Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8880661, 10 pages
https://doi.org/10.1155/2020/8880661

Research Article

An Early Warning Method of Distribution System Fault Risk Based on Data Mining

Yeying Mao,[1] Zhengyu Huang,[1] Changsen Feng,[2] Hui Chen,[1] Qiming Yang,[1] and Junchang Ma[1]

[1]*State Grid Suzhou Power Supply Company, Suzhou 215004, China*
[2]*College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China*

Correspondence should be addressed to Changsen Feng; fcs@zjut.edu.cn

Received 9 September 2020; Revised 20 September 2020; Accepted 20 November 2020; Published 7 December 2020

Academic Editor: Kai Wang

Accurate warning information of potential fault risk in the distribution network is essential to the economic operation as well as the rational allocation of maintenance resources. In this paper, we propose a fault risk warning method for a distribution system based on an improved RelieF-Softmax algorithm. Firstly, four categories including 24 fault features of the distribution system are determined through data investigation and preprocessing. Considering the frequency of distribution system faults, and then their consequences, the risk classification method of the distribution system is presented. Secondly, the K-maxmin clustering algorithm is introduced to improve the random sampling process, and then an improved RelieF feature extraction method is proposed to determine the optimal feature subset with the strongest correlation and minimum redundancy. Finally, the loss function of Softmax is improved to cope with the influence of sample imbalance on the prediction accuracy. The optimal feature subset and Softmax classifier are applied to forewarn the fault risk in the distribution system. The 191-feeder power distribution system in south China is employed to demonstrate the effectiveness of the proposed method.

1. Introduction

As the last step in the power industry, the distribution system is closely and directly connected to end-users [1, 2]. The stable operation of the distribution system is crucial to the reliable power supply for users [3–5]. However, the distribution system has a complex network structure and various equipment components and even worse, the external factors that cause distribution system faults are highly random. Since the causal relationship is nonlinear, the conventional fault prediction method based on the electrical mechanism is challenging to function. Therefore, exploring the potential risks in the distribution system operation process and furtherly taking corresponding measures have become a severe challenge to power supply companies [6, 7].

The concept of distribution system fault risk is put forward in [8], and the fault probability index system and fault consequence index system are established, respectively, from two dimensions of possibility and severity. In [9], time series, gray theory, and statistical analysis are applied to deeply analyse the distribution system's fault repair data and extract fault characteristics. However, the above methods focus on the definition of failure risk and the establishment of an index system, and the correlation analysis of failure outage factors and the establishment of the fault early warning model are less studied.

Data mining has recently been widely applied in the field of power systems due to its excellent computing performance and adaptability [10–13]. The fuzzy classification algorithm is developed in [14] to identify the fault cause. In [15], the correlation mining method between load characteristics and air temperature index is proposed based on the linkage analysis theory. Association rule mining is implemented in [16] to conduct fault cause analysis, which can effectively identify critical variables that have a substantial impact on fault. In [17], a novel framework of distribution system fault detection is presented to cope with the power system complexity and double-way power flows, and the

support vector data description method is adopted due to the limited available fault data. In [18], the fault line selection problem is transformed into a classification problem, including data preparation, training, classification, and evaluation.

For the feature dimension reduction, the characteristics and main algorithms of filtering, encapsulation, and embedded dimension reduction methods are summarized in [19, 20], and the robustness, prediction accuracy, and interpretability of different algorithms are compared. Literature [21] proposed a feature extraction method suitable for high-dimensional data. This method can extract feature vectors strongly related to classification but has poor performance in removing redundancy. In [22], principal component analysis is utilized to extract the main components from high-dimensional features through matrix transformation and eliminate the weak value information. Literature [23] conducts correlation recognition and redundancy removal for feature vectors based on correlation analysis theory.

Many researchers [24, 25] have explored the classification method and applied it to distribution system fault prediction [26–28]. Literature [29] presented a fault risk warning method of the distribution system based on improved support vector machine algorithm. Literature [30] focuses on meteorological factors and presents a distribution system fault classification prediction method combining AdaBoost and decision tree. Literature [31, 32] introduces the overall structure of a distribution operation analysis system and expands the application of the massive fault data to the fault risk level prediction and weak spot identification. However, the distribution system fault is an accidental event, and the proportion of fault data samples is far less than that of normal operation data sample; that is, the distribution system fault prediction is a typical unbalanced sample problem. Therefore, it is necessary to improve the model to adapt to the classification prediction of minority categories of samples.

Given the aforementioned consideration, a fault warning method based on improved RelieF-Softmax algorithm for the distribution system is proposed. Firstly, data acquisition and data preprocessing are carried out to determine the initial associated feature set of distribution system faults, and a fault risk classification method of the distribution system is proposed. Secondly, because of the RelieF algorithm's deficiency in the initial sampling and redundancy removal, an improved RelieF feature extraction method combined with the correlation coefficient method is proposed to screen out the fault optimal feature vector with the strongest correlation and minimum redundancy. Finally, a fault risk level prediction model of distribution system based on improved Softmax classifier is established to enhance unbalanced sample prediction accuracy. Taking 191 feeder lines in south China as examples, the analysis results demonstrate the effectiveness of the fault risk warning method proposed in this paper, which can provide crucial guiding significance to the operation practice.

The remainder of this paper is given as follows. Section 2 describes the preprocessing of distribution system data. The improved RelieF algorithm is presented in Section 3, and a fault risk warning method of the distribution system based on improved Softmax loss function is proposed in Section 4. The numerical results of the presented method are detailed in Section 5. The final section of the paper gives conclusions based on this study.

