in complex environment, makes the line loss analysis more refined, and improves the effectiveness of line loss management.

Acknowledgments

This paper is supported by Science and Technology Project of State Grid Headquarters, No.5600-20191916 8A-0-0-00.

References

[1] Li Ya, Liu Liping, Li Baiqing, et al, “Calculation method of transformer district line loss rate based on improved k-means clustering and BP neural network”, Chinese Journal of electrical engineering, vol. 36, no. 17, pp. 4543-4552, 2016.

[2] Deng Peng, Liu min, “research on line loss calculation of substation power grid based on improved clustering and RBF neural network”, Smart power, vol. 49, no.02, pp. 107-113, 2021.

[3] Yang Huihua, Wang Ke, Li Lingqiao, et al, “K-means clustering algorithm based on adaptive cuckoo search algorithm and its application”, Computer application, vol. 36, no.08, pp. 2066-2070, 2016.

[4] Wang Bo, Yu Xiangjun, “Parallel K-means clustering algorithm for adaptive cuckoo search”, Computer application research, vol. 35, no.03, pp. 675-679, 2018.

[5] Ding Zhongan, Gao Chen, Jiang Minmin, et al, “Distribution network line loss rate prediction model based on expert sample database and least squares support vector machine”, Hydropower and energy science, vol. 38, no.03, pp. 195-198 + 194, 2020.
[6] Zhong Xiaoqiang, Chen Jie, Jiang Minmin, et al, “Line loss analysis method in substation area based on deep learning”, Power grid technology, vol. 44, no.02, pp. 769-774, 2020.

[7] Deng Wei, Guo Yixiu, Li Yong, et al, “Distribution network loss prediction based on feature selection and stacking integrated learning”, Power system protection and control, vol. 48, no.15, pp. 108-115, 2020.

[8] RESHEF D N, RESHEF Y A, FINUCANE H K, et al, “Detecting novel associations in large datasets”, Science, vol. 334, pp. 1518–1524, 2011.