International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 7831-7849

https://doi.org/10.70517/ijhsa463671

Research on Quality Defect Diagnosis Model and Intelligent Operation and Maintenance Strategy of Distribution Grid Based on Cluster Analysis

Lijuan Yan[1], Ming Wang[2], Yan Zeng[2], Wensen Li[2] and Yu Zou[2,*]
[1] Guangxi Power Grid Co., Ltd, Nanning, Guangxi, 530022, China
[2] Qinzhou Power Supply Bureau of Guangxi Power Grid Co., Ltd, Qinzhou, Guangxi, 535000, China
Corresponding authors: (e-mail: ZY_1106@outlook.com).

Abstract With the increasing demand for electricity and the growing complexity of distribution networks, the identification and diagnosis of quality defects in power systems have become critical. Traditional fault diagnosis methods for distribution networks suffer from low accuracy and slow response time. In recent years, the application of data mining and artificial intelligence technologies has provided new ideas for power system fault diagnosis, especially in the diagnosis of quality defects in distribution networks. In order to improve the accuracy of fault diagnosis in distribution networks, this study proposes a quality defect diagnosis model for distribution networks based on hybrid clustering algorithm. The model first ensures the quality of data through data preprocessing, including data complementation, outlier processing, and data normalization; then, feature extraction is performed on the data through principal component analysis (PCA) dimensionality reduction to reduce the computational complexity. Finally, the clustering process is optimized by combining K-Means and hierarchical clustering algorithm to improve the accuracy of clustering results. The experimental results show that the accuracy of line loss anomaly identification of distribution network lines reaches 98.50% after using this model. In addition, by comparing with the traditional method, the optimized clustering algorithm has significant improvement in clustering time and error, the algorithm time is reduced by 26.5 seconds, and the average clustering error is reduced from 39.4832 to 7.8469. The model provides effective technical support for the intelligent operation and maintenance of the distribution network and has a better practical application value.

Index Terms distribution network, principal component analysis, quality defect diagnosis, hybrid clustering algorithm, line loss anomaly

I. Introduction

Cluster analysis is a quantitative method to study the problem of categorizing things with multiple elements [1]. Its basic principle is to quantitatively determine the affinity relationship between samples according to their own attributes, mathematically according to some similarity or difference index, and cluster the samples according to the degree of this affinity relationship [2]-[4]. In power systems, cluster analysis not only enables the diagnosis of quality defects in distribution networks, but also brings intelligent operation and maintenance strategies to distribution networks [5], [6].

Distribution network is the core link in the power system directly facing the power users, and the quality of its advantages and disadvantages closely affects the normal production and life [7]. With the expansion of the distribution network structure and its complexity, the probability of its failure is relatively increased, which makes the distribution network once there is a quality problem, it will have a great impact on the normal operation of people's daily life [8]-[10]. Therefore, the quality defect diagnosis of distribution network based on cluster analysis is particularly important, which can realize the quality defect diagnosis of distribution network in a short time in order to reduce the economic loss and improve the safety performance of distribution network [11]-[13].

At the same time, the current development and construction of smart distribution network operation and maintenance in China is not perfect, and it presents an unbalanced development in geographic location [14], [15]. Some areas are constrained by the degree of economic development, transportation and climatic environment, distribution network operation and maintenance work is relatively backward, or manual patrol mode of power system operation, not only consumes a lot of manpower, material and financial resources, and the efficiency and quality of the work is not outstanding, but also there are longer working hours, slower progress of the work progress problems [16]-[19]. Therefore, intelligent operation and maintenance under cluster analysis can collect and analyze the fault data of distribution network operation and maintenance, Internetize the overall distribution business, so as to realize

effective data collection and control, and then manage and control the whole distribution network operation and maintenance [20]-[22].

With the deep adjustment of the global energy structure and the acceleration of the smart grid construction, the distribution network, as an important link connecting the power source and the end user, has a direct impact on the operation stability and power quality on the security, economy and user satisfaction of the whole power system. However, the increasing complexity of the distribution network structure, the diversity of equipment, and the increased uncertainty of the operating environment make the identification of quality defects and fault diagnosis face great challenges. Traditional diagnostic methods relying on manual inspection or fixed thresholds can no longer meet the requirements of modern distribution systems for high precision, real-time and intelligence. In this context, data-driven intelligent analysis methods have become a research hotspot. Especially along with the development of intelligent sensors, big data acquisition and storage technology, a large amount of multi-source heterogeneous operation data has been accumulated during the operation of distribution networks, which provides a data basis for the intelligent identification of quality defects. Meanwhile, clustering analysis, as an unsupervised learning method, can automatically classify and pattern recognize the data, which is suitable for mining uncertainty and nonlinear features of fault information in distribution networks. Existing studies have shown that a single clustering algorithm is easily affected by initial parameters, noise interference and other factors in practical applications, resulting in unstable classification results and low diagnostic accuracy. Therefore, there is an urgent need to construct a hybrid clustering diagnostic model with high robustness and high accuracy, in order to improve the anomaly identification capability for key issues such as line line loss and provide effective support for intelligent operation and maintenance of distribution networks.

Based on the above problems, this paper proposes a hybrid clustering diagnostic method combining K-Means and cohesive hierarchical clustering, starting from data preprocessing, feature extraction and fusion, and taking into account the operating characteristics of distribution networks and line loss factors. By introducing supervised and unsupervised indicators to jointly determine the optimal number of clusters, the method improves the accuracy and stability of the clustering results; using principal component analysis to reduce the dimensionality of high-dimensional data, to realize the fusion of feature information, and to reduce the computational complexity; and constructing an anomaly detection mechanism to carry out a secondary screening of edge samples, which effectively improves the ability to identify minor anomalies or new types of faults. Finally, the model is applied to actual distribution network line loss data to verify its diagnostic effect and engineering applicability.

II. Diagnostic model of distribution network quality defects based on cluster analysis

In this chapter, based on the mining of fault characteristic information and data processing representation of distribution network, the main fault type of distribution network, i.e., power line fault, is investigated, and a line loss anomaly identification and diagnosis model based on hybrid clustering algorithm is proposed.

II. A. Distribution Network Fault Characterization Information Mining

II. A. 1) Fault mutation information

When different types of faults occur in the distribution network operation, the bus voltage near the fault point in the network is abnormal and shows different characteristics. The equivalent circuit before and after the fault is shown in Fig. 1, assuming that a fault occurs at the point between bus i and bus j, which is equivalent to injecting a fault current i_f at the fault location f. According to the superposition theorem, the current flowing in the network is the superposition of the fault current and the line current during normal operation. Assuming that the load power in the network does not change at the instant before or after the fault occurs, the moment of the sudden change in the current on the line connected to the bus i and the bus j will reflect the moment of the fault.

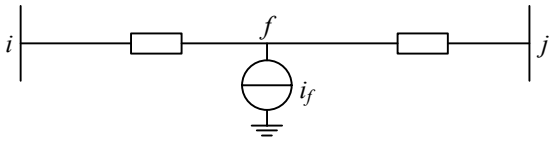

Figure 1: Equivalent circuit before and after fault occurrence

However, in the actual power system, the electrical parameters that change after a fault are not limited to voltage and current, such as active power and reactive power and other electrical parameters will also be affected by the fault and change. Only the collected current or voltage signal as the main criterion for detecting faults is relatively single, and may not be able to accurately detect faults. Therefore, it is necessary to mine the sudden change

information before and after the fault in the massive collected data, to transform the fault detection into anomaly detection, and then realize the fault early warning by using artificial intelligence technology.

II. A. 2) Raw data preprocessing

In view of the existence of vacant values and outliers in the collected distribution network data, in order to avoid the negative impact of poor data quality on the analysis and modeling process, it is necessary to preprocess the original data set. Data preprocessing includes three steps: data completion, outlier processing, and data normalization.

(1) Data Completion: Supplement the original data with missing records and a missing field in the records. In order to shorten the time and improve the accuracy of the supplemented data, linear regression is used to supplement the missing data. Taking the time axis as the horizontal axis and the sampled data as the vertical axis, linear fitting is performed on the non-null data points to derive the calculation formula, and then the time axis of the null points is substituted into the calculation formula to derive the estimated value of the null points.