2. Preprocessing of Distribution System Data

2.1. Data Collection. By investigating the widely used information management systems related to the distribution system of State Grid Corporation Company of China, six systems, including the distribution production management system, distribution automation system, electricity information collection system, geographical information system, 95598 customer service system, and marketing business management system, are selected to collect the operation data, equipment data, and historical fault information of the distribution system. The data classification and data sources are detailed in Table 1.

2.2. Data Preprocessing. Preprocessing the initial data is necessary to improve the prediction accuracy, generally including data cleaning, data transformation, data integration, and outlier diagnosis.

Data cleaning is to process the vacant and repeated values in the initial data to ensure the data set's integrity, consistency, and rationality. The empty values in the data samples can be eliminated or replaced by the mean or the median. The repeated value recognition rules could be set based on the logical relationship between the data samples. For example, if the feeder name and the power failure time of two samples are the same, it is considered that there may exist a repeated sample and one of them should be eliminated. Other rules can be presented according to the specific problems and decision-maker preferences.

Data transformation, mainly including standardized processing, data grading, and quantization, makes it easier to perform data analysis. The max-min method, z-score method, or decimal scaling method can be used for standardized data processing. For data such as rainfall, thunderstorm, and wind, continuous numerical values should be discretized and classified to highlight data differences.

Data integration is to integrate, summarize, and correlate data from multiple sources. Due to the diversity of data sources, it is necessary to cross-verify the data. For example, according to the historical faults data, the planned power failure part can be eliminated according to the power loss information of the user side, that is, the power failure only caused by the faults. Afterwards, this part's power failure information can be verified and compared with the power failure information recorded in the faults work order.

The outlier diagnosis effectively identifies and eliminates the wrong inputs and meaningless values that may appear in the initial data. Outliers may lead to a decrease in the accuracy of prediction results. Therefore, statistical methods, clustering, or graph-based methods should be applied to search and delete outliers.

Finally, the initial feature set could be obtained after a series of data preprocessing, as shown in Table 2.

Table 1: Related data of distribution system fault warning.

Category	Data sources	Specific information
Fault data	95598 customer service system, marketing business management system and distribution automation system	Power outage time, power outage frequency, name of feeder, substation of power outage feeder, number of affected households, feeder construction mode, power supply area classification
Meteorological information	China meteorological network	Rainfall, minimum temperature, maximum temperature, average temperature, humidity, wind speed, wind level, visibility, cloud cover, snowfall, thunderstorm day
Operation data	Electricity information collection system	Distribution transformer capacity, real-time load
Parameter data	Geographical information system and distribution production management system	The total length of the feeder, the length of the overhead section, the length of the cable, the operation time, and the substation

Table 2: Initial fault feature set of distribution system.

Category	Variable	Feature
Fault data	f_1	Fault risk level
	f_2	Monthly fault frequency
	f_3	Number of households affected by power failure
Meteorological information	f_4	Monthly cumulative rainfall
	f_5	Monthly mean temperature
	f_6	Monthly highest temperature
	f_7	Monthly extreme weather days
	f_8	Monthly mean humidity
	f_9	Monthly extreme humidity days
	f_{10}	Monthly gale days
	f_{11}	Monthly thunderstorm days
	f_{12}	Monthly snowy days
Operation data	f_{13}	Maximum monthly load
	f_{14}	Average monthly load
	f_{15}	Month classification
Parameter data	f_{16}	The geographical location classification
	f_{17}	Feeder construction mode
	f_{18}	Power supply area classification
	f_{19}	Length of overhead feeder section
	f_{20}	Length of cable feeder segment
	f_{21}	Total length of the feeder
	f_{22}	Number of feeder segment switches
	f_{23}	Number of feeder transformers
	f_{24}	Feeder operation time

2.3. Fault Risk Classification. Distribution system fault risk consists of the frequency of failure and the consequences of failure. Given the assessment indexes of power grid companies, the failure rate (the frequency of failure) of 100 km and the number of households affected by power failure (the influenced scope of fault) are selected as the basis for the classification of distribution system fault risk. Among them, the former is an important index used by power supply companies to assess each branch company's annual operation level, which reflects the fault frequency of distribution network feeders per unit length. The latter is similarly a standard index to measure the reliability of the power supply of distribution system, reflecting the scope of the influence of power failure.

Equation (1) is utilized to obtain the failure rate of the feeder per 100 km per month.

$$S_i = \frac{\sum_{j=1}^{n_f} f_{ij}}{L_i}, \tag{1}$$

where S_i represents the monthly failure rate per 100 km of the feeder i, f_{ij} denotes the state of the j_{th} power failure of the feeder i, and L_i is the length of the feeder i.

Equation (2) gives the mathematical definition of the number of households affected by the monthly power failure.

$$C_i = \sum_{j=1}^{n_f} \sum_{k \epsilon F_{ij}} n_{ij,k} t_{ij,k}, \tag{2}$$

where C_i is the number of households affected by the monthly power failure of the feeder i, n_f is the total number

of power failure accidents in that month, and F_{ij} is the set of transformers affected in the j_{th} power failure of the feeder i. $n_{ij,k}$ and $t_{ij,k}$ represent the number of households and power failure time of the k_{th} affected transformer in the j_{th} power failure of the feeder i, respectively.

Because of the calculated feeder failure rate per month of 100 km and the number of households affected by power failure, the power failure risk level of the distribution system can be divided into general, emergency, and severe. Referring to a city in south China in 2018, the annual failure rate is 2.502 times/ 100 km·year, and the number of households affected by the distribution system failure is 102,500 households. The fault risk classification is shown in Table 3. It should be pointed out that the threshold value of each level can be adjusted according to the actual situation of the local distribution system, and the higher risk level of any two indexes is taken as the result in the calculation.

