scientific reports

Check for updates

OPEN

Distribution network line loss analysis method based on improved clustering algorithm and isolated forest algorithm

Jian Li[1]✉, Shuoyu Li[2], Wen Zhao[1], Jiajie Li[1], Ke Zhang[1] & Zetao Jiang[1]

The long-term loss of distribution network in the process of distribution network development is caused by the backward management mode of distribution network. The traditional analysis and calculation methods of distribution network loss can not adapt to the current development environment of distribution network. To improve the accuracy of filling missing values in power load data, particle swarm optimization algorithm is proposed to optimize the clustering center of the clustering algorithm. Furthermore, the original isolated forest anomaly recognition algorithm can be used to detect outliers in the load data, and the coefficient of variation of the load data is used to improve the recognition accuracy of the algorithm. Finally, this paper introduces a breadth-first-based method for calculating line loss in the context of big data. An example is provided using the distribution network system of Yuxi City in Yunnan Province, and a simulation experiment is carried out. And the findings revealed that the error of the enhanced fuzzy C-mean clustering algorithm was on average − 6.35, with a standard deviation of 4.015 in the situation of partially missing data. The area under the characteristic curve of the improved isolated forest algorithm subjects in the case of the abnormal sample fuzzy situation was 0.8586, with the smallest decrease, based on the coefficient of variation, and through the refinement of the analysis, it was discovered that the feeder line loss rate is 7.62%. It is confirmed that the suggested technique can carry out distribution network line loss analysis fast and accurately and can serve as a guide for managing distribution network line loss.

Keywords Fuzzy C-Means, Isolated forest algorithm, Medium voltage distribution networks, Line loss analysis, Data processing

The Line Loss (LL) ratio is a crucial index for measuring the operating efficiency and economy of a power system. It represents the proportion of electrical energy lost due to the presence of components such as resistors and inductors during power transmission. The level of LL rate directly affects the safe and stable operation and economic benefit of the power grid. A current distribution network's LINE LOSS ANALYSIS (LLA) primarily relies on the expertise and professional judgement of specialists, which has a limited impact on improving the LL management level of the network[1,2]. Currently, there are two general routes for research on loss reduction in medium voltage distribution networks (MVDN) both domestically and internationally. The first is the study of power equipment, which aims to lower LL by producing more energy-saving equipment for cooperation. The study of LL after new energy is allowed access to the distribution network, the benchmarking value for LL, LL management, and LL causes are the key topics of the second area of research on theoretical LL[3]. In response to the country's calls for energy efficiency and emission reduction, the use of new energy technologies has gradually increased. In the case of electric power, the presence of numerous distributed photovoltaic power plants and distributed hydroelectric power plants has changed the direction of the original tidal currents, posing a new challenge for the distribution network. Although the efficiency and accuracy of LLA have increased significantly over the past few years due to advancements in power system technology and management, the distribution network LLA still faces some challenges as a result of the complexity of the distribution network's structure, load imbalance, and other factors[4,5]. Therefore, it is crucial to research a technique that can carry out LLA for distribution networks rapidly and accurately. This research innovatively proposes a data cleaning model based on the combination of ensemble learning and optimal clustering, and improves the shortcomings of ensemble

[1]Metrology Center, Guangdong Power Grid Co.,Ltd., Guangzhou 511545, China. [2]Power Supply Service, Dongguan Power Supply Bureau, Dongguan 523576, China. ✉email: honeyluyawahaha@163.com

learning and optimal clustering to enhance the accuracy and practicability of the data cleaning model. On this basis, based on a large number of existing data, a multi-level distribution network LL calculation model is constructed to obtain the fine LL under different data scenarios. Finally, according to the loss characteristic index and loss rate, the loss causes are determined, and the loss causes are identified for different types of feeders, so as to obtain the main reasons for the high loss rate.

The innovations of this research are as follows: (1) A fuzzy C-means (FCM) clustering algorithm based on random distributed delayed Particle Swarm Optimization (RODDPSO) algorithm is proposed, and the clustering center of FCM clustering algorithm is optimized to improve the accuracy of final data filling. (2) On the basis of the original isolated forest anomaly recognition algorithm, by calculating the coefficient of variation of load data, the abnormal subspace is screened to reduce the dimension, and the randomness of the algorithm is reduced by fixing the selection of cutting points, so as to improve the recognition accuracy of abnormal data. (3) The LL calculation model of the backward substitution method is established, which reflects the characteristics of data-driven, and provides real and reliable data support for the research of distribution network.

The contributions of this research are as follows: (1) To solve the problem of missing charge data in the distribution network, PSO algorithm is used to optimize the clustering center of the clustering algorithm, and the randomness and variable inertia weight of particles in the particle swarm optimization algorithm are added to avoid the PSO algorithm falling into the local optimal, and the accuracy of the final data filling is improved. (2) An improved isolated forest algorithm based on the coefficient of variation is proposed to solve the problem of low accuracy in identifying abnormal load data of distribution network caused by high-dimensional data and algorithm instability. (3) Taking full account of the advantages of multi-source data in data filling, the LL calculation model established in this paper is more accurate and more abundant than the traditional method.

The article develops the study through four parts, the first section provides a summary of current LLA research as well as isolated forest algorithm (IFA) and FCM clustering algorithms for distribution networks. The second part is the study of LLA modelling for MVDNs, the third part is the performance validation of the system designed for the study, and the fourth part is the conclusion.

The abbreviation and full name of this research design are shown in Table 1.

Related works

In an effort to reduce the LL rate, many academics have studied LL, one of the primary indicators of power supply firms. W. Hu established an LL assessment system. Firstly, the collected data were subjected to image processing, and then a reasonable LL interval calculation model was established based on convolutional neural network, based on which a loss reduction strategy was formed. After verification, the system can save electricity and improve economic efficiency[6]. Zhang proposed a LL prediction method based on a multidimensional information matrix and a multidimensional attention mechanism for the problem of high energy loss in low-voltage distribution networks. First, the distribution network characteristics and seasonal trend parameters are selected, and then the historical LL data is decomposed by the optimized variational mode decomposition method, and the model relationship between line loss index deviation and line loss deviation is constructed. Finally, the obtained data is input into the LSTNet network with dimensional attention mechanism. The results show that this method has weak hysteresis effect and high prediction accuracy[7]. Tang proposed a short-term LL prediction algorithm based on K-means-LightGBM. Firstly, a data quality evaluation system was established using the Hadoop platform, the feature dimensions with high correlation were normalized, the samples were classified by K-means clustering algorithm, and the model relationship between line loss index deviation and line loss deviation was constructed. Finally, it is verified that the algorithm has higher accuracy and is superior to the traditional algorithm[8]. In order to accurately diagnose abnormal LL, Liu proposed a hybrid clustering and long and short-term memory based scheme for abnormal LL detection in distribution networks. In this method, samples are classified by mixed clustering method, abnormal feeders are detected quickly, and abnormal feeders are predicted and substations under the jurisdiction of abnormal feeders are detected by long and short term memory method. It has been verified that the method can detect LL quickly and effectively[9]. In order to

Full name	Abbreviation
Line Loss	LL
LINE LOSS ANALYSIS	LLA
Medium voltage distribution networks	MVDN
Fuzzy C-means	FCM
random distributed delayed Particle Swarm Optimization	RODDPSO
Isolated forest algorithm	IFA
Particle Swarm Optimization	PSO
Improved isolated forestt	CV-iFores
Root Mean Square Error	RMSE
Mean Absolute Error	MAE
Standard Deviation	SD

Table 1. Correspondence table of abbreviations.

improve the LL calculation methods and management tools for distribution networks, Zhang's team established a simulation and analysis model for distribution networks based on the IEEE 34-node system after considering the impact of distributed PV access. The results indicated that when the access capacity of distributed power was too large, the LL of the system would increase[10].

