Scenario analysis of wind power output based on improved k-means algorithm

Weiyuan Wang¹, Fei Dou¹, Xuan Yu² and Gaowei Liu²,³
¹ State Grid Jiangsu Electric Power Co., LTD.; ² East China Electric Power Design Institute Co., LTD. of China Power Engineering Consulting Group ³Email: liugw2775@ecepdi.com

Abstract. To ensure the safety and reliability of grid operation, an accurate description of wind power output is crucial. In this paper, the Grey Wolf Optimization algorithm (GWO) optimized by the Cuckoo Search algorithm (CS) is proposed to improve the traditional k-means clustering algorithm and the improved k-means algorithm is applied to the scenario analysis of wind power output. The simulation results show that the error between the typical scene and the initial scene is only 4.56% when reduced by the improved k-means algorithm, so the typical scene not only maintains the temporal sequence of the initial scene, but also fits the output fluctuation and randomness of the initial scene well.

1. Introduction
The fluctuation and intermittent behavior of wind power has strong uncertainty. Its impact on power reliability, power quality, economy and social welfare becomes more prominent as its penetration rate increases [1]. The volatility of wind power can lead to voltage fluctuations and flicker. A large amount of distributed wind power is connected to the distribution network via converters, especially under high wind speed uncertainty, which has a significant impact on power quality issues such as surges, voltage drops, voltage flicker, harmonics, and transient power supply interruptions. Furthermore, the uncertainty of wind power increases the demand for spare capacity at all levels, especially Frequency Modulation (FM) and rotating spare [2-4]. Therefore, describing the intermittent output of wind power is an urgent problem to be solved in the power planning stage.

Scenario analysis theory is a commonly used method to describe random problems [5]. The typical scene set obtained through scenario reduction can retain its function of characterizing the probability distribution and randomness of random vectors. There have been some studies on scenario reduction methods. Literature [6] uses Fortet-Mourier type probability metric as an indicator to manage power load through scenario reduction; Literature [7] proposes an alternative procedure that effectively reduces the number of uncertain scenes in trade issues related to the electricity market; Literature [8] is based on the Wasserstein probability distance index, and uses the improved K-medoids parallel clustering algorithm for scenario fusion to obtain the typical scene set.

K-means algorithm [9-10] is a typical and common clustering algorithm that is widely used in scenario reduction. The traditional k-means algorithm may fall into a local optimum due to the random selection of the initial clustering center, which may not achieve an optimal solution globally, making the final clustering effect compromised.
In this paper, we propose an improved k-means clustering algorithm based on the CS-GWO algorithm. The general idea of it is to use the global search process of GWO to filter the optimal initial clustering center, and at the same time use CS to optimize GWO, in order to improve its global search capability and avoid falling into dimensional catastrophe and local optimization.

The main contributions of this paper are: (1) The global search capability of GWO and CS are used to improve the traditional k-means clustering algorithm, optimize the selection of initial clustering centers, and improve the accuracy of the clustering results to characterize the characteristics of initial random variables. (2) The improved k-means clustering algorithm is applied to scenario reduction to realize the scenario analysis of wind power output, and the effectiveness of the method is verified through examples.

2. Improved k-means clustering algorithm based on CS-GWO

2.1. Traditional k-means algorithm

The main idea of the k-means algorithm is to divide the \( n \) samples into categories according to their distance \( d \) from the feature space. The closer the distance (usually referred to as Euclidean distance), the closer the properties of the samples in the feature space, and the greater the similarity between the two samples, which can be grouped into one category. After several iterations of filtering calculations, a sample set of \( k \) types is finally formed. The objective function of k-means algorithm is to minimize the squared sum of intra-cluster errors, as shown in equation (1).

\[
\min \sum_{i=1}^{k} \sum_{x \in c_i} \text{dist}(x, \mu_i)^2
\]  

(1)

The ideal clustering result is a high similarity between similar intra-cluster scenes and a low similarity between inter-cluster scenes, which manifested in the feature space is the smallest Euclidean distance between intra-cluster scenes and the largest Euclidean distance between inter-cluster scenes. The traditional k-means algorithm has certain limitations in the selection of the initial clustering center, and it is easy to fall into the local optimum. So it is necessary to optimize the selection method of the initial clustering center to improve the quality of the k-means algorithm and achieve the optimal final clustering result.