3. Fault Feature Extraction of Distribution System Based on Improved RelieF Algorithm

The dimension and quality of input vectors directly affect the accuracy of classification prediction results. Too many input vectors may lead to model overfitting and operation efficiency reduction, and input collinearity will reduce the model's stability. Therefore, it is necessary to extract the optimal fault feature subset and screen the strongest correlation and the least redundant vectors, to improve the efficiency and accuracy of fault risk prediction.

3.1. RelieF Algorithm. RelieF algorithm is a typical filtering feature selection method to extract feature vectors with a significant distinguishing degree for target classification. In the RelieF algorithm, weight is assigned to each feature based on the distance measurement, so as to evaluate the ability of feature vectors to distinguish target categories.

The specific definition is as follows: for sample set D, randomly select sample s and find k nearest neighbors in the same kind of s, which are defined as H and k nonnearest neighbors, which are defined as M. In this paper, Euclidean metric is employed to measure the distance between samples. The calculation formula of the characteristic difference between samples is described as follows:

$$\text{diff}(a, X, Y) = \begin{cases} \dfrac{X(a) - Y(a)}{\max(a) - \min(a)} & \text{a is continuous,} \\ 0 \quad \text{a is discrete } X(a) \neq Y(a), \\ 1 \quad \text{a is discrete } X(a) = Y(a), \end{cases} \tag{3}$$

where diff (a, X, Y) represents the difference between sample X and sample Y on feature a.

The weight updating formula is defined as follows:

$$W_a = W_a - \sum_{i=1}^{k} \frac{\text{diff}(a, H_i, s)}{tk} + \sum_{M \notin \text{class}(s)} \frac{\left[(N(M)/(1 - N(\text{class}(s)))) \sum_{i=1}^{k} \text{diff}(a, M_i, s) \right]}{tk}, \tag{4}$$

where W_a denotes the weight value of feature a, k is the number of the nearest neighbor sample, t is the number of sampling, H_i and M_i represent the i_{th} nearest neighbor and non-nearest neighbor of the sample s, respectively, and class( ) is the ratio function of the sample number to the total sample number.

3.2. Improved RelieF Algorithm. Although the RelieF algorithm has no restrictions on data types and relatively high operating efficiency, it still has the following disadvantages:

(1) Considering that the initial random sampling is put back sampling, the selected sample may be too limited due to repeated sampling. Since the repeated sample does not provide new information for the classification and is an invalid input, the model results' accuracy may be affected.

(2) The algorithm has a weak ability to distinguish the redundant features, which leads to a considerable noise of the input features.

Given the consideration mentioned above, the RelieF algorithm is improved from two aspects. On the premise that the algorithm flow remains unchanged, the clustering algorithm is introduced to cluster the initial data, and a hierarchical sampling algorithm based on clustering is proposed. Due to the deficiency of redundancy elimination in the RelieF algorithm, the feature extraction method combining RelieF and the correlation coefficient method is presented to identify and eliminate redundant features effectively.

3.2.1. Hierarchical Sampling Based on K-Maxmin Clustering Algorithm. The K-maxmin distance method selects data point as far as possible as the clustering center based on Euclidean distance, so as to effectively avoid the situation that the initial clustering center may be too close when compared to the k-means method. The K-maxmin algorithm is with high efficiency and is unnecessary to determine the initial clustering number. Due to the page limit, the K-maxmin clustering algorithm process can refer to the literature [33], which will not be described here.

The K-maxmin clustering algorithm is introduced to cluster the initial feature set, and then stratified sampling is applied in line with the category proportion. The total sampling number M is distributed to all categories proportionally, and the number of sampling points of each category can be determined by the proportion of the category to the total sample. In this way, the low probability of local sampling in random sampling can be effectively avoided. Besides, each sampling is strictly limited to

Table 3: Classification of fault risk level in distribution system

Risk level	Status	Failure rate of 100 km	Households affected by power failure
1	General	=0	0
2	Emergency	(0, 0.208]	[0, 270]
3	Severe	≥0.208	≥270

nonrepeated sampling, which ensures that each sampling is assigned with new weight for the feature vectors, so as to significantly improve the classification effect of random sample points on the classification results.

3.2.2. Correlation Coefficient Method. Pearson coefficient is an indicator of the degree of linear correlation between variables, widely used in statistics. Its value is between (−1, 1). If it is positive, it means a positive correlation between two variables, or otherwise, it means a negative correlation. The higher the absolute value is, the higher the correlation will be. Pearson correlation coefficient calculation formula is described as follows.

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}, \tag{5}$$

where Cov (X, Y) is the covariance of sample X and sample Y, and σ_X and σ_Y are the variances of sample X and sample Y, respectively.

The correlation coefficient matrix can be obtained by calculating the correlation coefficient between each input vector. It is generally believed that the correlation is strong if the correlation coefficient is higher than 0.7, and the corresponding feature vector pair will be put into the redundant set. Suppose that the features extracted by RelieF algorithm show redundant features, the input vectors with a small weight value of redundant feature pairs will be eliminated, and only one feature vector is retained.

