Short-term Prediction of Distribution Network Faults Based on Support Vector Machine

Yuling Bai[1], Yunhua Li[2]
School of Automation Science and Electrical Engineering
Beihang University
Beijing, China
1:baiyuling@buaa.edu.cn,2:yhli@buaa.edu.cn

Yongmei Liu[3], Zhao Ma[4]
Power Distribution Research Department
China Electric Power Research Institute
Beijing, China
3:liukeyan@epri.sgcc.com.cn, 4: ma_zhao@hotmail.co.uk

***Abstract*— As the network end of power transmission, the distribution network (DN) directly determines the reliability of electricity energy supply. To predict failure accurately is important for increasing the repair efficiency of DN. Based on the failure data from DN in Beijing, the paper researches short-term DN failures prediction and proposes a fault judgment program based on weather and season factors. Failure is analyzed to determine its most important factors. Through support vector machine (SVM) algorithm and considering the relative meteorological factors, using the classification model predicts the number of failures in DN weekly, and establishes sub region classification forecasting model in week frequency with meteorological influence for DN failures prediction. Through the analysis for the number of DN failures data, we find the main influence factors are temperature, precipitation, wind and other meteorological factors. A short-term prediction program is tested lots of times with the data of DN failure. The practical data in Tongzhou district, Beijing, China, proved the effectiveness, precision and feasibility of the proposed method. The paper software used Matlab2014 and LIBSVM.**

***Keywords*— distribution network, fault classification, support vector machine, classification prediction.**

I. INTRODUCTION

With the development of industrialization, people tend to be more relied on power. As the power grid scale is expanding, the requirements of safety and stability to the power grid are getting higher. The rapid development of DN construction leads to its high-level scale. However DN load with the rapid development cannot meet the high reliability of power supply requirements. At present, the level of Chinese medium and low voltage DN automation is far less than those of the developed countries[1]. As the terminal network of the power transmission, DN straightly determines the electricity reliability and affects the social, economic, life and so on. Affected by various threats of the DN safety, its structure is relatively week especially under the poor weather condition, such as thunderstorms, winds and so on, which frequently rises the DN faults[2]. It is of great significance to improve the fault repair efficiency and reduce the economic loss caused by the DN fault under the means of arranging advanced specific examination, solving hidden faults, shortening the repair time and rational allocation of resources. Accurate and fast fault processing can not only shorten the repair time and reduce the work intensity of inspection line, but also help to deal with failure in time, reduce the damage to DN equipment during fault operation, improve the DN safety level and reduce the national economic losses finally. DN fault handling level is also the logo reflecting the comprehensive strength of regional power grid automation, operation management and research application and many other aspects.

In the early 21st century, America and European countries have begun the research and construction of smart grid. Through the realization of smart grid technology, Europe can reduce 15% of emissions in the next 12 years[3]. In 2010, Chinese government work report proposed 'Strengthen the Smart Grid Construction', which indicated that the smart grid construction had formed a consensus in our country. At present, many technologies in the DN fault are mainly focused on the fault diagnosis, fault location, grid reliability assessment and fault recovery, which are less used in prediction of DN fault number[4] [5] [6]. SVM methods have been applied to faults location and wind power prediction in DN[7] [8]. Although DN fault types are complex and plentiful, most of them happened commonly.

In this paper, the SVM classification prediction model is established and the short-term fault number of DN is forecasted by studying the influence of weather meteorological factors on the DN fault number. Based on the DN fault data recorded in Beijing, a classification forecasting model of meteorological and seasonal factors affecting the failure number is established by support vector machine multi-classification method. The factor that had the greatest impact on the failure number was determined and that of the least was deleted by analysis. And the classification prediction model of DN short-term failure frequency is established and the prediction of the DN fault number with high precision is realized.

II. DISTRIBUTION NETWORK FAULT ANALYSIS

A. Analysis of DN Fault Data

It is very important to predict the number of faults in the next period, shorten the troubleshooting time, reduce the economic loss caused by the fault of the DN and improve the

This paper was supported by State Grid Corporation of China Research Program (Grant Nos. EPRIPDKJ (2014)2863, PD-71-15-042).

978-1-5090-6161-7/17/$31.00 ©2017 IEEE

repair efficiency of DN. The reasons for the failure of DN can be divided into two categories: one is the external environmental factors. Weather factors such as low temperature freezing, heavy rain and wind, and seasonal factors such as tree growth, which led to the tree line faults; the other is the quality of the equipment itself, equipment aging and operation and maintenance issues.