It becomes quite challenging for people to extract accurate and truly relevant information from the power data due to interference in the data collection, transmission, and storage processes. C. C. Yi et al. aimed to address the issue of local optimization and error identification in traditional FCM clustering methods. They utilized the t-SNE method for reduction and initial clustering center selection, resulting in an improved FCM algorithm that significantly increased clustering accuracy[11]. Ke et al. proposed a high-precision intelligent prediction method based on back-propagation neural network and FCM clustering algorithm to predict the adsorption efficiencies of heavy metals with different biochar properties, and the phases classified the metal adsorption data by FCM algorithm[12]. The Minkowski distance and Chebyshev distance were combined as a measure of similarity in the FCM's clustering process, and then the principal component analysis was used to carry out the dimensionality reduction. S. Surono's team did this in order to address the issue that the FCM algorithm is easy to fall into local optimal solutions. The results showed that the method improved the accuracy of the clustering and optimized the objective function of the FCM[13]. In order to improve the accuracy of abnormal driving behavior monitoring, Wang et al. designed A driver abnormal behavior warning method based on the isolated forest algorithm. Through the analysis of abnormal driving behavior, XGBoost algorithm was used to extract the characteristics of abnormal driving behavior, and a detection model of abnormal driving behavior was established by constructing an isolated forest of abnormal driving behavior. The results show that the method can detect abnormal driving behavior with 98.6% accuracy[14]. A parameter distribution model for feature decomposition and error compensation correction of the specimen seating attitude was developed by N. Pan et al. using a multi-modal elastic-driven adaptive control method. The methodology increased the clustering's accuracy and optimized the FCM's objective function, according to the results. The anomaly detection signals were processed by the IFA, then the trajectory curve profile was extracted using a multi-scale alignment framework, and finally a parameter-sharing concatenated ternary deep learning model for feature tracking and data enhancement strategies was established[15].

In summary, the existing methods are basically to classify the samples, and then use the neural network to construct the model relationship between the line loss index deviation and the line loss deviation. However, in line loss calculation, the deviation coefficient of line loss index is used to calculate the line loss rate, and the data can be obtained is often less, so it is not suitable to use the power flow calculation algorithm which requires high data quantity. Therefore, this paper proposes a solution to solve the missing data and abnormal data, and alleviates the adverse impact of data quality problems on the distribution network analysis and calculation to the greatest extent. Then, based on the breadth-first backward generation accurate line loss analysis and calculation method, by improving the clustering algorithm and IFA, it is expected to quickly and effectively deduce the LL rate in the distribution network and the reasons for LL.

LLA modeling study of MVDN

The study establishes missing value filling based on improved clustering algorithm and outlier identification based on improved isolation forest algorithm to identify outliers and fill in missing values to reduce the impact of dirty data on subsequent LLA calculation. Then, the cleaned data are used to calculate the LL rate as a theoretical basis for distribution network planning. Finally, a feeder LL depletion identification model based on feeder classification is designed to identify the whole feeder LL depletion causes and eliminate irrelevant factors.

Missing data filling based on improved FCM algorithm

The data acquisition system of distribution network is a complex system composed of various sensors, transformers and software. The data acquisition system of distribution network includes dispatching system, production management system, measurement automation system, distribution geographic information system and marketing system. These data systems process data from a large number of different sources, and a large amount of dirty data is inevitably generated in the process of data transmission from tool to tool and from system to system[16,17].

To remove the impurities from the collected data, the data cleansing on the raw data need to be performed, the principle of which is shown in Fig. 1. By using statistical learning, machine learning, deep learning and other methods, with pre-set cleaning rules and strategies, the massive junk data is transformed into data that meets the needs and is of high quality. The degree of data cleaning depends on the adaptive ability of the cleaning methods, rules and strategies.

In the distribution system, the dirty data mainly comes from the dispatching system and the metering automation system, and the dirty data are mainly generated in the following three links: collection, transmission and storage. In the process of data collection, the main cause of dirty data is equipment failure. In the process of data transmission, the unreliable connection between the data acquisition device and the data transmission device, the aging of the transmission device, the instability of the transmission signal and the signal interference are also the main causes of dirty data. During the data storage process, the data storage module needs to convert the data after receiving the data signal from the sensor, and the data is easy to be abnormal during the conversion process. Data missing, data outlier (abnormal), data naming inconsistency, data illegality and data duplication are the most common types of dirty data in the distribution system. After processing the original data of the distribution network system with the research method, the complete, legal and good data with the same name are obtained.

Missing data, as a kind of junk data, has a large impact on the original data. The clustering algorithm is one of the most widely used methods for filling in the missing data based on similarities between the data, but it has some drawbacks, including that the number of clusters to use depends on experience, the cluster centre is prone to falling into local extremes during iteration, and the accuracy of the clustering for high-dimensional data will

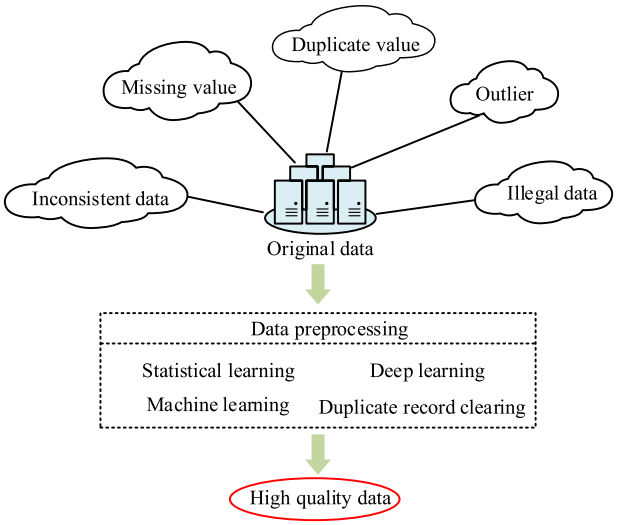


Figure 1. Basic framework of data preprocessing.

suffer, which will have an impact. RODDPSO algorithm replaces the traditional FCM clustering's centre self-renewal process with the clustering center's particle swarm optimization process. In order to acquire a more precise clustering centre for historical data, it is necessary to address the issue that the typical FCM clustering approach tends to slip into the local extreme value during the iterative process.

The selection of k-means algorithm K is difficult to grasp, and it is difficult to converge for non-convex data sets. If the types of data are not balanced, such as the amount of data is seriously unbalanced or the variance of the categories is different, the clustering effect is not good. And it adopts the iterative method, can only get the local optimal solution[18]. FCM takes into account the degree of membership of data points to clusters, and has better global optimization performance. Compared with other clustering algorithms such as K-means, FCM is insensitive to the initial center point and has faster convergence speed, which is suitable for large-scale data sets[19]. For the sample dataset containing $a \times b$, where a is the number of samples and b is the sample dimension, the objective function is minimized by continuously updating the clustering centre and the degree of affiliation of the FCM algorithm. Equation (1) contains the FCM algorithm's mathematical expression.

$$\begin{cases} \min J_{FCM}(U,V) = \sum_{i=1}^{a}\sum_{j=1}^{c} u_{ij}^{m}\left\|x_i - p_j\right\|^2 \\ \sum_{i=1}^{a} u_{ij} = 1, 1 \le j \le c \\ \sum_{j=1}^{c} u_{ij} \ge 0, 1 \le i \le a \\ u_{ij} \ge 1, 1 \le j \le c, 1 \le i \le a \end{cases} \tag{1}$$

In Eq. (1), U is the membership matrix, V denotes the matrix consisting of c cluster centre vectors of dimension b, m denotes the fuzzy factor, which generally takes the value of 2, u_{ij} denotes the element of the affiliation matrix that indicates the degree of affiliation of the ith sample belonging to the jth subclass; x_i denotes the data in the ith sample; p_j denotes the cluster centre of the jth subclass; and $\left\|x_i - p_j\right\|^2$ denotes the Euclidean distance between two vectors. Then the affiliation matrix and clustering centre matrix are updated according to Eq. (2) until the termination condition is satisfied.

$$\begin{cases} u_{ij} = \sum_{k=1}^{c}\left(\frac{\left\|x_i - p_j\right\|}{\left\|x_i - p_k\right\|}\right)^{-\frac{2}{m-1}} \\ p_j = \frac{\sum_{i=1}^{a} u_{ij}^{m}\cdot x_i}{\sum_{i=1}^{a} u_{ij}} \end{cases} \tag{2}$$

Since the Particle Swarm Optimization (PSO) algorithm is also prone to local optima, the study proposes linearly varying inertia weights, whose computational metric expression is shown in Eq. (3).

$$\omega = \omega_{\max} - (\omega_{\max} - \omega_{\min}) \times \frac{t}{t_{\max}} \tag{3}$$

In Eq. (3), ω denotes the inertia weight of PSO algorithm, $\omega_{\max}$ and $\omega_{\min}$ denote the maximum and minimum values of inertia weight, respectively, t denotes the current number of iterations, and $t_{\max}$ denotes the maximum number of iterations. The PSO algorithm needs to set a larger weight in order to speed up convergence at the beginning of iterations, and needs to set a smaller weight at the later stage of iterations in order to prevent the algorithm from skipping the optimal value.