3.3. Optimal Feature Extraction Method Based on the Improved RelieF Algorithm. The flow chart of the feature extraction method based on improved RelieF algorithm is depicted in Figure 1. Firstly, the initial data set is clustered by K-maxmin algorithm to realize stratified sampling and nonrepeated sampling. Secondly, the improved RelieF algorithm is applied to identify the feature vectors that can significantly distinguish the target classification. Finally, the correlation coefficient method is conducted to reduce the optimal feature subset's dimension, and the redundant feature vectors are furtherly eliminated to obtain the optimal feature subset.

4. Fault Risk Warning Method of Distribution System Based on Improved Softmax Loss Function

4.1. Improved Softmax. Softmax classification [34] is an extension of binary classification logistic regression to solve multiple classification problems. Its algorithm is based on Softmax regression, and the category with the highest output probability is the prediction category.

For input data with m dimensions $\{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, where x_i is the input vector, y_i is the corresponding category vector, and there are K categories; namely, y_i belongs to $\{1, 2, ..., K\}$. Softmax regression is used to estimate the probability that the input data belongs to each category. For any input vector, its prediction function can be expressed as

$$h_\theta(x) = P(y_i = j|x, \theta) = \frac{1}{\left(1 + e^{-\theta^T x}\right)},$$

$$h_\theta(x_i) = \begin{bmatrix} P(y_i = 1|x_i, \theta) \\ P(y_i = 2|x_i, \theta) \\ \vdots \quad \vdots \quad \vdots \\ P(y_i = K|x_i, \theta) \end{bmatrix} = \sum_{j=1}^{K} e^{\theta_j^{\mathrm{T}} x_i} \begin{bmatrix} e^{\theta_1^{\mathrm{T}} x_i} \\ e^{\theta_2^{\mathrm{T}} x_i} \\ \vdots \\ e^{\theta_K^{\mathrm{T}} x_i} \end{bmatrix}, \tag{6}$$

where $P(\ )$ represents the probability of occurrence within parentheses. $\theta = [\theta_1, \theta_2, ..., \theta_K]$ is the weight vector of $n \times K$, and n is the number of sample features. $\sum_{j=1}^{K} e^{\theta_j^{\mathrm{T}} x_i}$ is a normalized parameter that guarantees the sum of the probabilities to be 1.

The Softmax loss function is based on the logarithmic cross entropy theory, which can be expressed as

$$J(\theta) = -\frac{1}{m}\left\{\sum_{i=1}^{m}\sum_{j=1}^{K}\left[\mathrm{Ind}(y_i = j)\log\left(\frac{e^{\theta_j^{\mathrm{T}} x_i}}{\sum_{j=1}^{K} e^{\theta_j^{\mathrm{T}} x_i}}\right)\right]\right\}, \tag{7}$$

where Ind ($y_i = j$) is 0-1 indicating function, if true in parentheses, the value is 1, otherwise 0.

Combined with equations 6 and 7, the classification prediction problem can be transformed to solve the prediction function parameter, with equation (7) being minimized, so as to obtain the probability of different categories of the sample. The physical meaning of the loss function defined in equation (7) is to make the proportion of the correct classification samples as large as possible. Still, this function assumes no difference between the categories in the proper classification under the condition of sample data equilibrium. However, in the problem of fault early warning for the distribution system, the loss of high risk is mistaken for the low risk could be much bigger than the reverse. In other words, the correct classification of high risk is more important than the correct classification of low risk, and the low risk data account for a large proportion in the studied samples. Therefore, the loss function of equation (7) is improved to adapt to the proposed strategy in this paper.

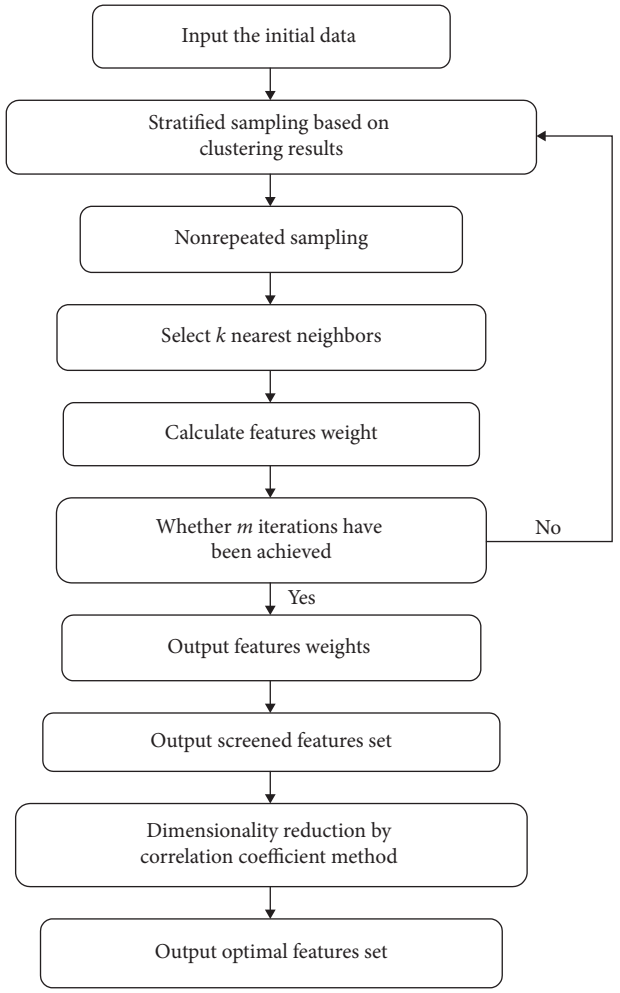

Figure 1: Flowchart of feature extraction method based on improved RelieF algorithm.

$$J(\theta) = -\frac{1}{m}\left\{\sum_{i=1}^{m}\sum_{j=1}^{K}\left[\alpha_j \mathrm{Ind}(y_i = j)\log\left(\frac{e^{\theta_j^{\mathrm{T}} x_i}}{\sum_{j=1}^{K} e^{\theta_j^{\mathrm{T}} x_i}}\right)\right]\right\} + \lambda\sum_{t=1}^{n}\sum_{j=1}^{K}\theta_{lj}^2, \tag{8}$$

where the first item is to measure the classification error, α_j is the category weight to adjust the sample imbalance degree and increase the weight of minority class being mistaken for other categories. The second item is the regularization function, where λ is the regularization parameter and called $L2$ norm. Regularization function can make it easy to obtain the optimal global solution while avoiding the training model's overfitting and improving the model's generalization ability.