(2) Abnormal value processing: Since the abnormal values are usually larger or smaller in magnitude than the normal values, and there is the possibility of replacing the normal values by a certain error code, statistical analysis can be used to determine the abnormal values. Convert the raw data into standard Z scores, statistically analyze the sampled data according to Eq. (1), and then substitute it back into Eq. (2) to obtain the corresponding X values:

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

$$X = \sigma Z + \mu \tag{2}$$

where, X is the sampled data, μ is the arithmetic mean, and σ is the variance.

(3) Data normalization: reduce the amount of data by selecting alternative, smaller data, and use a certain model to evaluate the data by storing only the parameters without storing the actual data, e.g., regression and log-linear models. At the same time, the distribution network sampling data are time series signals, the problem of time unsynchronization may occur during the data uploading process, and the data recording time is misaligned, which affects the subsequent fault analysis, and it is also necessary to first statute the data in chronological order.

Since the influence degree of each characteristic variable is closely related to its value range, all variables are pre-normalized according to equation (3):

$$\overline{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

where, x is the value of any of the 25 feature variables, $\overline{x}$ is the normalized value of the feature variable, and $x_{\max}$, $x_{\min}$ are the maximum and minimum of the actual range of the feature variable, respectively.

II. A. 3) Feature variable fusion

Considering the large dimensionality of the operational data collected from the distribution network, if the high-dimensional matrix is directly used for fault identification, it will cause great difficulties in computing and increase the computing cost. In order to gain a more intuitive insight into the characterization space of the dataset, parameter fusion and feature extraction are required for massive data and numerous characteristic variables. Principal Component Analysis (PCA) [23] is an unsupervised data dimensionality reduction method. The basic idea is to transform the data into a new coordinate system by linear transformation, and then utilize the idea of dimensionality reduction so that the first large variance of any data projection is on the first coordinate and the second large variance is on the second coordinate. Therefore, the eigenvector corresponding to the largest eigenvalue is the direction with the largest variance of the data, and the eigenvector corresponding to the smallest eigenvalue is the direction with the smallest variance of the data. The variance changes of the raw data in different directions reflect its intrinsic characteristics. When the power grid is in normal operation, the fluctuation of the data eigenvalue changes obtained by PCA dimensionality reduction is small. When the grid is in a sudden fault, the original data undergoes a sudden change, and the data feature quantity obtained by PCA dimensionality reduction also generates a large change, i.e., the data sample at this moment is inconsistent with the characteristics shown by the overall data sample, and deviates from other data samples in some directions, which means that the data sample is a fault moment sample. Therefore, in this paper, multidimensional feature variable fusion can be realized through PCA dimensionality reduction to characterize the real-time operation status of distribution network in the form of low-dimensional feature quantity.

II. B.Distribution network data processing representation

In this section, the processing and representation of distribution network data is introduced, firstly, the distance representation method is introduced, and different distance representation methods are used for different types of data. Secondly, the normalization method is introduced.

II. B. 1) Distance representation

(1) Marginal distance [24].

It is known that the standard form of a normal distribution for a single variable x can be written as:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)} \tag{4}$$

where μ and σ are the mean and standard deviation of the measurements x, respectively, the size of the vector of measurements is $1\times p$, and the standardization factor $1/\sigma\sqrt{2\pi}$ is designed to standardize the distribution such that the area under the curve is equal to 1. Extending the normal distribution to the multivariate case, its standard form becomes:

$$f(x)=\frac{1}{|\Sigma|^{1/2}\ (2\pi)^{p/2}}e^{-\frac{1}{2}(x_i-\mu)\Sigma^{-1}(x_i-\mu)^T} \tag{5}$$

where Σ is the covariance matrix of matrix X. From equations (4) and (5), it can be seen that in the multivariate case the variance σ^2 is given a replacement by the covariance matrix Σ. From the exponents of the multivariate standard normal distribution, the formula for the Marginal distance can be obtained. For a data matrix $X(n\times p)$, which contains n samples x_i, each sample is represented by p variables. The distance between the sample $x_i(1\times p)$ in the i th row of the data matrix and the row-mean vector $\bar{x}(1\times p)$ of the data matrix is computed as in equation (6):

$$MD_i=\sqrt{(x_i-\bar{x})\Sigma^{-1}(x_i-\bar{x})^T} \tag{6}$$

where i belongs to 1 to n. The above formula is the calculation method of Mahalanobis distance. The Mahalanobis distance can make the variables of different magnitudes have the same standard, which has an important role in principal component analysis, pattern recognition and so on.

(2) Pearson's correlation coefficient

Pearson correlation coefficient is usually used to calculate the degree of similarity between two samples [25]. The value of the correlation coefficient ranges from +1 to -1, the closer to +1, the greater the degree of positive correlation between the two samples, while the closer to -1, the greater the degree of negative correlation between the two samples, and the closer to 0, the lower the degree of correlation between the two samples. Knowing the two vectors $X=[x_1,x_2,\cdots,x_n]$ and the vectors $Y=[y_1,y_2,\cdots,y_n]$, the covariance between the two vectors and their respective standard deviations can be found:

$$cov(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \tag{7}$$

$$\sigma(X,Y)=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})} \tag{8}$$

The Pearson correlation coefficient can be obtained by dividing the covariance between two variables by the standard deviation of the two variables themselves:

$$\rho_{X,Y}=\frac{\mathrm{cov}(X,Y)}{\sigma_x\sigma_y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{9}$$

In certain data distributions, the covariance is not a good representation of the relationship between two variables, and utilizing the Pearson correlation coefficient provides a clearer representation of the correlation between the variables.

II. B. 2) Methods of quantitative standardization

The power grid belongs to a very large and complex network structure system with a variety of different equipment. From the smallest air switch in a house to the largest generator in a hydroelectric power plant and the transformer in a substation. With the introduction of the current RPMS system to the current latest WAMS system, various devices, variables of different magnitudes are recorded. It is necessary to process these data to remove the interference due to magnitude.

(1) Min-max normalization

Min-max normalization makes use of the attribute values of the samples themselves and performs a linear transformation of the samples, and the mapped attribute values will be in $[0,1]$. The i th attribute in the sample X is known to be x_i and the normalization transformation formula is shown in equation (10):

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

where $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the attribute values in the sample.

(2) Z-score normalization

Z-score standardization is to process the raw data through the mathematical expectation and standard deviation of each variable, and the standardization formula is shown in equation (11):

$$X^* = \frac{X - E(X)}{\sqrt{D(X)}} \tag{11}$$

where $E(X)$ and $D(X)$ are the mathematical expectation and variance of the random variable X, respectively. The mathematical expectation and variance of the random variable X^* after performing the Z-score are shown in equations (12) and (13):

$$E(X^*) = E\left[\frac{X - E(X)}{\sqrt{D(X)}}\right] = \frac{1}{\sqrt{D(X)}} E\left[X - E(X)\right] = 0 \tag{12}$$

$$D(X^*) = D\left[\frac{X - E(X)}{\sqrt{D(X)}}\right] = \frac{1}{D(X)} D\left[X - E(X)\right] = \frac{D(X)}{D(X)} = 1 \tag{13}$$

It can be seen that the Z-score normalization method will change the mean of the random variable to 0 and the variance to 1. This method can be used in the case where the minimum and maximum values of the random variable are unknown, and if the data needs to be classified, clustered, or dimensionality reduced using PCA, the Z-score normalization will have a better effect compared to the Min-max normalization.

II. C.Basic fault types in the grid

The common power grid system is made up of a combination of several important parts, including the five main components of power generation, power transformation, power transmission and distribution, and power consumption. This section will provide a brief overview of the common types of faults that occur during grid operation and the general reasons for their occurrence.

II. C. 1) Power line failures

Power lines are the bridges between power grids at all levels connecting power stations, substations and users. Power lines can be categorized into distribution lines and transmission lines according to different usage scenarios. The lines connecting the users and the power supply side are the distribution lines, and the general voltage ranges from low voltage 380V lines to high voltage 10kV lines.

Usually, power line faults can be categorized into two forms: disconnection faults and short-circuit faults. The main causes of disconnection faults are:

(1) Line breaks due to force majeure such as bad weather, natural disasters, or man-made damage.

(2) Due to the aging and corrosion of the connectors between the lines caused by the connectors loose and fall off.

(3) Poor contact of the connection parts or line short circuit caused by the heat of the line short circuit.

The main reasons for short circuit failure are:

(1) Due to bad weather and natural disasters caused by contact between the lines or line grounding short circuit.
(2) Line aging caused by the insulation layer breakage and failure of the line short circuit.
(3) Staff operating errors or human damage caused by short circuit.
(4) Line insulation or insulation connectors on the pollution channel caused by creepage or flashover phenomenon.