In order to improve the accuracy of the LL rate calculation, the existing grid companies generally set the load data collection once every 15 min, so the daily load profile is composed of 96 points, while in the PSO algorithm, the dimensions of the particles and the cluster centers are both 96. The RODDPSO algorithm differs from the conventional PSO algorithm in that it optimizes the cluster centre matrix. As a result, the particle in the RODDPSO algorithm is a three-dimensional array of size $c \times 1 \times d$.

In Eq. (4), one particle holds the c clustering centre, and during the RODDPSO iteration, the affiliation degree is changed in real time according to the position data of the best clustering centre. By resolving the fitness value of the two particles as the FCM algorithm's objective function until the end of the RODDPSO iteration, a more precise clustering centre position and affiliation degree is obtained. The formula for updating particle velocity and position can be obtained through particle swarm search behavior. To obtain Eq. (4), a random distributed delay term is introduced in the velocity update process.

$$\begin{aligned} v_i(t+1) = {} & \omega v_i(t) + c_1 r_1 \left(p_{best(i)}(t) - x_i(t)\right) + c_2 r_2 \left(g_{best(i)}(t) - x_i(x)\right) \\ & + m_1(\xi) c_3 r_3 \sum_{\tau=1}^{N} \alpha(\tau) \left(p_{best(i)}(t-\tau) - x_i(t)\right) + m_g(\xi) c_4 r_4 \sum_{\tau=1}^{N} \alpha(\tau) \left(g_{best(i)}(t-\tau) - x_i(t)\right) \end{aligned} \tag{4}$$

In formula (4), t represents the current number of iterations; c_1 and c_2 are individual learning factors and social learning factors, respectively. c_3 and c_4 are learning factors of distributed delay terms, whose values are c_1 and c_2, respectively. N indicates the upper limit of the distributed delay item. $\alpha(\tau)$ represents a vector N where each element is selected from 0 to 1; $r_i (i = 1, 2, 3, 4)$ is a random number uniformly distributed in [0.1]; $m_1(\xi)$ and $m_g(\xi)$ represent distributed delay term intensity factors determined by evolutionary state ξ.

The degree of affiliation is the degree of similarity between the sample data and the clustering centre, so it can be used to compensate for missing data in the daily load profile, and the accuracy of the affiliation will affect the filling effect. In practical applications, it has been found that there are two common characteristics of missing data: one is randomness and the other is long time series. The missing of these two types of data will affect the accuracy of the affiliation to some extent, while in the random case, the affiliation of this type of data is negligible. However, if it is missing for a long period of time, or even for a whole day, then there is a lack of reliable data for subordination calculations. In this condition, the grid should be pre-filled based on the power data in the feeder and the regularity of the daily load variation of the transformer. Equation (5) expresses the mathematical relationship between the degree of affiliation for each cluster center and the missing data. It is important to consider the effect of multiple degrees of affiliation on the missing data from an overall perspective.

$$x_{ij} = \sum_{k=1}^{c} u_{ik} \cdot p_{kj} \tag{5}$$

In Eq. (5), x_{ij} denotes the j th dimension data in the i th sample, u_{ik} denotes the affiliation of the i th sample belonging to the k th clustering centre, and p_{kj} denotes the j th dimension data in the k th clustering centre.

Figure 2 depicts the overall flow of the updated FCM algorithm, which fills in the missing data to produce a complete daily load curve. First, the power supply company's data platform is used to extract the daily historical load data, the format is processed, and the degree of missing data is determined. If the missing value is small, the number of categories can be determined directly; however, if the missing value is large, pre-populating the data with electricity data can be used. After classifying the historical load data and initializing the RODDPSO algorithm's parameters, the clusters was calculated. On this basis then the position equation and velocity equation of the particles are updated. Once the iteration termination condition is determined to be satisfied or not, the result is output. If the condition is not satisfied, the steps are repeated until it is.

Improved IFA-based anomaly data identification

In addition to the frequent missing data phenomenon in the raw data of MVDN, data anomalies also occur frequently. When it comes to missing data, it can usually be identified with the naked eye. However, identifying abnormal data manually can be challenging. Furthermore, anomalous data can introduce bias in engineers' data interpretation and calculations, which can negatively impact the effectiveness and financial gains of power grid firms. IFA is a popular outlier detection algorithm that isolates outliers from conventional observations by building multiple random isolation trees. The average number of comparisons required to isolate a given observation can then be used as a measure of its outlier. IFA is particularly well suited for working with large data sets. It has linear time complexity and is computationally more efficient due to the use of subsampling[20]. The existing data anomaly detection methods are mainly based on the description of normal samples, giving the region of a normal sample in the feature space, and the samples not in this region are regarded as abnormal[21]. The main disadvantage of these methods is that the anomaly detector only optimizes the description of the normal sample,

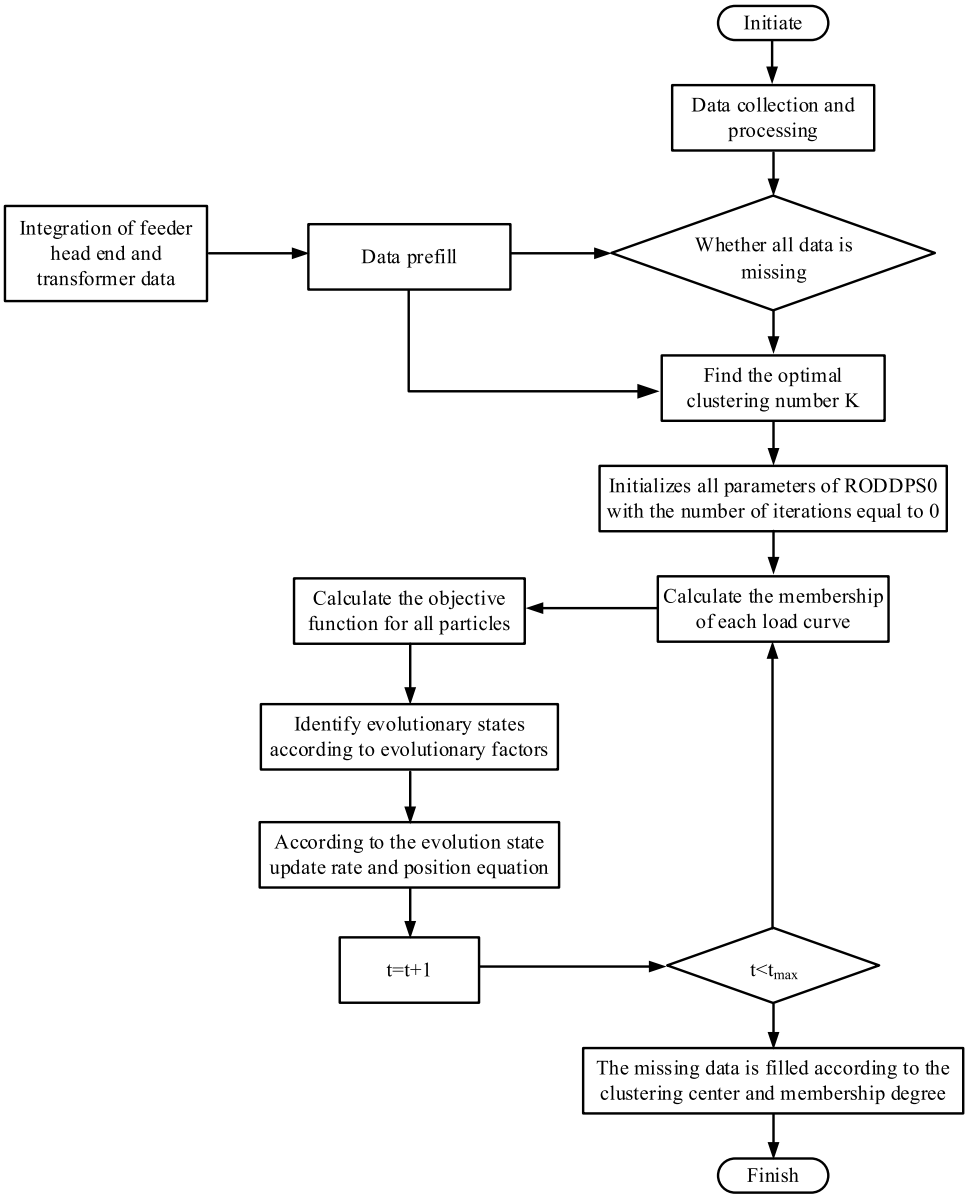


Figure 2. Flow chart of improved FCM algorithm.

not the description of the abnormal sample, which can cause a large number of false positives, or only detect a small number of anomalies.