The gradient descent method is a common optimization method to solve the maximum or minimum value of a function. Hence, it is employed to train the Softmax classifier. The partial derivative of equation (8) can be expressed as

$$\frac{\partial}{\partial\theta_j}J(\theta) = -\frac{1}{m}\left\{\sum_{i=1}^{m}\alpha_j x_i\left[\mathrm{Ind}(y_i = j) - h_\theta(x_i)\right]\right\} + \frac{2}{\lambda}\theta_j. \tag{9}$$

According to equation (10), update theta with each iteration:

$$\theta_j := \theta_j - \delta\frac{\partial}{\partial\theta_j}J(\theta), \tag{10}$$

where δ is the iteration step.

4.2. Fault Risk Warning Method of Distribution System. The data-driven fault warning method can be divided into three stages: data acquisition and preprocessing, feature extraction, and risk level prediction. The research idea and risk warning process of this paper are illustrated in Figure 2.

Data acquisition and preprocessing. Collect fault data, operation data, parameter data, and meteorological information and perform data cleaning, integration, and outliers diagnose as required. Determine the fault risk classification based on the failure rate of 100 km and the number of households affected by power failure. Consequently, the initial sample set can be obtained, where each row represents a sample, each column denotes a feature, and the last column represents the fault risk level.

Feature extraction. Feature extraction includes two aspects: remove weak correlation and redundancy. The improved RelieF algorithm and correlation coefficient method are presented to obtain the optimal feature subset with the strongest correlation and minimum redundancy.

Risk level prediction. The improved Softmax is proposed to train the training set and learn the mapping relationship between the fault influencing factors and the fault risk level of the distribution system. Based on this learning model, the fault risk level of test samples can be reasonably predicted.

5. Case Study

A total of 191 feeders and their data from January 2018 to December 2018 in a southern city are collected as training samples to predict the monthly feeder fault risk level from January 2019 to June 2019.

5.1. Data Preprocessing. The failure data, operation data, ledger data, and meteorological data of 191 feeders collected from January 2018 to December 2018 are processed by the method in Section 2. Taking each feeder as a unit, 24 fault features of 4 categories are obtained, as shown in Table 2. Among them, f_1 fault risk level can be determined comprehensively in line with f_2 and f_3. After preprocessing, the initial data set is obtained with 2292 samples, including 2154 samples of class I, 95 samples of class II, and 43 samples of class III.

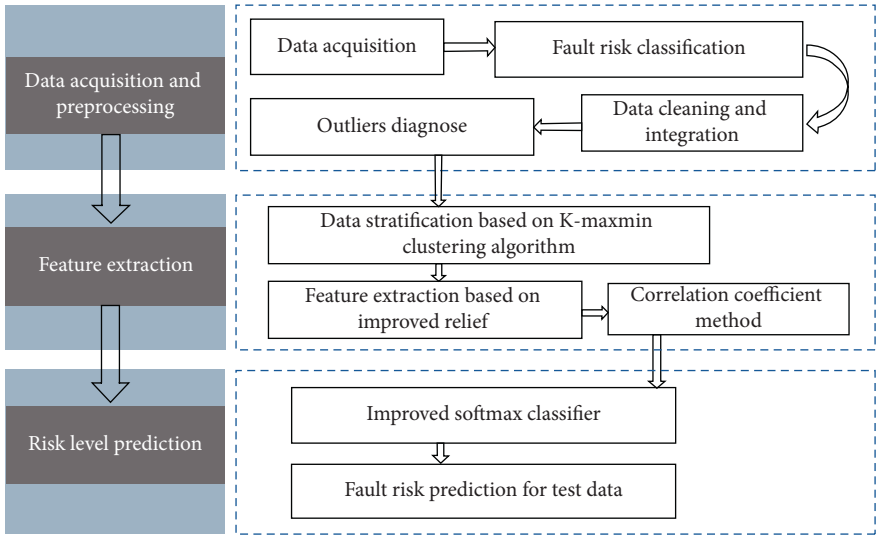

Figure 2: Flowchart of fault risk warning method for distribution system.

5.2. Analysis of Feature Extraction Results. The number of categories obtained based on K-maxmin clustering algorithm is 5, and the sampling proportion of each cluster can be determined according to the ratio of the sample number of each cluster in the total sample. Afterwards, the improved RelieF algorithm is applied to extract key fault feature, in which the number of sampling is 30, the nearest neighbor number is 8, and the number of iterations is 20. The calculation result of the feature weight is depicted in Figure 3. The dotted line in Figure 3 is the average value of weight 0.127, which is also the threshold value of feature weight screening. As can be seen from Figure 4, 10 feature weights are lower than the threshold, so the weak correlation features that are eliminated are $\{f_{23}, f_{20}, f_8, f_{22}, f_{24}, f_4, f_{12}, f_{19}, f_{18}, f_{17}\}$. According to the Pearson correlation coefficient method as mentioned earlier, there are three strongly correlated vector pairs (f_5, f_6), (f_7, f_9), and (f_7, f_{11}), and therefore f_{11}, f_9 and f_6 are eliminated, and the optimal feature set obtained is $\{f_7, f_{10}, f_{15}, f_5, f_{21}, f_{13}, f_{14}, f_{16}\}$, as shown in Table 4.

