IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 20, 2020, accepted October 21, 2020, date of publication October 28, 2020, date of current version November 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034365

Power System Fault Classification and Prediction Based on a Three-Layer Data Mining Structure

YUNLIANG WANG[1], XIAODONG WANG[1], YANJUAN WU[1], AND YANNAN GUO[2]

[1]Tianjin Key Laboratory for Control Theory and Applications in Complicated Systems, Tianjin University of Technology, Tianjin 300384, China
[2]Tianjin Tianda Qiushi Electric Power High Technology Company Ltd., Tianjin 300392, China

Corresponding author: Yanjuan Wu (wuyanjuan12@126.com)

This work was supported in part by the Tianjin Science and Technology Plan Project under Grant 18ZXYENC00100, and in part by the State Grid Chongqing Electric Power Company Science and Technology Project Funding under Grant SGTYHT/17-JS-199.

ABSTRACT In traditional fault diagnosis methods in power systems, it is difficult to accurately classify and predict the types of faults. With the emergence of big data technology, the fault classification and prediction methods based on big data analysis and processing have been applied in power systems. To make the classification and prediction of the fault types more accurate, this paper proposes a hybrid data mining method for power system fault classification and prediction based on clustering, association rules and stochastic gradient descent. This method uses a three-layer data mining model: The first layer uses the K-means clustering algorithm to preprocess the original fault data source, and it proposes to use self-encoding to simplify the data form. The second layer effectively eliminates the data that have little impact on the prediction results by using association rules, and the highly correlated data are mined to become the regression training data. The third layer first uses the cross-validation method to obtain the optimal parameters of each fault model, and then, it uses stochastic gradient descent for data regression training to obtain a classification and prediction model for each fault type. Finally, a verification example shows that compared with a single data mining algorithm model, the proposed method is more comparative in terms of the data mining, and the established power system fault classification and prediction model has global optimality and higher prediction accuracy, which has a certain feasibility for real-time online power system fault classification and prediction. This method reduces the disturbances from low-impact or irrelevant data by mining the fault data three times, and it uses cross-validation to optimize the multiple regression parameters of the regression model to solve the problems of low accuracy, large errors and easily falling into a local optimum, given the conduct of fault classification and prediction.

INDEX TERMS Association rules, data mining, K-means, machine learning, power system fault, stochastic gradient descent algorithm.

I. INTRODUCTION

To ensure the reliability and stability of the power system, predicting power faults in advance and making the corresponding preventive measures can effectively prevent the occurrence of power accidents and reduce economic losses. Short-circuit faults are relatively common faults in the distribution lines of power systems, and they can easily cause other corresponding electrical faults; therefore, their hazards are large. To prevent the occurrence of short-circuit faults, the following steps can be taken: (1) combining big data knowledge with machine learning algorithms, (2) mining fault historical data to find the correlation and potential laws, and then (3) building a predictive model through training data. These steps demonstrate an important and valuable research direction.

Early power fault diagnosis methods have mainly used protection devices at all levels to work. The staff can determine the fault location based on the real-time voltage data and the status of the alarm device by patrolling and inspecting the electrical equipment. The shortcomings of this traditional diagnostic method are lower efficiency and higher cost. Thus, a new mathematical analysis model has drawn the attention of researchers, and through the improvement of equipment

The associate editor coordinating the review of this manuscript and approving it for publication was Jihwan P. Choi.

functions, the control and protection ability of the power system was improved [1], [2]. Although the performance of these new mathematical analysis models and equipment is enhanced, the intelligence, interaction and automation of the equipment are not sufficient. It is possible to judge the occurrence of the fault and take protective actions for the power system in time, but it cannot predict the type of the fault, and thus, the adopted protective measures could cause protection failure due to inappropriate choices, and even enlarge the fault loss. Therefore, it is necessary to further study the prediction of the power system fault types, which will help the operators to take correct protection and remedial measures in time to minimize the fault loss.

Compared with early fault diagnosis methods, artificial intelligence diagnosis methods have been applied in the fields of fault diagnosis and prediction, such as fuzzy diagnosis methods [3], diagnosis methods based on genetic algorithms [4], [5], fault diagnosis methods using expert systems [6], [7], methods based on neural networks [8]–[10], and diagnosis methods using the support vector machine (SVM) [11], [12]. The effective use of these artificial intelligence technology methods has been superior to early diagnosis methods to a certain extent.

However, with a power system that generates massive amounts of data every moment, the traditional artificial intelligence diagnosis method cannot process the big data systematically, and the accuracy of the system fault diagnosis results cannot be further improved, which affects the efficiency of the diagnosis. The emergence of the data mining methods [13], [14] improved the performance of the fault diagnosis to a large extent. Data mining is a cutting-edge technology of data analysis, which can quickly obtain valuable information from various types of data. The functions are mainly the following: 1) Automatically predicting trends and behaviors. 2) Association analysis can find hidden associations in the data. 3) Clustering can enhance people's understanding of the similarities among things. 4) Deviation detection can look for meaningful differences between the observations and the reference values. However, most of the data mining diagnosis methods are implemented using a single algorithm model. For example, one study [15] proposed a fault diagnosis method based on decision trees for vehicle test data mining. Since the decision tree ignores the correlations of the attributes in the vehicle test data set, overfitting is prone to occur. Another study [16] developed a social network analysis management framework for the industry environmental risks using association rules based on frequent patterns, which is suitable for discrete data, but it is more difficult to implement, and its performance will decrease on some data sets. Therefore, the models achieved by a single algorithm are not ideal.

Some researchers began to pay attention to the improvement and optimization of the selected algorithms [17]. Some optimization algorithms used to solve the optimal solution problem of the algorithm model have been applied, which mainly include the gradient descent method [18], Newton method [19], and the meta-heuristic algorithm [20]–[25]. Gradient descent is one of the most commonly used methods when solving for the model parameters of machine learning algorithms, especially unconstrained optimization problems. Newton's method provides a method for solving nonlinear optimization problems whose convergence rate is fast, but each iteration requires solving a complex Hessian matrix. Meta-heuristic algorithms are based on an intuitive or empirical construction, which can give a feasible solution to the problem for an acceptable calculation time and space when the degree of deviation of the feasible solution from the optimal solution might not necessarily be predicted in advance. However, it cannot guarantee that the global optimal solution will be obtained absolutely, and it often falls into a local optimum on some problems. As a result, the hybrid data mining method, which combines multiple algorithms, has emerged. One study [26] proposed a power system line trip fault prediction method based on an long-short term memory (LSTM) network and SVM. Another study [27] proposed an optimized neural network fault diagnosis strategy for heating systems based on data mining, which used an association rule mining method to optimize the selection of the feature sets. A data driven modeling method for an aeroengine aerodynamic model that combined stochastic gradient descent (SGD) and support vector regression was proposed [28]. In addition, one study [29] proposed a port cargo throughput prediction method based on empirical mode decomposition (EMD) recurrent neural network and adaptive grouping algorithm. Another study [30] proposed a similarity grouping-guided neural network modeling method for maritime time series prediction. The experiments on both port cargo throughput and vessel traffic flow have illustrated its superior performance in terms of prediction accuracy and robustness. It can be seen that the fault diagnosis and prediction model of the hybrid data mining method is excellent and exceeds other methods.

Cluster analysis as one of the most important research branches in the field of data mining, which classifies clustered objects according to their own characteristics. Cluster analysis has been widely used in software engineering, machine learning, statistics, image analysis, web clustering engines and text mining. Association rules, as an inductive learning algorithm, have a strong ability to discover certain rules and associations in the data. As the representative algorithm of association rules, Apriori [31] uses a layer-by-layer search strategy to traverse the solution space. SGD is often used to train various machine learning models due to its fast learning rate and online update [32]. When addressing big data, SGD has a small number of calculations in a single iteration, and thus, the convergence speed is significantly higher than that of other algorithms. The optimization efficiency is better than that of the classic algorithm, and therefore, the application of SGD in data regression training is extended to many different fields.

Based on the above-mentioned considerations, this paper proposes a hybrid data mining algorithm based on K-means clustering, Apriori association rules and SGD to classify

and predict power system faults. The hybrid algorithm performs three-layer mining on the fault data to establish different fault prediction models: Firstly, K-means clustering and self-coding are used to preprocess the raw data. Then, the association rules filter the samples for the second layer of the data mining. Finally, SGD is used for the data regression training and completes the third layer of the data mining. This mining mode solves some of the current problems faced by data mining. Firstly, it reduces the interference between the complex data and avoids obtaining results from local optimization. Secondly, the complementary functions between the algorithms ensure the integrity of the data mining. Thirdly, the method adjusts parameters according to the different fault prediction models, in such a way that the fault prediction model has good robustness and fault tolerance, which can be applied to various actual fault prediction scenarios. Compared with the single algorithm model, the proposed method has greatly improved the accuracy and reliability of power system fault classification and prediction, which can be used to optimize parameters online and can be applied to different operating states.

The paper originally proposes the three-layer data mining structure, each layer structure has a special data mining function, and cooperate with each other to complete the classification and prediction work of the power system fault types. The main contributions are outlined as follows:

1) The clustering algorithm and self-encoding were used to preprocess complex source data, which classifies the source data and simplifies the form of the classified data.
2) The method uses association rules to filter the samples in advance and classifies them according to the type of fault, which increases the correlations in the data.
3) The cross-validation method finds the optimal parameters that correspond to different fault models, and then, stochastic gradient descent is used to train the fault models, which improves the accuracy of the power system's fault prediction.
4) A multi-layer data mining model based on K-means, association rules and stochastic gradient descent is built, which improves the completeness of the data mining.

The remainder of this paper is organized as follows: the description of the problem is presented in Section II. The proposed algorithm model framework and the theory of each part are explained in Section III. Then, in the fourth section, the whole test example is introduced, and the results are verified. Finally, the fifth section concludes the study.

II. PROBLEM STATEMENT

Short-circuit faults are very common faults in power systems, which can cause large-scale power outages. When faults occur, the power protection components can decide only whether to act according to the current operating conditions, but they can fail to determine what type of fault has occurred, which affects the timely handling of the fault. Therefore, this paper uses three layers of data mining on the original data of the power system short-circuit faults, and it establishes a fault classification and prediction model (FCPM) to predict whether a fault is about to occur and to predict the type of fault that will occur.

III. FAULT CLASSIFICATION AND PREDICTION METHOD

This section will introduce the structure and implementation process of the proposed method, and the mathematical model of each algorithm will be introduced in detail.

A. OVERALL METHOD ARCHITECTURE

This paper proposes a fault classification and prediction method based on K-means clustering, association rules and SGD. The source samples are the node voltage data after a certain fault occurs in the power system. The fault types are mainly single-phase ground fault (SPGF), two-phase phase-phase fault (TPPF), two-phase ground fault (TPGF) and three-phase fault (TPF). After the data collection is completed, the source sample library is shown as follows: $\{\boldsymbol{AQ}, \boldsymbol{G_i}\}$, where, $\boldsymbol{AQ} = \{\boldsymbol{X_1}, \boldsymbol{X_2} \ldots \boldsymbol{X_i}\}$ is the voltage data set in the source sample library, and $\{\boldsymbol{G_i}\}$ is the fault type of the fault node, $\boldsymbol{G_i} \in \{1, 2, 3, 4\}$, where $\boldsymbol{G_i} = 1$ is SPGF, $\boldsymbol{G_i} = 2$ is TPPF, $\boldsymbol{G_i} = 3$ is TPGF, and $\boldsymbol{G_i} = 4$ is TPF.