The occurrence of abnormal data destroys the normal periodicity and continuity of power load to a certain extent. In order to accurately understand the distribution of abnormal data and the change amplitude of abnormal data, it is necessary to describe the distribution characteristics of data[22]. According to the characteristics and actual conditions of power load data, neither the central tendency measurement nor the distribution shape measurement can accurately describe the distribution characteristics of abnormal data. The range, standard deviation and coefficient of variation in the discrete trend measure can reflect the distribution of abnormal data, but when comparing different periods or different populations of the same population, the range and standard deviation are lack of comparability, while the coefficient of variation eliminates the above defects and has a wider application range[23]. Therefore, this paper selects the coefficient of variation as the criterion for screening the abnormal subspace.

Massive high-dimensional data sets are used in the study as the research object. The discrete degree measure function of the coefficient of variation is combined, and an improved isolated forest (CV-iForest) algorithm based on the coefficient of variation is proposed as a solution to the issues of low reliability of the high-dimensional data sets and high randomness of the iForest algorithm. In order to clearly express the relationship between various parts of the CV-iForest anomaly detection model, the general framework of CV-iForest is given in Fig. 3. The anomaly detection model consists of four layers: input layer, data pruning layer, data mining layer, and anomaly

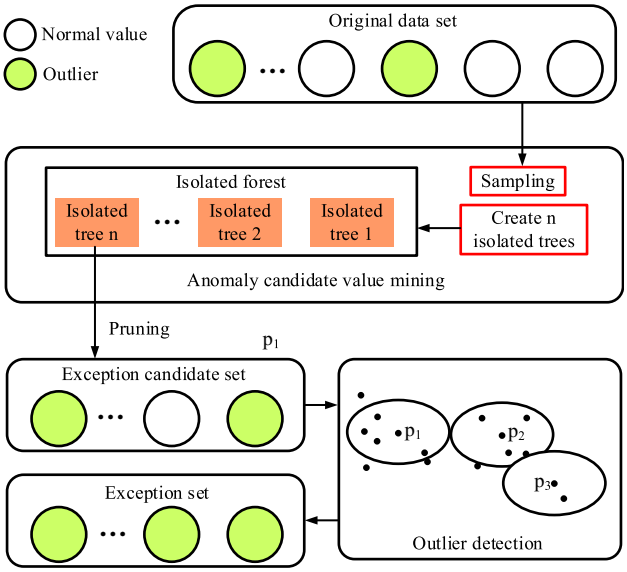


Figure 3. Improved iForest model mechanism diagram.

detection layer, which are respectively responsible for downscaling and pruning the data, extracting the anomaly candidate set, calculating the coefficient of variation, and outputting the anomaly detection results.

The core idea of isolated forest is to cut the abnormal data continuously, because the density of abnormal data is much smaller than the normal data clusters, so the abnormal data can be "isolated" with fewer cuts. In a binary tree structure, abnormal data is indicated by cuts closer to the root node, while normal data remains in deeper positions. In order to better demonstrate the principle that samples assign outlier score values by the average path length from all trees to the root node, the research draws a diagram of outlier score distribution, as shown in Fig. 4.

In the iForest algorithm, there are two key training parameters: the sub-sample size s and the number of isolated trees n. Its learning process focuses on constructing orphaned trees and forming isolated forests by using existing data. Firstly, s subsamples are randomly selected from the dataset D of size k dimension m to form the training samples $D_i = \{d_1, d_2, d_3, \cdots, d_n\}(s \leq k)$. Then the separation dimension is randomly selected from the training sample D_i and a cut point C is selected on the great and small interval of this separation dimension, all data samples greater than or equal to C are classified into the right branch of the isolation tree, and the remaining portion is classified into the left branch, and the step is repeated until the samples cannot be cut any more or reach the limited height of the tree. Finally, repeat the above steps all the time to construct multiple isolation trees to form an isolation forest.

After training, since outliers are generally isolated in the first few rounds of isolation and their average path lengths are relatively short, the outlier's anomaly score is calculated based on the average path length of the sample to determine whether the sample is anomalous or not.

Following that, based on the average path length $c(s)$ of sample x, the anomaly score of sample x may be derived; its computation expression is presented in Eq. (6).

$$score(x, s) = 2^{-\frac{E(pathL(x))}{c(s)}} \tag{6}$$

In Eq. (6), $E\left(pathL(x)\right)$ is the average value of the path length of sample x in the forest. Since the selection of dimensions and segmentation points in the training phase of the iForest algorithm is random, which leads to its

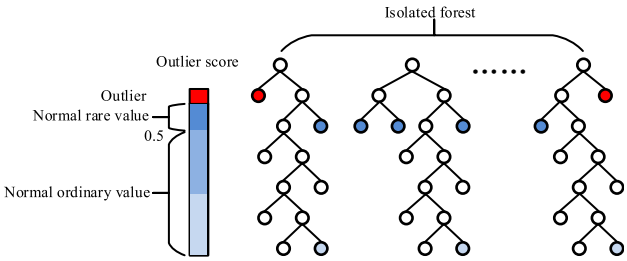


Figure 4. Schematic diagram of anomaly score.

poor global stability, and for high-dimensional loaded data, some of the dimensional information is still unused after its modeling, which leads to its low reliability. Therefore, the study improves the iForest algorithm based on data reduction and novel isolation strategies.

In the process of data dimensionality reduction, the coefficient of variation is dimensionless, and the coefficient of variation of each dimension is calculated in order to eliminate variables with high dispersion, that is to say, to eliminate the anomalous subspace. This is done by calculating the coefficient of variation C_V for each dimension in the dataset and filtering out the anomalous subspace $W = \{w_1, w_2, \cdots, w_i\}$ based on $C_{V\ \min} \leq C_V \leq C_{V\ \max}$ (where $i = a \times m$, a are the critical coefficients with values in the range of [0, 1], and m are the dimensions of dataset D. Where the coefficient of variation C_V is calculated in Eq. (7).

$$C_V = \frac{\sigma}{\mu} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}/\mu \tag{7}$$

In Eq. (7), σ denotes the standard deviation of the sample, μ denotes the number of data in the sample, n denotes the mean value of the sample, and x_i denotes the i th value in the sample. The anomalous subspace resulting from the data dimensionality reduction is used as a new training set, and the isolation strategy is followed to improve the isolation effect of a single isolated tree during the isolation tree construction process. The study proposes two isolation strategies. One strategy involves randomly selecting the isolation dimension from the anomalous subspace. The samples can be placed into the left and right branches of the isolation tree by selecting an isolation dimension and using the midpoint of the largest interval between neighboring data on that dimension as the isolation point.

Computational and analytical model construction for MVDNLL

MVDN differs from high voltage transmission grids in its ability to ignore the conductance of conductors and transformers to ground[24]. As a result, simpler approaches are typically used to calculate LL, such as the power method, the equivalent resistance method, the maximum current method, and the root mean square current method. However, these LL calculation methods can lead to a single LL calculation result and low calculation accuracy due to limitations such as missing measurement data[25]. Therefore, the study introduces a breadth-first based forward back generation LL calculation method in the context of big data. The method is based on the actual feeder topology and the load data of each transformer, transformers are equated with impedance, and the forward back generation method is used to calculate the losses of lines and transformers[26]. Given the complexity and numerous branches of the current MVDN feeder line topology, the forward-back generation method is less efficient. To address this, we conducted topology identification using the original algorithm and stratified it to enable layered calculation of LL. In the feeder topology, the connection relationship between the nodes is represented by the node association matrix; based on this, the root node of the feeder is the first layer node of the node hierarchy matrix; then the nodes that are connected to the root node but have not been written into the node hierarchy matrix are found, and then they are written in the second layer node in the node hierarchy matrix, and so on, until all the nodes are written in the node hierarchy matrix[27].