The result of feature extraction in Table 4 reflects that the features directly related to the fault are retained. Among them, the maximum monthly load, the average monthly load, the month classification, and the geographical location of the feeder correspond to the load characteristics, geographical characteristics, and time characteristics of the fault, respectively, and they have a relatively obvious and direct correlation with the feeder fault. In terms of meteorological data and ledger data, the monthly extreme weather days, monthly gale days, and the total length of feeder are retained, and some redundant indexes are effectively eliminated. According to the distribution system's actual operation in different areas, the optimal feature set obtained may also be different.

It is necessary to conduct sensitivity analysis. The sampling number and the nearest neighbor number are adjusted to 80 and 10, respectively. The calculation results show little difference from Table 4, indicating that the improved RelieF algorithm is relatively stable, and there is no

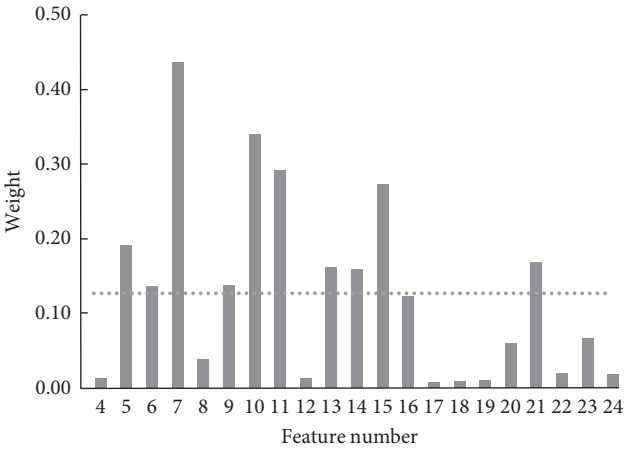

Figure 3: Feature weights.

1088	8	2	99.09%
1	39	0	97.5%
0	0	8	100.00%
99.91%	82.98%	80.00%	99.04%

Figure 4: Obfuscation matrix of predicted model.

Table 4: Optimal features subset.

Variable	Weight	Feature
f_7	0.43588	Monthly extreme weather days
f_{10}	0.33952	Monthly gale days
f_{15}	0.27240	Month classification
f_5	0.19035	Monthly mean temperature
f_{21}	0.16770	Total length of feeder
f_{13}	0.16135	Maximum monthly load
f_{14}	0.15948	Average monthly load
f_{16}	0.12939	The geographical location classification

need to increase the number of iterations to improve the performance.

5.3. Early Warning of Distribution System Fault Risk. Eight optimal features are extracted to train the Softmax classifier, and the monthly failure risk levels of 191 feeders from January to June 2019 are predicted. In order to measure the processing ability of the model to unbalanced samples, the definitions of commonly used accuracy and recall rate are slightly adjusted in this paper, and the classification accuracy T_{pr} and recall rate T_{re} for minority classes are proposed. H_a, A_r and confusion matrix are also introduced as evaluation indexes of the model.

The classification accuracy T_{pr} and recall rate T_{re} can be defined as follows:

$$T_{\text{pr}} = \frac{T_2 + T_3}{F_2 + F_3 + T_2 + T_3}, \quad T_{\text{re}} = \frac{T_2 + T_3}{F_1 + T_2 + T_3}, \tag{11}$$

where T_2 and T_3 are the numbers of correct classification in the target categories II and III, respectively, and F_1, F_2 and F_3 are the numbers of mistaken classification in the target category in the three categories, respectively.

H_a is the weighted harmonic value of precision rate and recall rate, which can evaluate the overall performance of the model, and its results are more focused on the classification performance of minority category.

$$H_a = \frac{\left(1 + \beta^2\right) T_{\text{re}} T_{\text{pr}}}{\beta^2 T_{\text{re}} + T_{\text{pr}}}, \tag{12}$$

where β is the weight coefficient, is positive, and represents the relative relationship between model recall and precision. In this paper, the value is set to 1.

A_r is the proportion of correctly classified samples to all samples, which can measure the model's overall classification performance and can be expressed as

$$A_r = \frac{T_1 + T_2 + T_3}{T_1 + T_2 + T_3 + F_1 + F_2 + F_3}. \tag{13}$$

As to the Softmax classifier initialization, the weight attenuation parameter is 0.002, and the corresponding value of α for each category is inversely proportional to the ratio of category sample size, which is set as 1, 20, and 50, respectively. The gradient descent learning rate is set as 0.1, and the number of iterations is 500.

5.3.1. Prediction Results. The model prediction results are shown in Figure 4. Each row represents the actual category, and each column denotes the prediction category. The diagonal elements represent the number of samples correctly classified, and the off-diagonal elements represent the number of samples incorrectly classified. The last horizontal row and last vertical column represent the classification accuracy and recall rate of various categories. The last element represents the overall prediction accuracy rate of the model.

Figure 4 shows that the recall rate of all levels of 1146 test samples is 99.09%, 97.5%, and 100.00%, respectively, the accuracy rate is 99.91%, 82.98%, and 80.00%, and the overall classification accuracy rate is 99.04%. The classification accuracy of the first category is relatively high, which is because the number of samples in the first category is large, which corresponds to the samples without faults or with less impact, and the learning performance is good. The classification accuracy of category III is low, mainly because the total number of such samples is small, and any misclassification will greatly impact the results.