2.2. Improved k-means algorithm

2.2.1. Grey Wolf Optimization (GWO).Grey wolf Optimization (GWO) [11] is a group search intelligence algorithm proposed by simulation based on the predation behavior of grey wolves in the biological world. GWO divides wolf packs into different classes. In the whole process, \( \alpha \) wolf, \( \beta \) wolf and \( \delta \) wolf dominate hunting (optimization), and \( \omega \) wolf moves under the leader's instruction. The mathematical model is shown in equations (2)-(6).

\[
X(t + 1) = X_p(t) - A \left| C \cdot X_p(t) - X(t) \right|
\]  

(2)

\[
A = 2\alpha \cdot L_f - \alpha
\]  

(3)

\[
C = 2 \cdot L_f
\]  

(4)

\[
D_\alpha = |C_1 \cdot X_a - X|
\]  

(5)

\[
D_\beta = |C_2 \cdot X_\beta - X|
\]  

(6)

\[
D_\delta = |C_3 \cdot X_\delta - X|
\]  

\[
X_1 = X_a - A_\alpha \cdot D_\alpha
\]  

(7)

\[
X_2 = X_\beta - A_\beta \cdot D_\beta
\]  

(8)

\[
X_3 = X_\delta - A_\delta \cdot D_\delta
\]  

(9)

\[
X(t + 1) = (X_1 + X_2 + X_3) / 3
\]  

(10)
where $A$ and $C$ are vectors of coefficients of the model, $X$ and $X_p$ are vectors of grey wolf and prey positions respectively, $t$ represents the number of current iterations; the value of $a$ decreases linearly from 2 to 0; $l_1, l_2 \in \text{random}[0,1]$; $D_{\alpha}, D_{\beta}$, and $D_\delta$ represent the distance between the next grey wolf to be updated and the three dominant wolves respectively.

2.2.2. Cuckoo Search algorithm (CS). Cuckoo Search algorithm [12] is also a bionic heuristic search algorithm, which is based on the parasitic characteristics of cuckoo nest search, combined with the Levy flight characteristics in the flight search process. CS has a strong global optimization capability. The flight path of the cuckoo to find the most suitable nest for parasitism is consistent with the Levy flight, which can be expressed by equation (7):

$$X_{i+1} = X_i + a \odot \text{Levy}(\beta)$$

where $a$ is the random factor; $\text{Levy}(\beta)$ describes the Levy flight and is simulated using the Mantegna method. The formula for the Mantegna method is shown in equation (8):

$$s = \frac{u}{|v|^\beta}$$

where $u \sim N(0, \sigma^2)$, $v \sim N(0, 1)$, $\sigma = \left\{ \Gamma(1+\beta)\sin(\frac{\pi \beta}{2})/2^{\frac{\beta-1}{2}}\beta(1+\beta)^{\frac{1}{2}} \right\}^{\frac{1}{2}}$.

When the host bird finds the cuckoo’s parasitic egg with a probability of $P_h$, the cuckoo relocates the parasitic nest. This process can be described in a stochastic manner, as shown in equation (9):

$$X_{i+1} = X_i + r \odot \text{Heaviside}(P_h - \delta) \odot (X_i - X_j)$$

where $r$ and $\delta$ are random numbers, $\text{Heaviside}(\cdot)$ is the jump function, $X_i$ and $X_j$ are the nests at other locations.

2.2.3. Improvement of k-means algorithm based on CS-GWO. After many iterations of GWO, the parameters that determine its position tend to fall into the local optimum, resulting in the algorithm failing to achieve the global optimum. CS relies on Levy flight and cuckoo search to effectively achieve the global range of search, and can avoid the local optimum caused by a single parameter. Therefore, the two algorithms can be fused, using GWO to optimize the choice of the initial clustering center and then CS to address the shortcomings of GWO falling into the local optimum.