Taking the IEEE31 node system as an example, whose node hierarchy is shown schematically in Fig. 5. The branch currents are calculated by back generation layer by layer, and the node injection currents and branch currents are calculated as shown in Eq. (8).

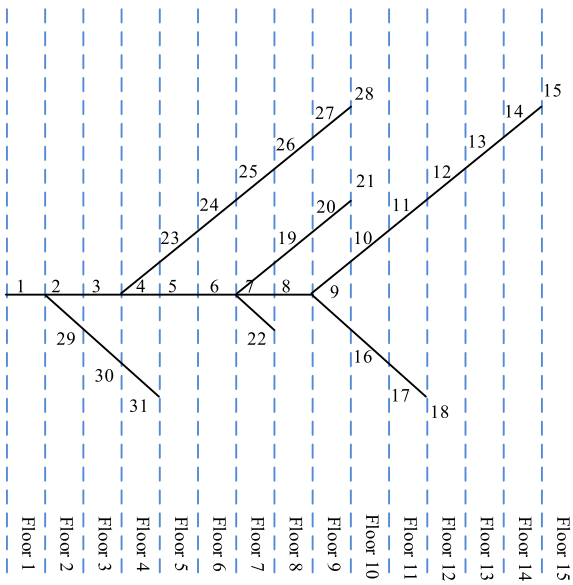


Figure 5. IEEE 31 node layer diagram.

$$\begin{cases} I_j^k = \left(P_{Load} - jQ_{Load}\right)/V_j^{k-1} \\ I_l^k = I_j^k + \sum_{m \in M} I_m^k \end{cases} \tag{8}$$

In Eq. (8), I_j^k denotes the injected current of the end node j of the branch l, P_{Lood} denotes the active power of the branch, Q_{Lood} denotes the reactive power of the branch, V_j^{k-1} denotes the voltage of the last node of the end node j of the branch l. I_l^k denotes the branch current of the branch, I_m^k denotes the current of m, the lower branch of branch l, and M denotes the set of all the lower branches that are directly connected to node j. Then, starting from the first node to the last node layer by layer, the voltage of node j is calculated as shown in Eq. (9).

$$V_j^k = V_i^k - Z_l I_l^k \tag{9}$$

In Eq. (9), V_j^k denotes the voltage of the end node j of the branch l, V_i^k denotes the voltage of the beginning node i of the branch l, and Z_l denotes the impedance value of the end node j of the branch l. Repeat the above calculation steps continuously until the convergence condition $\max \left|V^k - V^{k-1}\right| \leq \varepsilon$ is satisfied.

The computed voltage and current at each branch can be used to determine the line's overall loss. The specific calculation is shown in Eq. (10).

$$\Delta P_l = 3 \times \sum_{i=1}^{L} I_i^2 R_i \tag{10}$$

In Eq. (10), L denotes the number of branches of the feeder, I_i^2 denotes the current amplitude flowing through the branches, and R_i denotes the resistance of the branches. To simplify the calculation, the study has treated the transformer as an element with only resistance and reactance, and calculated its equivalent impedance using Eq. (11).

$$\begin{cases} R = \dfrac{P_k V_N^2}{S_N^2} \times 10^3 \Omega \\ X = \dfrac{V_k\% \cdot V_N^2}{100 S_N} \times 10^3 \Omega \end{cases} \tag{11}$$

In Eq. (11), P_k indicates transformer short-circuit loss in kW. V_N is the rated voltage of the transformer in kV. S_N indicates the rated capacity of the transformer in kVA. And $V_k\%$ indicates the impedance voltage percentage. Transformer loss includes variable loss and fixed loss brought by equivalent resistance, so the transformer loss calculation equation can be expressed as Eq. (12).

$$\Delta P = P_0 + P_R \tag{12}$$

In Eq. (12), P_0 denotes the fixed losses of the transformer and P_R denotes the losses of the transformer equivalent resistance. The feeder LL rate is the proportion of LL and transformer losses in the feeder to the power supply, where the power supply is the sum of feeder LL consumption and power sales. Its calculation equation is shown in Eq. (13).

$$\lambda = \frac{\Delta P_l T + \Delta P_T T}{\Delta P_l T + \Delta P_T T + W_{es}} \times 100\% \tag{13}$$

In Eq. (13), λ denotes the feeder LL rate, T denotes is the duration of feeder power sales, and W_{es} denotes the total power sales of the feeder. To identify the cause of LL of feeders, the research first collects and preprocesses the network loss index parameters. And then based on the standardised data, logistic regression analysis is carried out on the whole grid to identify the important factors affecting the grid loss and exclude irrelevant factors. On this basis, the logistic regression method is used to identify the causes of feeder LL consumption in different regions from the perspective of power supply zoning. By categorizing the new feeders and predicting the LL high and low using the logistic regression model of the category to which they belong, the causes of losses were identified.

LLA model performance validation

The study has calculated the LL of the power supply line in Yuxi city in Yunnan Province as an example, and the feeder has been analyzed in detail. The experimental data is based on Electricity Load Diagrams data set in UCI database, which contains active load data of 370 users from 2017 to 2021. The data collection interval is once every 15 min, so the data dimension is 96. All users' load data from 2011 was selected as the experimental data to be used by the algorithm. Out of the 365 samples from 2017, the study randomly generated 37 abnormal samples (10% of the total). The abnormal samples are divided into two categories to detect the algorithm's effect: obvious outliers and fuzzy outliers.

Effectiveness of Missing Data Filling

To study the filling accuracy of the data filling algorithm in two cases, partially missing and completely missing, where the partially missing section is set in the first half of the daily load profile with a defect rate of 12.5%, which

means that the load information is missing continuously for three hours, Table 2 shows the basic information of this section of the feeder.

The study employs a transformer's daily active load in this feeder as an example. It clusters and analyses the active load in two states using the FCM clustering method, and it derives the results in Fig. 6 based on the ideal number of clusters. The horizontal axis is the 96 load data collection points in one day, and the vertical axis is the instantaneous active power of certain points at a certain moment. The cluster analysis reveals that while the two types of load profiles share similar trends, the second type exhibits significantly larger and more volatile peaks and troughs than the first. In order to verify the accuracy of the data filling algorithm when partial and all data are missing, the study takes the active load of Yuxi City in Yunnan Province in July 2020 as an example data. In the original data, part of the missing data is inserted in the front section of the daily load curve, and the missing rate is 12.5%, that is, the load data is missing for 3 h continuously. To assess the feasibility of data filling methods, commonly used methods include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Standard Deviation (SD). The quality of the selected method can be determined by comparing the RMSE, MAE, and SD of the missing values with the actual values. In general, a smaller RMSE and MAE indicate higher precision in filling. A smaller SD indicates smoother filling values, and no filling value should have a particularly large difference.

In order to test the filling effect, the improved FCM (IFCM) algorithm was compared with the Back Propagation neural network from literature[28], the conventional FCM procedure, and the cluster mean filling algorithm, each used to fill in two missing data examples. Its absolute error curve is given in Fig. 7 along with these comparisons. The revised FCM method, which has a higher accuracy compared to other algorithms, has an average error of -6.35kW for partial missing data and a maximum error of − 10.63kW for whole missing data. The comparison of the basic FCM algorithm indicates that the cluster center optimized by RODDPSO more accurately reflects the overall characteristics of the data. The filling accuracy of BPNN is the lowest due to the selected feeders being located in rural areas of class D power supply zones. These areas do not have any noticeable regularity in load changes and are susceptible to the influence of environmental factors. Building the neural network will be difficult without sufficient historical data. If there is a lack of historical data, it is difficult to complete the establishment of the neural network.

