A Classified Warning Method for Heavy Overload in Distribution Networks Considering the Characteristics of Unbalanced Datasets

Guohui Ren*
Yuncheng Power Supply Company of State Grid Shanxi Electric Power Company, Yuncheng, Shanxi, 044000, China

***Abstract*—In order to achieve heavy overload warning and capacity planning for the distribution network, it is necessary to classify the heavy overload warning of the distribution network. A distribution network with heavy overload classification warning method based on imbalanced dataset feature extraction is proposed. Screening the feature indicator set related to distribution network overload, constructing a hierarchical prediction framework for distribution network load situation, combining information such as power distribution points, road construction, municipal planning, and power load distribution to form distribution network capacity planning and line renovation plans. Based on K-means clustering, the undersampling method is used to extract features from the unbalanced dataset of distribution network overload classification, using decision trees as the basic learning unit. It includes multiple decision trees trained by Bagging integrated learning theory and random subspace method. The random forest algorithm is used to realize the feature detection and distribution network capacity planning of distribution network weight overload grading, and the grading early warning of distribution network weight overload is realized according to the capacity planning results. Tests have shown that this method has good accuracy in predicting electrical loads and can effectively solve the problem of excess capacity caused by light or no load, improving the ability of heavy overload warning and capacity planning in the distribution network.**

Keywords*—Imbalanced data; feature extraction; distribution network; overload classification warning

I. INTRODUCTION

With the development of energy [1-3], research on the security and load stability of power distribution network networking has always received attention. In distribution network networking, it is necessary to build a graded warning model for heavy overload in the distribution network, screen the set of characteristic indicators related to heavy overload in the distribution network, and construct a graded prediction framework for the load situation of the distribution network. This can not only achieve short-term warning of heavy overload risk in the distribution network, but also predict and distinguish no-load and light load lines. Studying the classification and early warning method for heavy overload in the distribution network is of great significance in improving the power supply capacity and economic benefits of the distribution network. By establishing a distribution network load prediction model, scientific capacity planning and line transformation are carried out to address the problem of heavy overload in the distribution network, providing technical support and support for improving the reliability of power supply, emergency response ability, and customer service level of the distribution network. The study of a graded warning model for heavy overload in the distribution network has important practical significance in promoting capacity planning and line optimization and renovation design.

In conducting relevant research on the phenomenon of heavy overload in distribution transformers, the method of feature analysis using imbalanced datasets is used to collect historical operation data of distribution networks, power outage repair work orders, transformer load data, and meteorological data. The above data are mostly imbalanced datasets, and data mining is conducted based on the aforementioned multi-source heterogeneous data to screen the feature indicator set related to heavy overload in distribution networks, Building a hierarchical prediction framework for the load situation of the distribution network can not only achieve short-term warning of the risk of heavy overload in the distribution network, improve the summer warning ability during peak hours, but also identify no-load and light load lines, providing a solution for later capacity planning. Starting from the actual situation of the distribution network, integrating multi-source heterogeneous data of the distribution network, proposing a distribution network overload warning and capacity planning technology based on historical data of the distribution network, completing the establishment of a hierarchical prediction model for the distribution network load situation, and forming a distribution network capacity planning and line transformation plan.

However, due to the low probability of distribution transformer overload, it is often difficult to obtain sufficient effective data for early data analysis and later model establishment, which also increases the difficulty for classification models to accurately predict power outages. In addition, many factors that affect power outages are difficult to present and obtain in the form of data values, which to some extent limits the establishment of feature index systems, making it difficult for prediction models to fully consider all influencing factors, thus increasing the difficulty of feature engineering and data preprocessing work. In the classification and warning of heavy overload in the distribution network, it is necessary to establish a distribution network load level prediction model based on imbalanced datasets to achieve early warning of heavy overload in the distribution network. It is expected to reduce the line outage rate index by more than 5%. In the analysis of imbalanced feature sets, how to combine

*Corresponding Author

information such as power distribution, road construction, municipal planning, and power load distribution after establishing a load prediction model, The formation of targeted distribution network capacity planning and line renovation plans is still an urgent problem to be solved. For example, Huang Yuanfang [4] et al. proposed a distribution transformer heavy overload risk warning method that takes load uncertainty into account. This method uses quantile regression algorithm of gated cycle unit to predict the load level of distribution transformer at different subpoints. In addition, utility function is used to describe the severity of heavy overload accidents suffered by distribution transformers. Combined with the power system risk theory, the potential heavy overload risk level of distribution transformers is assessed, and the risk warning is realized according to the evaluation results. Shi Changkai [5] et al. studied a method for predicting the load load of the Spring Festival distribution based on BP network and grey model. According to the particularity and regularity of the Spring Festival power load, this method uses fuzzy clustering method to divide the Spring Festival holiday period. Based on BP neural network and grey prediction system, the prediction model of the daily maximum load of the Spring Festival distribution is established, and the rated parameters of the distribution are combined. Judging whether the configuration is overloaded or not by analysing the prediction result. However, the warning accuracy of the above method is low, resulting in poor application effect.