The number of DN failure is about 3300 times in Beijing in 2015. The analysis found that it was mainly divided into 8 types of failures, where user impact and external force are the main fault types. The failure of users and DN operators has little relationship, so follow-up research is no longer paying too much attention on it. These 8 fault types accounted for 67% of the total faults, as shown in Fig.1. In these 8 fault types, what related to weather factors accounted for 63%, and what related to season accounted for 8%. The analysis shows that external environmental factors is the main cause of DN faults.

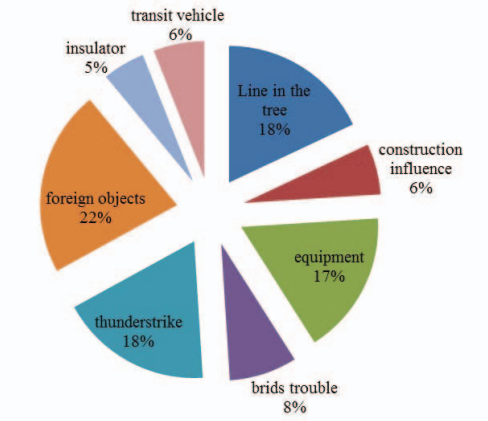

Fig. 1. Exampl percentage analysis of distribution category in DN.

The analysis of the faults data found that: Meteorological factors have the greatest impact and it is quantified. The faults that occurred by thunderstorms, precipitation and extreme weather are more prominent. The faults occurred by equipment, tree growth and bird damage increased in rain days, and tree line faults also increased in windy days. It can be seen that climate change also affects the occurrence of other types faults.

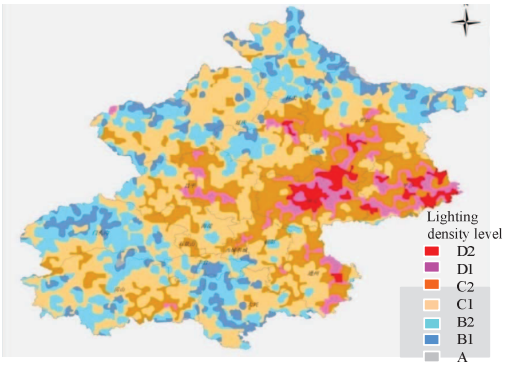

Fig. 2. Lightning density distribution and lightning failure

Some failures occur cyclically. Most of the bird faults occur in March, April and May. Most of the lightning strike faults are concentrated in May, June, July, August and September, especially in July and August. Figure 2 shows the year lightning density map in Beijing, lightning density is significantly different in different areas, the deeper the color in Fig.2 indicates the greater the lightning density. Lightning distribution in different regions are different, so it is also different that DN faults density resulting from lightning.

Figure 3 indicates that the fault number exists obvious seasonal differences, more thunderstorm season. There are more faults emerged in thunderstorm season.

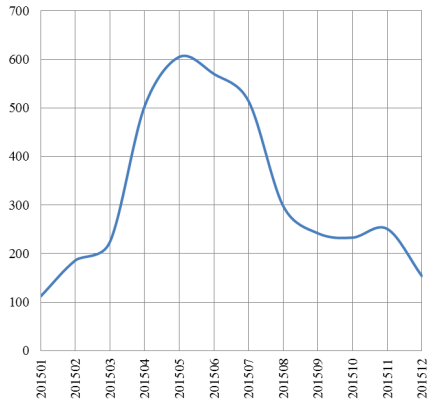

Fig. 3. The trend of the failure occurred in different months

Figure 4 is Comparison of the number of failures in different years and different regions. In Fig.4, the left side of each group is the DN faults number in 2014, the right side of each group is the DN faults number in 2015. Contrast the two sets of data，most of the faults number in each group in 2014 is larger than the same group in 2015 except Chengqu, Changping and Pinggu, which are similar.

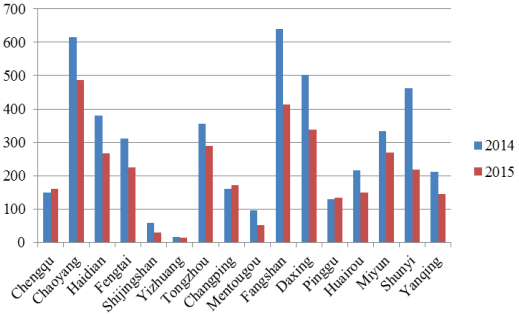

Fig. 4. Comparison of the number of failures in different years and different regions

Figure 5 shows the comparative analysis of the DN fault types in Huairou and Ghengqu.There are different main reasons for the occurrence of faults in different area，so it is necessary to set up different models for different regions.

