Fault Prediction of Distribution Network Based on Support Vector Machine

Wen-si HUANG¹, Xin LU¹, Ye LIU², Quan CHEN², Mo-han QI³,* and Hong-jun GAO³

¹State Grid Xintong Yili Technology Co. Ltd., Beijing, China
²State Grid Jiangsu Electric Power Co. Ltd, Jiangsu, China
³College of Electric Engineering, Sichuan University, Chengdu, China

Keywords: Distribution network, Fault prediction, Particle swarm optimization, Support vector machine, Fault level.

Abstract. With the increasing demand for reliability of distribution network in modern society, fault prediction plays an important role in operation, maintenance and reliability improvement of distribution network. In view of the fact that there are many factors affecting the distribution network fault and there are strong uncertainties, a fault prediction method based on support vector machine (SVM) is proposed in this paper. Firstly, the influencing factors of distribution network faults are analyzed. Then, the optimal input variables are obtained by feature selection of the relevant variables. Finally, based on the proposed fault classification method, SVM with particle swarm optimization (PSO) parameters are used to predict distribution network fault. The validity of the proposed method is verified by simulation experiments with actual data.

Introduction

With the development of the global energy Internet and the access of various new energy sources and micro-grids, the interconnection between power grids becomes more complex, and the influence factors of power grid faults also increase. Through the fault prediction of the power network, the staff can react to the power system in time to prevent the occurrence of the fault. Distribution network faults account for a large proportion in power system. Therefore, fast and accurate fault prediction of distribution network is very important to the stability and reliability of the whole power system.

In the traditional research on distribution network reliability, the reliability model is established through the reliability evaluation of distribution network, and finally the reliability index of distribution network is obtained. There are also distribution network reliability prediction methods such as grey prediction and regression prediction [1]. However, these methods are based on reliability index, and do not have the prospects of distribution network failure. Fault prediction is used in the reliability study of existing distribution network to meet the demand of power system. In the fault prediction of distribution network, the prediction model is usually built based on different feeder data. However, there are many factors affecting the fault of distribution network. It is the key to select the input characteristic variables of the fault prediction model scientifically and effectively. Therefore, this paper carries out in-depth analysis from data sources to feature variables selection, so as to make the performance of building the model better.

In the existing research on fault prediction of distribution network, various methods are proposed. There are expert system methods [2], [3], Bayesian networks methods [4], [5], Petri nets [6], [7], [8] A power network monitoring method based on rough set is proposed. [9] A fault detection and classification method based on artificial neural network (ANN) and wavelet transform and resolution analysis technology is proposed. [10] Different types of faults in power system are classified by using electromagnetic transient program and ANN, and a fault location determination method is proposed. [11] A fault prediction method for distribution network based on support vector machine (SVM) is
proposed, but the scientific selection of penalty parameter $c$ and kernel function parameter $\sigma$ in SVM is not considered. If not selected properly, the prediction performance will be poor.

This paper presents a data mining based fault prediction method for distribution network. Firstly, the influencing factors and data sources of distribution network fault prediction are analyzed, and data preprocessing methods and feature selection methods are proposed. Secondly, the distribution network fault is classified according to the fault frequency and fault range. Finally, the support vector machine with particle swarm optimization (PSO) is used to build the prediction model. The validity of the proposed method is verified by an example analysis.

**Framework of the Model**

1) Distribution network fault prediction involves many factors. Firstly, the influencing factors of distribution network fault prediction are analyzed, then the data sources are investigated, and the data obtained are preprocessed.

2) Selecting the input feature of the prediction model. Then a fault prediction model is built by using SVM with PSO optimization parameters.

The frame diagram of the method presented in this paper is shown in Fig. 1.