In response to the current problems, combined with big data analysis and processing technology, a distribution network overload classification warning method based on imbalanced dataset feature extraction is proposed. Firstly, a hierarchical data acquisition model of distribution network heavy overload is established, and the feature index set related to distribution network heavy overload is screened. Combined with the unbalanced learning algorithm of inter class correlation, the random forest algorithm is used to realize the feature detection of distribution network heavy overload classification and the distribution network capacity planning. According to the capacity planning results, the distribution network heavy overload classification warning is realized. Then extract the inter class feature quantities of heavy overload in the distribution network, and based on the feature extraction results, achieve graded warning of heavy overload in the distribution network. Finally, simulation testing was conducted to demonstrate the superior performance of the method proposed in this paper in improving the ability of distribution network overload classification warning and planning. This method improves the early warning performance, and it has certain feasibility and effectiveness.

II. Overall Architecture and Data Sampling of Distribution Network Overload Warning and Capacity Planning

A. Overall Architecture of Distribution Network Capacity Planning

In order to realize the hierarchical early warning design of distribution network heavy overload, the capacity planning model of the hierarchical early warning of distribution network heavy overload is constructed. Through the parameter analysis of the distribution network heavy overload early warning model, the KNN algorithm is used to process the missing values and outlier of the distribution network multi-source heterogeneous data. For the missing value processing of distribution network heavy overload early warning and capacity planning, based on the characteristics of distribution network data periodicity and continuity, KNN algorithm is used to realize the study of distribution network heavy overload early warning and capacity planning, find out the corresponding position of missing data in other cycles, so as to fill the missing data [6-7]. For the outlier processing of distribution network heavy overload early warning and capacity planning. Firstly, the STL-ESD technology, which combines the time series decomposition algorithm and the single sample multiple outlier detection algorithm, is used to detect the outlier in the distribution network load data, and the KNN algorithm is used to replace the distribution network heavy overload early warning and capacity planning outlier to ensure the completeness of the distribution network heavy overload early warning and capacity planning data.

On the basis of analyzing the data of heavy overload warning and capacity planning in the distribution network [8-9], a penalty based feature selection algorithm is used to analyze the power outage characteristics of the distribution network. Based on the fuzzy feature selection method, a feature subset of heavy overload warning and capacity planning in the distribution network is established. During the training process of the target model, the feature selection of heavy overload warning and capacity planning in the distribution network is carried out simultaneously, that is, feature selection is taken as a part of the model.

Using the K-means clustering [10-12] based undersampling method, the non-outage class dataset in the distribution network heavy overload warning and capacity planning dataset is undersampled. This method divides the imbalanced dataset into a majority class (non-outage dataset) and a minority class (outage dataset). Then, the clustering algorithm is used to cluster the multi class dataset of the distribution network heavy overload warning and capacity planning, and the random undersampling model parameters are obtained, Finally, the undersampling model parameters for distribution network overload warning and capacity planning are obtained. Based on the above analysis, the overall structure of the distribution network overload warning and capacity planning is shown in Fig. 1.

Finally, a hierarchical prediction model for distribution network load is established using the Adaboost ensemble algorithm, which combines several weak classifiers into a strong classifier to improve the performance of hierarchical prediction based on distribution network load [13-14].

The undersampling method for power outage datasets based on K-means clustering mainly has two processes. The first process is to cluster the majority class dataset (non-power outage dataset) using K-means clustering method, dividing the dataset into K clusters; The second process is to conduct random undersampling in each cluster according to the density distribution. Specifically, it is ordered according to the size of the data variance in each cluster. First, the cluster with small

difference is subject to random undersampling at a certain sampling rate. After sampling, the multi class dataset and the few class dataset are combined to obtain a new balanced dataset.

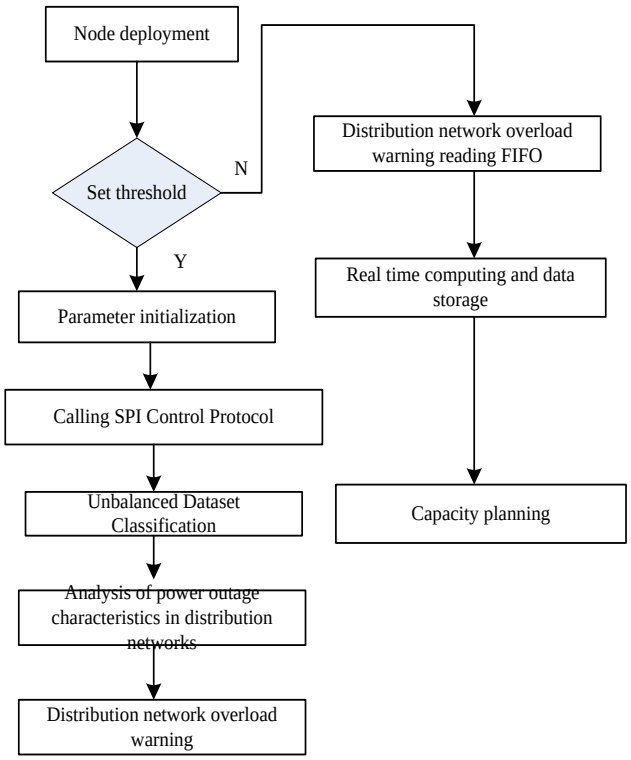


Fig. 1. Distribution network overload warning and capacity planning.