In the short-circuit fault, there are mainly single-phase short-circuit, two-phase short-circuit and three-phase short-circuit, of which single-phase short-circuit is the most common and three-phase short-circuit is the rarest. For fault diagnosis, how to accurately identify the fault in the case of few samples is also an important research content.

II. C. 2) Substation failures

The substation is in a key position in the whole power grid system, it is used to change the voltage to adapt to the different needs of each line, step up to meet the needs of long-distance transmission, step down to meet the needs of users for normal use.

Transformer common faults are:

(1) export short-circuit failure, in the transformer occurs external short-circuit resulting in a very short period of time to produce more than the rated current dozens of times the current, if not dealt with in a timely manner will have serious consequences for the winding.

(2) Insulation failure, the transformer is filled with insulating oil, with the transformer for a long time and aging, the insulating oil may produce decomposition and other conditions leading to insulation failure.

(3) Discharge failure, due to some conductors, connectors or switches and other parts of the contact is poor resulting in abnormal discharge phenomenon, in serious cases may cause damage to the equipment and lead to short circuit and other secondary failure.

II. C. 3) Failure of the converter station

The conversion of current in high-voltage DC transmission is realized by the converter station. The converter station has the function of rectification and inversion, which can convert the alternating current and direct current in the power grid to each other. The inverter is a kind of equipment that converts direct current into alternating current. The common failure of inverter is commutation failure. Rectifier and inverter role is the opposite, is the AC power will be converted to DC power of a device. Common failures are unipolar blocking and bipolar blocking.

II. D.Hybrid Clustering Algorithm

In this section, the hybrid clustering algorithm K-Means-AHC is proposed by optimizing the design for the problems of traditional clustering algorithms.

II. D. 1) Selection of optimal number of clustering centers

In order to solve the problem that the value of the number of clustering centers K is set empirically in the traditional K-Means clustering algorithm [26], this section proposes a combination of supervised and unsupervised indicators to select the optimal number of clustering centers by combining the comprehensive indicator data.

The supervised metrics evaluate the clustering effect by calculating the degree of conformity of the clustering results with the benchmark data, and in the training process of the algorithm, for the dataset $X=\{X_1,X_2,\cdots,X_m\}$, it is assumed that the clusters obtained through clustering are divided into $C=\{C_1,C_2,\cdots,C_n\}$, the collected data is divided into $P=\{P_1,P_2,\cdots,P_n\}$, the sample points are compared two by two to get the supervised metrics calculation formula as:

$$S_{sp}=\frac{a}{a+b+c} \tag{14}$$

where a denotes that two samples belong to the same cluster in C and are in the same group in P. b denotes that two samples belong to the same cluster in C but not the same group in P. c indicates that the two samples are in different clusters in C but in the same group in P.

The unsupervised metric evaluates the clustering effect by calculating the profile coefficient, which for any sample point in the dataset is calculated as:

$$S_{usp}(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{15}$$

$$S_{usp} = \frac{1}{m}\sum_{i=1}^{m} S_{usp}(i) \tag{16}$$

where $a(i)$ is the average distance from data point i to all its data points of the same category, and $b(i)$ is the average distance from data point i to its non-similar data points.

Combining the above two indicators yields the composite indicator data, calculated as:

$$S = \frac{S_{sp} + S_{usp}}{2} \tag{17}$$

The value of the composite index S is between $[0,1]$, the larger the value, the better the clustering, the value of K when S obtains the maximum value is used as the optimal number of clustering centers of the algorithm.

II. D. 2) Initial Cluster Center Selection

In order to solve the influence of the selection of initial cluster centers on the algorithm results in the traditional K-Means clustering algorithm, a hybrid clustering algorithm combining K-Means and cohesive hierarchical clustering [27] is proposed.

The hierarchical clustering algorithm is constructed as a clustering tree based on a specified inter-cluster distance measure and a certain termination condition. First, each sample in the dataset is treated as a cluster, and then the distance between the clusters is calculated, and the clusters with the closest distances are aggregated together, and the hierarchical clustering is completed when the specified clustering objective is reached.

There are various ways to calculate the distance between clusters, and in this section, the calculation is carried out through the average distance formula, which can maximize the avoidance of noise interference and improve the accuracy of clustering. When the sample $x_i \in C_i$, $x_j \in C_j$, the calculation formula is:

$$dist(C_i, C_j) = average\left\{dist(x_i, x_j)\right\} \tag{18}$$

Depending on the number of different clustering centers K determined, the algorithm can coalesce the dataset into a specified number of clusters by one operation without the need for multiple operations.

In summary, the hybrid clustering algorithm flow is shown in Fig. 2.

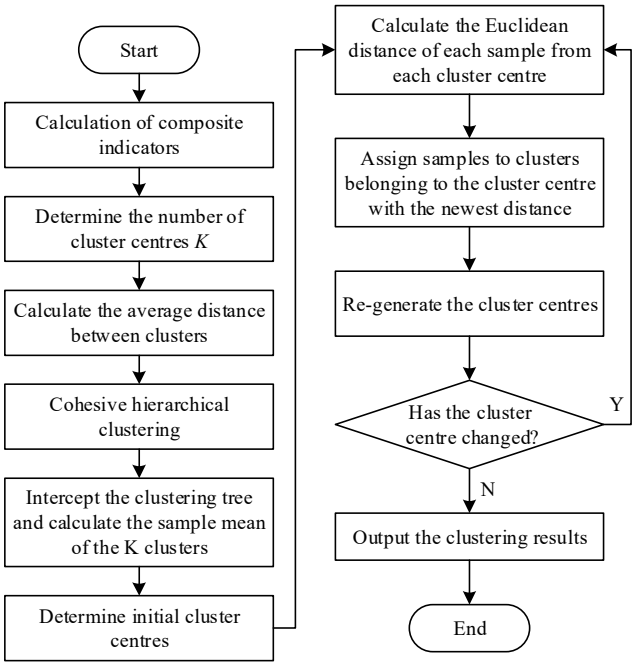

Figure 2: Flow chart of hybrid clustering algorithm

II. E. Diagnostic model for line line loss anomaly identification based on hybrid clustering

Since power line faults are the main fault types faced by distribution networks among the fault types mentioned in the previous section, a line line line loss anomaly identification and diagnosis model is constructed in this section in conjunction with a hybrid clustering algorithm, i.e., the improved K-means clustering algorithm, to realize the diagnosis of quality defects in distribution networks.

II. E. 1) Distribution network line line loss anomaly detection method and process

The flow of the line line loss identification and diagnosis method based on hybrid clustering algorithm proposed in this chapter is shown in Fig. 3.

(1) Collect line historical operation data and perform data preprocessing. The relevant line loss data, equipment ledger data, user file data, etc. are collected through various data collection and management systems of the power company.

(2) Since different feature data have different scales and orders of magnitude, in order to facilitate the calculation of the algorithm and improve the accuracy of the clustering algorithm, the indicator collection data are preprocessed.

(3) Based on the hybrid clustering algorithm, feature variables are extracted, clustered and analyzed for the operation data of multiple distribution lines. There are many factors related to the line loss rate of the distribution network, and by constructing the line loss optimization feature indicator system, it is used as the input variable of the clustering model.

(4) According to the clustering results, line loss anomaly detection is carried out, and the next step of analysis is implemented for the detected possible anomalous lines.

(5) Output the abnormal diagnosis results for field inspectors' reference, and compare them with the actual field inspection results, and feedback the comparison results to construct the abnormal line loss sample set.

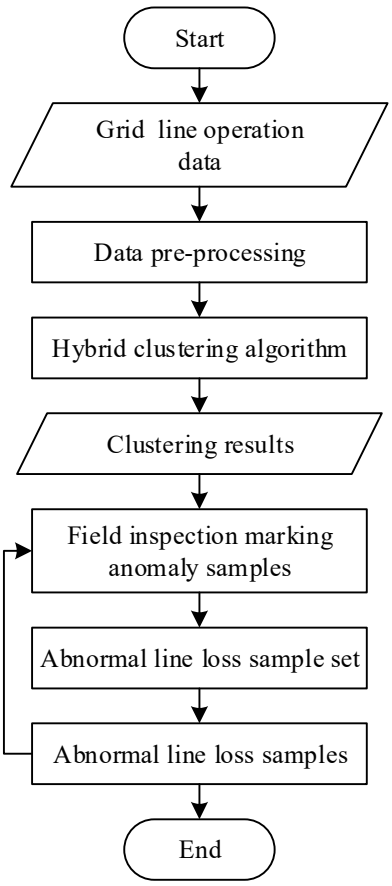

Figure 3: Line loss identification and diagnosis process based on hybrid clustering algorithm

II. E. 2) Line loss anomaly diagnosis model based on hybrid clustering algorithm

The line line loss anomaly diagnosis model of distribution network based on hybrid clustering algorithm constructed in this paper is shown in Fig. 4.