![Framework diagram of the proposed mode.](image)

**Data Preprocessing and Feature Selection**

**Analysis of Influencing Factors**

Different feeders have different power supply ranges and different equipment operating environments, which lead to different factors of feeder failure. There are three main factors leading to feeder failure.
1) Operating factors
The load state in the feeder supply area is the main factor affecting the operation of distribution network. Different load states will have different effects on the equipment within the feeder range. For example, long-term overload will adversely affect the equipment and cause equipment aging, thus increasing the risk of distribution network failure. Operating factors include average monthly load of feeder, maximum monthly load, fault time, total fault load, etc.

2) Equipment factors
The main equipment factor is the equipment status on the feeder. Equipment factors include the average operating time of each equipment (fuses, load switches, transformers, branch lines, insulated wires, etc.) or the number or length of various equipment.

3) External factors
External factors are mainly affected by weather factors. For example, long-term exposure to sunshine and rain can change the strength of equipment or affect the service life of equipment, resulting in distribution network failure. External factors include monthly average temperature, monthly maximum temperature, monthly minimum temperature, monthly precipitation, monthly sunshine time, etc.

Data Preprocessing
The fault prediction method proposed in this paper is based on the historical fault data of distribution network. However, there are many factors affecting distribution network faults, and there are many fault data sources involved. Combined with the actual situation, this paper mainly collects data from three distribution network management information systems. They are production management system including distribution network fault data, distribution geographic information system including distribution network equipment account data, and intelligent public distribution transformer monitoring system including distribution transformer load level.

The data collected are normalized by the following formula:

\[
x_{\text{norm}} = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}
\]

Where, \(x_i\) is the original data, \(x_{\text{max}}\) and \(x_{\text{min}}\) are the maximum and minimum values in the data sequence, respectively; \(x_{\text{norm}}\) is the normalized value.

Feature Selection
In this paper, Relief F algorithm is used as feature selection algorithm to select model input vectors effectively. This algorithm is a filter multi-category feature selection algorithm. Quantitative analysis of different features is carried out by calculating feature weights. The algorithm has many advantages, such as no limitation of data type, strong anti-noise ability, simple and efficient calculation. The concrete principle of Relief F algorithm is shown in the literature [12].

Risk Classification and Assessment Indicators
Risk Classification
Distribution network fault involves not only the number of blackouts, but also the scope of the fault. In this paper, the outage frequency and fault range based on statistical probability are used to classify the fault grade. Among them, the fault range is expressed by the total load of the outage. Because the total load of different feeder binding is different, the sum of the load of different feeder faults in the same month is divided by the total checking capacity of the feeder to reflect the extent of the fault range. The ratio of fault load \(L_i\) is expressed as follows:
\[ L_i = \sum_{j=1}^{n} \frac{S_{ij}}{S_{iN}} \]  

Where, \( n \) denotes the total number of faults in the month; \( S_{iN} \) denotes the checking capacity of feeder \( i \)th; \( S_{ij} \) denotes the fault load of feeder \( i \)th in the month \( j \). In order to consider both the number of outages and the scope of faults, a weighted method is used to establish the fault classification index \( f \) of distribution network:

\[ f_i = \lambda \cdot N_i + L_i \]  

In the formula, \( N \) is the number of faults; \( \lambda \) is the weight coefficient, and this paper takes 0.25; \( f_i \) is the index of feeder line \( i \)th. The classification of different fault levels is as follows:

\[
F_i =
\begin{cases}
1, & \text{if } f_i < 1 \\
2, & \text{if } 1 \leq f_i < 3 \\
3, & \text{if } 3 \leq f_i
\end{cases}
\]

Where, \( F_i \) represents the failure level. If the fault level is 1, it indicates that the feeder is in normal condition. In general, the number of faults is up to 2 times and \( L_i \) is less than 50%. If the fault level is 2, the risk of the feeder fault state is high. In general, the number of faults is 3 to 8 times and \( L_i \) is less than 100%. If the fault level is 3, it indicates that the risk of the feeder fault state is very high. In general, the number of faults is more than 8 times and \( L_i \) is more than 100%.