B. Data Feature Sampling

Given the training sample set $\{(x_i, t_i)\}$, for distribution network capacity planning and line renovation, where $x_i = \{x_{i1}, x_{i2}, \cdots, x_{in}\} \in R^n, t_i = \{t_{i1}, t_{i2}, \cdots, t_{im}\} \in R^m$, including L hidden layer nodes for distribution network capacity planning and line transformation data feature detection, the distribution network weight overload activation function is f(x), in the case of a single output node, the output function of the distribution network weight overload early warning ELM is:

$$f_L(x) = \sum_{i=1}^{L} \beta_i h_i(x) = h(x)\beta \quad (1)$$

Where, $\beta = [\beta_1, \dots, \beta_L]^T$, T represents the output weight vector between the distribution network overload warning hidden layer and the output node for L hidden layer nodes, $h(x) = [h_1(x), \dots, h_L(x)]$ represents the hidden layer output vector of the distribution network overload warning input x, that is, $h(x)$ maps the input distribution network overload warning data from the d-dimensional input space to the L-dimensional fuzzy dynamic feature space. Combined with the distribution of power supply points, road construction, and municipal planning, a heterogeneous algorithm is used to establish a clustering model. The minimum training decision function for the internal temporal features of the sample is obtained as follows:

$$Minimize: \|\boldsymbol{H\beta} - \boldsymbol{T}\|^2 \; and \; \|\boldsymbol{\beta}\| \quad (2)$$

Where, H is the dynamic allocation matrix of hidden layer output of cluster, which is expressed as:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h}(\boldsymbol{x}_1) \\ \boldsymbol{h}(\boldsymbol{x}_2) \\ \vdots \\ \boldsymbol{h}(\boldsymbol{x}_n) \end{bmatrix} = \begin{bmatrix} h_1(\boldsymbol{x}_1) & \dots & h_L(\boldsymbol{x}_1) \\ h_1(\boldsymbol{x}_2) & \dots & h_L(\boldsymbol{x}_2) \\ \vdots & \vdots & \vdots \\ h_1(\boldsymbol{x}_n) & \dots & h_L(\boldsymbol{x}_n) \end{bmatrix} \quad (3)$$

The process clustering learning of distribution network heavy overload warning is an unsupervised learning process. Select the K-mean standard SVM [15-16] classification model that meets the integration conditions, and get the maximum classification interval between the two classes of $2/\|\beta\|$. This norm actually controls the complexity of the function in the ELM feature space. Using a visual clustering point graph overlay analysis method, the disputed samples are temporarily classified into multiple clustering clusters, and two types of ELM models with single output are defined as follows:

$$\text{Minimize}: L_{\text{ELM}} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{C}{2}\sum_{i=1}^{N} {\varepsilon_i}^2$$

$$\text{Subject to: } h(x_i)\boldsymbol{\beta} = t_i - \varepsilon_i, i = 1, \dots, N. \quad (4)$$

Where, $\frac{1}{2}\|\beta\|^2$ represents the structural risk of power outage data, $\frac{C}{2}\sum_{i=1}^{N} {\varepsilon_i}^2$ represents the empirical risk of distribution network load.

Based on the KKT principle, the k-fold cross validation method is used to partition the data and transform the heavy overload classification warning problem of the distribution network into a dual optimization problem:

$$L_{ELM} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + \frac{C}{2}\sum_{i=1}^{N} {\varepsilon_i}^2 - \sum_{i=1}^{N} \alpha_i(h(x_i)\boldsymbol{\beta} - t_i + \varepsilon_i) \quad (5)$$

According to equation (5), the sliding time window method is used to obtain the KKT constraint conditions:

$$\frac{\partial L_{ELM}}{\partial \boldsymbol{\beta}} = 0 \rightarrow \boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i h(x_i)^T = \boldsymbol{H^T}\alpha \quad (6)$$

$$\frac{\partial L_{ELM}}{\partial \varepsilon_i} = 0 \rightarrow \alpha_i = C\varepsilon_i, i = 1, \dots, N \quad (7)$$

$$\frac{\partial L_{ELM}}{\partial \alpha_i} = 0 \rightarrow h(x_i)\boldsymbol{\beta} - t_i + \varepsilon_i = 0, i = 1, \dots, N. \quad (8)$$

Where, $\alpha = [\alpha_1, \dots, \alpha_N]^T$, where the standard deviation of response data fluctuation is used for data compression to obtain each Lagrange multiplier α_i. For the i-th training sample.