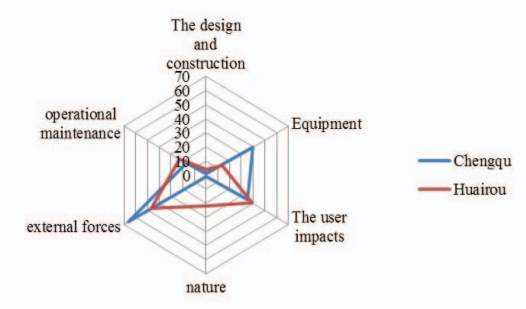

Fig. 5. Comparison of fault distribution in different areas

Based on the above analysis, it can be seen that different regions need to establish a separate model for analysis and forecast. Figure 6 shows the number of permanent DN failures in 100 km. Combining with the number of annual failures in each region of Fig.4, Tongzhou and Chaoyang which have high failure is selected finally as examples for classification prediction.

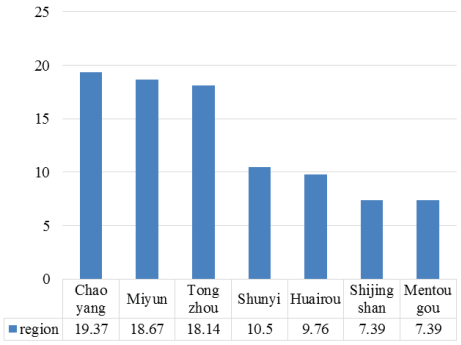

Fig. 6. Number of permanent failures in 100 km range in different areas

In addition, analysis found that the insulator failure rate is high in external factors, and P15, P20 needle insulators are more prone to failure than other insulators. Insulator failure more occurred in July and August, so it is recommended that reserves more insulator components in this time for repair.

B. *Analysis of Correlation of DN Faults Number and Meteorological*

The Bayesian classifier is used to determine the specific cause of faults by the given weather, time and the performance of the fault. The analysis used 906 data in the r language. The accuracy of the cross validation test is about 67%, which means that: 1) The frequency of faults occurrence and the climate has a strong relationship. 2) Faults occur with seasonal, such as bird damage, lightning failure.

III. SHORT-TERM PREDICTION METHOD FOR DN FAULTS

In order to predict the short - term DN faults accurately to improve the efficiency of DN maintenance, it is necessary to establish a short - term faults classification forecasting model. SVM method is selected to establish the classification forecasting model through the comparison of methods. Through the analysis of fault data and influencing factors, the input and output of this model are determined.

There are many kinds of prediction algorithms, which are time series, neural network and support vector machine widely used. The time series forecasting method is ideal for short-term prediction with relatively uniform sequence variation. Its advantage is that its historical data and workload requirement are small, but time series is less accurate in nonlinear prediction. The neural network method has been widely used in the nonlinear field, but this method is easy to fall into the local optimum and its convergence speed is slow. The support vector machine prediction method solves the problem which is easy to fall into the local minimum point. SVM is characterized by small training sample data, strong generalization ability and high prediction accuracy.

A. *The Working Principle of SVM*

The main idea of SVM is to establish a classification hyper-plane as decision surface, so that the isolation edge between positive and negative cases is maximized. It can be solved on SVM that classification(c-SVC, n-SVC), regression (e-SVR, n-SVR) and distribution estimation (one-class-SVM). SVM can effectively solve many kinds of problems: cross-validate the selection parameters, weight the unbalanced samples and Probability estimates of class problems. DN faults data is not large, but its dimension is higher, which is the advantages of SVM[10]. Therefore, this paper presents a DN faults prediction method based on LIBSVM SVM.

The SVM method is based on the VC theory and structural risk minimization principle of Statistical Learning Theory (SLT). According to the limited sample information, this method seeks the best compromise between the training sample learning accuracy and learning ability in order to obtain the best generalized ability.

VC dimension is a measure of the function class, which also means the complexity of the problem. The higher the VC dimension, the more complex a problem. It can be seen that SVM solving problem has nothing to do with dimension of the sample, which makes SVM very suitable to solve classification problem.

Machine learning is essentially the approximation of the real problem. Risk is the accumulation of the error between the hypothesis and the real solution, so the empirical risk(w) is the difference between the classification result of the classifier based on the sample data and the real result. Previous machine learning methods regard minimized empirical risk as the goal, but it cannot guarantee has no error in the real text.