Evaluation Indicators

In this paper, Kappa statistical index [13] is used as the evaluation index of distribution network fault prediction. Kappa statistical index is obtained by calculating the error matrix. The matrix is a matrix of \( n_s \) rows and \( n_s \) columns, where rows represent classification points and columns represent prediction points. The calculation method is as follows:

\[
K = \frac{N_s \sum_{i=1}^{r} x_{ir} - \sum_{i=1}^{r} x_{ir} \cdot x_{ic}}{N^2 - \sum_{i=1}^{r} x_{ir} \cdot x_{ic}}
\]

Where, the total number of samples of \( N_s \); \( r \) is the number of rows of a matrix; \( x_{ir} \) is the value of the diagonal line of a matrix; \( x_{ir} \) and \( x_{ic} \) represent the sum of the \( i \)th row and the \( i \)th column, respectively. The accuracy of fault prediction can be obtained from the formula. \( K \) value is between 0 and 1. The larger \( K \) value, the higher the accuracy of fault prediction in distribution network.

Fault Prediction Method for Distribution Network Based on Hybrid Algorithms

SVM

When the number of samples is small, using SVM algorithm has better performance than other algorithms. SVM algorithm maps samples from low-dimensional space to high-dimensional space in a non-linear way, and then transforms them from high-dimensional space to low-dimensional space through kernel function. In this way, the computational burden can be reduced and the optimal solution can be obtained.

The training samples are \( \{x_i, y_i \}, \ldots, \{x_i, y_i \} \). By transforming the data into high-dimensional space and linear regression, the original non-linear fitting problem is transformed into linear regression problem. The fitting function is expressed as follows:
\[ f(x) = \omega \phi(x) + b \]  
Where, \( \omega \) is the weight vector; \( x \) is the input vector; \( b \) is the fitting deviation. By training the samples several times, the following formula can be minimized. Follows:

\[ R(f) = c \sum_{i=1}^{n} \Gamma(f(x_i) - y_i) + \frac{1}{2} \omega^2 \]  
\[ \Gamma(f(x) - y_i) = \frac{1}{n} \left[ \left| f(x_i) - y_i \right| - \varepsilon \right] \left\{ \begin{array}{l} \left| f(x_i) - y_i \right| \geq \varepsilon \\ \left| f(x_i) - y_i \right| < \varepsilon \end{array} \right. \]  

Where, the first term is the empirical error of the optimization problem, and the second term \( \omega^2/2 \) is a positive programming parameter. \( \Gamma(.) \) is the cost parameter to compensate the normalization parameter and empirical error; \( c \) is the penalty factor, which has an important impact on the training effect; and \( \varepsilon \) is the parameter of the loss function of the optimization problem. By introducing relaxation variables \( \xi \) and \( \xi^* \), formula (9) is transformed into:

\[ \min \frac{1}{2} \omega^2 + c \sum_{i=1}^{n} (\xi + \xi^*) \]  
\[ s.t \xi^* \left\{ \begin{array}{l} y_i - \omega \phi(x) - b \leq \varepsilon + \xi \\ \omega \phi(x) + b - y_i \leq \varepsilon + \xi \end{array} \right. \]  

Transform the above problems into dual ones:

\[ \left\{ \begin{array}{l} \max \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) K(x_i, x_j) \\ s.t \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \leq \alpha_i, \alpha_i^* \leq c \end{array} \right. \]  

Where, both \( \alpha_i \) and \( \alpha_i^* \) are introduced Lagrange multipliers. Finally, according to the above formula, the SVM model can be obtained as follows:

\[ f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(x_i, x_j) + b \]  
\[ k(x_i, x_j) \] is a kernel function, generally a Sine function or a Sigmoidal function and a Radial basis function. In this paper, the radial basis function is used as the kernel function:

\[ k(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\sigma^2) \]  