Randomly divide K into k sets with similar numbers. When the number of training samples is small (i.e. N<L), (6) and (7) are introduced into equation (8). From the above formula, it can be inferred that:

$$\beta = H^T\left(HH^T + \frac{I}{C}\right)T \quad (9)$$

Similarly, when the training sample is large (i.e. N>L), it can be inferred that the secondary learner contains the feature values extracted by the primary learner:

$$\beta = \left(H^TH + \frac{I}{C}\right)H^TT \quad (10)$$

Similarly, when the training sample is large (i.e. N>L), it can be inferred that the secondary learner contains the feature values extracted by the primary learner:

$$f(x) = h(x)\beta = h(x)H^T\left(HH^T + \frac{I}{C}\right)T \text{ or}$$

$$f(x) = h(x)\beta = h(x)\left(H^TH + \frac{I}{C}\right)H^TT \quad (11)$$

Based on the above analysis, a data collection and feature analysis model for heavy overload classification warning in the distribution network is constructed, combined with feature clustering and feature detection of imbalanced datasets, to achieve fuzzy sampling of overload data.

III. Classification Warning and Planning Algorithm for Heavy Overload in Distribution Network

A. K-means Clustering based Feature Clustering Algorithm for Unbalanced Data in Distribution Networks

Using heterogeneous ensemble learning methods and K-means clustering based sampling methods can avoid deleting too much information on a certain data distribution and prevent data distortion caused by undersampling unevenness. The algorithm flow of the undersampling method for power outage datasets based on K-means clustering is as follows:

Step 1: preprocess the data set of unbalanced heavy overload hierarchical early warning of the original distribution network, including missing value processing based on KNN algorithm and outlier processing based on STL-ESD algorithm, and then select the characteristics to obtain the characteristic data set D with the labels of power failure (y=1) and non-power failure (y=0).

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}, x_i \in R, y_i \in \{0,1\} \quad (12)$$

Step 2: Split the unbalanced dataset D of the distribution network overload classification warning, and select 80% of it as the training dataset D_1, 20% for test dataset D_2.

Step 3: Training dataset D for the imbalanced dataset of the distribution network overload classification warning_ Unstop dataset T in D_1 _ Cluster 0 and randomly select k cluster centroids as: $\mu_1, \mu_1, \dots, \mu_k \in R^n$.

Step 4: For$(x_i, y_i) \in T_0$. Calculate the cluster to which the distribution network overload classification warning imbalanced dataset belongs,

$$c^{(x_i)} = \arg\min_j ||x_i - \mu_j||^2, j = 1,2, \dots, k \quad (13)$$

Where, $c^{(x_i)}$ is the cluster to which photovoltaic characteristic data such as device parameters (x_i, y_i) belongs.

Step 5: For each cluster j, recalculate the centroid of the unbalanced dataset cluster for the distribution network overload classification warning.

$$\mu_j' = \frac{\sum_{(x_i,y_i)\in T_0} I(c^{(x_i)}=j)x_i}{\sum_{(x_i,y_i)\in T_0} I(c^{(x_i)}=j)} \quad (14)$$

Step 6: Calculate the maximum movement distance of the cluster center in the imbalanced dataset of the distribution network overload classification warning $d = \max(||\mu_j' - \mu_j||_2)$.

If $d > \varepsilon$, update $\mu_j = \mu_j'$, skip to step 7 for execution.

Step 7: Scale each cluster of the above clustering results α Perform random undersampling to obtain the distribution network overload classification warning imbalanced dataset, balanced training dataset D_0'.

B. Distribution Network Overload Classification Warning Imbalanced Dataset Diversity Scheduling Warning

The specific algorithm principle of using AdaBoost algorithm to build a distribution network overload classification warning imbalanced dataset scheduling is as follows:

Step 1: Input the training dataset of the distribution network overload classification warning imbalance dataset $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}, x_1 \in X, y_1 \in Y = (-1, +1)$, $(i = 1,2, \dots, N)$.

Step 2: Assuming that the weights of each sample in the dataset are initially uniform, the initial weights of the common modeling parameters for the imbalanced distribution network overload classification warning dataset in each sample are:

$$\omega_{1i} = \frac{1}{N} \quad (15)$$

Step 3: Set the maximum number of iterations [17] (i.e. the maximum number of linear combination weak classifiers) T: $t = 1,2,3, \dots, T$, to obtain the initial clustering objective function of the distribution network overload classification warning imbalanced dataset.

Step 4: When training the t-th weak classifier, the weighted weight [18-19] of the i-th sample's distribution network overload classification warning imbalanced dataset is ω_{ti} with $\sum_{i=1}^{N} \omega_{ti} = 1$, the classifier of the distribution network overload classification warning imbalanced dataset trained is represented as $G_t(x)$.

Step 5: Calculate the imbalanced dataset D of the distribution network overload classification warning in $G_i(x)$ Classification error rate on i (x) ε_t:

$$\varepsilon_t = P[G_t(x_i) \neq y_i] = \sum_{i=1}^{N} \omega_{ti} I[G_t(x_i) \neq y_i] \quad (16)$$

Where, I (·) is the indicator function, and the mutual information I (·) of the unbalanced data set of the distribution network heavy overload hierarchical early warning is equal to 1. Classifier $G_t(x)$ The classification error rate of t (x) on the weighted training dataset D is equal to that of $G_t(x)$; the sum of the weights of the misclassified samples in t (x).