The overall architecture of the three-layer data mining method is shown in Fig. 1. The proposed method integrates three data mining algorithms: K-means clustering, Apriori association rules and SGD. In the process of three-layer data mining, the K-means method and self-coding method are used to preprocess the raw data, simplify the data form, reduce the complexity of the data set, and accelerate the data processing speed. After using the Apriori algorithm to mine the data for the second time, the relevant samples are sorted out according to the fault type for regression training, which can prevent the SGD from falling into a local optimum due to using random data samples and improves the accuracy of the regression training.

B. THE FIRST LAYER OF THE DATA MINING PROCESSING METHODS AND RULES

After obtaining the source samples, the K-means clustering algorithm clusters the source samples and preprocesses the data. Moreover, a data encoding rule is proposed to encode the clustered data samples and simplify the data form, which cooperates with K-means clustering to conduct first-layer data mining and the sorting of samples to obtain sample library I. The specific methods and rules are as follows:

1) K-MEANS CLUSTERING METHOD

The K-means clustering method in this paper includes three main aspects: the Euclidean distance is used to classify the data samples; The criterion function is used to judge whether the sample clustering is completed; and the number of best classification clusters is determined by comparing contour coefficients.

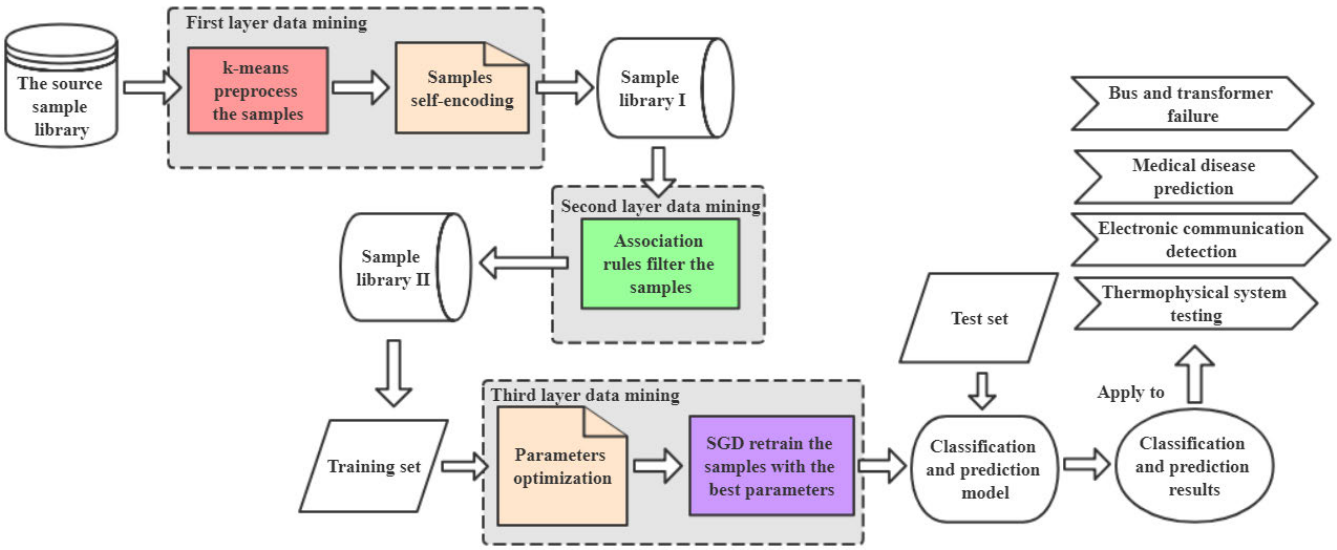


FIGURE 1. The three-layer data mining process of the proposed method.

a: EUCLIDEAN DISTANCE JUDGMENT METHOD

The K-means clustering method classifies the samples according to the Euclidean distance between the data sample and the center of each cluster, and they are classified into the cluster with the minimum Euclidean distance. The Euclidean distance is calculated as in formula (1):

$$d\left(X, Y^j\right) = \sqrt{\left(x_1 - y_1^j\right)^2 + \left(x_2 - y_2^j\right)^2 + \ldots + \left(x_n - y_n^j\right)^2} = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i^j\right)^2} \tag{1}$$

where $X = (x_1, x_2, \ldots, x_n)$ is any unclassified sample in n-dimensional space that corresponds to the elements (the voltage data on the non-faulty node) in the $\boldsymbol{AQ}$ of the source sample library. $Y^j = (y_1^j, y_2^j, \ldots, y_n^j)$ is the center of the j_{th} cluster. When classifying the samples for the first time, any sample can be randomly selected as the cluster center.

b: CRITERION FUNCTION

The average of all samples in each cluster is used to update the cluster center, and the criterion function is used to determine whether the cluster center stops updating. The criterion function is to minimize the sum of the squared errors between the samples in the cluster and the cluster center, which is shown in formula (2):

$$\min \sum_{j=1}^{K} \sum_{x_i^j \in X^j, y_i^j \in Y^j} \left(x_i^j - y_i^j\right)^2 \tag{2}$$

where Y^j is the j_{th} cluster center, y_i^j is the i_{th} element data in Y^j, K is the number of the clusters, X^j is any samples in the j_{th} cluster, and x_i^j is the i_{th} element data in X^j.

When the criterion function of formula (2) converges, which is when the cluster center does not change significantly, the cluster center stops updating. At this time, the sample classification into K clusters is completed.

c: THE CONTOUR COEFFICIENT

To obtain the optimal number of clusters in K-means clustering in the first-layer data mining, the method of calculating the contour coefficients of different clusters is used. Then, by comparing those contour coefficients, the number of clusters with the largest contour coefficient is found to be the optimal number of clusters. For each sample of a cluster, the contour coefficient calculation method is shown in formula (3):

(1) First, the cluster cohesion α_k is calculated. (The average distance from x to all other points in the cluster to which it belongs).
(2) Then, the separation degree b_k between the cluster and the other clusters is calculated. (The average distance between x and all points that are not in the same cluster).
(3) Lastly, the contour coefficient S_k is calculated. (The difference between α_k and b_k is divided by the larger of the two).

$$S_k = \frac{b_k - \alpha_k}{\max\left(b_k, \alpha_k\right)} \tag{3}$$

The value of the contour coefficient is in the range $[-1, 1]$. The closer it is to 1, the larger the value of S_k is. The average value of the contour coefficients of all samples is used as the contour coefficient under the current cluster number K. The larger the contour coefficient is, the farther the distance between the clusters, and the better the classification effect. Therefore, the K value with the largest contour coefficient

is taken to be the optimal number of clusters for the source sample library.

2) K-MEANS CLUSTERING RULES

The rules for clustering $\boldsymbol{AQ}$ in a source sample library using the K-means clustering method are shown in Fig. 2: To obtain the optimal number of clusters, the enumeration method is used to increase K from 2. When the number of clusters is K, the clustering rules are described as follows:

1) Firstly, K samples are randomly selected as the initial cluster center.
2) According to formula (1), the distance between each sample and the center of each cluster is calculated, and each sample is classified into the cluster with the minimum Euclidean distance.
3) The average of all samples in each cluster is taken as the new cluster center, and the criterion function is calculated to determine whether the minimum is reached. If the minimum of the criterion function is not reached, then return to 2). This process will be repeated until the criterion function of formula (2) reaches the minimum.
4) The contour coefficient under the value of K is calculated according to formula (3), which is compared with that of the completed clusters, and the value of K that corresponds to the maximum is taken as the cluster number; then, the clustering of $\boldsymbol{AQ}$ is completed. If the maximum value of the contour coefficient does not appear, the value of K will be updated and returned to 1) to continue the clustering.

When the source samples are clustered in the case of the different fault types, the optimal K values of the clusters on the different nodes are different. All of the source samples in different nodes must go through the above process to determine their respective optimal K and complete the clustering of the source samples on every node. After the source samples are clustered, the samples in each cluster have some similarities.

3) SAMPLE SELF-CODING AND CODING LIBRARY FORMULATION RULES

Although the samples of each cluster after clustering have a certain similarity, the data form is not simple enough to handle. Therefore, after the $\boldsymbol{AQ}$ of the source sample library is clustered, self-encoding is performed on the classified samples to simplify the data form. To keep the important attribute information of the encoded data, such as the node that the sample belongs to and the cluster that the sample belongs to, the rules are formulated as follows: For the each sample in the $\boldsymbol{AQ}$ after clustering, the node (T) to which it belongs is queried first, and then, the cluster (W) that it belongs to is queried, and the final coding form is *T0W*. For example, the *T0W* is 103, which represents that the sample is the voltage data sample classified into the third cluster on the first node, where $T \in N$ (N is a natural number, which represents the

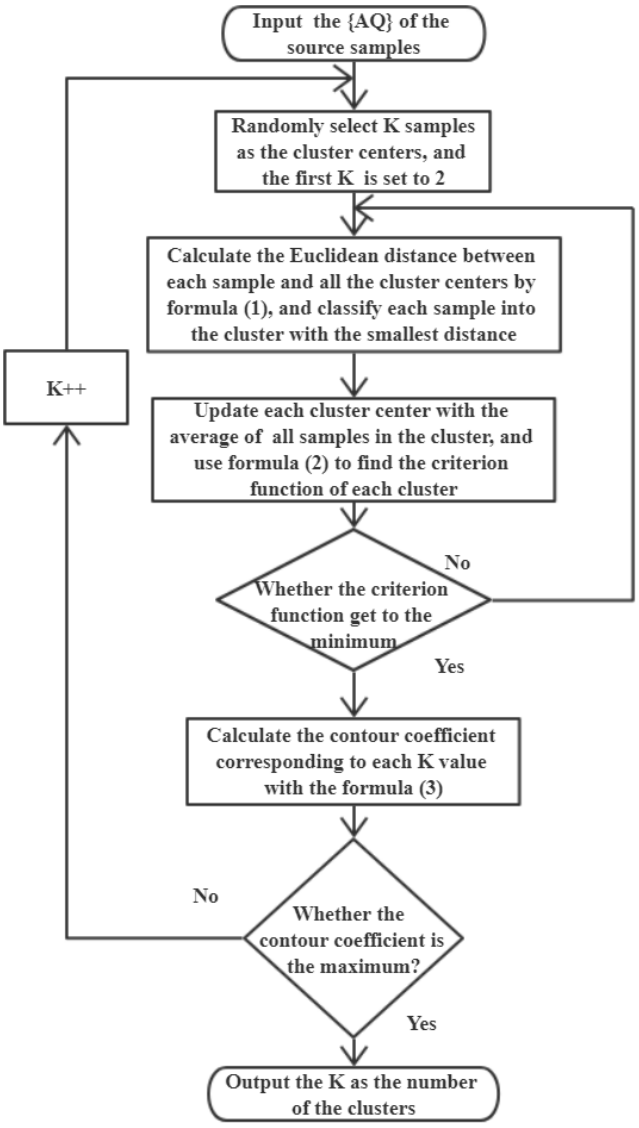


FIGURE 2. The clustering preprocessing process of the samples.

node number except for the faulty node), $0 < W \leq K$ and $W \in N$. The coding rules are shown in Fig. 3:

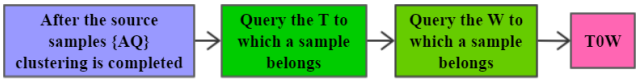


FIGURE 3. Self-encoding rules.

$\boldsymbol{AQ}$ is recorded as $\boldsymbol{BQ}$ after the clustering and the self-encoding. After the source samples are clustered and self-encoded, the sample data have a concise form, which is easier to manage.

After processing the source samples through K-means clustering and self-encoding, the source sample library $\{\boldsymbol{AQ}, \boldsymbol{G_i}\}$ is transformed into the sample library I $\{\boldsymbol{BQ}, \boldsymbol{G_i}\}$, and the first-layer data mining is completed. It digs out the inner connections of the unlabeled different data samples in

the source samples, makes the data sample in the same cluster as relevant as possible, and prepares for the second-layer data mining.