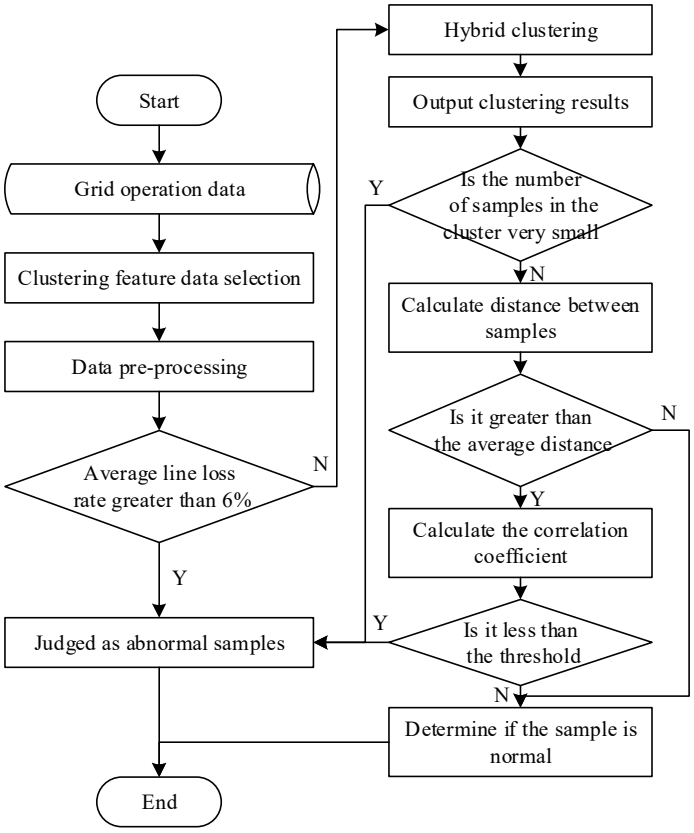

Figure 4: Flow chart of line loss anomaly detection clustering model

(1) Selection of clustered feature data

By analyzing and comparing various factors related to line loss rate, this paper selects five clustered feature data, namely, power factor, power supply, average value of line loss rate, coefficient of variation of line loss rate, and meter opening record, and the determined clustered feature data take into full consideration of the various factors of line loss of the distribution network, which can scientifically and comprehensively reflect the characteristics of different line loss rate data.

(2) Characteristic data preprocessing

Collect the 31-day operation data of N feeders, i.e., the clustered feature data determined by step (1). In order to facilitate the calculation and improve the accuracy of the clustering results, the influence factor collection data is preprocessed.

(3) Perform hybrid clustering

a) Select the optimal clustering center number K value by combining supervised and unsupervised indicators with comprehensive indicator data, i.e., it is hoped that the data set will be clustered to obtain K sets.

b) Select initial clustering centers by combining K-Means and hierarchical clustering.

c) For all data points, calculate their Euclidean distances from each clustering center separately, and classify the data point into the category where the center with its smallest distance is located.

d) After categorizing all the data into good sets, there are a total of K sets. Then recalculate the clustering center for each set.

e) When the distance between the new clustering center and the initial clustering center is not greater than a set threshold, it is determined that the clustering has achieved the desired effect and the algorithm ends.

f) If the distance between the new clustering center and the initial clustering center changes a lot, then continue to iterate c) ~ e) steps.

g) Output clustering results.

(4) Line loss abnormality diagnosis

According to the clustering results, diagnose whether there is line loss abnormality in distribution network lines, and the decision-making process is divided into the following three steps:

a) According to the current standards set by the power company, lines with an average line loss rate of more than 6% over a period of time are judged as line loss abnormal lines.

b) For lines with average line loss rate less than 6%, hybrid cluster analysis is performed to obtain the clustering results. Since the number of abnormal line loss samples in the dataset is a small number, the cluster $C_n(n=1,2,\cdots,k)$ containing very few samples in the clustering result $C=\{C_1,C_2,\cdots,C_k\}$ is regarded as an anomalous class, and the samples contained in C_n are determined as the lines that may have abnormal line loss.

c) For clusters containing a large number of samples in the clustering result, the samples located at the edge position within the cluster are determined as anomalies by Euclidean distance and correlation coefficient. The Euclidean distance d between each sample x in the cluster and the clustering center y is calculated by equation (19), and the samples whose distance is larger than the average Euclidean distance are classified as edge samples:

$$d=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2} \tag{19}$$

The correlation coefficient between the line loss time-series data curve of the edge sample and the line loss time-series data curve of the cluster clustering center to which it belongs is calculated by Eq. (20), and the larger the correlation coefficient is, the greater is the similarity between the two curves:

$$\rho=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{20}$$

III. Model application experiment and result analysis

In this chapter, the proposed hybrid clustering-based line line loss anomaly identification and diagnosis model for distribution networks is applied to verify its diagnostic effect.

III. A. Experiments on the application of hybrid clustering algorithm

III. A. 1) Case studies

According to the research method of cohesive hierarchical clustering optimization k-means clustering, the distribution network line loss data is selected for clustering division. Firstly, the horizontal coordinate is set as electricity sales, and the vertical coordinate is the line loss rate, and the line loss data are coarsely clustered, and the obtained coarse clustering results are shown in Fig. 5. Because of the normalization of the data, the results in the figure are shown in the range of horizontal and vertical coordinates from 0 to 100.

Each different color clustering cluster represents a different clustering circle, in which the pentagrams of each color represent different clustering centroids, and the blue solid circle in each cluster is the region where the strong correlation points are located, and the area between the peripheral larger hollow circle and the solid circle is the region where the weak correlation points are located. The cohesive hierarchical clustering algorithm clusters the line loss data into 10 classes, and the number of line loss points within each cluster is different. In the clustering results, there are more data points in the bright yellow and dark yellow regions, which can be seen that users with less electricity consumption and lower line loss rate account for the majority, and there are fewer data points in the light red, dark cyan, and tan regions, which can be seen that users with more electricity sales and lower line loss and users with less electricity sales and higher line loss account for the minority. However, from Figure 5, it can also be seen that the Canopy algorithm is not reasonable enough to divide the boundary values, and the number of data points in the clustering circles that are close to each other is large, so k-means clustering is needed to do further clustering.

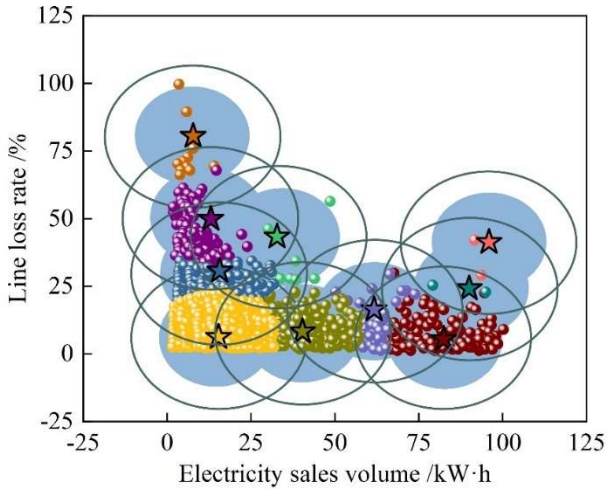

Figure 5: Agglomerative hierarchical clustering simulation

Next, the cohesive hierarchical clustering center and the number of clusters are used as the initial clustering center and the initial number of clusters of the k-means algorithm, and the k-means algorithm is used to do further fine clustering of the line loss data. k-means clustering results are shown in Fig. 6. The line loss data points are divided into 10 clusters, different color points represent different clusters, and the five-pointed star represents the clustering center of each cluster.