Based on classification error rate ε_t Computational basis classifier $G_t(x)$ Weight of t (x):

$$\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t} \quad (17)$$

Training data distribution weights for the imbalanced dataset of distribution network overload classification warning after the t+1 iteration$\omega_{t+1,i}$ Update:

$$\omega_{t+1,i} = \begin{cases} \frac{\omega_{ti}}{Z_t} e^{-\alpha t}, G_t(x) = y_i \\ \frac{\omega_{ti}}{Z_t} e^{\alpha t}, G_t(x) \neq y_i \end{cases} \quad (18)$$

If the imbalanced dataset samples of the distribution network overload grading warning are correctly classified [20], its weight will decrease; on the contrary, if misclassified, the weight will increase.

$$Z_t = \begin{cases} \sum_{i=1}^{N} \omega_{t,i} e^{-\alpha t}, G_t(x_i) = y_i \\ \sum_{i=1}^{N} \omega_{t,i} e^{\alpha t}, G_t(x_i) \neq y_i \end{cases} \quad (19)$$

Then output the final distribution network overload classification warning imbalanced data classification output:

$$G(x) = \text{sign}[\sum_{t=1}^{T} \alpha_t G_t(x) \quad (20)$$

By processing the missing values and outlier in the data set, combined with the analysis results of the characteristics of the unbalanced data set, the system realizes the early warning of the distribution network heavy overload hierarchical early warning unbalanced data set diversity dispatching, focuses the main theories of distribution network capacity planning on the early load forecasting, and improves the early warning stability and dynamic analysis capability.

IV. Experimental Testing and Result Analysis

A. Evaluation Index System

Establish a distribution network load level prediction model based on imbalanced datasets to achieve early warning of heavy overload in the distribution network. It is expected to reduce the line outage rate index by more than 20%. In the experiment, an evaluation index system is set up, and Gini index is given. Gini index can measure the impurity of nodes, and its formula is:

$$GINI(t) = 1 - \sum_j p^2(j/t) \quad (21)$$

In the formula: t is the branch attribute of the distribution network load level evaluation node; $p(j/t)$ represents the proportion of the target category of the distribution network load level in node t. The Gini standard definition for the distribution node t of the distribution network overload classification warning imbalanced dataset is as follows:

$$GINI(s,t) = p_L GINI(t_L) + p_R GINI(t_R) \quad (22)$$

The division standard for the imbalanced dataset of distribution network overload classification warning is to minimize GINI (s, t).

The least squares deviation is commonly used to measure the heavy overload classification warning ability of the regression tree allocation network, and the fitting error formula of node t is:

$$Err(t) = \frac{1}{n_t} \sum_{D_t} (y_i - k_t)^2 \quad (23)$$

In the formula: n is the number of instances in node t; k_t is the average of the target values of instances in each node:

$$k_i = \frac{1}{n_i} \sum_{n_i} y_i \quad (24)$$

The least squares deviation standard for dynamic nodes in the distribution network overload classification warning imbalanced dataset divided by attribute values s is defined as:

$$Err(s,t) = \frac{n_{tL}}{n_t} Err(t_L) + \frac{n_{tR}}{n_t} Err(t_R) \quad (25)$$

In order to simplify the calculation process in the computer and avoid multiple traversals of attribute values, the above equation is simplified, and the hierarchical scheduling error of unbalanced data for distribution network overload classification warning can be obtained as:

$$Err(s,t) = \frac{S_L^2}{n_{tL}} + \frac{S_R^2}{n_{tR}} \quad (26)$$

Wherein,

$$S_L = \sum_{D_{iL}} y_i, \; S_R = \sum_{D_{iR}} y_i \quad (27)$$

By conducting hierarchical detection of the distribution network weight process, capacity planning and line transformation, analyzing the irrelevant or redundant information in the distribution network dataset, and designing partition standards to maximize $Err(s,t)$.

B. Result Analysis

Matlab is used for simulation test, and 270 normal distribution sample points are given in the unbalanced data set of the distribution network data set. These sample points are divided into two categories. The normal distribution N (u, Σ) respectively:

Class 1: N: ([1; 2], [0.23 0; 0 0.57]), a total of 70 points; Class 2: N: ([2.21; 0], [0.65 0; 0 0.88]), a total of 200 points;

Randomly generate 100 noise data using SVM, ELM, and ELM_ The CIL algorithm performs this and provides an imbalanced dataset sample sequence as shown in Fig. 2.