C. THE SECOND LAYER OF THE DATA MINING PROCESSING METHODS AND RULES

Because there are some potential laws between the voltage at the node and the fault types in the power system, the association rules are used in the second-layer data mining to find out the samples that are highly correlated with a certain fault type. Training the FCPM with these highly correlated samples will greatly improve the accuracy of the FCPM.

1) APRIORI ASSOCIATION RULE METHOD

The Apriori algorithm is an association rule algorithm that is based on mining frequent item sets: the elements in $\boldsymbol{BQ}$ and $\boldsymbol{G_i}$ in sample library I were correspondingly combined into a whole sample library M: $\{\boldsymbol{Z_1}, \boldsymbol{Z_2}, \ldots, Z_i\}$, and each row of the sample library M was taken as a sample group. The association rules for frequent item sets are used to find the association between two or more samples in the sample group. By calculating the support, the confidence, and the lift of these frequent item sets, the correlation degree between the samples is measured, and the non-empty sets that meet the requirements of the support, the confidence and the lift are selected.

Assuming that $\boldsymbol{Z_x}$ and $\boldsymbol{Z_y}$ are non-empty sets of M, the support, the confidence and the lift are calculated as follows:

a: SUPPORT

Support is the probability of $\boldsymbol{Z_x}$ and $\boldsymbol{Z_y}$ appearing simultaneously.

$$\text{Support}\left(Z_x \rightarrow Z_y\right) = P\left(Z_x \cap Z_y\right) \tag{4}$$

b: CONFIDENCE

Confidence is the probability that $\boldsymbol{Z_y}$ appears at the same time when $\boldsymbol{Z_x}$ appears.

$$\text{Confidence}\left(Z_x \rightarrow Z_y\right) = P(Z_x \cap Z_y)/P(Z_x) \tag{5}$$

c: LIFT

Lift represents the ratio of the probability of $\boldsymbol{Z_y}$ appearing at the same time that $\boldsymbol{Z_x}$ appears and the probability of $\boldsymbol{Z_y}$ appearing.

$$\text{Lift}\left(Z_x \rightarrow Z_y\right) = \frac{P(Z_x \cap Z_y)}{P(Z_x)P(Z_y)} \tag{6}$$

2) THE SECOND-LAYER MINING RULES BASED ON APRIORI ASSOCIATION RULE

The Apriori association rule method is used to conduct the second-layer data mining of $\boldsymbol{BQ}$ in the sample library I:

1) Firstly, the minimum support and the minimum confidence are set, and the sample library M is scanned to find all of the frequent N item sets. (N increases from 1.)
2) The candidate N+1 item sets are found by connecting and pruning based on the frequent N item sets (N + 1 = 2, 3 . . .).
3) By scanning the sample library M, all of the non-empty sets larger than the minimum support in the candidate N+1 item set are found as the frequent N+1 item sets.
4) If the frequent N+1 item sets are empty sets, then the confidence and the lift of the rules composed of all of the frequent item sets are calculated, and the rules that meet the minimum confidence and that have a lift greater than 1 are found to be the strong association rules. Otherwise, return to 2) to search the higher order frequent item sets.

The sample sets that satisfy the strong association rules constitute the association library; then, all of the sample sets related to $\boldsymbol{G_i}$ are extracted, where the samples are sorted out according to the fault types. These samples form the sample library II: $\{\boldsymbol{CQ_j}, \boldsymbol{G_j}\}$, where $\boldsymbol{G_j}$ is the j_{th} fault type, and CQ_j is the strong association sample sets that correspond to $\boldsymbol{G_j}$. The difference between the source sample library, the sample library I and the sample library II is as follows: the source sample library and the sample library I are the same in their dimension and in the number of samples, and the source sample library standardizes the form of the samples through clustering preprocessing and self-encoding to form sample library I. After association mining, the associated library obtained from sample library I is very large, but only the samples related to $\boldsymbol{G_i}$ are extracted to form sample library II, and thus, the data size of sample library II is much smaller than that of the complete association library.

After the association rules mining, the samples are highly correlated in their attributes, and the information associated with the fault types is stored, which is helpful for mining valuable results during the SGD data regression training. In this way, the result deviation caused by data redundancy is avoided, and the performance and accuracy of the regression analysis are improved.

D. THE THIRD LAYER OF THE DATA MINING PROCESSING METHODS AND RULES

After the first two layers of data mining, K-means clustering and Apriori association rules have mined the strong correlation samples that correspond to the different types of power system faults. The third layer of the data mining uses these strong association samples of sample library II to establish the FCPM for each fault type, and it achieves the goal of fault classification and prediction. To accelerate the prediction speed and further improve the prediction accuracy, the cross-validation method is used to obtain the optimal parameters in each fault prediction model. Then, the SGD obtains the solution of the optimal parameters for each fault prediction model by performing regression training on the strong association samples. The specific description is as follows:

1) FAULT CLASSIFICATION AND PREDICTION MODEL BASED ON STOCHASTIC GRADIENT DESCENT

SGD is an iterative optimization algorithm that is often used to solve and optimize model parameters of machine learning algorithms. SGD is a deformed form of the gradient descent algorithm, which has been successfully applied to text classification [33] and large-scale sparse machine learning problems in natural language processing [34], [35]. The gradient is to obtain the partial derivative of the unknown parameters of a multivariate function and obtain the vector composed of these partial derivative functions. When all of the partial derivatives in the gradient are 0, the optimal solution of the model parameters can be obtained. SGD uses only one sample per iteration. When processing large-volume samples, only a small number of samples can be used to iterate the model parameters to obtain the optimal solution. Therefore, SGD has the advantage of having a fast training speed.

a: PREDICTION MODEL FUNCTION

Given sample library II: $\{\boldsymbol{CQ_j}, \boldsymbol{G_j}\}$, assuming that the weight coefficients of the samples at each node are linear, a linear model function is obtained:

$$f\left(CQ_j\right) = w^T\left(CQ_j\right) + b \tag{7}$$

where w is the model parameter vector, and b is the intercept. $w^T CQ_j$ is the inner product of $\boldsymbol{CQ_j}$ and w.

b: THE PARAMETER OPTIMIZATION METHOD BASED ON SGD OF THE FAULT PREDICTION MODEL

i) LOSS FUNCTION

The loss function is used to estimate the difference between the actual value $\boldsymbol{G_j}$ and the model predicted value $f(\boldsymbol{CQ_j})$ that corresponds to the sample, which is expressed by $L(\boldsymbol{G_j}, f(\boldsymbol{CQ_j}))$. This article uses the following two loss functions: the SVM type loss function is shown in formula (8), and the logistic regression type loss function is shown in formula (9):

Hinge: equivalent to SVM classification:

$$L\left(G_j, f\left(CQ_j\right)\right) = \max\left(0, 1 - G_j f\left(CQ_j\right)\right) \tag{8}$$

Log: equivalent to Logistic regression:

$$L\left(G_j, f\left(CQ_j\right)\right) = \log\left(1 + \exp\left(-G_j f\left(CQ_j\right)\right)\right) \tag{9}$$

ii) RISK FUNCTION

The risk function is the expectation of the loss function, and it is also called the empirical risk:

$$\mathrm{Er} = \frac{1}{n}\sum_{i=1}^{n} L\left(G_j, f\left(CQ_j\right)\right) \tag{10}$$

Although the objective function is to minimize the empirical risk, because of learning historical data and the complexity of the functions, it could lead to overfitting of the prediction results. Therefore, the structural risks is used to avoid over-fitting:

$$\mathrm{Sr} = \alpha R\left(w\right) \tag{11}$$

where α is a hyperparameter. By setting α to reduce the parameter scale, the purpose of model simplification is achieved, which means that the model has better generalization ability. The regular item $R(w)$ is used to measure the complexity of the loss function, and it limits the parameters of the loss function. The regular items $R(w)$ mainly include $L1$ regularization and $L2$ regularization:

$$L1 = \sum_{j=1}^{m} |w|_j = \|w\|_1 \tag{12}$$

$$L2 = \frac{1}{2}\sum_{j=1}^{m} w_j^2 = \|w\|_2^2 \tag{13}$$

where $L1$ regularization can produce a sparse weight matrix, which can be used for feature selection. $L2$ regularization can prevent the model from overfitting by reducing the weight coefficient. To a certain extent, $L1$ can also prevent overfitting, but the effect is not as good as $L2$.

c: THE OPTIMIZED OBJECTIVE FUNCTION

The smaller the empirical risk and structural risk are, the better the model fit; as a result, the final objective optimization function is

$$\min : E\left(w, b\right) = \frac{1}{n}\sum_{j=1}^{n} L\left(G_j, f\left(CQ_j\right)\right) + \alpha R\left(w\right) \tag{14}$$

SGD considers a set of training samples each time to find the true gradient of the objective optimization function. For each set of samples, the iterative model parameters are updated by the update rule given by formula (15):

$$w \leftarrow w - \eta\left(\alpha\frac{\partial R\left(w\right)}{\partial w} + \frac{\partial L\left(w^T\left(CQ_j\right) + b, G_j\right)}{\partial w}\right) \tag{15}$$

where η is the learning rate of the step size in the control parameter space. To prevent the parameter w from oscillating near the solution, η is decreased according to the following formula (16):

$$\eta^{(t)} = \frac{1}{\alpha\left(t_0 + t\right)} \tag{16}$$

where t is the time step, and t_0 is the initial step size, which is the same as the initial value of the weight by default; additionally, α and t jointly affect the learning rate.

d: K-FOLD CROSS-VALIDATION PARAMETER OPTIMIZATION METHOD

The K-fold cross-validation method is used to find out the optimal parameter group (loss function L, hyperparameter α, regular term $R(w)$ and iteration number N); then, the optimal model parameter w is solved by the iteration calculation of SGD. The strong correlation samples that correspond to a certain fault type in sample library II are used to train the parameter group, and the solution with the highest cross-validation score under the fault type is regarded as the optimal solution of the parameter group. The optimal value of the parameter group and its cross-validation scores that

correspond to different faults are different. The method steps are as follows:

1) Firstly, all of the samples in the j_{th} sample set (CQ_j, G_j) are divided into K parts in equal proportion, and each part is used as the cross-validation set; the other K-1 parts are used as the training set.
2) After completing the cross-validation K times, the average of the correct rate over K times for the cross-validation results is used as the cross-validation score.
3) By comparing the cross-validation scores, the parameter group (L, R, α, N) with the highest score is selected as the optimal parameter set.
4) The optimal model parameter w is solved by substituting the optimal parameters (L, R, α, N) into formula (14) and (15) for the current iteration.

2) THE THIRD-LAYER DATA MINING RULES BASED ON THE STOCHASTIC GRADIENT DESCENT METHOD

The SGD optimization algorithm performs third-layer data mining on sample library II, as shown in Fig. 4:

1) Firstly, a prediction model function is established.
2) The K-fold cross-validation method optimizes the four parameters (L, R, α, N) of the SGD under the prediction model function.
3) By comparing the cross-validation scores, the optimal parameter set is determined.
4) The training set is retrained under the optimal parameter group to obtain the FCPM.
5) Finally, the test set is used to test the performance of the FCPM.

The optimal model parameter w can be obtained through the optimal loss function and the optimal regular terms; then, the fitting law of the samples in $\boldsymbol{CQ_j}$ to the fault result in $\boldsymbol{G_j}$ is found, in such a way that the optimization model can classify and predict the faults from the new data.

E. ALGORITHMIC MODEL EVALUATION: CONFUSION MATRIX AND ROC CURVE

Model evaluation can more intuitively see the quality of the model based on the corresponding indicators. The confusion matrix and the ROC curve are used in this article to evaluate the results.