**PSO**

In the PSO algorithm, each particle updates its speed and position according to the individual optimal particle \( P_i \) and the group optimal particle \( g \). The velocity and position update formulas are as follows:

\[ v_{i,d}(t+1) = \omega v_{i,d}(t) + c_1 r_1 \left[p_{i,d}(t) - x_{i,d}(t)\right] + c_2 r_2 \left[p_{g,d}(t) - x_{i,d}(t)\right] \]  
\[ x_{i,d}(t+1) = x_{i,d}(t) + v_{i,d}(t+1) \]  

Where, \( t \) represents the current number of iterations; \( v \) and \( x \) represent the velocity and position of a particle, respectively. \( \omega \) is the inertial weight, which reflects the relationship between the current
particle and the velocity of the previous generation of particles. \( c_1 \) and \( c_2 \) are accelerating factors, the former reflects local information exchange, the latter reflects global information exchange, \( r_1 \) and \( r_2 \) are random numbers between \([0,1]\).

In this paper, PSO is used to optimize the penalty parameter \( c \) of SVM and the parameter \( \sigma \) of kernel function. The evaluation index in Section 4.3 is taken as the objective function of PSO. The specific process is shown in the framework diagram of Section II.

**Case Studies**

In this paper, the actual data of a city's distribution network are selected for the experiment of distribution network fault prediction. The distribution network has 84 feeders. Select the data needed for fault prediction of distribution network from January 2012 to December 2015 from production management system, distribution geographic information system and intelligent public distribution transformer monitoring system.

The data processing method proposed in this paper is used to process the original data, and then feature vector selection method is used to filter the feature vectors. The results are shown in the following table:

| Lable                  | Feature selection results                                      |
|------------------------|----------------------------------------------------------------|
| **Operational factors**| Average monthly load of feeders                               |
|                        | Monthly maximum load of feeder                                 |
|                        | Average commissioning time of load switch                      |
|                        | Average commissioning time of fuses                           |
|                        | Average commissioning time of segmented cables                 |
|                        | Time of fault                                                  |
|                        | Number of failures in the current month                         |
|                        | Number of failures in the previous month                       |
|                        | Number of failures in the first two months                     |
|                        | Total fault load                                               |
|                        | Substation where feeder is located                              |
| **Equipment factors**  | Number of transformers                                          |
|                        | Length of sectional insulated conductor                         |
|                        | Number of fuses                                                 |
| **External factors**   | Monthly mean temperature                                        |
|                        | Monthly maximum temperature                                     |
|                        | Monthly minimum temperature                                     |
|                        | Monthly average precipitation                                   |
|                        | Classification of Monthly Thunderstorm Days                    |
|                        | Classification of Monthly Gale Days                             |

From the results, it can be seen that after feature selection, the total number of input features required by the model is 20, which are strongly correlated with distribution network faults. In the selected distribution network area of this paper, the annual summer is the high incidence period of distribution network faults. Select May 2013 to May 2014 as training samples, June 2014 as test samples, July as prediction samples. The fault level distribution in each sample is shown in the following table:
Table 2. Number of fault levels in each sample.

| Type of sample   | Number of Level 1 | Number of Level 2 | Number of Level 3 |
|------------------|-------------------|-------------------|-------------------|
| training sample  | 874               | 263               | 48                |
| Test samples     | 75                | 10                | 2                 |
| Prediction sample| 89                | 17                | 4                 |

In order to verify the effectiveness of the proposed method, the prediction model based on artificial neural network and the prediction model based on C 4.5 decision tree are used as comparison methods, which are recorded as model 1 and model 2 respectively, and the method presented in this paper is recorded as model 3. The predicted results of each model are as follows:

**Figure 2. Prediction results of different models.**

The vertical coordinates in the figure indicate the prediction accuracy. From the forecasting results, it can be seen that there are obvious differences in the forecasting results for three different grades of faults. The one-level fault occurrence frequency is the highest, and the prediction results of three different models are better, among which the prediction accuracy of the method proposed in this paper is the highest. Secondly, the two-level fault has a high accuracy. The lowest prediction accuracy is three-level fault. In the forecasting sample, the number of three-level faults occurs four times, and the number of times is less. In addition, there are many factors inducing three-level faults, which makes the forecasting more difficult.