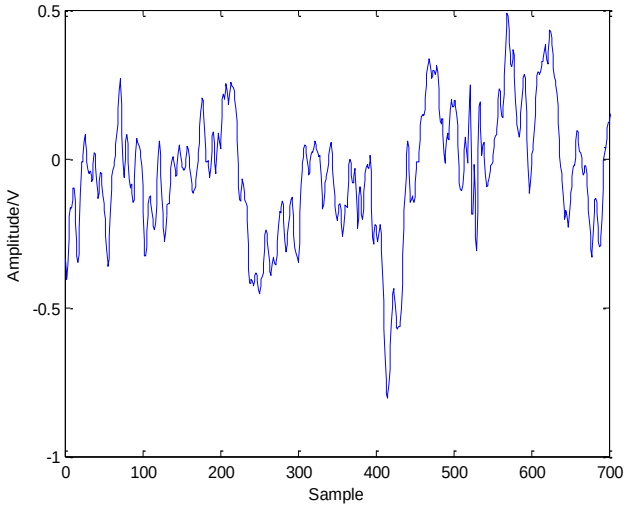


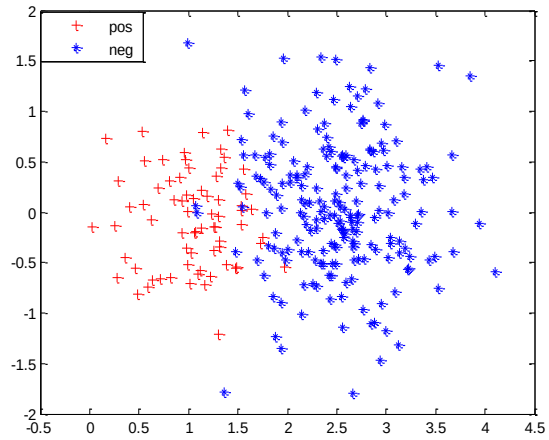
Fig. 2. Sample sequence of imbalanced dataset.

Using the data in Fig. 2 as the test object, different methods were used to obtain a noise free dataset and reference [4] method, reference [5] method, and ELM after adding the noise set_ The CIL classification warning results are shown in Table I.

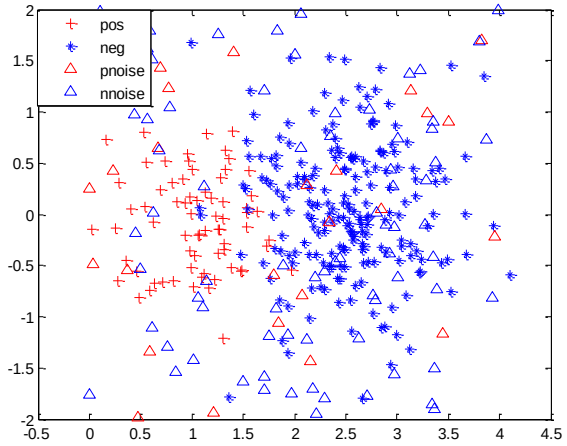
TABLE I. CLASSIFICATION WARNING RESULTS FOR NOISE DATASETS

	Reference [4] Method	Reference [5] Method	Proposed method
Accuracy	93.324	93.686	95.412
SE	88.235	84.029	95.235
SP	95.098	97.059	99.118
GM	91.598	90.314	94.676

According to the analysis of the hierarchical warning results in Table I, the classification and early warning results of the noise data sets of the three methods are all good, but the classification and early warning results of the proposed method are more accurate, the lowest value is 94.676, while the lowest value of the literature method is 88.235 and 84.029, which is more than 6 points higher than that of the proposed method. The clustering results of imbalanced dataset features are shown in Fig. 3.



(a) Normal distribution of datasets



(b) Dataset distribution after adding noise

Fig. 3. Clustering results of unbalanced dataset features.

By analyzing Fig. 3, it can be seen that the method proposed in this paper has a good ability to plan for hierarchical warning of heavy overloads in the distribution network by clustering the features of the imbalanced dataset. According to the historical operation data of the distribution network (active load rate, three-phase imbalance, defect records, fault records, etc.), distribution transformer account information, meteorological environment data, geographical environment data, user scale, and other multi-source heterogeneous data, first process the missing values and outlier in the data set, and carry out the distribution network heavy overload classification warning for each data set. The comparison results are shown in Fig. 4 to 6, the overall testing accuracy of the reference method is slightly lower than that of this method, the early warning accuracy of different data sets of the proposed method is higher than that of the literature method. The maximum value of the proposed method is 94.80, while that of the literature method is 87.48, which is more than 7 points higher than that of the proposed method. However, due to its significantly faster learning speed than support vector machines, the algorithm proposed in this paper has significant advantages in large-scale imbalanced data classification of sample sets.

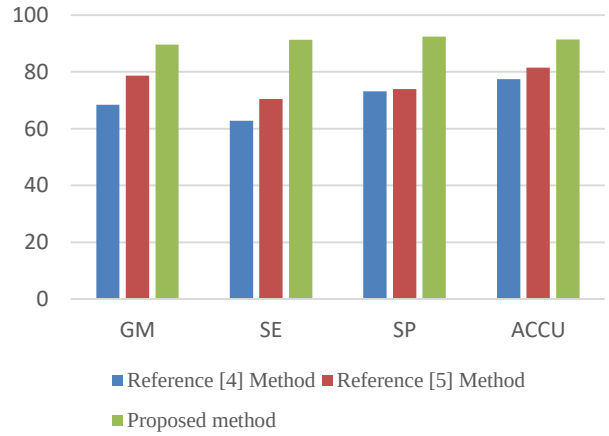


Fig. 4. Comparison of the results of various grading warning methods in Pima India.