1) CONFUSION MATRIX

The confusion matrices are also called the probability tables or the error matrices. This type of matrix is a specific matrix that is used to visualize the performance of the algorithm. The calculation formula of the overall model accuracy of FCPM, the precision of each fault type, the recall rate, and the F1 score are as follows:

Assuming that the test sample set has a total of S samples:

1) Accuracy: the ratio between the number of correct predictions and the total number of predictions:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{S}} \tag{17}$$

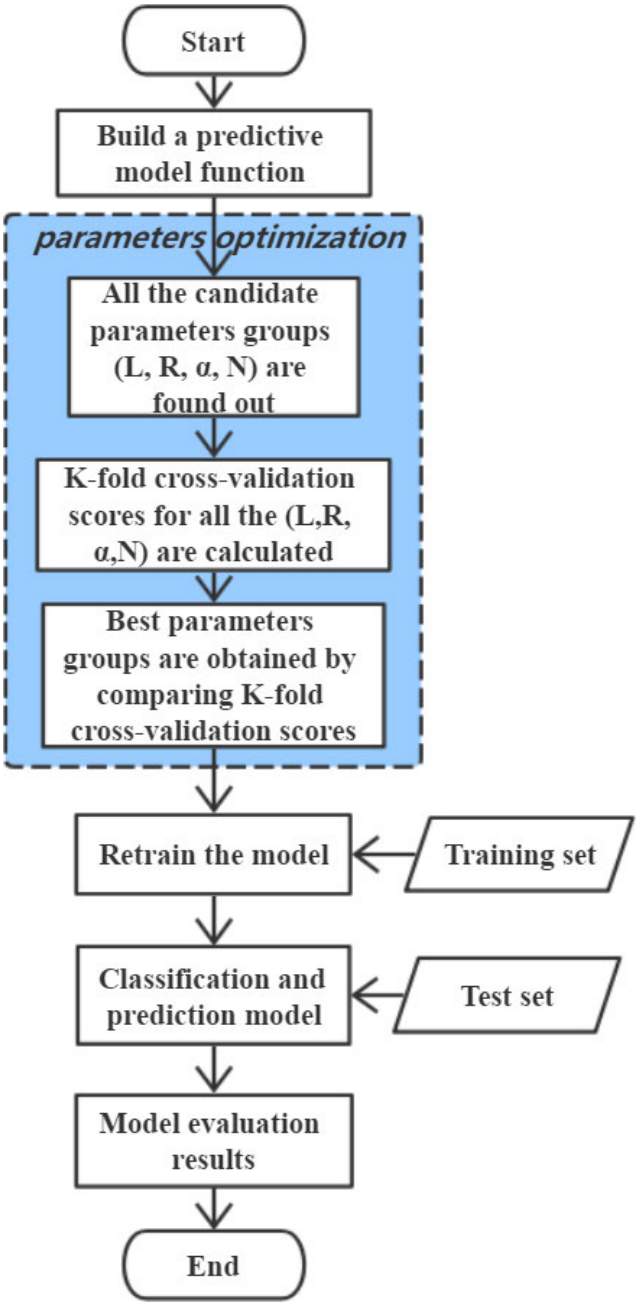


FIGURE 4. The third-layer data mining based on SGD.

2) Precision: the ratio of the correct positive number to the true and false positives number:

$$\text{Pr ecision} = \frac{\text{TP}}{\text{TP+FP}} \tag{18}$$

3) Recall: the ratio of the correct positive number to the true and false negatives number:

$$\text{Re call} = \frac{\text{TP}}{\text{TP+FN}} \tag{19}$$

4) F1: Harmonic average of the Precision and the Recall.

$$\text{F1} = \frac{2 * \text{Pr ecision} * \text{Re call}}{\text{Pr ecision} + \text{Re call}} \tag{20}$$

The multi-class classification confusion matrix of the model is converted into a binary classification confusion matrix to calculate the above indicators. Each type of fault is considered separately from the other three types of fault. The three-phase fault ($G_j = 4$) is taken as an example:

TABLE 1. The meaning of TP, TN, FN and FP in the confusion matrix.

TP	TN	FN	FP
true value G_j=4 predicted value G_j=4	true value G_j≠4 predicted value G_j≠4	true value G_j=4 predicted value G_j≠4	true value G_j≠4 predicted value G_j=4

where TP, TN, FN, and FP in formulas (17) (18) (19) are the number of samples that meet the above.

According to these performance indicators of the FCPM, it can be compared with other methods to find the advantages and disadvantages of the method's performance.

2) ROC CURVE

The Receiver Operating Characteristic Curve (ROC) is an important and common model evaluation method to judge the classification results. The ROC space defines the false positive rate (FPR) as the X axis and the true positive rate (TPR) as the Y axis.

TPR: The rate of being correctly judged to be positive among all of the actually positive samples.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{21}$$

FPR: the rate of being falsely judged to be positive among all of the actually negative samples.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{22}$$

Given a classification model, a coordinate point (X=FPR, Y=TPR) can be calculated from the true and predicted values of all of the samples. In a model, the coordinates (FPR, TPR) under different thresholds are drawn in the ROC space, which becomes the ROC curve of the specific model.

F. STATISTICAL TEST AND ALGORITHM TIME COMPLEXITY

To judge about the significance of the results, the statistical test method is added to the discussion. In addition, in consideration of the effectiveness of the proposed method, the time complexity and the computational running time are also discussed.

1) F TEST

The statistical test method used in this paper is the F test, which tests the overall significance of the linear regression equation. The multiple variables in the model are used to judge the significance of the impact, and the following assumptions are constructed:

$$H_0 = \beta_1 = \beta_2 = \ldots = \beta_n = 0 \tag{23}$$

$$H_1 = \exists i \in \{1, 2, \ldots, n\}, \quad s.t.\ \beta_i \neq 0 \tag{24}$$

Then, the test statistics are constructed as follows:

$$y_i = \beta x_i + \varepsilon_i \tag{25}$$

where x_i is the sample vector, y_i is the predicted value vector, β is the variable coefficient, and ε_i is the difference between the average value of a single sample and the average value of the overall sample.

Regression sum of squares:

$$\mathrm{SSR} = \sum_{i=1}^{n} \left(y_i - y^a\right)^2 \tag{26}$$

Sum of squared residuals for regression:

$$\mathrm{SSE} = \sum_{i=1}^{n} \left(y_i - y\right)^2 \tag{27}$$

Then, the F statistic is constructed:

$$\mathrm{F} = \frac{\mathrm{SSR}/p}{\mathrm{SSE}/\left(n - p - 1\right)} \quad (\mathrm{F} \geq 0) \tag{28}$$

where y is the actual value that corresponds to the sample vector, y^a is the average value of y, p is the degree of freedom, and n is a small number of samples extracted from the sample library.

The F value is used to test and measure the overall significance level of the model. When the F statistic is close to zero, it proves that the original hypothesis H_0 holds, which means that the overall significance level of the model is low. The larger the F statistic is, the higher the significance level of the model, which proves that the model fits well and the model is built successfully.

2) BIG O NOTATION

The more statements that are executed in the algorithm, the more time it takes for the computation. The number of executions of a statement in an algorithm is called the time frequency, which is denoted as $V(n)$, where n is the number of samples. If there is an auxiliary function $f(n)$ such that when n approaches infinity, the limit value of $V(n)/f(n)$ is a constant that is not equal to zero, then $f(n)$ is said to be a function of the same magnitude as $V(n)$, and thus, it is denoted as $V(n) = O(f(n))$, which is called the time complexity.

The calculation method is called Big O notation, whose derivation rules are as follows: 1) $O(1)$ represents the time complexity of all constant functions. 2) The time complexity of other functions retains only the highest order, and its coefficient is 1.

G. REALIZATION PROCESS OF HYBRID ALGORITHM PREDICTION METHOD

The flowchart of the method's implementation is shown in Fig. 5:

IV. EXAMPLES

The calculation examples in this section are compiled and run in the jupyter notebook of Anaconda with the help of some SK-learn toolkit functions in the Windows 7 environment.

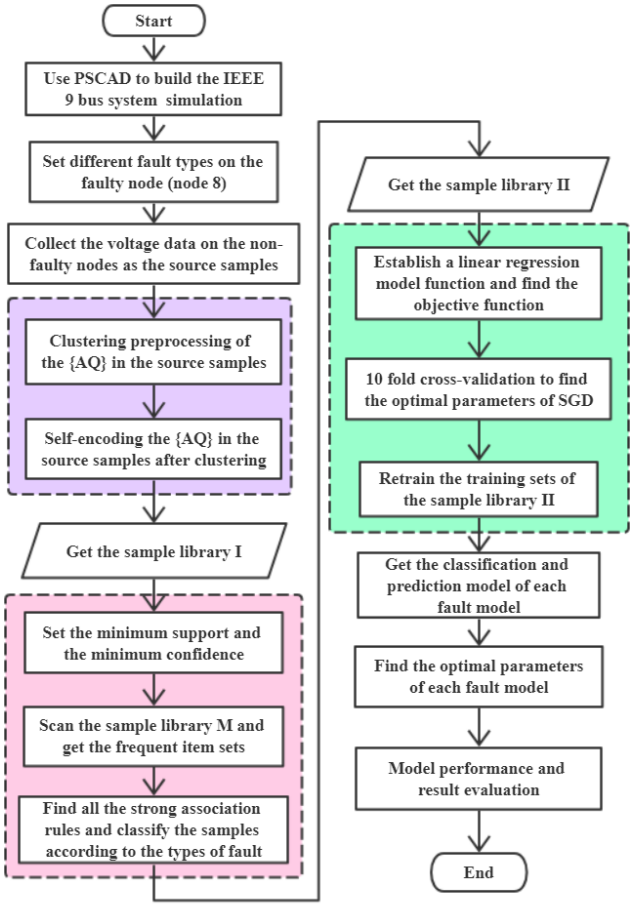


FIGURE 5. Overall method flow.

A. SIMULATION ESTABLISHMENT

The U.S. Western Power Grid WSCC 9 bus system is taken as an example, and some parameters were modified to establish a simulation model of the power system. The classification and prediction of common short-circuit faults in power grids is studied. As shown in Fig. 6, the fault simulation model is established using PSCAD, which is electromagnetic transient simulation software [36]. Firstly, the node load and network parameters of the IEEE 9 bus system were set, where node 8 is set as the faulty node, and a universal meter on each node was installed to obtain the real-time voltage data. Then, the occurrence of the faults was controlled through time fault logic, to ensure the timeliness of addressing faults in practice, and the time for the occurrence of a fault is set to 0.2 s. The control type is external control, a dial was set that can change the fault type by manual interactive control, and the number on the dial corresponds to a certain fault. For example, if the value of the dial is 1, it corresponds to SPGF, and the dial was linked to the control panel user interface, which changes the fault type by changing the digital position on the control panel. When a type of fault occurs, the voltage data are collected separately. This simulation mainly collects the data within the visible range of the waveform graph. Faults in multiple time periods are set, and the time of the voltage sample collection is the same after all of the faults.

B. DATA COLLECTION AND PREPROCESSING

When different fault types occur at 8 nodes, the voltage data on the other nodes are collected. After the simulation is completed, the waveform is obtained, and the voltage data are established in the data table. The calculation example is shown in Fig. 6, where more than 20,000 sets of data on the different fault types are used as source sample data to train the different fault models. When the three-phase short-circuit fault occurs at node 8, the voltage waveforms of node 7 and node 8 are shown in Fig. 7. The horizontal axis of the figure is the time, and the vertical axis is the voltage value.