It can be seen that the bright yellow cluster is the area with the least amount of electricity sold and the lowest line loss rate. The dark blue cluster is the area with the highest electricity sales and lower line loss rate. The brownish yellow cluster is the region with lower electricity sales and higher line loss rate. The traditional fixed-threshold method is used to determine line loss anomalies, and the brownish-yellow clusters that exceed the set line loss rate thresholds are usually judged as abnormal areas, while the bright yellow clusters and dark blue domains are regarded as normal areas. However, the characteristics of electricity consumption vary from region to region, and a one-size-fits-all judgment of line loss anomalies can result in large errors. For example, areas within the brown-yellow clusters may have some households within the station area with power theft, resulting in a low line loss rate despite the large amount of electricity sold, while areas in the brown-yellow domain may be heavy industrial plants with a high line loss rate in the industrial production process. Therefore, it is necessary to cluster and divide the line loss data to learn the law of line loss within the same cluster of data, so as to realize the anomaly diagnosis for the future data in this region.

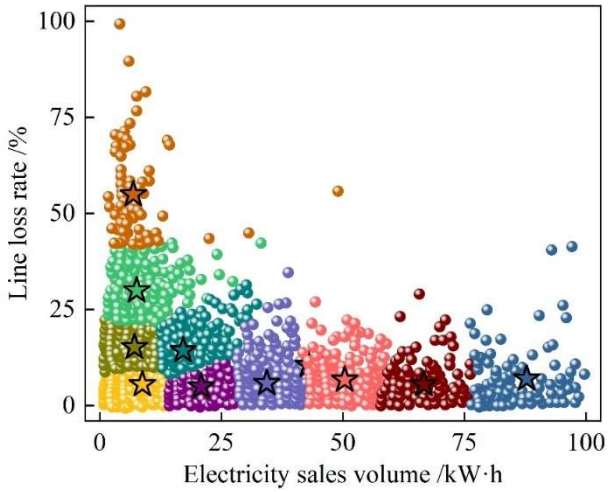

Figure 6: K-means clustering results

III. A. 2) Validation of clustering effect

In order to verify the final implementation of K-means clustering optimized by the cohesive hierarchical clustering algorithm, the contour coefficient method is selected for verification. Contour coefficient is an evaluation method to

verify the effect of clustering, which is derived from the degree of cohesion and separation of the data after clustering. The degree of cohesion of the data is the average value of the Euclidean distance from the sample points to other points within the same cluster, and the degree of separation of the data is the average Euclidean distance from the sample points to all points in different clusters. The contour coefficient was calculated as:

$$S(i) = \frac{U(i) - V(i)}{\max\{V(i), U(i)\}} \tag{21}$$

where $S(i)$ is the contour coefficient of sample point i, $U(i)$ is the separation degree of sample point i, and $V(i)$ is the cohesion degree of sample point i. It can be seen that the value of contour coefficient is between -1 and 1, the more the contour coefficient tends to 1 represents that the degree of cohesion and separation are relatively better, which also proves that the clustering effect is better. The same line loss data is handed over to the single K-means algorithm, so that the number of clusters is from 2 to 18 for clustering division, and the statistics of clustering time length, clustering average error and contour coefficients after the completion of each clustering.

The results of the contour coefficient method to validate the number of clusters K are shown in Fig. 7, which shows the contour coefficients computed by the single K-means algorithm at different numbers of clustering clusters. Although the contour coefficients are maximum when the number of clustering clusters is 2, it does not follow the business logic. Therefore, when the number of clustering clusters is chosen to be 10 and the contour coefficients are larger, it represents a better clustering effect as well, which proves the correctness of the results of the cohesive hierarchical clustering optimization K-means clustering.

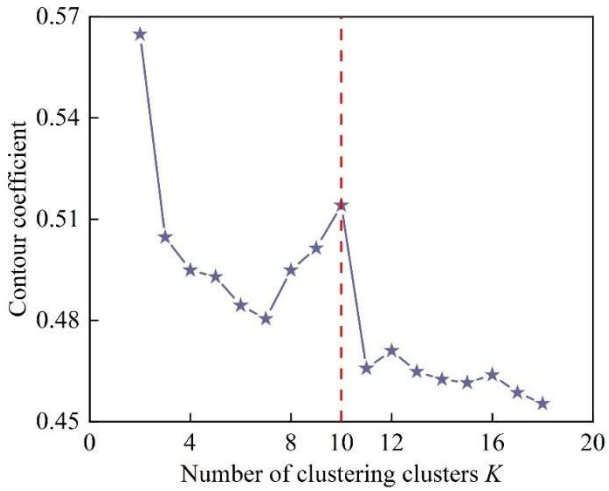

Figure 7: The number of clusters *K* is verified by the contour coefficient method

The cohesive hierarchical clustering optimization K-means clustering algorithm is compared with single K-means clustering in terms of algorithm duration, clustering average error and profile coefficient, and the results of clustering algorithm comparison are shown in Table 1.

Using cohesive hierarchical clustering to optimize the K-means clustering algorithm not only accelerates the speed of the algorithm, but also improves the clustering effect, the algorithm length is reduced by about 26.5 s, the average clustering error is reduced from 39.4832 to 7.8469, and the profile coefficient is increased by 0.0851. Using cohesive hierarchical clustering to firstly carry out a coarse clustering of the data, to obtain the number of cluster clusters K, and the approximate K clustering centers, and then use K-means for further fine clustering.K-means is weak against noise, and it is beneficial to remove smaller clusters directly through cohesive hierarchical clustering to prevent interference. The center of each cluster selected by cohesive hierarchical clustering is more scientific as the clustering center of K-means. Doing K-means clustering for the content of each cluster after cohesive hierarchical clustering also reduces the number of similar calculations.

Table 1: Comparison of clustering algorithms

Clustering algorithm	Algorithm duration /s	Average clustering error	Contour coefficient
k - means	38.5724	39.4832	0.5122
Optimized K-means	12.0361	7.8469	0.5973

III. B. Verification experiment of the effect of new anomaly recognition and diagnosis

III. B. 1) Verification Experiment on the Effectiveness of New Anomaly Recognition

In order to verify the effectiveness of the proposed improved K-means clustering algorithm for the identification and diagnosis of new anomalies in distribution network line losses, this paper randomly draws 4 single anomalies and 11 double anomalies as new anomalies from 250 types of anomalies, and the rest of the 235 categories as known anomalies, and retrains the samples of known anomalies of the 235 categories, to obtain the new anomaly diagnostic model $M^{(1)}$, which is not trained for the 15 categories of anomalies drawn from the model. And there is no training in this model for the drawn samples of 15 classes of anomalies, so for the model $M^{(1)}$ these 15 classes of anomalies are considered to be new anomalies that can not be recognized and diagnosed correctly.

In order to experimentally verify the recognition effect of the clustering algorithm on the 15 classes of new anomalies, the 15 classes of new anomalies are first diagnosed with the trained new model $M^{(1)}$, and after outputting the anomalies Y_i, then the sample set corresponding to the anomalies Y_i is clustered to obtain the corresponding center of mass $\bar{\mu}_j^*$ and threshold V_i, then calculate the distance D_i from the 15 new anomaly feature information sets to the clustering center of mass $\bar{\mu}_j^*$, and judge whether it is a new anomaly or not by comparing the magnitude of D_i and V_i, if D_i is greater than V_i then it is a new exception and vice versa. There are 15 classes of new anomalies and 600 samples, and the identification of 15 classes of new anomalies is shown in Table 2. Among them, number 1-4 are single anomaly types and number 5-15 are double anomaly types.

Observing the data, it can be seen that the recognition accuracy rate of single anomaly type is 100%, and the recognition effect is very good. The recognition accuracy rate of double anomaly is 97.95%, the effect is not as good as single anomaly, and there are three types of recognition errors. Combined with the distribution network linear line loss anomaly inference knowledge base can be seen, the recognition rate will be reduced when the anomaly feature information is more complex, and the comprehensive recognition accuracy rate is 98.50%, which is in line with the engineering needs.

Table 2: Identification results of 15 new anomalies

Abnormal number	Total sample size	Identify the number of successful samples	Identify the number of incorrect samples	Recognition accuracy rate /%	Accumulate the recognition accuracy rate /%
1	40	40	0	100.00	100.00
2	40	40	0	100.00	100.00
3	40	40	0	100.00	100.00
4	40	40	0	100.00	100.00
5	40	40	0	100.00	100.00
6	40	40	0	100.00	100.00
7	40	40	0	100.00	100.00
8	40	38	2	95.00	99.38
9	40	40	0	100.00	99.44
10	40	40	0	100.00	99.50
11	40	37	3	92.50	98.86
12	40	40	0	100.00	98.96
13	40	36	4	90.00	98.27
14	40	40	0	100.00	98.39
15	40	40	0	100.00	98.50

III. B. 2) Verification experiments on the effect of new anomaly diagnosis

In this paper, 40 samples are collected for each type of anomaly, firstly, the samples of 15 types of new anomalies are processed as follows, 2 samples are randomly selected from 40 samples of each type of anomalies in each round as the new anomalies of the actual inputs, and then 2 samples are randomly selected from the 38 samples remaining in each type of anomalies as the two samples that are manually introduced based on the new anomaly feature dataset, and finally, the 4 samples are oversampled to get 40 samples.