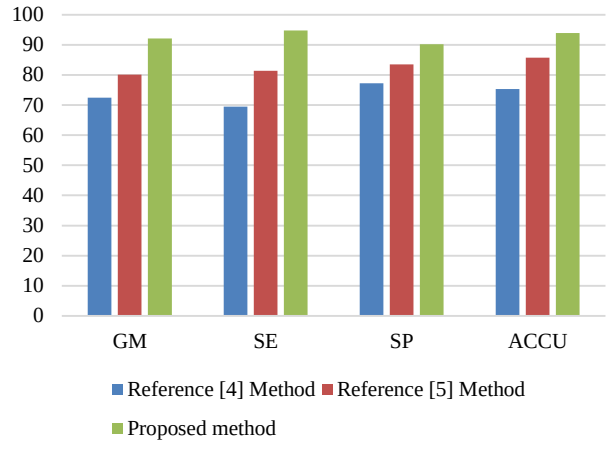


Fig. 5. Comparison of the results of various grading warning methods in transfusion.

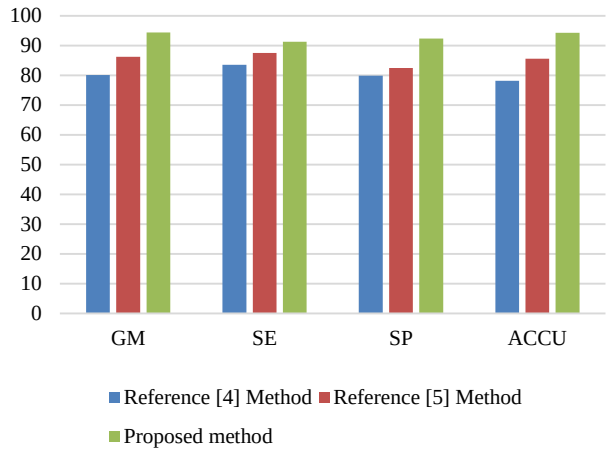


Fig. 6. Comparison of the results of various grading warning methods in Transfusion Haberman.

C. Results and Discussion

By setting up the load forecasting model of distribution network, scientific capacity planning and line reconstruction are carried out to improve the power supply reliability, emergency response ability and customer service level of distribution network. By processing the missing values and outliers in the data set, the problems of missing and abnormal data caused by various factors in the load data of distribution network are solved, the quality of the data set is improved, and the subsequent model training is supported. By constructing new features and selecting the optimal feature subset, the problem of poor performance of prediction model caused by information irrelevance or redundancy in distribution network data set is solved. The problem of poor prediction accuracy of a few classes due to unbalance of data is solved by undersampling. The AdaBoost classification algorithm is used to solve the problem of building a hierarchical forecasting model for the load condition of distribution network. By collecting distribution network historical operation data, power outage repair work order, transformer load data and meteorological data, data mining is carried out based on the above multi-source heterogeneous data. The characteristic index set related to distribution network overload is selected, and the hierarchical forecasting framework of distribution network load is constructed. To realize short-term early warning of heavy load risk of distribution network in order to improve summer peak warning, emergency response ability and customer service level. Extract effective information from heterogeneous data from multiple sources, construct new features, and form a feature database. The feature selection method is used to filter out the optimal feature subset and undersample the unbalanced data set to reduce the unbalance degree of the data set. Finally, AdaBoost and other classification methods are selected to establish a hierarchical forecasting model of distribution network load condition to realize heavy overload warning. The load forecasting model is established to predict the total load and maximum load in the future, and the capacity planning and line reconstruction plan of the distribution network are formed by combining the information of distribution, road construction, municipal planning and load distribution, etc., so as to improve the power supply capacity and economic benefits of the distribution network. The load forecasting model is established by using regression analysis and other load forecasting methods to predict the total load and maximum load in the future. At the same time, considering the distribution point, road construction, municipal planning, power load distribution and other factors, the capacity planning and line transformation of the distribution network. The analysis shows that the heavy overload classification early warning method adopted in this paper has good accuracy for power load prediction, effectively solves the problem of excess capacity caused by light load or no load, and improves the capacity of heavy overload early warning and capacity planning of distribution network. Compared with the literature method, the accuracy of this method is much higher than the literature method.

V. Conclusions

Distribution network occasionally faces heavy overload phenomenon, and if the phenomenon is not warned in time, it is easy to affect the normal operation of distribution network. Therefore, in order to achieve accurate warning of heavy overload of distribution network, it is necessary to classify the heavy overload warning of distribution network. Therefore, a new classification and early warning method of heavy overload of distribution network based on feature extraction of unbalanced data set is studied. In this method, the characteristics of distribution network are defined, and K-means clustering and undersampling methods are introduced to extract features from unbalanced data sets of distribution network overload classification. At the same time, decision tree algorithm and random forest algorithm are adopted to build a heavy overload early warning method to achieve early warning. After the design of the method is completed, the performance of the proposed method is analyzed through experiments. It can be seen from the experimental results that the proposed method has high accuracy in early warning and effectively improves the early warning ability of heavy overload in the distribution network.