The first-layer clustering preprocessing is to randomly select K as the initial number of the classification clusters, and after the criterion function converges, the contour coefficient method is used to find the best K value of the samples on each node. When four types of short-circuit faults occur at 8 nodes, all of the voltage data samples at node 2 in the voltage data source sample library are extracted, and then, they are clustered according to the clustering process mentioned in section III.B. Finally, the contour coefficient method is used to match the best K value of the voltage data samples at node 2, and the result is shown in Fig. 8. It can be seen that when $K = 10$, the contour coefficient S_k is the local maximum, and therefore, the K value of the cluster number of the voltage data samples at node 2 is set to 10, and all of the voltage data samples are divided into 10 clusters. The voltage data samples at other nodes are also divided into the most suitable number of clusters according to the above method, where the data samples in each cluster have large similarity after clustering.

After the data preprocessing of all nodes is completed, the classified data samples are encoded through encoding rules. For example, when the amplitude of the voltage data at node 2 ranges from 108.26 to 226.82, the optimal K value is 10, and thus, the voltage data samples are divided into 10 clusters. The sample value ranges of each cluster are [108.26,113.45], [113.46,151.40], [151.41,164.36], [164.37,177.85], [177.86,182.45], [182.46,185.66], [185.67,188.49], [188.50,192.15], [192.16,206.17], and [206.18,226.82]. When the voltage of a certain classified data sample is 158.60, the value falls in the third cluster of the voltage range at node 2, and thus, the data sample is recorded as 203. The voltage data samples at the other nodes are also encoded in the same way to obtain sample library I. The self-encoding form of some samples in sample library I is shown in Table 2:

C. ASSOCIATION MINING AND REGRESSION TRAINING

After the source samples' clustering is preprocessed, the Apriori algorithm is used to mine the association rules for the data samples of sample library I: Firstly, the minimum support is set to 0.3, and the minimum confidence is set to 0.7. The sample sets in sample library I whose support degree is greater than 0.3 are determined to be the frequent itemsets. Then, all of the frequent itemsets that meet the conditions of the confidence being greater than 0.7 and the lift being greater

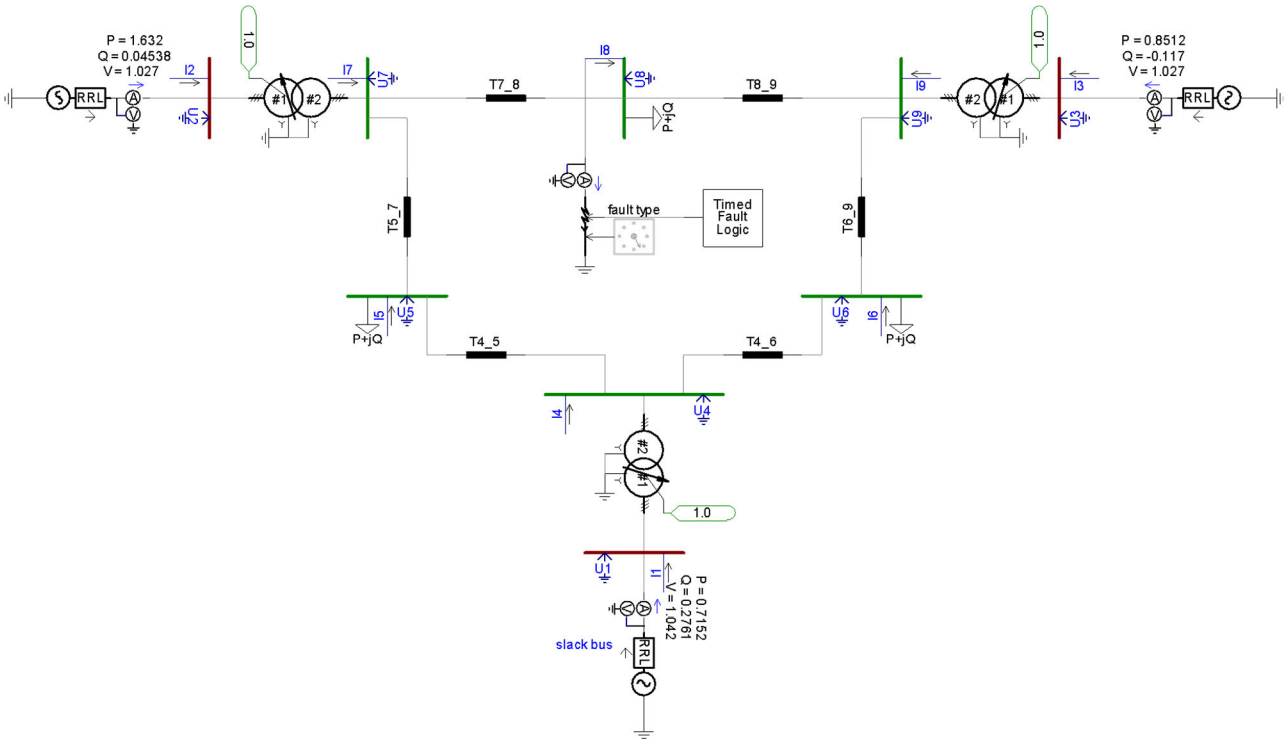


FIGURE 6. IEEE 9 bus system fault simulation model.

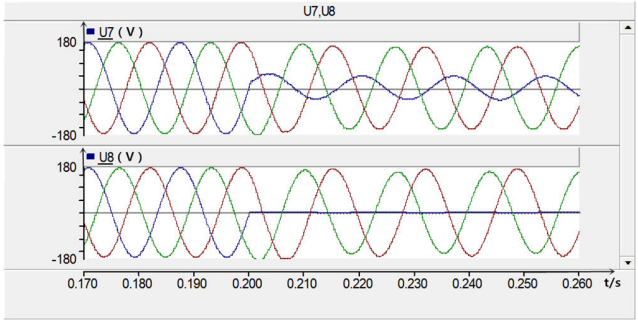


FIGURE 7. The voltage waveform of node 7 and node 8.

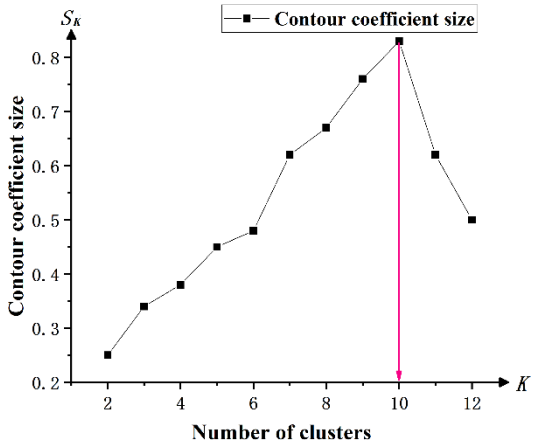


FIGURE 8. The best K of the voltage samples of node 2 (three-phase fault).

than 1 are screened out as samples of the strong association rules. The mining result of the frequent itemsets is partially listed in Table 3:

The brackets and arrows in the second column of Table 3 represent that the preceding event could cause the subsequent event to occur. The corresponding values of the support degree, the confidence degree, and the lift degree are given in the third, fourth and fifth columns, respectively. For example, the strong association rule of $\{101,206,906\}\rightarrow\{3\}$ in the seventh row and second column of Table 3 is that when the voltage data at node 1 are in the first cluster, the voltage at node 2 are in the sixth cluster, and the voltage at node 9 is in the sixth cluster, which could cause two-phase ground faults at the faulty node. All of the sample groups with strong association rules are sorted from sample library II according to the type of faults. The self-encoding form of those samples in sample library II that are highly related to single-phase ground faults is shown in Table 4:

Before the training on sample library II, to avoid the sensitivity of the proposed method to the parameter value, the 10-fold cross-validation method is used to optimize the parameters of L, R, α in formula (14) and hyperparameter N, and the model is retrained by the training set under the optimal parameters to obtain the optimal model; finally, the model result is verified by the test set. The specific implementation is as follows:

1) All of the samples of a certain fault in sample library II are subjected to 10-fold cross-validation. All of the

TABLE 2. The self-encoding form of some samples in sample library I.

Group	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 9	Fault type
1	105	201	304	409	505	606	702	908	3
2	102	203	305	407	502	602	701	909	4
3	102	201	306	402	506	601	701	905	1
4	106	202	304	405	501	608	708	905	1
5	103	204	305	405	501	607	702	906	2
6	105	204	301	408	504	604	706	905	1
7	104	202	306	406	503	606	707	907	2
8	106	203	305	407	502	605	708	908	2
9	102	203	302	402	507	607	704	903	2
10	103	204	305	408	502	604	705	905	1
11	101	207	306	401	507	604	701	902	2
12	106	205	302	405	505	603	707	902	2
13	105	206	304	409	505	605	705	903	2
14	105	203	305	407	503	605	703	908	1
15	104	204	306	406	504	606	702	905	1
16	104	202	306	405	504	607	701	905	1

TABLE 3. Partial rules mined from the frequent itemsets.

Group	Sample group	Support	Confidence	Lift
1	{905}→{1}	0.35	0.749	2.14
2	{306}→{1}	0.42	0.749	1.78
3	{205,407}→{1}	0.33	0.850	2.56
4	{306,905}→{1}	0.41	0.846	2.06
5	{205,606,701}→{2}	0.55	0.884	1.61
6	{606}→{2}	0.44	0.738	1.74
7	{403,505}→{2}	0.45	0.749	1.68
8	{101,206,906}→{3}	0.53	0.874	1.64
9	{305,406}→{3}	0.57	0.765	1.34
10	{105,406,502}→{3}	0.55	0.889	1.62
11	{405,506,901}→{3}	0.5	0.889	1.61
12	{205,506}→{3}	0.45	0.714	1.59
13	{105,604,908}→{4}	0.36	0.659	1.83
14	{108,303}→{407}	0.45	0.683	1.52
15	{402,503,607}→{105}	0.48	0.702	1.46

TABLE 4. The self-encoding form of the some samples of sample library II.

Group	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 9	Fault type
1	104	204	306	406	504	606	702	905	1
2	104	203	305	407	503	605	701	906	1
3	104	202	306	405	504	607	701	905	1
4	103	202	306	405	503	607	702	905	1
5	103	204	305	405	503	607	702	905	1
6	106	204	305	405	504	604	701	905	1
7	105	205	306	406	503	606	702	907	1
8	105	205	305	407	503	605	703	908	1
9	105	203	306	406	504	607	702	905	1
10	106	204	305	405	503	604	702	905	1
11	103	202	306	407	504	604	701	905	1
12	103	204	306	405	503	605	702	907	1
13	105	202	304	406	503	605	701	905	1

samples are divided into 10 parts, each of which is used as a cross-validation set in turn, and the other 9 parts are used as a training set. The samples were trained 10 times in total. The parameters with the highest cross-validation score are taken as the optimal parameters.

2) Sample library II is divided into training set A and test set B.
3) Training set A retrains the model under the optimal parameters.
4) Finally, test set B tests the model and obtains the results.

To prove that the regression training accuracy of the sample library after the clustering and association rule mining is higher than after only clustering (without mining by association rules), sample library I is also subjected to 10-fold cross-validation. For example, in the parameter optimization process of the single-phase short-circuit fault model, after each group of parameters is substituted into the SGD algorithm program, the cross-validation scores of sample library I and sample library II that correspond to these solutions are listed, as shown in Table 5. Then, the cross-validation scores are compared, and the optimal solution of the parameters (L, R, α, N) of each fault model is selected. It can be seen that the cross-validation score is the highest at 0.556 during the regression of sample library I, and the corresponding optimal parameters group is (*Log*, *L*1, *0.1*, *1000*). During the regression of sample library II, the cross-validation score is the highest at 0.788, and the corresponding optimal parameters group is (*Log*, *L*2, *0.1*, *500*). In addition, the cross-validation scores of sample library II is higher than those of sample library I under the same parameter groups.