Three classical data filling methods, linear interpolation, mean filling and mode filling, were selected in this study. RMSE, MAE and SD were used to evaluate the performance of FCM algorithm, and 30 experimental results were statistically analyzed. The evaluation indexes obtained were shown in Fig. 8. Different letters in the figure indicate significant difference between the same index ($p < 0.05$). In Fig. 8, the SDs of the improved FCM algorithm for the two data missing cases are 4.015kW and 10.156kW respectively, which is significantly different

Power supply partition	D
Cable length /km	9.63
Length of overhead line /km	10.36
Total line length /km	19.45
Power supply radius /km	5.21
Number of transformers	65
Transformer capacity /MVA	43.26
Number of public transformers	3
Common transformer capacity /MVA	0.52
Number of special transformers	62
Dedicated transformer capacity /MVA	42.69

Table 2. Feeder basic information.

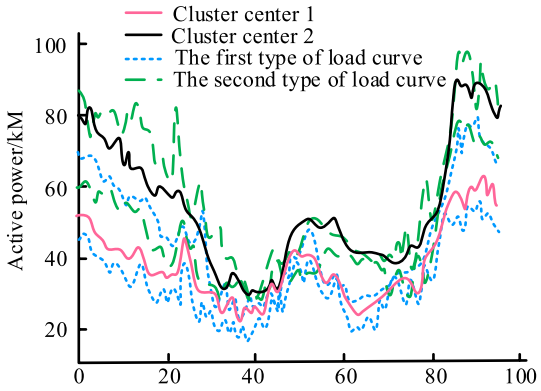


Figure 6. Clustering effect of historical active load curve.

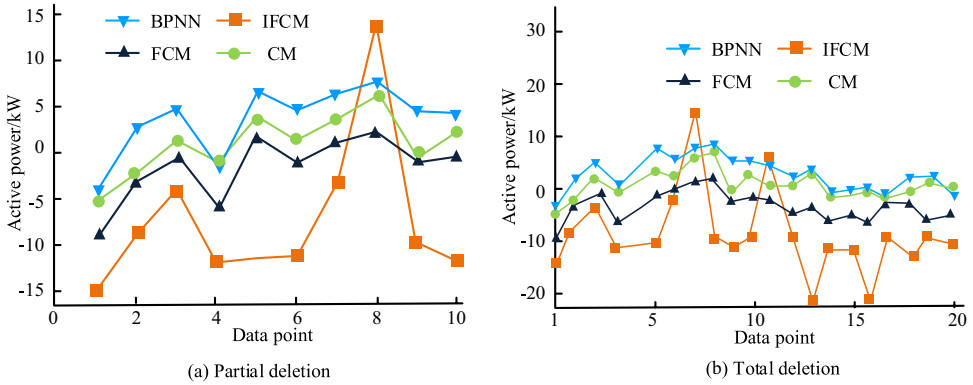


Figure 7. Active power absolute error curve.

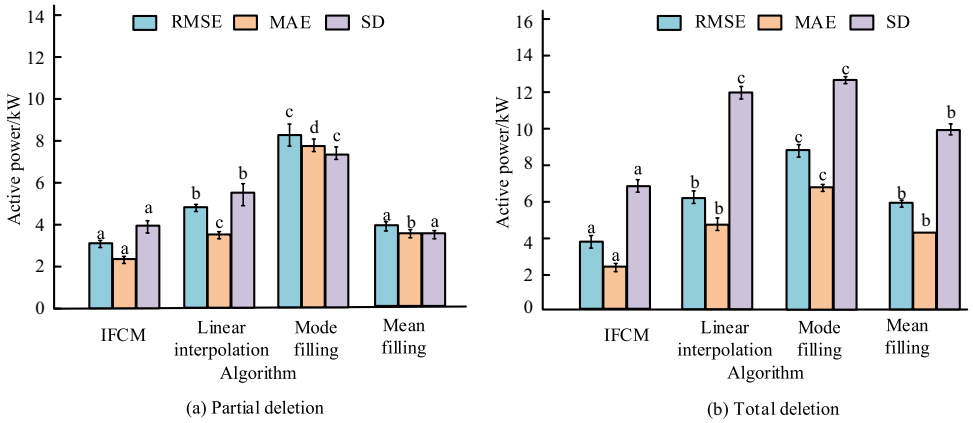


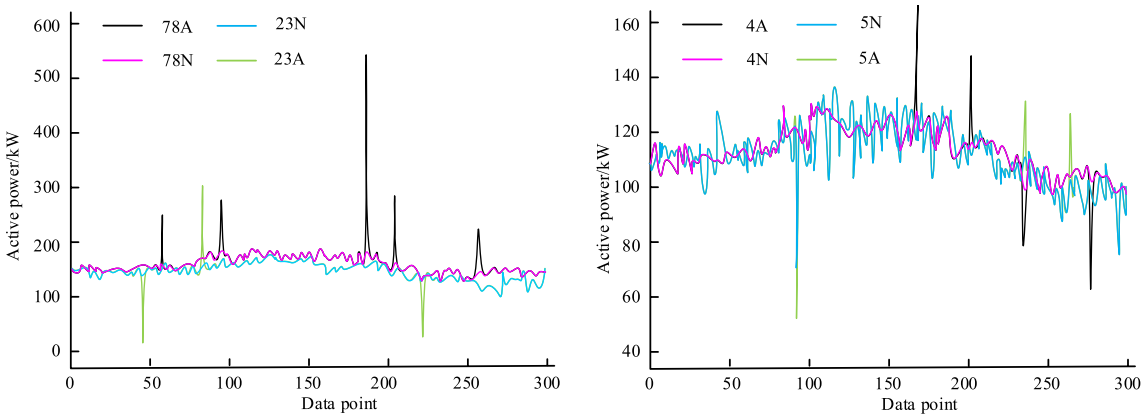
Figure 8. Error evaluation index.

from the other three methods ($p < 0.05$), indicating smoother filling value. The MAE and RMSE were 3.154kW and 2.416kW, respectively, when partial data were missing. The MAE and RMSE with total data missing were 5.635kW and 3.529kW, respectively, the smallest in the two missing cases. From partial missing to complete missing, the error evaluation index of the improved FCM algorithm changes the least, which indicates that the robustness of the proposed algorithm is better than that of other algorithms.

Effects of abnormal data identification

The aberrant samples are split into two categories of clear and fuzzy anomalies to test the algorithm's efficacy. The HBOS algorithm, the LOF algorithm, the K-means algorithm, and the iForest algorithm—all of which are implemented in Python—are all compared with the CV-iForest algorithm suggested in the study to confirm the efficacy and superiority of the algorithm in this work. The study calculates the coefficient of variation for each dimension, ranks the coefficients of variation, and filters the top $a \times m$ dimensions as anomalous subspaces. The graphic shows that, in comparison to the data in the 23rd dimension, the data in the 78th dimension has more spikes, more pronounced spikes, and a larger coefficient of variation. Figure 9 shows the data in the first two dimensions with the largest coefficient of variation in the case of blurring of the anomalous samples. Compared to the 23rd dimensional data, the 5th dimensional data has more spikes and the degree of spikes is more pronounced and the coefficient of variation is larger.

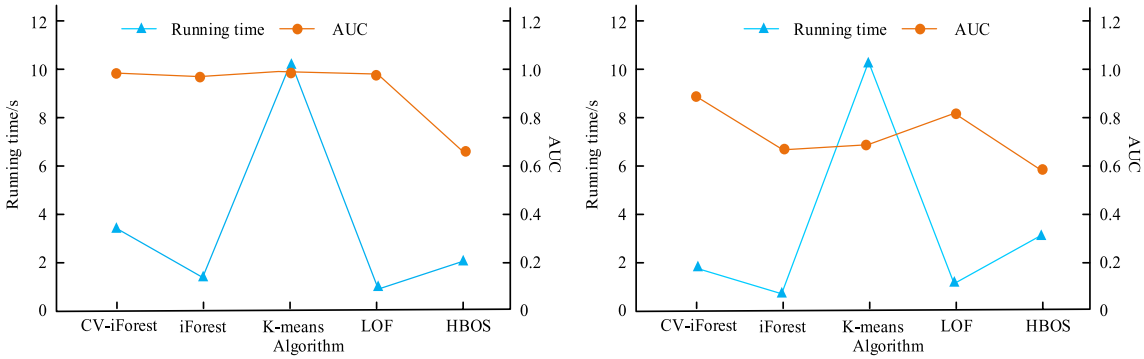
Figure 10 displays the evaluation findings for the experimental choice of the subject characteristics Area Under Curve (AUC) value as the evaluation index of the running effect in order to assess the algorithm's detection accuracy. In the table, for the abnormal samples with obvious outliers, the AUC value of CV-iForest algorithm is 0.9971, which is the highest among several algorithms, but the running speed is slower, which is still within the acceptable range. LOF algorithm has the fastest running speed, which reaches 0.7189 s, and the AUC value is second only to that of the CV-iForest algorithm, but it is not applicable to load data that does not have the attribute of density. K-means algorithm runs the slowest, which is not applicable to load data without density. The K-means algorithm is the slowest, indicating that it is not suitable for high dimensional load datasets with large amount of data, and the HOBS algorithm has the lowest AUC value of 0.6629, which indicates that the load data do not comply with the assumptions of data distribution of this statistical method. In cases of abnormal sample blurring, the iForest algorithm maintains the highest AUC value of 0.8586 with the smallest decrease, while the LOF algorithm demonstrates the second-best adaptive ability in abnormal scenarios, following only the CV-iForest algorithm. The K-means algorithm and iForest algorithm show the most significant decrease