Acknowledgment

This work is supported by the Science and Technology Project of State Grid Shanxi Electric Power Company “Research and application of distribution network equipment and power management technology based on IoT perception” (5205M0220004).

References

[1] CHEN Jinpeng, HU Zhijian, CHEN Weinan, et al. Load prediction of integrated energy system based on combination of quadratic modal decomposition and deep bidirectional long short-term memory and multiple linear regression[J]. Automation of Electric Power Systems, 2021, 45(13):85-94.

[2] WANG Xuan, WANG Shouxiang, ZHAO Qianyu, et al. A multi-energy load prediction model based on deep multi-task learning and ensemble approach for regional integrated energy systems[J]. International Journal of Electrical Power & Energy Systems, 2021(126):106583.

[3] LIANG Zhi, SUN Guoqiang, LI Hucheng, et al. Short-term load forecasting based on VMD and PSO optimized deep belief network[J]. Power System Technology, 2018, 42(2):598-606.

[4] HUANG Yuanfang, LIU Yunkai, ZHENG Shimming, et al. Distribution Transformer Heavy Overload Risk Warning Considering Load Uncertainty[J]. Power System and Clean Energy, 2021, 37(10):17-24.

[5] SHI Changkai, YAN Wenqi, ZHANG Xiaohui, et al. Chinese New Year load forecasting based on BP network and Grey Model[J]. Journal of Electric Power Science and Technology, 2016, 31(3):140-145.

[6] KERMANI Mehran Mozaffari, AZARDERAKHSH Reza. Reliable Architecture-Oblivious Error Detection Schemes for Secure Cryptographic GCM Structures[J]. IEEE Transactions on Reliability, 2019, 68(4):1347-1355.

[7] ANASTASOVA Mila, Azarderakhsh Reza, Kermani Mehranmozaffari. Fast Strategies for the Implementation of SIKE Round 3 on ARM Cortex-M4[J]. Institute of Electrical and Electronics Engineers (IEEE), 2021, 68(10):4129-4141.

[8] ZOU Changyue, RAO Hong, XU Shukai, et al. Analysis of resonance between a VSC-HVDC converter and the AC grid[J]. IEEE Transactions on Power Electronics, 2018, 33(12):10157-10168.

[9] SUN Jian. Impedance-based stability criterion for grid-connected inverters[J]. IEEE Transactions on Power Electronics, 2011, 26(11):3075-3078.

[10] ZHOU Yu, SUN Hongyu, ZHU Wenhao, et al. Segmented sample data selection method based on K-means clustering[J]. Application Research of Computers, 2021, 38(6):1683-1688.

[11] GUO Jing, GENG Haijun, WU Yong. Research on K-means clustering algorithm based on bacterial population optimization[J]. Journal of Nanjing University of Science and Technology, 2021, 45(3):314-319.

[12] ZHOU Xiangzhen, LI Shuai, SUI Dong. Data-driven optimization of K-means clustering algorithm based on quantum artificial bee colonies[J]. Journal of Nanjing University of Science and Technology, 2023, 47(2):199-206.

[13] LI Yunfeng, HE Zhiyuan, PANG Hui, et al. High frequency stability analysis and suppression strategy of MMC-HVDC systems (Part I): stability analysis[J]. Proceedings of the CSEE, 2021, 41(17):5842-5855.

[14] ZHU Jizhong, DONG Hanjiang, LI Shenglin, et al. Review of data-driven load forecasting for integrated energy system[J]. Proceedings of the CSEE, 2021, 41(23):7905-7924.

[15] WANG Zhibin, XIAO Yanjiao, WANG Jue, et al. Based on convolutional neural network and SVM lightning monitoring and early warning[J]. Journal of Natural Disasters, 2022, 31(1):219-225.

[16] WAN Wei, LIU QQ, SUN Hongchang, et al. Electricity unusual behavior of early warning method[J]. Journal of Harbin University of Science and Technology, 2022, 27(4):53-62.

[17] MEHRAN Mozaffari Kermani, ARASH Reyhani Masoleh. A Low-Power High-Performance Concurrent Fault Detection Approach for the Composite Field S-Box and Inverse S-Box[J]. IEEE Transactions on Computers, 2011, 60(9):1327-1340.

[18] LI Ran, SUN Fan, DING Xing, et al. Ultra short-term load forecasting for user-level integrated energy system considering multi-energy spatio-temporal coupling[J]. Power System Technology, 2020, 44(11):4121-4131.

[19] LUO Fengzhang, ZHANG Xu, YANG Xin, et al. Load analysis and prediction of integrated energy distribution system based on deep learning[J]. High Voltage Engineering, 2021, 47(1):23-32.

[20] Negnevitsky M, Nguyen D H, Piekutowski M. Risk assessment for power system operation planning with wind power penetration[J]. IEEE Transactions on Power Systems, 2015, 30(3):1359-1368.