**Conclusion**

A novel fault prediction method for distribution network is presented in this paper. On the basis of fault classification, a fault prediction model is built by using SVM and PSO. The experimental results show that: (1) Compared with the other two existing distribution network prediction models, the proposed model has better performance. (2) For one-level and two-level faults, the proposed method has a fairly high prediction accuracy, while the prediction accuracy of three-level faults is slightly lower.

**Acknowledgement**

This work was supported by Science and Technology Project of State Grid Corporation of China (Research and Application on Power Grid Planning Data Automatic Collection and Intelligent
References

[1] Cheng Pengfei, He Jiang, Ma Yan. Prediction method for reliability underlying parameters of 10 kV city distribution network [J]. Guangdong Electric Power, 2014, 7(08): 94-97 +101.

[2] M. T. Veljko, R. T. Predrag, and M. D. Zeljko, “Expert system for fault detection and isolation of coal-shortage in thermal power plants,” in Proc. Conf. Control Fault-Tolerant Syst., Nice, France, Oct. 2010, pp. 666–671.

[3] C. Nan, F. Khan, and M. T. Iqbal, “Abnormal process condition prediction (fault diagnosis) using G2 expert system,” in Proc. Can. Conf. Elect. Comput. Eng., Vancouver, BC, Canada, Apr. 2007, pp. 1507–1510.

[4] Z. Yongli, H. Limin, and L. Jinling, “Bayesian networks-based approach for power systems fault diagnosis,” IEEE Trans. Power Del., vol. 21, no. 2, pp. 634–639, Apr. 2000.

[5] Y. Brandt, B. W. Jervis, and Y. Maiden, “Circuit multi-fault diagnosis and prediction error estimation using a committee of Bayesian neural networks,” in Proc. IEE Colloq. Test. Mixed Signal Circuits Syst., London, U.K., Oct. 1997, pp. 1–7.

[6] Z. Jiang, Z. Li, N. Wu, and M. Zhou, “A Petri net approach to fault diagnosis and restoration for power transmission systems to avoid the output interruption of substations,” IEEE Syst. J., vol. 12, no. 3, pp. 2566–2576, Jul. 2018.

[7] S. K. Yang, “A condition-based failure-prediction and processing-scheme for preventive maintenance,” IEEE Trans. Rel., vol. 52, no. 3, pp. 373–383, Sep. 2003.

[8] X. Xu and J. F. Peters, “Rough set methods in power system fault classification,” in Proc. Can. Conf. Elect. Comput. Eng., Winnipeg, MB, Canada, May 2002, pp. 100–105.

[9] W. Li, A. Monti, and F. Ponci, “Fault detection and classification in medium voltage DC shipboard power systems with wavelets and artificial neural networks,” IEEE Trans. Instrum. Meas., vol. 63, no. 11, pp. 2651–2665, Nov. 2014.

[10] S. Bhattacharya, “Fault detection on a ring-main type power system network using artificial neural network and wavelet entropy method,” in Proc. Int. Conf. Comput., Commun. Autom., Noida, India, May 2015, pp. 1032–1037.

[11] D. Thukaram, H. P. Khincha and H. P. Vijaynarasimha, “Artificial neural network and support vector Machine approach for locating faults in radial distribution systems,” in IEEE Transactions on Power Delivery, vol. 20, no. 2, pp. 710-721, April 2005.

[12] Wei Na, Wang Tao, Research into the feature selection method by combining Relief F and support vector machine [J]. Computer Applications and Software, 2008, 25(1): 283-285.

[13] Ye Shengyong. Study on power systems transient stability assessment based on machine learning method [D]. Chengdu: Southwest Jiaotong University, 2010.