Each sample $S_{j,k}$ consists of anomaly feature information set and anomaly type as shown in equation (22):

$$S_{j,k} = \{X, Y\} \tag{22}$$

where j is the corresponding number of the anomaly type, k is the kth sample of the jth type of anomaly, $j = 0,1,2,\cdots,250$; $k = 0,1,2,\cdots,40$.

Taking the anomaly “insufficient output of the test device” as an example, this paper chooses the trend of the feature information data of the anomaly “insufficient output of the test device” to have a more intuitive understanding of the feature information, which has 40 sample points, each sample point has 96 time The data has 40 sample points, each with 96 time dimensions, and the characteristic information data obtained in the laboratory is shown in Fig. 8, and the characteristic information data obtained from oversampling is shown in Fig. 9. In order to facilitate the observation of the data characteristics, are the data before normalization processing, the data outline is different, where the unit of abnormal characteristic quantity is 1, the unit of abnormal sampling current value is A, and the unit of abnormal sampling voltage value is V.

As can be seen from Fig. 8, the feature information dataset obtained in the laboratory is divided into three categories at the sampling value data, mainly because there are differences in the three current voltages when simulating the three cases of phase A ground fault, phase B ground fault and phase C ground fault of the line. Comparison of Fig. 8 and Fig. 9 shows that the sample obtained by randomly selecting a sample for oversampling has only one simulated case, and the data error is slightly larger than that obtained from the experimental simulation.

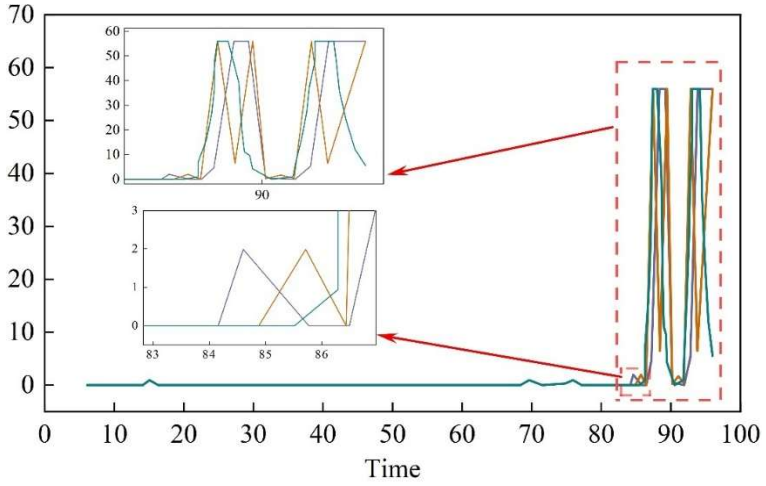

Figure 8: Feature information data obtained from the experiment

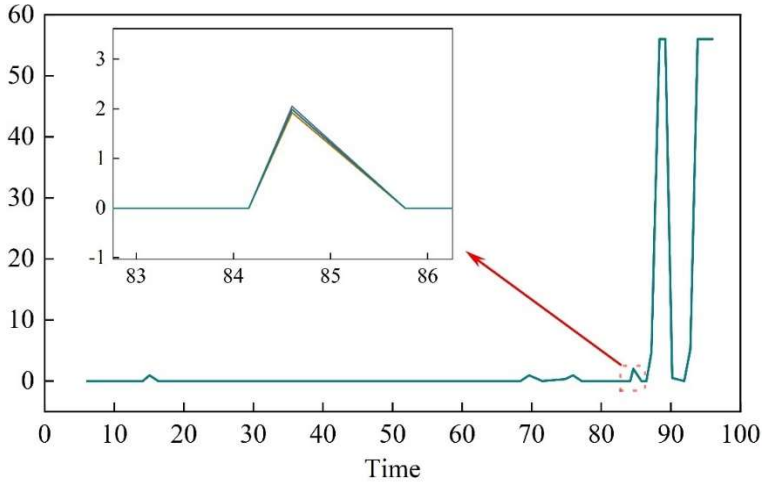

Figure 9: Feature information data obtained through oversampling

Eventually 600 samples are obtained using oversampling, 15 samples from each class of anomalies are randomly selected and added to the training set for model updating, and 12 samples from each class of anomalies obtained from laboratory testing are randomly selected and added to the test set. The model $M^{(1)}$ is updated with the new training set, and the new model $M^{(2)}$ is obtained, and then the model $M^{(2)}$ is used to test the model on the 180 new anomaly samples, and the curve of the change of Accuracy during the training process of the model M and $M^{(1)}$ is shown in Figure 10. It can be seen that the model $M^{(1)}$ converges faster, and the stabilized posterior value

is slightly lower than that of model M. And the number of samples in the training set is reduced, speeding up the convergence speed while slightly reducing the accuracy.

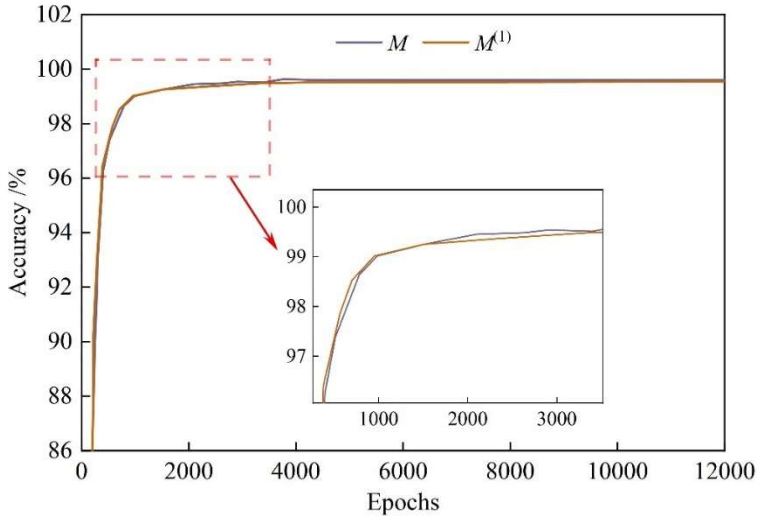

Figure 10: Accuracy curves of the two models

The results of $M^{(2)}$ tested on 180 samples of new anomalies are shown in Table 3, and the training time of model M and $M^{(1)}$ and $M^{(2)}$ are shown in Table 4. The updated model $M^{(2)}$ is good at diagnosing new anomalies with an accuracy of 99.72%, which meets the engineering requirements. And the training time of $M^{(1)}$ is 13.43h, which is slightly lower than that of $M^{(2)}$ and M. After the model structure is determined, the training time mainly depends on the size of the training set and the number of iterations of training.

Table 3: The diagnostic effect of model $M^{(2)}$ on new abnormalities

Model	MAE	Accuracy /%	F1-score /%
$M^{(2)}$	0.0059	99.72	92.14

Table 4: The training time of the three models

Model	M	$M^{(1)}$	$M^{(2)}$
Training time /h	14.72	13.43	14.71

It can be obtained that the method based on the improved K-means clustering and oversampling in this paper can effectively identify and diagnose new anomalies appearing in the testing process, and the diagnostic accuracy is high, which effectively improves the scalability of the abnormal diagnosis method of distribution network line loss.

IV. Design of Smart Distribution Grid Condition Monitoring System Based on Electricity IoT

Based on the distribution network quality defect diagnosis model, this chapter designs an intelligent distribution network condition monitoring system based on power IoT to realize the intelligent operation and maintenance of distribution network.

IV. A. Overall system architecture

The overall architecture of the smart distribution network condition monitoring system designed in this paper is shown in Fig. 11, which adopts a layered architecture consisting of a perception layer, a network layer, a data processing layer and an application layer. The workflow of the system is as follows: the sensing layer collects multi-source heterogeneous data from distribution equipment through various types of sensors, and transmits them to the cloud through the network layer. The data processing layer stores, cleans and fuses the massive monitoring data to extract features and diagnose faults. The application layer provides decision support for distribution network scheduling and operation and maintenance based on the data analysis results.