It can be seen from Table 5 that under the mathematical model of SGD, the optimal loss functions of sample library I and sample library II both selected logistic regression. Because the amount of classification calculation in Logistic regression is less and the storage resource is less, the training data can be quickly integrated into the model. Compared with SVM, it is easy to obtain the probability scores of the samples. For the optimal regularization items, the sample library I chooses $L1$, because the features between the samples are not obvious in sample library I, and thus, the features must be sparse, which reduces the number of weight parameters and the complexity of the model. Sample library II selects $L2$, which reflects that the features between the samples in sample library II have a certain similarity after the association rules, and therefore, the complexity of the model is reduced only by reducing the value of the weight. In addition, $L2$ can also be combined with logistic regression to solve multicollinearity problems. Both choices for α are 0.1, which reflects the same degree of simplification of the parameter scale. For the optimal number of iterations N, sample library I iterated 1000 times, while sample library II iterated only 500 times. It can be seen that the cross-validation training of sample library II has a short convergence time, fast fitting speed, and lower model complexity.

To further explain that the data mining process using the clustering and association rules is more accurate than that using only clustering, training set A in sample library I and training set A in sample library II are used to retrain the model separately under each group of parameters, and the SGD test

TABLE 5. Cross-validation scores of the sample library I and the sample library II in the single-phase short-circuit fault model.

Group	Alternative parameters	Cross-validation score (Sample library Ⅰ)	Cross-validation score (Sample library Ⅱ)
1	*Hinge, L1, 0.1, 500*	0.475	0.565
2	*Hinge, L1, 0.1, 1000*	0.325	0.614
3	*Hinge, L1, 0.05, 500*	0.378	0.588
4	*Hinge, L1, 0.05, 1000*	0.356	0.656
5	*Hinge, L1, 0.01, 500*	0.348	0.505
6	*Hinge, L1, 0.01, 1000*	0.454	0.586
7	*Hinge, L2, 0.1, 500*	0.325	0.686
8	*Hinge, L2, 0.1, 1000*	0.316	0.624
9	*Hinge, L2, 0.05, 500*	0.346	0.446
10	*Hinge, L2, 0.05, 1000*	0.445	0.554
11	*Hinge, L2, 0.01, 500*	0.418	0.528
12	*Hinge, L2, 0.01, 1000*	0.356	0.596
13	*Log, L2, 0.1, 500*	0.456	0.788
14	*Log, L2, 0.1, 1000*	0.435	0.628
15	*Log, L2, 0.05, 500*	0.215	0.436
16	*Log, L2, 0.05, 1000*	0.248	0.548
17	*Log, L2, 0.01, 500*	0.336	0.578
18	*Log, L2, 0.01, 1000*	0.236	0.536
19	*Log, L1, 0.1, 500*	0.345	0.546
20	*Log, L1, 0.1, 1000*	0.556	0.698
21	*Log, L1, 0.05, 500*	0.315	0.535
22	*Log, L1, 0.05, 1000*	0.326	0.554
23	*Log, L1, 0.01, 500*	0.328	0.624
24	*Log, L1, 0.01, 1000*	0.384	0.588

scores of each group of parameters are compared. As shown in Fig. 9(a), it can be seen that sample library I obtains the highest SGD test score at the 20_{th} group of the parameters (*Log*, *L*1, 0.1, 1000), which is 0.71. Sample library II obtains the highest SGD test score at the 13_{th} group of the parameters (*Log*, *L*2, 0.1, 500), which is 0.82. Furthermore, the parameters of the other fault models have also been optimized, and the SGD test scores of sample library I and sample library II are obtained; the results are shown in Fig. 9(b), 9(c) and 9(d).

It can be seen from Fig. 9 that under the same parameters, the SGD test score of sample library II is always higher than that of sample library I. This finding occurs because the parameter optimization directly uses SVM or SGD or other algorithms in sample library I, which is slow and can easily fall into a local optimum. However, the samples in sample library II, which were processed by the clustering and association rules, are more closely related to each other, and thus, the parameter optimization speed is faster, and the best classification point or the best classification line or the best classification surface will be found accurately. Moreover, the proposed method optimizes the unknown parameters in advance by adopting the cross-validation method, which greatly accelerates the training speed of the fault prediction model and improves the accuracy of the model.

Through the obtained optimal parameter set (L, R, α, N), the optimal solutions of the model parameters w of different fault types are solved by SGD iterations according to formulas (14) and (15), which are shown in Table 6, where a positive value indicates a positive correlation that makes the variable and the dependent variable change in the same direction; a negative value indicates a negative correlation that makes the variable and the dependent variable change in the opposite direction.

TABLE 6. The model parameter *w* of the different fault types.

Single-phase ground		two-phase phase-phase		two-phase ground		three-phase fault model	
node1	6.3923	*node1*	4.2268	*node1*	5.0025	*node1*	4.6894
node2	1.7931	*node2*	3.4426	*node2*	3.2284	*node2*	4.6473
node3	4.2512	*node3*	2.6687	*node3*	3.1131	*node3*	3.2485
node4	-0.2413	*node4*	0.3695	*node4*	0.6123	*node4*	-0.2356
node5	4.7168	*node5*	-3.2456	*node5*	-2.5535	*node5*	3.4236
node6	-2.5159	*node6*	3.2258	*node6*	3.5721	*node6*	-2.4264
node7	-2.0162	*node7*	-2.1145	*node7*	-3.0361	*node7*	-4.2364
node9	-3.4614	*node9*	-3.2536	*node9*	-4.2364	*node9*	-4.2525

D. TEST SET VERIFICATION

To measure the performance of the fault prediction model obtained by the proposed method, the test set of the source sample library is selected to test the model. There are 2658 group samples, which are randomly selected from the source sample library to participate in the test, and the confusion matrix is used to evaluate the accuracy of the test results and the precision of each fault type. The result of the confusion matrix of the test is shown in Fig. 10, in that the predicted and actual value distributions of each fault type can be found, where each row represents the real fault type, and each column represents the fault type predicted by the model. The darker the color is, the larger the number of samples.

The test results of the test set show that the overall accuracy of the model is 93.8%. The precision of each fault model is the following: the single-phase ground fault is 94.7%, the two-phase phase-phase fault is 91%, the two-phase ground fault is 95.4% and the three-phase fault is 93%. Therefore, the accuracy of the FCPM of sample library II under the optimal parameters is high, and the prediction precision of each fault model is also high.

E. STATISTICAL TEST AND COST EFFECTIVENESS OF THE PROPOSED METHOD

Because the classification prediction model is a multiple linear regression model, the F test mentioned in section III.F.1) is used to test the significant difference of the model and whether the selection of multiple parameters in the model is appropriate. 100 sets of samples were selected for the F test, and 8 non-faulty nodes decided the degree of freedom $p = 8$.

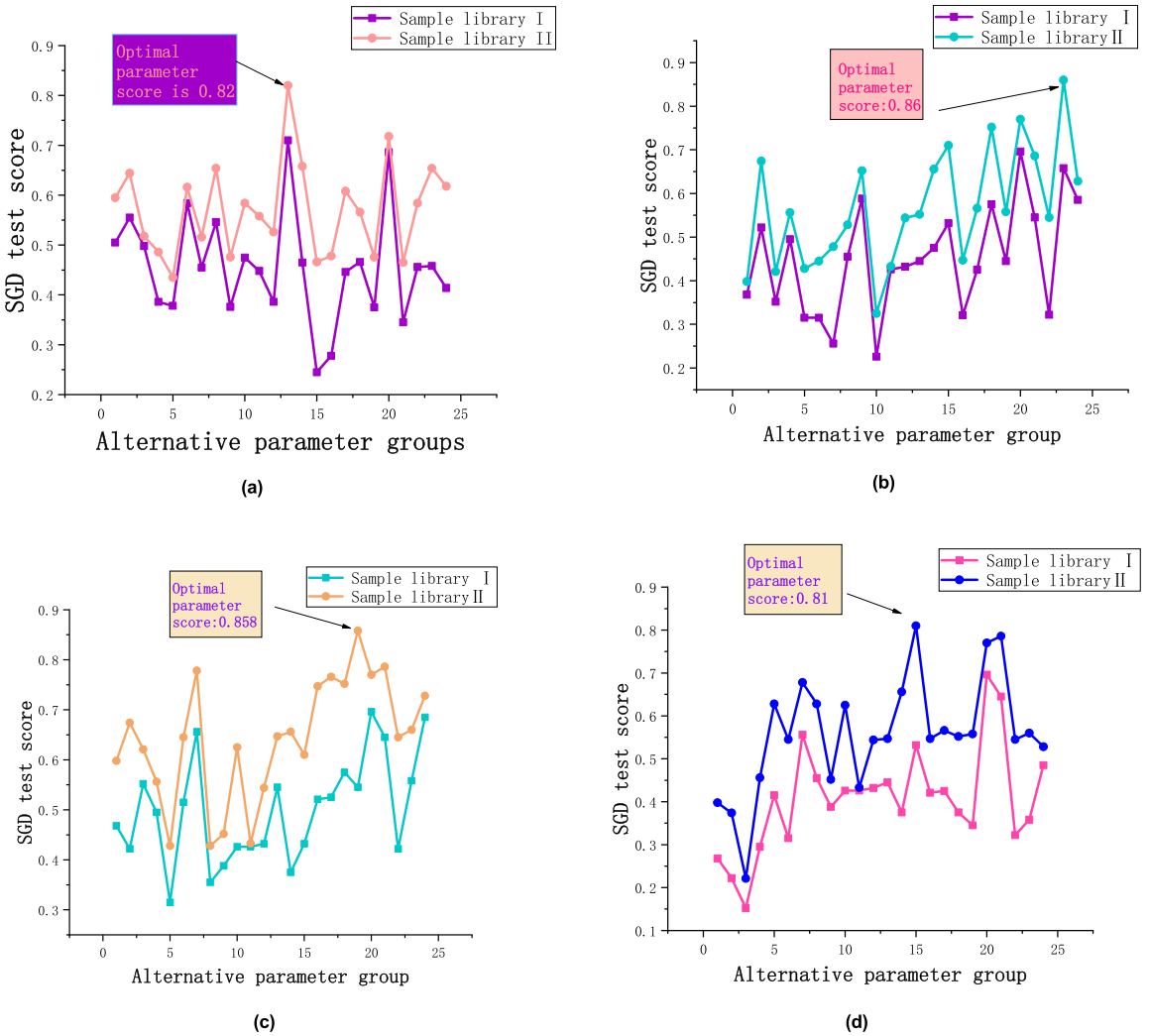


FIGURE 9. SGD test scores of the different fault models. (a) Single-phase ground fault model. (b) two-phase phase-phase fault model. (c) two-phase ground fault model. (d) three-phase fault model.

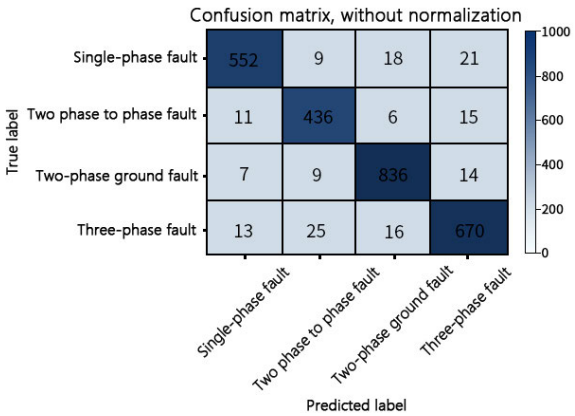


FIGURE 10. Confusion matrix for various types of faults.