In the case of unbalanced samples, the model's recall rate is reasonable, and the recall rate of categories II and III with serious risk level is similar to that of category I, which indicates that the model can effectively identify high risk faults of the distribution system. It can be concluded from the confusion matrix that the misclassification in the prediction model mainly focuses on category I into II or III, indicating that the model focuses more on the recall rate of categories II and III with high risk, which is due to the higher cost of misclassification of the high risk categories.

5.3.2. Comparison of Different Prediction Methods. Four indexes including T_{re}, T_{pr}, H_a and A_r are utilized to compare the prediction results of the training set and test set of the improved RelieF-Softmax (denoted as A), improved Softmax (denoted as B without feature extraction), and improved RelieF (denoted as C without improvement in Softmax loss function). The predicting results are shown in Table 5.

It can be found from Table 5 that Case A, based on the improved RelieF-Softmax algorithm, performs well in both the training set and the test set, and its performance is better compared to Case B and Case C, indicating that the improved model has better generalization ability. Besides, it can be seen that feature extraction can improve classification performance to a certain extent by comparing Case A and Case B. Further, by comparing the training set and the test set of Case B, it can be concluded that the training classification effect of the model is significantly better than that of the test set, which is because too much redundant input vector will make the model passively learn too much information, and the complex model will lead to the reduction of generalization performance.

By comparing Case A and Case C, it can be found that the recall rate of Case A is considerably higher than that of

Table 5: Predicting results of different cases.

Algorithm	T_{re}	T_{pr}	H_a	A_r
A (training set)	0.9855	0.8242	0.8977	0.9865
A (test set)	0.9792	0.8246	0.8952	0.9904
B (training set)	0.9565	0.7765	0.8571	0.9817
B (test set)	0.9375	0.6716	0.7826	0.9791
C (training set)	0.8986	0.7654	0.8267	0.9799
C (test set)	0.8750	0.7636	0.8155	0.9869

Case C, indicating that the former can effectively identify high risk categories with better classification performance. For the distribution system fault prediction problem, there are various influencing factors and complex correlations. The distribution system fault data itself belongs to a minority of samples, so it is necessary to adopt the improved RelieF-Softmax algorithm to improve the model's adaptability and prediction efficiency.

6. Conclusions

In this paper, a fault risk warning method of distribution system based on improved RelieF-Softmax algorithm is proposed, and the following conclusions can be drawn.

Compared with the single dimension reduction method, the feature extraction method based on the improved RelieF algorithm can effectively overcome the deficiency, which is the fact that the traditional RelieF algorithm cannot remove the redundancy, reduce the dimension of features, and therefore improve the classification performance. The failure data in a year are analysed based on data mining technology, and the fault risk level of the distribution system is predicted. The model can effectively identify minority category samples and avoid misclassification of high risk samples that lead to severe consequences. The proposed method can provide a scientific basis for maintenance and repair resource configuration for the distribution system.

This paper aims to put forward the thinking of the fault risk early warning method for the distribution system. Due to the differences in regional information level, distribution system operational means, and the distribution system data acquisition method, the regional features should be considered in the concrete analysis. In the future, the relevant data of distribution system in an all-round way should be extracted as thoroughly as possible, and the distribution system falt features need to be identified through association mining and other efficient means, such as deep learning, and the accuracy and efficiency of the fault early warning can be enhanced further.

Data Availability

The data can be accessed from the distribution system of State Grid Corporation Company (http://www.js.sgcc.com.cn/sz/).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was funded by State Grid Suzhou Power Supply Company. This work was supported by the World Class Urban Distribution Network Demonstration Project of Suzhou Historic District.

References

[1] L. Wang, R. Yan, F. Bai, T. K. Saha, and K. Wang, "A distributed inter-phase coordination algorithm for voltage control with unbalanced PV integration in LV systems," *IEEE Transactions on Sustainable Energy*, p. 1, 2020.

[2] L. Wang, R. Yan, and T. K. Saha, "Voltage regulation challenges with unbalanced PV integration in low voltage distribution systems and the corresponding solution," *Applied Energy*, vol. 256, 2019.

[3] T. Morstyn, A. Teytelboym, and M. D. McCulloch, "Designing decentralized markets for distribution system flexibility," *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 2128–2139, 2018.

[4] B. Sultana, M. W. Mustafa, U. Sultana, and A. R. Bhatti, "Review on reliability improvement and power loss reduction in distribution system via network reconfiguration," *Renewable and Sustainable Energy Reviews*, vol. 66, pp. 297–310, 2016.

[5] Y. Xu, C. C. Liu, K. P. Schneider, F. K. Tuffner, and D. T. Ton, "Microgrids for service restoration to critical load in a resilient distribution system," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 426–437, 2018.

[6] A. Arif, Z. Wang, J. Wang, and C. Chen, "Power distribution system outage management with co-optimization of repairs, reconfiguration, and DG dispatch," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4109–4118, 2018.

[7] Z. Lin, D. Duan, Q. Yang et al., "Data-driven fault localization in distribution systems with distributed energy resources," *Energies*, vol. 13, no. 1, p. 275, 2020.

[8] S. Y. Ge, Z. H. Zhu, H. Liu et al., "Comprehensive evaluation method for distribution network fault risk," in *Proceedings of the 4th Annual General Conference of The European Political Science Association (CSU-EPSA)*, vol. 26, no. 7, pp. 40–45, June 2014, in Chinese.