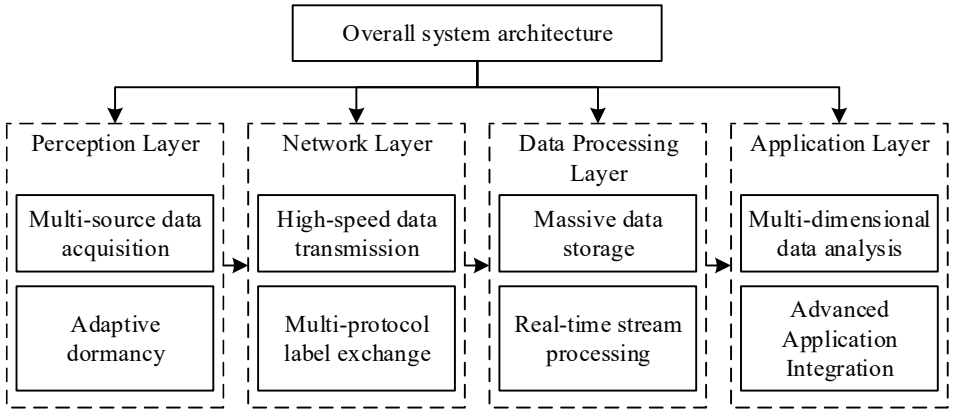

Figure 11: Basic composition and functions of the system

IV. B. System Hierarchy

IV. B. 1) Perception layer

The sensing layer transmits the data collected by each device, such as temperature, current, voltage, and environmental parameters, to the network layer in real time through the RS-485 bus. In the data transmission process, a multi-device conflict avoidance mechanism based on Modbus-RTU protocol is adopted. When the number of devices is n, the waiting time of each device obeys the exponential distribution with parameter $\lambda = \frac{n}{t}$, i.e., probability density:

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0 \tag{23}$$

where t denotes the bus idle time. By reasonably setting the value of λ, the probability of bus conflict can be effectively reduced and the data transmission efficiency can be improved. In addition, the sensing layer introduces an adaptive sleep mechanism based on asynchronous time-division multiplexing, which dynamically adjusts the wake-up period of the device according to the frequency of data changes, which reduces the energy consumption of the device and prolongs the life cycle of the system while guaranteeing the monitoring quality.

IV. B. 2) Network layer

The network layer plays the role of connecting the sensing layer with the data processing layer in the system, and is responsible for efficiently and reliably transmitting the massive data collected in the sensing layer to the cloud. This layer is mainly composed of industrial Ethernet switches, 4G routers and other communication devices.

Inside the network layer, for the characteristics of large geographic span and complex environment of the distribution network, the Multiprotocol Label Switching (MPLS) technology is adopted to realize the fast grouping and forwarding of data. Let the packet size be L bits and the link transmission rate be R bps, the transmission delay T of a single packet can be expressed as:

$$T = \frac{L}{R} \tag{24}$$

By reasonably setting the L and R parameters, the transmission delay can be minimized while guaranteeing the transmission quality to improve the real-time performance of the network. In addition, the network layer adopts a traffic scheduling mechanism based on software-defined networking (SDN) to dynamically optimize the data transmission path according to the link state, which avoids link congestion and reduces the packet loss rate while improving the robustness of the network.

IV. B. 3) Data processing layer

The data processing layer is the core of the whole system, bearing the heavy responsibility of data storage, calculation and analysis. This layer is based on distributed computing architecture and is configured with high-performance server clusters and massive storage arrays. Among them, the RH2288H V3 server based on Intel Xeon E7-8890v4 processor is adopted, and a single node can provide up to 3TB of memory capacity. The DS8870 model was selected for the storage array, supporting 32Gb fiber interface and a single cabinet capacity of up to

1.64PB. The data processing layer uses Hadoop as the basic framework, and seamlessly integrates real-time streaming processing with offline batch processing by deploying components such as Spark and Storm.

Real-time data is aggregated to the Storm cluster through Kafka at a rate of 2.6 million entries/s, and after a series of pre-processing such as cleaning and conversion, the distribution network load prediction model is constructed based on the least-squares method, i.e:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{25}$$

where y is the predicted load, x_i is the influence factor, and β_i is the parameter to be estimated, and the prediction result is finally stored in Redis cache for application layer to call.

The offline data is then introduced by Sqoop tool and synchronized to Hive data warehouse every hour, combined with Impala to achieve sub-second query response, providing data support for advanced applications such as distribution network trend calculation and state estimation. Meanwhile, the data processing layer also integrates the machine learning library Mahout, which mines the fault occurrence laws embedded in massive historical data by constructing prediction models such as SVM and Random Forest.

IV. B. 4) Application layer

The application layer is located in the top layer of the system, which is directly facing the business applications such as power distribution network dispatching, operation and maintenance, etc. This layer is based on the B/S architecture, and builds a visual human-computer interaction interface. The application layer introduces advanced Web 3D rendering technologies, such as WebGL and three.js, and combines them with a high-precision GIS platform to realize the realistic rendering of three-dimensional scenes of the distribution network, which greatly improves the usability of the system. In addition, the layer also integrates Pentaho, an open-source business intelligence suite under the Hadoop ecosystem, which supports OLAP, data mining and other multi-dimensional analysis functions. Users can easily generate various types of statistical analysis reports through a Web browser to assist management decisions.

The functional process of the application layer can be summarized as three links: data acquisition, indicator calculation and visual presentation. Take the load forecasting function as an example, first of all, get the forecast value from Redis by calling API, and then combine it with the actual load data to calculate the forecast deviation ΔP, that is:

$$\Delta P = \frac{\left| P_p - P_r \right|}{P_r} \times 100\% \tag{26}$$

where P_p is the predicted load and P_r is the actual load, the units are kW. when the deviation exceeds the threshold, the system automatically triggers a warning, generates an alarm event, and labels the abnormal load nodes on the GIS platform, which guides the on-duty personnel to locate the problem quickly. In addition, the system also provides a series of advanced application functions, such as trend calculation and state estimation, by calling MATLAB engine, executing the corresponding algorithms in the background and pushing the calculation results to the front-end interface via WebSocket, realizing a seamless connection between the calculation and presentation.

IV. C. Experimental verification of system simulation

IV. C. 1) Simulation experiment program and evaluation indexes

In order to evaluate the performance of this system, a complete set of simulation experiment program is designed in this paper. Firstly, 12 station areas and 60 10kV distribution lines are selected as test objects in the distribution network of a typical rural area in region S. IoT sensing devices such as intelligent distribution substation terminals, distribution automation terminals, fault indicators, and environmental monitors are deployed and accessed to this system for data collection, analysis, and management. At the same time, three key indicators of power supply reliability, power quality and operation and maintenance efficiency are selected as the basis for evaluating the system performance, including system average outage time SAIDI, system average outage frequency SAIFI, voltage pass rate, reactive power compensation efficiency, fault location time, etc., and the corresponding thresholds and weighting coefficients are set to construct a comprehensive evaluation model. In addition, different fault scenarios and load curves are designed to simulate the actual operating conditions of the distribution network, and to investigate the system's response capability and robustness under abnormal conditions.

IV. C. 2) Experimental results and analysis

In order to verify the actual effect, this paper carries out a 6-month simulation experiment and collects a large amount of operational data and performance indicators, and the comparative results of power supply reliability

indicators, power quality indicators and operation and maintenance efficiency indicators are shown in Tables 5 to 7, respectively.

As can be seen from Table 5, the power supply reliability indicators SAIDI and SAIFI are reduced by 50.28% and 60.81% respectively after the adoption of this system, which greatly improves the continuity and stability of power supply in rural areas, mainly due to the system's unified real-time sensing and intelligent analysis capabilities, which can timely detect and dispose of all kinds of faults and avoid the occurrence of large-scale power outages.

Table 6 shows that the system has also achieved significant results in improving the voltage compliance rate and reactive power compensation efficiency, which guarantees the power quality of rural users. In addition the system improves the reactive power compensation efficiency by 9.36 percentage points through precise scheduling and optimized allocation, which significantly reduces the reactive power loss of the line.

Observing Table 7, it can be seen that this system also greatly improves the intelligent level and working efficiency of distribution network operation and maintenance. For example, the fault location time is shortened from 22.98 min in the traditional mode to 5.46 min, a reduction of 76.24%, which means that the emergency personnel are able to respond quickly and locate the fault accurately, which significantly shortens the outage time. At the same time, the system also improves the completion rate of maintenance tasks by 12.68 percentage points by optimizing the operation and maintenance strategy and resource allocation, which guarantees the timeliness and reliability of operation and maintenance work.