In combination with CS, after initializing the wolf pack position, the Levy flight can be used to update the positions of the dominant wolves $\alpha$, $\beta$ and $\delta$ to ensure that their updated positions are optimal within the current iteration step, i.e. to determine the initial host nest. Afterwards, consider the case where the prey finds the wolf pack during the hunt, i.e., the host finds the cuckoo’s parasitic eggs, and randomly update the grey wolf’s position again to avoid falling into the local optimal. According to the ideas above, the CS-GWO algorithm flows as follows:

step1. Set the maximum number of iterations $T$ and other parameters, select the adaptation function, and initialize the positions of $\alpha$ wolf, $\beta$ wolf and $\delta$ wolf.

step2. Calculate the value of the wolf pack adaptation function, perform stratification, and update the positions of $\alpha$ wolf, $\beta$ wolf and $\delta$ wolf.

step3. Use Levy flight to update the positions of $\alpha$ wolf, $\beta$ wolf and $\delta$ wolf, and update wolf pack positions using the GWO algorithm.

step4. Determine if the prey has found the wolves, if so, randomly update the location of the grey wolves, if not, go to step5.

step5. Repeat step2–step4 until the maximum number of iterations $T$ is reached, then output the final optimal position of $\alpha$ wolf.

Therefore, the CS-GWO algorithm can be used to find the optimal initial clustering center for the k-means algorithm to obtain the best clustering effect.
3. The process of scenario analysis of wind power output

3.1. The process of scenario selection
The output of wind power is obviously intermittent and fluctuating, and is greatly influenced by the season in the long time scale. Assume that the initial scene set is \( U = \{U_1, U_2, ..., U_N\} \), and the individual scene \( U_n (n = 1, 2, ..., N) \) is one day's output data. If there are \( m \) samples in a day, then \( U_1, U_2, ..., U_N \) is an \( N \times m \)-dimensional vector, so the scene set \( U \) can be represented by an \( N \times m \) matrix.

The corresponding initial scene sets are obtained after the preliminary division of \( U \) according to the seasons. The initial scene sets of different seasons are reduced by the improved k-means algorithm to obtain the \( k \) types of scenes, and then the typical scene set \( \{C_1, C_2, ..., C_k\} \) is obtained after merging various types of clustering centers \( C_1, C_2, ..., C_k \). Since the temporal order and probability distribution characteristics of the scene data are maintained during the reduction and merging process, \( C \) can reflect the temporal order characteristics and probability density distribution characteristics of the original scene set with a small number of samples.

3.2. Indicator for evaluating the effectiveness of clustering
The value of scene category \( k \) is crucial when selecting typical scenes through clustering. The silhouette coefficient can be introduced as a metric to evaluate the effect of clustering. In the optimal case, the intra-cluster distance is minimum and the inter-cluster distance is maximum, that is, each cluster after clustering is sufficiently compact and sufficiently dispersed between clusters. The formula for calculating the silhouette coefficient for individual sample points is shown in equation (10):

\[
SC_i = \frac{q - p'}{\max\{p', q'\}}
\]  

where \( p' \) is the average distance from a point within a cluster to other points within the cluster, \( q' \) is the average distance from that point to all other points within the cluster outside the cluster.

The mean value of the \( SC \) of all sample points is called the average clustering silhouette coefficient \( SC_k \), which can be used to evaluate the clustering effect. Its calculation formula is shown in (11):

\[
SC_k = \frac{1}{n} \sum_{i=1}^{n} SC_i
\]

where \( k \) is the number of corresponding clustering centers and \( n \) is the number of sample points.

3.3. The flow of scenario analysis based on improved k-means algorithm
When using the CS-GWO algorithm to find the optimal initial clustering center for k-means, the adaptation function of the GWO algorithm is set to a function related to the evaluation of the k-means clustering effect. In this way, the optimal position of \( \alpha \) wolf obtained after many iterations can be used as the optimal initial clustering center for k-means to obtain the best clustering effect. According to the analysis above, the flow of scenario analysis of wind power output is shown in Figure 1.