(a) Abnormal samples are obvious

(b) Abnormal samples are blurred

Figure 9. Abnormal samples are obvious.



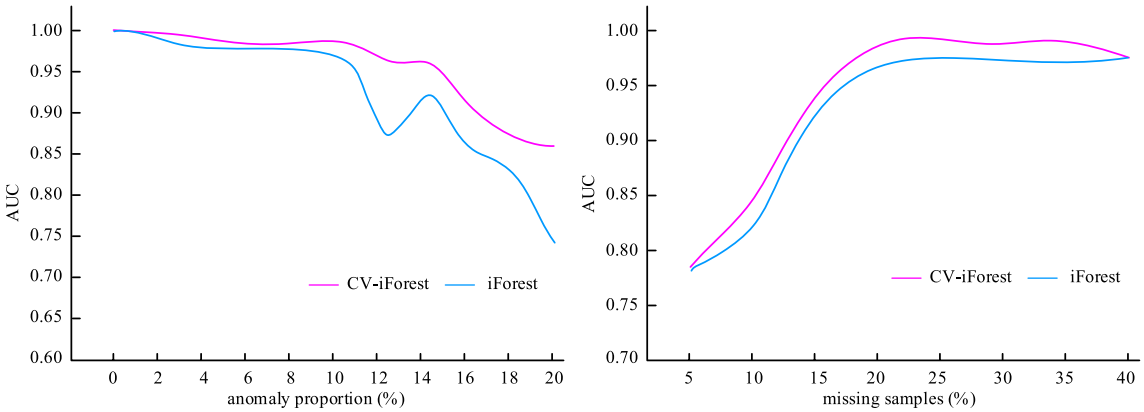
(a) Abnormal samples are obvious

(b) Abnormal samples are blurred

Figure 10. Algorithm operation effect.

in AUC value, while the HBOS algorithm has the lowest AUC value in both scenarios and also experiences a noticeable decrease.

The study evaluates the stability of the model under the Load Diagrams dataset by varying the dataset outliers rate and missing samples for comparison experiments. In Fig. 11a, the model's performance is fluctuating as the outliers rate continues to increase, and the model's performance is the best when the outliers rate is between 0 and 10%. In general, when the proportion of non-normal samples increases, the distribution among the sample categories tends to be consistent, which results in a classification model with high accuracy. However, the



(a) The experimental results of the model performance change with the proportion of anomalies

(b) Experimental results of model performance with the missing samples

Figure 11. Analysis of experimental results of model stability.

iForest model takes advantage of the sparsity of the outliers, which is more likely to aggregate when the outliers are greater than 10%, leading to an increase in the number of outlier decompositions, which in turn affects the effectiveness of the model. As shown in Fig. 11b, the performance of the model increases significantly from 5 to 20% base classifiers, suggesting a progressive rise in the variability of the base classifiers, increasing the stability of the final model. The model performance did not always improve when missing samples increased from 20 to 40%, which may be related to the creation of weak classifiers with accuracy lower than 0.5.

LL Calculation and Cause Identification

Taking a power supply line in a prefecture-level city in Yunnan Province as an example, the line loss is calculated and compared with the model proposed by Liang et al.[29], and Table 3 shows the refined analysis results of this feeder. LL is the loss caused by problems such as line material. LL ratio is the proportion of LL to the total LL. Public distribution loss is the loss caused by public distribution. Public distribution loss ratio is the proportion of public distribution loss to the total LL. From the table, it can be seen that the LL rate of this feeder is 7.62%, which exceeds the LL rate standard of Class D power supply sub-district, and the LL rate is unqualified. The LL reaches 58,159.89kW, the loss of public distribution substation reaches 134.896kW, and the LL accounts for 99.77%, which indicates that the loss of this feeder is basically borne by the line, and the loss of the dedicated distribution substation is borne by the users themselves. Due to the large access load, most of the distribution transformers have low voltage phenomenon and 5 branch circuits have heavy overload phenomenon. Liang et al. 's method belongs to the power flow algorithm, and its main idea is to establish the power flow equation of the station area distribution network. For different circuits, different calculation models should be established according to the parameters, which have low generality and great dependence on the quality of parameters. Compared with the model proposed in this study, the overall performance is still lower than that of the model proposed in this study, although there is little difference in each index.

The study has analyzed five branches of this feeder where heavy overload phenomenon exists and the results obtained are shown in Table 4. The table shows that two branches of the feeder have severe overloading phenomenon, which is caused by the failure of the branch type to match the load current, and according to the wire cross-section requirements in the MVDN Planning Technical Guidelines, these two branches should be replaced with 240mm^2 diameter wires.

Conclusion

The study suggests a data cleaning model based on the improved FCM clustering algorithm and IFA in accordance with the two common data quality problems occurring in the power load data in order to enhance the grid's ability to use energy and to encourage a change in the management style of the power supply company. A model for calculating and analyzing LL in the context of electric power big data is proposed. The traditional distribution network LL calculation method is not applicable and has flaws such as unclear identification of LL causes in distribution network feeders. The results show that the average error of the improved FCM algorithm is the smallest among several algorithms. The average error and standard deviation of the improved FCM algorithm are -6.35kW and 4.015kW when part of data is missing, and the average error and standard deviation of the improved FCM algorithm are -10.63kW and 10.156kW when all data is missing. The MAE and RMSE were 3.154kW and

Key index	Research model	Liang et al.[29]
Line loss rate /%	7.62	7.03%
Line loss/kW	58,159.89	57,231.19
Line loss ratio	99.77%	97.32%
Common distribution loss /kW	134.896	139.647
Common distribution loss ratio	0.23%	0.26%
Low voltage distribution ratio	66%	63%
Number of heavy overload branches	5	4
High loss variation ratio	0%	0%

Table 3. Feeder Refinement Analysis Results.

Fore end node	End node	Main trunk or not	Daily average load rate/%	Loss rate /%
1	2	Yes	93.94	0.74
2	3	Yes	93.58	0.74
6	7	Yes	143.33	23.45
7	8	Yes	84.69	22.39
8	13	Yes	136.62	50.03

Table 4. Feeder Heavy Overload Branch Information.

2.416kW, respectively, when partial data were missing. When the data is incomplete, it is 5.635kW and 3.529kW respectively. From partial missing to total missing, the error evaluation index of the improved FCM algorithm changes the least, indicating that compared with the classical data filling method, the proposed algorithm is relatively robust. In the calculation of the original line loss data of a power supply line in a prefecture-level city in Yunnan province, compared with Liang et al. 's method, RODDPSO algorithm and cv-forest algorithm are used in this study to deal with abnormal or missing data, and the nonlinear relationship between input and output can be found only by learning a small number of data samples. There is no need to establish upper and lower power constraints on nodes with abnormal or missing data, and the line loss calculation results can be obtained quickly and accurately by artificial intelligence algorithm. The AUC value of cv-forest algorithm is 0.9971 for abnormal samples with clear outliers, and 0.8586 for fuzzy anomaly samples, in which the accuracy of fuzzy anomaly samples decreases the least. Further research shows that the loss rate of the feeder is 7.62%, and two branches are seriously overloaded. The power supply radius and bare conductor resistance are the key factors leading to the high loss rate. In this study, RODDPSO algorithm and cv-forest algorithm were used to process and calculate the line loss data. Although the results are good, the computing resources required are large. Therefore, different neural networks can be considered for lightweight processing to improve the efficiency and accuracy of the algorithm.