Comprehensive analysis of the above, the simulation experiment results fully proved the significant advantages of this system in improving the power supply reliability, power quality and operation and maintenance efficiency of rural distribution networks, which is of great significance for promoting the intelligent development of rural power grids.

Table 5: Comparison of power supply reliability indicators

Indicator	This system	Traditional mode	Improvement rate
SAIDI (min/household · year)	178.9	359.8	50.28%
SAIFI (times/household · year)	2.9	7.4	60.81%

Table 6: Comparison of power quality indicators

Indicator	This system	Traditional mode	Improvement rate
Voltage qualification rate (%)	99.15	95.44	3.89%
Reactive power compensation efficiency (%)	96.37	88.12	9.36%

Table 7: Comparison of operation and maintenance efficiency indicators

Indicator	This system	Traditional mode	Improvement rate
Fault location time (min)	5.46	22.98	76.24%
Completion rate of maintenance tasks (%)	98.13	87.09	12.68%

V. Conclusion

Intelligent operation and maintenance of distribution networks is of great significance to improve the operation efficiency and reliability of power systems. The quality defect diagnosis model of distribution network based on hybrid clustering algorithm proposed in this study has been experimentally verified and achieved remarkable results. The model achieves 98.50% accuracy in abnormality diagnosis, which greatly improves the fault identification ability of distribution network. By comparing with the traditional K-means clustering algorithm, the optimized hybrid clustering algorithm is optimized in terms of clustering time and error, the average clustering error is reduced from 39.48 to 7.85, and the contour coefficient is improved from 0.5122 to 0.5973. In addition, in terms of the new anomaly identification, the model reaches 99.72% in identification accuracy, which proves its strong anomaly detection capability. Through this innovative method, the accuracy and efficiency of quality defect diagnosis in distribution networks have been greatly improved, providing solid technical support for intelligent operation and maintenance.

Funding

Supported by Science and Technology Project of Guangxi Power Grid Co., Ltd.: Research on Automatic Acceptance Technology of Distribution Network UAV Based on Visual Fusion Tracking and Image Recognition (040600KC24010001).

References

[1] Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2022). Statistical power for cluster analysis. BMC bioinformatics, 23(1), 205.

[2] Scitovski, R., Sabo, K., Martínez-Álvarez, F., & Ungar, Š. (2021). Cluster analysis and applications (pp. 08-23). Cham: Springer.

[3] Brusco, M. J., Singh, R., Cradit, J. D., & Steinley, D. (2017). Cluster analysis in empirical OM research: survey and recommendations. International Journal of Operations & Production Management, 37(3), 300-320.

[4] Thrun, M. C. (2018). Approaches to cluster analysis. In Projection-based clustering through self-organization and swarm intelligence: Combining cluster analysis with the visualization of high-dimensional data (pp. 21-31). Wiesbaden: Springer Fachmedien Wiesbaden.

[5] Siraj, S. E. B., Sing, T. Y., Raguraman, R., Marimuthu, P. N., & Nithiyananthan, K. (2016). Application of Cluster Analysis and Association Analysis Model Based Power System Fault Identification". European Journal of Scientific Research, 138, 16-28.

[6] Coletta, G., Vaccaro, A., Villacci, D., & Zobaa, A. F. (2018). Application of cluster analysis for enhancing power consumption awareness in smart grids. In Application of Smart Grid Technologies (pp. 397-414). Academic Press.

[7] Mishra, S., Das, D., & Paul, S. (2017). A comprehensive review on power distribution network reconfiguration. Energy Systems, 8, 227-284.

[8] Mirshekali, H., Dashti, R., Keshavarz, A., Torabi, A. J., & Shaker, H. R. (2020). A novel fault location methodology for smart distribution networks. IEEE Transactions on smart grid, 12(2), 1277-1288.

[9] Dashtdar, M. (2018). Fault location in distribution network based on fault current analysis using artificial neural network. International Journal of Electrical and Computer Sciences (IJECS), 1(2), 18-32.

[10] Gururajapathy, S. S., Mokhlis, H., & Illias, H. A. (2017). Fault location and detection techniques in power distribution systems with distributed generation: A review. Renewable and sustainable energy reviews, 74, 949-958.

[11] Florencias-Oliveros, O., Agüera-Pérez, A., González-de-la-Rosa, J. J., Palomares-Salas, J. C., Sierra-Fernández, J. M., & Montero, Á. J. (2017, April). Cluster analysis for Power Quality monitoring. In 2017 11th IEEE International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG) (pp. 626-631). IEEE.

[12] Bíscaro, A. A. P., Pereira, R. A. F., Kezunovic, M., & Mantovani, J. R. S. (2015). Integrated fault location and power-quality analysis in electric power distribution systems. IEEE Transactions on power delivery, 31(2), 428-436.

[13] Nsaif, Y. M., Lipu, M. H., Ayob, A., Yusof, Y., & Hussain, A. (2021). Fault detection and protection schemes for distributed generation integrated to distribution network: Challenges and suggestions. IEEE Access, 9, 142693-142717.

[14] Ge, L., Li, Y., Li, Y., Yan, J., & Sun, Y. (2022). Smart distribution network situation awareness for high-quality operation and maintenance: a brief review. Energies, 15(3), 828.

[15] Wang, Z., Zhao, C., Shi, Y., & Kang, Z. (2025). Intelligent operation and maintenance system design based on power distribution networks. In Equipment Intelligent Operation and Maintenance (pp. 829-838). CRC Press.

[16] Yan, Y., Liu, Y., Fang, J., Lu, Y., & Jiang, X. (2021). Application status and development trends for intelligent perception of distribution network. High Voltage, 6(6), 938-954.

[17] Kazmi, S. A. A., Shahzad, M. K., Khan, A. Z., & Shin, D. R. (2017). Smart distribution networks: A review of modern distribution concepts from a planning perspective. Energies, 10(4), 501.

[18] Ghadi, M. J., Ghavidel, S., Rajabi, A., Azizivahed, A., Li, L., & Zhang, J. (2019). A review on economic and technical operation of active distribution systems. Renewable and Sustainable Energy Reviews, 104, 38-53.

[19] Wang, Y. B., Wang, Z., Zhao, D. Y., & Zhang, H. F. (2015). Intelligent Operation and Maintenance of Substations Based on Internet of Things (IoT) Technology. Applied Mechanics and Materials, 742, 708-716.

[20] Jureedi, N. V. V. K., Rosalina, K. M., & Kumar, N. P. (2020). Clustering analysis and its application in electrical distribution system. International Journal Of Recent Advances in Engineering & Technology, 8, 38-43.

[21] Dong, F., Hou, Y., Li, W., & Wang, Y. (2022). Intelligent decision-making of distribution network planning scheme with distributed wind power generations. International Journal of Electrical Power & Energy Systems, 136, 107673.

[22] Li, D., Huo, J., Wang, Y., Li, J., & Jin, H. (2025). The Operation and Maintenance Strategy of Smart Grid based on Intelligent Perception and Optimization Algorithm. Scalable Computing: Practice and Experience, 26(2), 907-915.

[23] Na Wang, Feng Li & Xiaodong Fan. (2025). Quality assessment of Bupleurum chinense by coupling Belousov-Zhabotinsky oscillation with principal component analysis. International Journal of Electrochemical Science,20(7),101033-101033.

[24] Hongfei Zu, Jing Zhu, Xinfeng Wang, Xiang Zhang, Ning Chen, Gangxiang Guo & Zhangwei Chen. (2025). Research on self-adaptive grid point cloud down-sampling method based on plane fitting and Mahalanobis distance Gaussian weighting. Neurocomputing,634,129746-129746.

[25] Spyros Tserkis, Syed M. Assad, Ping Koy Lam & Prineha Narang. (2025). Quantifying total correlations in quantum systems through the Pearson correlation coefficient. Physics Letters A,543,130432-130432.

[26] Jun Lu, Tingjin Luo & Kai Li. (2025). A forward k-means algorithm for regression clustering. Information Sciences,711,122105-122105.

[27] original/au/au-aff. (2018). Movie Recommender System Using Two Way Filtering and Agglomerative Hierarchical Clustering. Journal of Computational and Theoretical Nanoscience,15(6),2269-2272.