The variance analysis of the regression equation is shown in Table 7:

It is calculated that F=87.58, which refers to the F-value table with a significance level of 0.05 and a confidence level of 95%. When $p = 8$, $F_p = 1.94$, $F > F_p$. Therefore, there are significant differences between the variables and dependent variables in this model, and the model is constructed reasonably.

TABLE 7. The variance analysis of the regression equation.

Types of sum of squares	Degree of freedom	F value
SSR	p=8	87.58
SSE	n-p-1=91	

To measure the computational complexity of the model algorithm, the Big O notation method mentioned in section III.F.2) is used to calculate the time complexity of the algorithm. The time complexity of each method is shown in Table 8:

The time complexity of the overall model is O (n^2), and the commonly used time complexity from small to large is O(1)<O(n)<O(nlogn)<O(n^2)<O(n^3)<O(2^n)<O(n!), thus, the

TABLE 8. The time complexity of each method.

algorithm	time complexity
K-means	$O(Kn^2)$
Self-coding	O(n)
Apriori	$O(4n^2)$
Cross-validation	O(10n)
SGD	O(n)
Overall model	$O(n^2)$

time complexity of the overall model is feasible. In addition, the overall running time of the program is 3.7 s. The overall cost effectiveness of the proposed method is not large, which guarantees the timeliness of the fault classification and prediction.

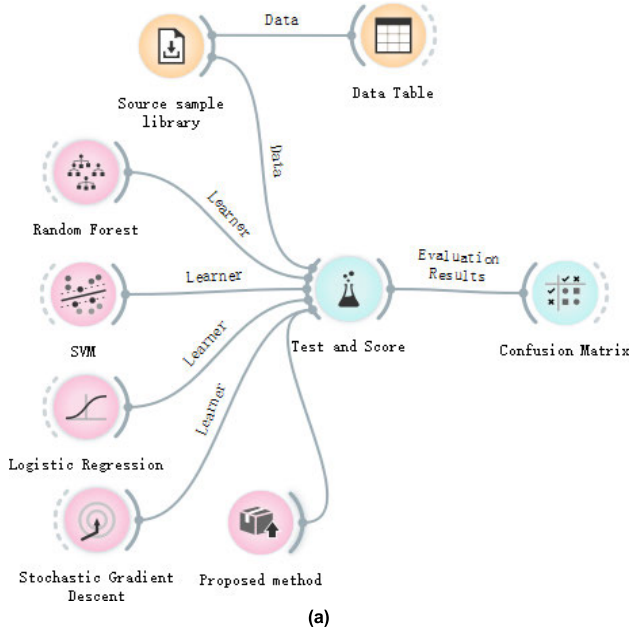
F. METHODS COMPARISON AND VERIFICATION

The proposed method is compared with other methods to accomplish its algorithm verification, which further illustrates the effectiveness and scalability of the proposed method. The regression algorithms involved in the comparison include logistic regression [37], SVM [38], [39], random forest [40], [41], SGD and the proposed method. The constructed model is shown in Fig. 11(a), and the test scores of the performance indexes are shown in Fig. 11(b).

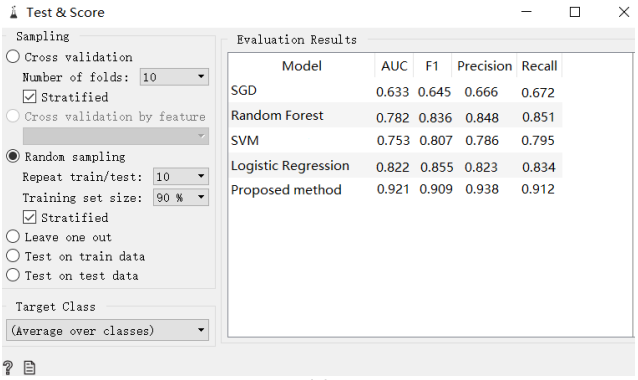
To compare the performances of several algorithms, the ROC curve of each algorithm is shown in Fig. 12. The larger the area under the ROC curve (AUC) is, the better the performances of the algorithms. In addition, the F1 scores, the precision and the recall rate of the algorithms are shown in Fig. 13.

As can be seen from Fig. 12, the AUC values of the proposed method, logistic regression, random forest, SVM, and SGD are 0.921, 0.822, 0.782, 0.753, 0.633, respectively, which proves that the proposed method has the best classification and prediction effect.

In addition, it can be seen from Fig. 13 that the performance index scores of all of the algorithms sorting from the largest to the smallest are as follows: the proposed method, logistic regression, random forest, SVM and SGD. Before the parameter optimization and the association rules, the F1 value, the precision rate and the recall rate of the model are 64.5%, 66.6%, and 67.2%, respectively. After the parameter optimization and the association rules, the F1 value, the precision rate and the recall rate of the model are 90.9%, 93.8%, and 91.2%, respectively. The experimental results show that compared with logistic regression, SVM and random forest, the performance indexes of the proposed method are significantly better. The reason is that there is a common problem with SVM, random forest and logistic regression: When faced with a small amount of data and ambiguous features, it cannot classify or perform regression very well, and thus, the performance will be defective. However, the proposed method clarifies the features in advance through clustering



(a)



(b)

FIGURE 11. Model building and the scores of each evaluation index.

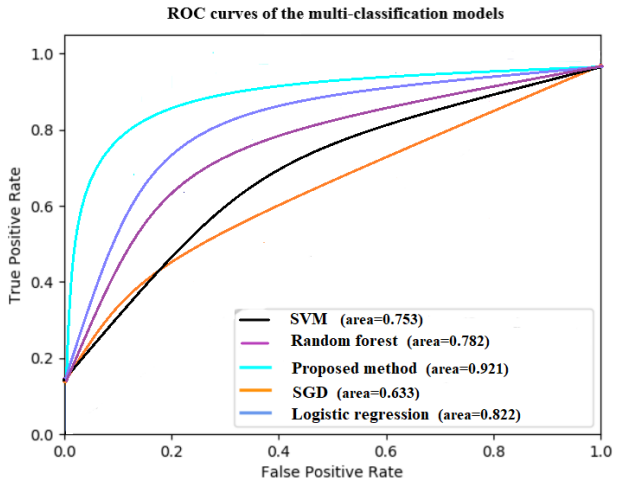


FIGURE 12. Comparison of the ROC curve of the different regression algorithms.

and association rules. The parameter optimization makes the model results better, and thus, the performance indexes are better than the above three algorithms. In addition, the

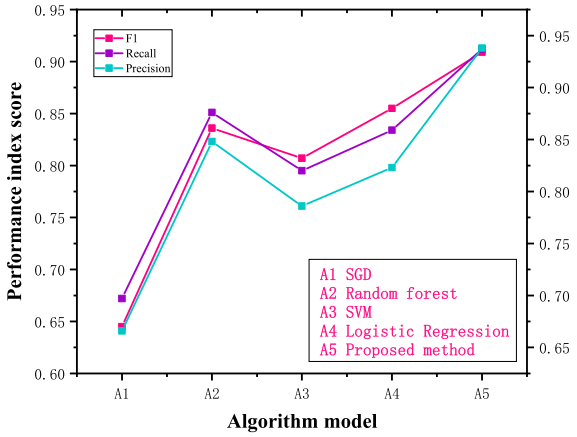


FIGURE 13. The F1 score, precision and recall of the algorithm model.

proposed method also proves that the performance index scores of the SGD when directly used for regression is not high. However, after the source samples are processed by the clustering preprocessing and association rules, SGD after parameter optimization is used to train the processed samples again, and the performance index scores are significantly improved. The reason is that the loss function used by SGD each time is only determined by a small batch of data, and the loss function is different from the real complete set loss function; thus, the gradient of its solution also contains a certain degree of randomness. At a saddle point or local minimum point, it will oscillate and jump, and thus, the result is that the prediction accuracy is not high. Applying clustering and association rules to filter the samples in advance can reduce the irrelevance between small batches of data and reduce the shock. Therefore, the performance indexes of SGD are improved based on the fast data training speed of the clustering preprocessing and association rules.

G. PRACTICAL APPLICATIONS

To prove the scalability of the proposed method, a short-circuit fault classification and prediction test is carried out with the help of a power grid cooperation project in a certain urban area distribution network. The distribution network has 8 generators, 12 transformers, and 56 data collection points. According to the method in the article, in the distribution network, more than 20,000 sets of voltage data on the different fault types are used as source sample data, the amount of the source sample data and the types of the fault are the same as the WSCC 9 bus system. In this experiment, the voltage data on the 56 collection points in the distribution network affect the fault types together, the data at each collection point have been processed by the algorithm model for a total of 210 times, therefore, the computational cost of the classification and prediction model in the distribution network is $56 \times 20000 \times 210 = 0.235$B. In addition, the computational cost of the classification and prediction model in the WSCC 9 bus system is $9 \times 20000 \times 78 \approx 0.014$B. This experiment was also done on a personal computer. After the test is completed, the accuracy of the fault classification, the runtime and the computational cost of two systems are shown in Table 9:

TABLE 9. Comparison of fault classification accuracy, runtime and computational cost in the different systems.

Systems	Accuracy	Runtime(s)	Computational cost(B)
WSCC 9 bus system	93.8%	3.7	0.014
Practical application system	89.2%	11.6	0.235

Among them, B stands for billion, which is the unit of the number of the times the computer runs. It can be seen from the table that, compared with the WSCC 9 bus system, the accuracy of the fault classification results has little change, being 93.8% and 89.2% respectively. The computational cost is 0.235B, which proves that the proposed method can be computationally efficient to be used in practical applications. It is concluded that with the increase of the system nodes, the runtime and the computational cost will increase, which is because the runtime and the computational cost are related to the complexity of the actual system, such as the number of nodes, the number of line branches, etc. Therefore, it can be predicted that in a more complex actual system, the computational cost will be greater.

V. CONCLUSION

Through multiple data mining methods, including clustering, association rules, cross-validation optimization and SGD, the proposed method identifies the data samples that are strongly related to the specific faults and determines the potential laws for building a more accurate fault classification and prediction model. Moreover, through the existing operating data, the proposed method can predict which type of fault will occur soon, and it plays an important role in the classification and prediction of fault types. For the specific fault models, the proposed method uses an optimization algorithm to determine the optimal parameters of the fault models after clustering and association mining of the training samples; then, the fault models are obtained from training samples under the optimal parameters, and thus, the effect of the classification and prediction is better than other methods mentioned in this paper.

The proposed method processes the source data in advance, avoiding the low accuracy of the fault classification and prediction model due to the low-impact or irrelevant data, and it can realize the fault classification and prediction of the power system in time and accurately. Otherwise, the proposed method can be widely applied to the fault classification and prediction of various busbars, transformers, transmission lines in the power system and the classification and prediction of the other systems that involve multi-attribute classification. In addition, it can also be extended to medical

disease prediction, electronic communication fault detection and other fields.

However, the proposed method has some limitations: through the experiment test on the distribution network in an urban area, it is found that the proposed method cannot achieve rapid and real-time fault prediction in practical applications, and the timeliness needs to be further improved. In addition, this paper only considers the voltage data of each node in the same time period under different fault conditions, without further analysis the high-dimensional data sample composed of voltage data of each node, current data of each branch and time. Therefore, the future research work will focus on the practical engineering applications of the multi-dimensional data of power systems for fault classification and prediction.