Fundings

The research is supported by: Project source: key science and technology projects of China Southern Power Grid Corporation. Project Title: Research and application of LL analysis and diagnosis and loss reduction and carbon reduction technology for new power systems. [Project No. 035900KK52220006 (GDKJXM20220254); When the paper formed by the research and development results of this project is published, it must be indicated that "China Southern Power Grid Corporation Science and Technology Project Funding [Project No. 035900KK52220006 (GDKJXM20220254)].

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 25 December 2023; Accepted: 23 July 2024
Published online: 22 August 2024

References

1. Danjuma, M. U., Yusuf, B. & Yusuf, I. Reliability, availability, maintainability, and dependability analysis of cold standby series-parallel system. *JCCE.* **1**(4), 193–200 (2022).
2. Saeed, M., Ahmad, M. R. & Rahman, A. U. Refined pythagorean fuzzy sets: Properties set-theoretic operations and axiomatic results. *JCCE.* **2**(1), 10–16 (2022).
3. Choudhuri, S., Adeniye, S. & Sen, A. Distribution alignment using complement entropy objective and adaptive consensus-based label refinement for partial domain adaptation. *AIA.* **1**(1), 43–51 (2022).
4. Oslund, S., Washington, C. & So, A. Multiview robust adversarial stickers for arbitrary objects in the physical world. *JCCE.* **1**(4), 152–158 (2022).
5. Wang, X., Cheng, M. & Eaton, J. Fake node attacks on graph convolutional networks. *JCCE.* **1**(4), 165–173 (2022).
6. Hu, W., Guo, Q., Wang, W., Wang, W. H. & Song, S. H. "Loss reduction strategy and evaluation system based on reasonable line loss interval of transformer area. *Appl. Energ.* **306**(15), 123–133. https://doi.org/10.1016/j.apenergy.2021.118123 (2022).
7. Zhang, Z. Y., Yang, Y., Zhao, H. & Xiao, R. Prediction method of line loss rate in low-voltage distribution network based on multi-dimensional information matrix and dimensional attention mechanism-long-and short-term time-series network. *IET Gener Transm DIS* **16**(20), 4187–4203. https://doi.org/10.1049/gtd2.12590.Aug (2022).
8. Tang, Z. *et al.* Research on short-term low-voltage distribution network line loss prediction based on Kmeans-LightGBM. *J. Circuit Syst. Comp.* **31**(13), 135–146. https://doi.org/10.1142/S0218126622502280 (2022).
9. Liu, K. Y., Jia, D. L., Kang, Z. J. & Luo, L. Anomaly detection method of distribution network line loss based on hybrid clustering and LSTM. *J. Electr. Eng. Technol.* **17**(2), 1131–1141. https://doi.org/10.1007/s42835-021-00958-4 (2022).
10. Zhang, L. *et al.* Distribution network line loss calculation method considering distributed photovoltaic acces. *J. Phys. Conf. Ser.* **2488**(1), 63–72. https://doi.org/10.1088/1742-6596/2488/1/012057 (2023).
11. Yi, C. C., Tuo, S., Tu, S. & Zhang, W. T. Improved fuzzy C-means clustering algorithm based on t-SNE for terahertz spectral recognition. *Infrared Phys. Technol.* **117**(9), 214–225. https://doi.org/10.1016/j.infrared.2021.103856 (2021).
12. Ke, B., Nguyen, H., Bui, X., Bui, H. & Nguyen-Thoi, T. "Prediction of the sorption efficiency of heavy metal onto biochar using a robust combination of fuzzy C-means clustering and back-propagation neural network. *J. Environ. Manage.* **293**(9), 214–225. https://doi.org/10.1016/j.jenvman.2021.112808 (2021).
13. Surono, S. & Putri, R. D. A. Optimization of Fuzzy C-means clustering algorithm with combination of Minkowski and Chebyshev distance using principal component analysis. *Int. J. Fuzzy Syst.* **23**(1), 139–144 (2021).
14. Wang, A. J. & Zhang, F. A driver abnormal behavior warning method based on isolated forest algorithm. *ATS* **3**(12), 55–66 (2023).
15 Pan, N., Jiang, X., Pan, D. & Liu, Y. Study of the bullet rifling linear traces matching technology based on deep learning. *J. Intell. Fuzzy Syst.* **40**(4), 16–22. https://doi.org/10.3233/JIFS-189617 (2021).
16 Long, X. M., Chen, Y. J. & Zhou, J. Development of AR experiment on electric-thermal effect by open framework with simulation-based asset and user-defined input. *Artif. Intell. Appl.* **1**(1), 52–57 (2022).
17. Yastrebov, A., Kubus, L. & Poczeta, K. Multiobjective evolutionary algorithm IDEA and k-means clustering for modeling multi-dimenional medical data based on fuzzy cognitive maps. *Nat. Comput.* **22**(3), 601–611 (2023).
18. Shi, H., Wang, P., Yang, X. & Yu, H. An improved mean imputation clustering algorithm for incomplete data. *Neural Process Lett.* **54**(5), 3537–3550 (2022).
19. Yang, Q. F. *et al.* HCDC: A novel hierarchical clustering algorithm based on density-distance cores for data sets with varying density. *Inf. Syst.* **114**(5), 1–14 (2023).
20. Sebastian, B., Philipp-Jan, H. & Katharina, M. Randomized outlier detection with trees. *JDSA* **13**(2), 91–104. https://doi.org/10.1007/s41060-020-00238-w (2022).

21. Shao, N. & Chen, Y. Abnormal data detection and identification method of distribution internet of things monitoring terminal based on spatiotemporal correlation. *Energies* **15**(6), 2151–2164. https://doi.org/10.3390/en15062151 (2022).
22. Liang, J. F., Li, W., Zhao, Y. P., Zhou, Y. & Zou, Q. W. A risk identification method for abnormal key data in the whole process of production project. *Int. J. Data Min. Bioin.* **24**(3), 1–3. https://doi.org/10.1504/IJDMB.2022.130345 (2022).
23. Wang, Y., Zhang, X. Y. & Liu, H. F. Intelligent identification of the line-transformer relationship in distribution networks based on GAN processing unbalanced data. *Sustainability* **14**(14), 624–647. https://doi.org/10.3390/su14148611 (2022).
24. Fu, J. *et al.* A novel optimization strategy for line loss reduction in distribution networks with large penetration of distributed generation. *Int. J. Elec. Power* **150**(8), 1091121–10911216 (2023).
25 Liu, X. Automatic routing of medium voltage distribution network based on load complementary characteristics and power supply unit division. *Int. J. Elec. Power.* **133**(2), 106467.1-106467.13. https://doi.org/10.1016/j.ijepes.2020.106467 (2021).
26. Liu, K. *et al.* Energy loss calculation of low voltage distribution area based on variational mode decomposition and least squares support vector machine. *MPE* **2021**(33), 8530389.1-8530389.11. https://doi.org/10.1155/2021/8530389 (2021).
27. Dashtdar, M. *et al.* Improving voltage profile and reducing power losses based on reconfiguration and optimal placement of UPQC in the network by considering system reliability indices. *Int. T Electr. Energy* **31**(11), e13120.1-e13120.29. https://doi.org/10.1002/2050-7038.13120 (2021).
28. Min, Y. C., Chai, H. K., Huang, Y. F., Wei, D. C. & Jia, Y. P. Artificial intelligence generated synthetic datasets as the remedy for data scarcity in water quality index estimation. *Water Resour. Manag.* **37**(15), 6183–6198. https://doi.org/10.1007/s11269-023-03650-6 (2023).
29. Liang, C. *et al.* Line loss interval algorithm for distribution network with DG based on linear optimization under abnormal or missing measurement data. *Energies* **15**(11), 4158. https://doi.org/10.3390/en15114158 (2022).

Author contributions

J. L. and S. L. collected the samples. W. Z. and J. L. analysed the data. K. Z. andZ. J. conducted the experiments and analysed the results. All authors discussed the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.