Therefore, the concept of structural risk minimization is introduced and generalization error bounds is introduced. As formula (1) shows, it means that the real risk should be characterized by two parts. One is the empirical risk which represents the error of the classifier in the given sample. The other is the confidence risk which Represents how much we can trust the classifiers classification in unknown text. Confidence risk is related to two quantities. One is the number of samples. The greater the number of samples, the more accurately the result, the less confidence risk. The other is VC

dimension of the categorical function. Its impact on real risks is instead of sample number.

$$R(\mathrm{w}) = \mathrm{Re}\,mp(w) + \phi(n/h) \tag{1}$$

R (w) is the real risk, Remp (w) is the empirical risk, and Φ(n/h) is the confidence risk. The goal of statistical learning is to minimize the sum of experience risk and confidence risk.

SVM prediction is divided into classification prediction and regression prediction. As a very important task in data mining, the purpose of classification is to learn a classification function or a classification model that can map data items in a database to one of a given category. At present, the judgment method of classification research results can be carried out from three aspects[11]:

- Prediction accuracy (accuracy of non-sample data).
- Computational complexity (time and space complexity when the method is implemented).
- Pattern simplicity (it is desirable that the decision tree is small or the rules are small in the same effect case).

B. Distribution Process of Short - term DN Faults Number

The external environment factors caused by seasonal changes in weather are easier to find quantitative data , but equipment problems due to equipment type, the beginning of the use, different performance parameters whose data records are difficult to quantify to input into the classification forecasting model. The faults caused by weather and season nearly 60%, so classification forecast model uses weather and season data as the input, fault categories as the output.

1) Model establishment

SVM model establishment contains data extraction, calibration, data pretreatment and training. The training sets is used to predict the classification of the test sets. The algorithm flow is shown in Fig.7.

Fig. 7. Comparison of fault distribution in different areas

According to the characteristics and DN failure experience, model initially select temperature (minimum temperature, maximum temperature and daily average temperature) and precipitation (thunderstorms, no rain, light rain, showers, heavy rain, heavy rain, Middle and high) as the influencing factors to analyze the correlation with the occurrence of failure.

The number of faults in DN has seasonal rules. The influence of the temperature on the number of distribution faults in different seasons is different. Season can be judged from temperature: The 0 following is winter, the highest temperature for a week above 35°C is summer and the middle is spring and autumn. In winter, the temperature factors are recorded in the model at the lowest temperature. In summary, the temperature factors are influenced by the highest temperature. In the spring and autumn, the temperature factors are influenced by the average temperature.

From Table 1, we can see that the correlation coefficient of the influence of different season temperature on the number of faults is above 0.5, the correlation is high. Summer is positively correlated, and the other season is negatively correlated.

TABLE I. THE CORRELATION COEFFICIENT OF THE EFFECT OF TEMPERATURE ON THE NUMBER OF FAULTS

Temperature	Correlation coefficient
The lowest temperature	-0.62
The highest temperature	0.72
The average temperature	-0.54

Experiments show that rainfall has a great influence on the fault. The final model is temperature, rainfall, wind and extreme weather as the independent variables and the short-term DN faults number as the dependent variable and used the radial basis function(RBF) to analyze and forecast. The extreme weather in the independent variable is expressed in the form of a dummy variable:

$$X_{\text{extrewether}} = \begin{cases} 0, & \text{not extrewether} \\ 1, & \text{rainstrom and so on} \end{cases} \tag{2}$$

Figure 8 shows the graphical representation of 52 groups data. The first is classification labels of DN fault. Zero represents fault days is small. One represents fault days is in the frequent range. Two represents the faults frequent occurrence. Attrib1 to attrib11 is the data distribution of influencing factors(thunder shower, rainstorm, days whose wind is greater than 4, days whose minimum temperature is less than 0 or maximum temperature is greater than 35°C, daily maximum temperature, daily minimum temperature, days whose wind less than or equal to 3, intermediate wind power days, small precipitation days, medium precipitation days and large precipitation days successively).