![Figure 1. The process of scenario analysis.](image-url)
4. Simulation and analysis

A 365-day actual wind power output dataset for an onshore 100MW wind farm in Wyoming, USA for 2018 was obtained by querying the National Renewable Energy Laboratory website, with a sampling step of 1 hour. In this paper, a total of 92 days in the summer (June to August) with 2208 sampling points are selected for scenario reduction. The wind power output data in summer is standardized and the output curve is shown in Figure 2. Using the clustering method mentioned above, the scenario reduction is carried out by evaluating the indicator $SC$ and selecting the clustering center $k=6$ to reduce the summer output curve of the wind farm shown in Figure 2. The relationship between the indicator $SC$ and the number of clustering centers $k$ is shown in Figure 3.

The results of the improved clustering method proposed in this paper are compared with the clustering results of the traditional k-means algorithm, as shown in Figure 4 and Figure 5. Figure 4 shows the clustering results of the traditional k-means algorithm. It can be seen from Figure 4 that the difference between Category 1 and Category 2 output curves is not obvious. Although the output curve of Category 1 in Figure 4 is similar to that of Category 6 in Figure 5, the degree of curve aggregation is lower than the latter. Also in Figure 4, the low number of force curve scenes in Category 6 is similar to the partial force scenes in Category 5, indicating that there is little distinction between the two categories. On the whole, the traditional k-means algorithm has poor clustering effect.
From Figure 5, it can be seen that improved k-means algorithm based on CS-GWO can effectively achieve clustering. Among them, the first category of output curve shows the phenomenon of output increase after 12:00, which is related to the increase of wind speed at night. But the overall output is small, the output is not higher than 0.5 p.u., and the output is close to zero during the period from 0:00 to 12:00, so the wind power utilization value is not high. The second category of output curve has less fluctuation throughout the day, and the 24-hour output is close to 0. The third category of output curve shows obvious anti-peak characteristics, reaching the output trough during 6:00-12:00, and gradually approaching the peak during 12:00-18:00 when the output increases, and the output fluctuates greatly throughout the day, which is easy to affect the stability of the power grid. The fourth category of output curve also presents obvious anti-peak characteristics, but it is different from the third category of output curve in that the output is larger in the time period of 0:00-6:00 and fluctuates more throughout the day. The overall output of category 5 output curve is larger, the time of output more than 0.5 p.u. is more than 50% of the 24 hours in a day. At the same time, it can realize full power continuous output in 14:00-20:00, which belongs to the high output situation. The sixth category of output curve has obvious anti-peak characteristics, reaching the all-day trough at 6:00-12:00, then rising continuously to reach the peak, but the all-day output does not exceed 0.5 p.u.. At this time, the wind power output is weak, which is a low output situation.

The cumulative distribution function is used to evaluate whether the typical scenes selected after scene reduction can effectively represent the probabilistic nature of the initial scenes. The typical scenes selected by improved k-means algorithm based on CS-GWO are shown in Figure 6, and the probability of each typical scene is shown in Table 1. The cumulative probability distribution curves for the initial scenes and the typical scenes selected by the traditional and improved k-means algorithm are shown in Figure 7.

Table 1. Probability of different typical scenes.

| Type of scene | 1    | 2    | 3    | 4    | 5    | 6    |
|---------------|------|------|------|------|------|------|
| Probability   | 0.17 | 0.16 | 0.33 | 0.08 | 0.08 | 0.18 |

Figure 6. Six typical scenes selected by improved k-means algorithm.

Figure 7. The cumulative distribution curves of the typical scene and the initial scene.