APPENDIX

The nomenclature of the article:

Terms	Abbreviations
support vector machine	**SVM**
long-short term memory	**LSTM**
stochastic gradient descent	**SGD**
empirical mode decomposition	**EMD**
fault classification and prediction model	**FCPM**
single-phase ground fault	**SPGF**
two-phase phase-phase fault	**TPPF**
two-phase ground fault	**TPGF**
three-phase fault	**TPF**
structural risks	**SR**
empirical risk	**ER**
loss function	**L**
receiver operating characteristic	**ROC**
true positive	**TP**
true negative	**TN**
false positive	**FP**
false negative	**FN**
true positive rate	**TPR**
false positive rate	**FPR**
regression sum of squares	**SSR**
sum of squared residuals for regression	**SSE**
area under receiver operating characteristic curve	**AUC**

REFERENCES

[1] G. M. Ali and S. A. Al-Mawsawi, "Multiple UPFCs mathematical model enhancing multi-machine power system control," in *Proc. 10th Jordanian Int. Electr. Electron. Eng. Conf. (JIEEEC)*, May 2017, pp. 1–4.

[2] Q. Wang and P. Qiu, "The application of equipment overheating and arcing fault warning and protection systems of switchgear in power systems," in *Proc. IEEE Innov. Smart Grid Technol.-Asia (ISGT Asia)*, May 2019, pp. 1135–1137.

[3] L. Song, H. Wang, and P. Chen, "Step-by-step fuzzy diagnosis method for equipment based on symptom extraction and trivalent logic fuzzy diagnosis theory," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 6, pp. 3467–3478, Dec. 2018.

[4] L. Bessissa, L. Boukezzi, D. Mahi, and A. Boubakeur, "Lifetime estimation and diagnosis of XLPE used in HV insulation cables under thermal ageing: Arithmetic sequences optimised by genetic algorithms approach," *IET Gener., Transmiss. Distrib.*, vol. 11, no. 10, pp. 2429–2437, Jul. 2017.

[5] D. Kumar, I. Kamwa, and S. R. Samantaray, "Multi-objective design of advanced power distribution networks using restricted-population-based multi-objective seeker-optimisation-algorithm and fuzzy-operator," *IET Gener., Transmiss. Distrib.*, vol. 9, no. 11, pp. 1195–1215, Aug. 2015.

[6] L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," *IEEE Access*, vol. 7, pp. 54007–54014, 2019.

[7] T. K. Saha and P. Purkait, "Investigation of an expert system for the condition assessment of transformer insulation based on dielectric response measurements," *IEEE Trans. Power Del.*, vol. 19, no. 3, pp. 1127–1134, Jul. 2004.

[8] S. Chen, H. Ge, J. Li, and M. Pecht, "Progressive improved convolutional neural network for avionics fault diagnosis," *IEEE Access*, vol. 7, pp. 177362–177375, 2019.

[9] G. Rigatos, A. Piccolo, and P. Siano, "Neural network-based approach for early detection of cascading events in electric power systems," *IET Gener., Transmiss. Distrib.*, vol. 3, no. 7, pp. 650–665, Jul. 2009.

[10] H. Malik and S. Mishra, "Artificial neural network and empirical mode decomposition based imbalance fault diagnosis of wind turbine using TurbSim, FAST and simulink," *IET Renew. Power Gener.*, vol. 11, no. 6, pp. 889–902, May 2017.

[11] J. Liu, Z. Zhao, C. Tang, C. Yao, C. Li, and S. Islam, "Classifying transformer winding deformation fault types and degrees using FRA based on support vector machine," *IEEE Access*, vol. 7, pp. 112494–112504, 2019.

[12] F. N. Rudsari, A. A. Razi-Kazemi, and M. A. Shoorehdeli, "Fault analysis of high-voltage circuit breakers based on coil current and contact travel waveforms through modified SVM classifier," *IEEE Trans. Power Del.*, vol. 34, no. 4, pp. 1608–1618, Aug. 2019.

[13] H. Lei, H. Yifei, and G. Yi, "The research of business intelligence system based on data mining," in *Proc. Int. Conf. Logistics, Informat. Service Sci. (LISS)*, Jul. 2015, pp. 1–5.

[14] S. D. Mohaghegh, "Reservoir simulation and modeling based on artificial intelligence and data mining (AI&DM)," *J. Natural Gas Sci. Eng.*, vol. 3, no. 6, pp. 697–705, 2011.

[15] H. L. Han, H. Y. Ma, and Y. Yang, "Study on the test data fault mining technology based on decision tree," *Procedia Comput. Sci.*, vol. 154, pp. 232–237, Jan. 2019.

[16] F. Ciarapica, M. Bevilacqua, and S. Antomarioni, "An approach based on association rules and social network analysis for managing environmental risk: A case study from a process industry," *Process Saf. Environ. Protection*, vol. 128, pp. 50–64, Aug. 2019.

[17] M. Krishnan and G. Jabert, "Detection of soil borne pathogens in coffee plantations by modified k-means clustering," in *Proc. Int. Conf. Opt. Imag. Sensor Secur. (ICOSS)*, Coimbatore, India, Jul. 2013, pp. 1–8.

[18] S. Debnath and M. Saeedifard, "Simulation-based gradient-descent optimization of modular multilevel converter controller parameters," *IEEE Trans. Ind. Electron.*, vol. 63, no. 1, pp. 102–112, Jan. 2016.

[19] A. Garcés, "On the convergence of Newton's method in power flow studies for DC microgrids," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5770–5777, Sep. 2018.

[20] A. A. El-Fergany and M. El-Arini, "Meta-heuristic algorithms-based real power loss minimisation including line thermal overloading constraints," *IET Gener., Transmiss. Distrib.*, vol. 7, no. 6, pp. 613–619, Jun. 2013.

[21] L. Yang, S. L. Ho, and W. N. Fu, "Design optimizations of electromagnetic devices using sensitivity analysis and Tabu algorithm," *IEEE Trans. Magn.*, vol. 50, no. 11, pp. 1–4, Nov. 2014.

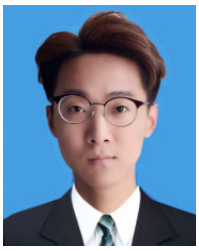
[22] H. Jia, J. Li, W. Song, X. Peng, C. Lang, and Y. Li, "Spotted hyena optimization algorithm with simulated annealing for feature selection," *IEEE Access*, vol. 7, pp. 71943–71962, 2019.

[23] A. Y. Abdelaziz, R. A. Osama, and S. M. El-Khodary, "Reconfiguration of distribution systems for loss reduction using the hyper-cube ant colony optimisation algorithm," *IET Gener., Transmiss. Distrib.*, vol. 6, no. 2, pp. 176–187, 2012.

[24] Z. Wang, Y. Fu, C. Song, P. Zeng, and L. Qiao, "Power system anomaly detection based on OCSVM optimized by improved particle swarm optimization," *IEEE Access*, vol. 7, pp. 181580–181588, 2019.

[25] X. Zhou, Z. Wang, D. Li, H. Zhou, Y. Qin, and J. Wang, "Guidance systematic error separation for mobile launch vehicles using artificial fish swarm algorithm," *IEEE Access*, vol. 7, pp. 31422–31434, 2019.
[26] S. Zhang, Y. Wang, M. Liu, and Z. Bao, "Data-based line trip fault prediction in power systems using LSTM networks and SVM," *IEEE Access*, vol. 6, pp. 7675–7686, 2018.
[27] Y. Guo, G. Li, H. Chen, J. Wang, M. Guo, S. Sun, and W. Hu, "Optimized neural network-based fault diagnosis strategy for VRF system in heating mode using data mining," *Appl. Thermal Eng.*, vol. 125, pp. 1402–1413, Oct. 2017.
[28] L.-H. Ren, Z.-F. Ye, and Y.-P. Zhao, "A modeling method for aero-engine by combining stochastic gradient descent with support vector regression," *Aerosp. Sci. Technol.*, vol. 99, Apr. 2020, Art. no. 105775.
[29] Y. Li, R. W. Liu, Z. Liu, and J. Liu, "EMD-based recurrent neural network with adaptive regrouping for port cargo throughput prediction," in *Proc. Int. Conf. Neural Inf. Process.*, 2018, pp. 499–510.
[30] Y. Li, R. W. Liu, Z. Liu, and J. Liu, "Similarity grouping-guided neural network modeling for maritime time series prediction," *IEEE Access*, vol. 7, pp. 72647–72659, 2019.
[31] M. Tian, L. Zhang, P. Guo, H. Zhang, Q. Chen, Y. Li, and A. Xue, "Data dependence analysis for defects data of relay protection devices based on *apriori* algorithm," *IEEE Access*, vol. 8, pp. 120647–120653, 2020.
[32] G. Cui, J. Guo, Y. Fan, Y. Lan, and X. Cheng, "Trend-smooth: Accelerate asynchronous SGD by smoothing parameters using parameter trends," *IEEE Access*, vol. 7, pp. 156848–156859, 2019.
[33] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD," in *Proc. 4th Int. Conf. Inf. Technol., Comput., Electr. Eng. (ICITACEE)*, Semarang, Indonesia, Oct. 2017, pp. 45–49.
[34] F. Kabir, S. Siddique, M. R. A. Kotwal, and M. N. Huda, "Bangla text document categorization using stochastic gradient descent (SGD) classifier," in *Proc. Int. Conf. Cognit. Comput. Inf. Process. (CCIP)*, Mar. 2015, pp. 1–4.
[35] X. Zhang, N. Gu, R. Yasrab, and H. Ye, "GT-SGD: A novel gradient synchronization algorithm in training distributed recurrent neural network language models," in *Proc. Int. Conf. Netw. Netw. Appl. (NaNA)*, Oct. 2017, pp. 274–278.
[36] X. Han and H. Zhang, "Power system electromagnetic transient and electromechanical transient hybrid simulation based on PSCAD," in *Proc. 5th Int. Conf. Electr. Utility Deregulation Restruct. Power Technol. (DRPT)*, Changsha, China, Nov. 2015, pp. 210–215.
[37] Z. Zhang and Y. Han, "Detection of ovarian tumors in obstetric ultrasound imaging using logistic regression classifier with an advanced machine learning approach," *IEEE Access*, vol. 8, pp. 44999–45008, 2020.
[38] N. Yang and Y. Wang, "Identify silent data corruption vulnerable instructions using SVM," *IEEE Access*, vol. 7, pp. 40210–40219, 2019.
[39] X. Yuan, Z. Liu, Z. Miao, Z. Zhao, F. Zhou, and Y. Song, "Fault diagnosis of analog circuits based on IH-PSO optimized support vector machine," *IEEE Access*, vol. 7, pp. 137945–137958, 2019.
[40] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3434–3445, Dec. 2016.
[41] B. Ni, S. Yan, M. Wang, A. A. Kassim, and Q. Tian, "High-order local spatial context modeling by spatialized random forest," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 739–751, Feb. 2013.

YUNLIANG WANG received the M.S. degree in power systems and automation from Tianjin University, Tianjin, China, in 1988. He is currently a Professor with the School of Electrical and Electronic Engineering, Tianjin University of Technology. His current research interests include intelligent control, data mining, multi-motor coordinated control, microcomputer control, and power electronics technology.



XIAODONG WANG was born in Handan, Hebei, China. In 2018, he joined the School of Electrical and Electronic Engineering, Tianjin University of Technology, where he is currently a Graduate Student. His major is electrical engineering. His main research interests include smart grid, artificial intelligence, data mining, and power system fault prediction.

YANJUAN WU received the M.S. and Ph.D. degrees in power systems and automation from Tianjin University, Tianjin, China, in 2005 and 2013, respectively. She is currently an Associate Professor with the School of Electrical and Electronic Engineering, Tianjin University of Technology. Her current research interests include intelligent control, data mining, smart grids, and grid optimization and control.

YANNAN GUO was born in Chifeng, Inner Mongolia. He received the master's degree in electrical and electronic engineering from the Tianjin University of Technology. He currently works with Tianjin Tianda Qiushi Electric Power High Technology Company Ltd. His main research interests include power system automation, smart grid, and power fault detection.