[9] Y. M. Zeng, *Analysis and Application of Electric Rush Repair Data*, Xiamen University, Xiamen, China, 2018, in Chinese.

[10] A. S. Zamzam, X. Fu, and N. D. Sidiropoulos, "Data-driven learning-based optimization for distribution system state estimation," *IEEE Transactions on Power Systems*, vol. 34, no. 6, pp. 4796–4805, 2019.

[11] Y. Zhou, Y. Huang, J. Pang, and K. Wang, "Remaining useful life prediction for supercapacitor based on long short-term memory neural network," *Journal of Power Sources*, vol. 11, no. 4, pp. 2687–2697, 2020.

[12] Y. Zhou, Y. Wang, K. Wang et al., "Hybrid genetic algorithm method for efficient and robust evaluation of remaining useful life of supercapacitors," *Applied Energy*, vol. 260, 2020.

[13] H. Jiang, Y. Zhang, E. Muljadi, J. J. Zhang, and D. W. Gao, "A short-term and high-resolution distribution system load forecasting approach using support vector regression with hybrid parameters optimization," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3341–3350, 2018.

[14] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution Fault Cause identification with imbalanced data using the data mining-based fuzzy classification E-Algorithm," *IEEE*

Transactions on Power Systems, vol. 22, no. 1, pp. 164–171, 2007.

[15] R. Ma, X. Zhou, Z. Peng et al., "Data mining on correlation feature of load characteristics statistical indexes considering temperature," in *Proceedings of the International Conference on Computer Science and Environmental Engineering (CSEE)*, vol. 35, no. 1, pp. 43–51, Florence, Italy, May 2015.

[16] M. Doostan and B. H. Chowdhury, "Power distribution system fault cause analysis by using association rule mining," *Electric Power Systems Research*, vol. 152, pp. 140–147, 2017.

[17] Z. Lin, D. Duan, Q. Yang et al., "One-class classifier based fault detection in distribution systems with distributed energy resources,," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP 2018)*, IEEE, Anaheim, CA, USA, November 2018.

[18] Z. Shao and L. Wang, "Fault line selection method for distribution system based on big data and feature classification," in *2017 China International Electrical and Energy Conference (CIEEC)*, October 2017.

[19] Z. Q. Li, J. Q. Du, B. Nie et al., "Summary of feature selection methods," *Computer Engineering and Applications*, vol. 55, no. 24, pp. 10–19, 2019.

[20] Z. Zeng, H. Zhang, R. Zhang, and Y. Zhang, "A mixed feature selection method considering interaction," *Mathematical Problems in Engineering*, vol. 2015, Article ID 989067, , 2015.

[21] L. X. Zhang, *Study on Feature Selection and Ensemble Learning Based on Feature Selection for High-Dimensional Datasets*, Tsinghua University, Beijing, China, 2014, in Chinese.

[22] D. Jain and V. Singh, "An efficient hybrid feature selection model for dimensionality reduction," *Procedia Computer Science*, vol. 132, pp. 333–341, 2018.

[23] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[24] S. S. Gururajapathy, H. Mokhlis, H. A. B. Illias, and L. J. Awalin, "Support vector classification and regression for fault location in distribution system using voltage sag profile," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 12, no. 4, pp. 519–526, 2017.

[25] G.-T. Xia, C. Li, K. Wang, and L.-W. Li, "Structural design and electrochemical performance of PANI/CNTs and MnO2/CNTs supercapacitor," *Science of Advanced Materials*, vol. 11, no. 8, pp. 1079–1086, 2019.

[26] K. Wang, L. Li, Y. Lan, P. Dong, and G. Xia, "Application research of chaotic carrier frequency modulation technology in two-stage matrix converter," *Mathematical Problems in Engineering*, vol. 2019, Article ID 2614327, , 2019.

[27] O. W. Chuan, N. F. Ab Aziz, Z. M. Yasin, N. A. Salim, and N. A. Wahab, "Fault classification in smart distribution network using support vector machine," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 3, pp. 1148–1155, 2020.

[28] K. Wang, L. Li, W. Xue et al., "Electrodeposition synthesis of PANI/MnO2/graphene composite materials and its electrochemical performance," *International Journal of Electrochemical Science*, vol. 12, no. 9, pp. 8306–8314, 2017.

[29] K. Y. Liu, X. Z. Wu, C. Shi, and D. Jia, "Fault risk early warning of distribution network based on data mining," *Electric Power Automation Equipment*, vol. 38, no. 5, pp. 148–153, 2018.

[30] W. Zhang, W. X. Sheng, K. Y. Liu, and S. Du, "A Prediction method of fault risk level for distribution network considering correlation of weather factors," *Power System Technology*, vol. 42, no. 8, pp. 2391–2398, 2018.

[31] W. Zhang, W. Sheng, S. Du et al., "Architecture and technology implementation of massive data based distribution network operation analysis system," *Automation of Electric Power Systems*, vol. 44, no. 3, pp. 147–155, 2020.

[32] M. Pignati, L. Zanni, P. Romano, R. Cherkaoui, and M. Paolone, "Fault detection and faulted line identification in active distribution networks using synchrophasors-based real-time state estimation," *IEEE Transactions on Power Delivery*, vol. 32, no. 1, pp. 381–392, 2017.

[33] Y. Zhang and L. Chen, "Research on feature extraction method based on clustering and PCA fusion," *Computer Engineering and Applications*, vol. 46, no. 11, pp. 148–150, 2010.

[34] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.