It can be seen from Figure 7 that the cumulative probability curves of the initial scene and the typical scene follow the same trend and are very close to each other, indicating that the difference between them is very small, and thus the typical scene can be a good fit to describe the wind power fluctuations of the initial scene. The area between the cumulative probability curves of the initial scene and the typical scene is used to quantitatively describe the clustering effect of the k-means algorithm, that is, the error rate of the clustering result is equal to the error area divided by the total area under the cumulative probability curve of the initial scene. It is calculated that the error rate between the typical scene selected by the traditional k-means algorithm and the initial scene is 9.45%, while the error rate between the typical scene selected by the improved k-means algorithm based on CS-GWO and the initial scene is 4.46%. In addition, the error rates between the initial scene and the typical scene
selected by improved k-means algorithm based on only GWO and only CS are calculated, which are 6.73% and 7.98%, respectively. Thus, the error rate of the clustering results of the CS-GWO based k-means algorithm is 52.8% lower than that of the traditional k-means algorithm, and therefore the clustering effect is significantly improved and is better than the clustering effect improved by only GWO or CS. In summary, compared with the traditional k-means clustering algorithm, the improved k-means clustering algorithm based on CS-GWO can reduce scenes more effectively, and the reduced and merged typical scenes can maintain the temporal sequence of the initial scenes, and can fit the output fluctuations and randomness of the initial scenes well.

5. Conclusions
In this paper, an improved k-means clustering algorithm based on CS-GWO algorithm is proposed and applied to the scenario analysis of wind power output. The main conclusions are as follows:

1) The Cuckoo Search algorithm can optimize the Grey Wolf Optimization algorithm to avoid the local optimum due to a single parameter and thus achieve the global optimum.

2) The improved k-means algorithm based on CS-GWO optimizes the selection of initial clustering centers, improves the accuracy of the clustering results in characterizing the initial random variable characteristics, and reduces the error of the clustering results by 52.8% compared to the traditional k-means algorithm.

3) The proposed improved algorithm selects a typical scenario with an error of only 4.46% from the initial scene, thus the typical scene can effectively describe the wind power while maintaining the probability distribution and timing characteristics of the initial output.

References
[1] Ciupgeanu D A, Lzroiu G, Barelli L 2019 Wind energy integration: Variability analysis and power system impact assessment Energy 185
[2] Chen W H, Chen C Y, Huang C Y, et al. 2017 Power output analysis and optimization of two straight-bladed vertical-axis wind turbines Applied Energy 185(1) 223-232
[3] Wang Bin, Dong Xinzhou, Pan Zhencun, et al. 2011 Analyzing voltage sag propagation characteristics of distribution systems with wind power Automation of Electric Power Systems, 35(10) 53-58. (in Chinese)
[4] Xue Yusheng, Lei Xing, Xue Feng, Yu Chen, Dong Chaoyang, Wen Fushuan, Ju Ping 2014 A Review on Impacts of Wind Power Uncertainties on Power Systems Proceedings of the CSEE, 34(29) 5029-5040
[5] Hsia P, Samuel J 1994 Formal Approach to Scenario Analysis IEEE Software 11(2) 33-41
[6] Heitsch H 2003 Werner Römisch. Scenario Reduction Algorithms in Stochastic Programming Computational Optimization & Applications 24(2) 187-206
[7] Morales J M, Pineda S, Conejo A J, et al. 2009 Scenario Reduction for Futures Market Trading in Electricity Markets IEEE Transactions on Power Systems 24(2) 878-888
[8] Wang Qun, Dong Wenzhe, Yang Li 2015 A Wind Power/Photovoltaic Typical Scenario Set Generation Algorithm Based on Wasserstein Distance Metric and Revised K-medoids Cluster Proceedings of the CSEE 35(11) 2654-2661 (in Chinese)
[9] Hartigan J A, Wong M A 1979 A K-Means Clustering Algorithm Applied Stats 28(1)
[10] Wang Baoshi, Duan Zhiqiang, Qu Denghui 2010 Application of k-means Algorithm Based on Effective Index to Identifying Bad Data of the Power System Northeast Electric Power Technology 31(03) 16-18+31
[11] Mirjalili S, Mirjalili S M, Lewis A 2014 Grey Wolf Optimizer Advances in Engineering Software 69 46-61
[12] Amir Hossein Gandomi, Xin-She Yang, Amir Hossein Alavi 2013 Erratum to: Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems Engineering with Computers 29(2) 245-245