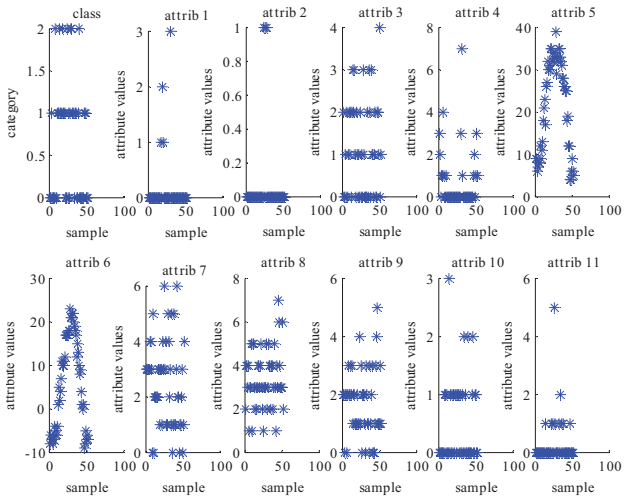

Fig. 8. Factors of DN faults

2) Selection of Model Kernel Function

SVM uses kernel functions to map data to high-dimensional space, which makes it linearly separable in high dimensional space for low dimensional linear inseparable problem. In generally, the inner product typed of nonlinear transformation is chosen as kernel function.

The selection of kernel function determines SVM performance. The radial basis function(RBF) is the most widely used kernel function in classification prediction. In addition to RBF kernel function, the other kinds of kernel functions are polynomial function and hyperbolic tangent function in neural network, and their expressions are as follows:

$$K(X_i, X_j) = \exp\left(-\gamma \left\| X_i - X_j \right\|\right)^2 \tag{3}$$

$$K(X_i, X_j) = (\left\langle X_i, X_j \right\rangle + 1)^d \tag{4}$$

$$K(X_i, X_j) = \tanh(\eta < X_i, X_j > + \theta) \tag{5}$$

In these formulas, X represents the independent variables, that is, the meteorological factors in this paper. Parameter γ the kernel parameter. Parameter d is the maximum times of the polynomial kernel function.

RBF(formula 3) is a real-valued function that only depends on the distance from the origin or it can be the distance to any point c, where point c is generally chosen as the center point. RBF kernel function can approximate any nonlinear function and deal with the system which is difficult to parse the regularity. RBF with good generalization ability and have a quick learning convergence speed. It can be seen as a high-dimensional space Surface fitting (approximation) problem. Learning is to find a multidimensional space to match the surface of the training data and then practice the surface to deal with a new classification data number.

RBF structure consists of input layer, hidden layer and output layer. Input layer links up to the external environment. Hidden layer conducts the non-linear transformation from the input space to the hidden space. Output layer is linear to provide input mode response for the input layer.

$$\begin{cases} \min\limits_{a} \dfrac{1}{2} \sum\limits_{i,j=1}^{n} y_i y_j a_i a_j K(X_i, X_j) - \sum\limits_{i=1}^{n} a_i \\ s.t. \sum\limits_{i=1}^{n} a_i y_i = 0,\ 0 \le a_i \le C \end{cases} \tag{6}$$

The essence of SVM classification is to construct the optimization plane by formula (6). X is the input vector. Y is the output vector. Parameter a_i is a Lagrangian multiplier which is used to restrain function and associated with the original function. It is used to equal to the equation by matching the number of variables, so as to find the extreme value of the original function. $K(X_i, X_j)$ is RBF in this paper. The role of C is to adjust the range of the confidence interval of the learning machine in the determined data subspace. There are two important parameters: the penalty factor C and the kernel parameter γ. The optimal parameter group C and γ determines the accuracy of the SVM classifier. The optimized parameter C is changed in the different data subspaces and the kernel parameter γ is to change the mapping function, that is, to change the subspace distribution complexity of the sample data. The maximum VC dimension of the linear classifier determines the minimum error of linear classification[12]. In order to find the optimal parameter group C and γ, the exhaustive method is used to select the parameters.

IV. Short - term Faults Prediction and Result Analysis

A. Classification and Prediction Experimental Results

The number forecast of short-term DN faults is predicted by 52 weeks data in 2015. 7 groups to 9 groups are randomly selected as the test sets and the remaining 45 groups to 43 groups data are training sets. In this section, the blue star marks are the model predictions and the red round labels are true.

1) Two categories

The predictions of short-term faults is divided into 2 categories (1 time to 9 times, 10 times and more). As the fig.9 (a) shows, the fifth group's actual classification is same as prediction results, that is, the fifth week's actual faults is more than 10 times, and its predicted faults is also more than 10 times. Accuracy = 100% (9/9).

2) Three categories

The predictions of short-term faults is divided into three categories (5 times, 6 times to10 times , 10 times and more). As the fig.9 (c) shows, accuracy is 71.4286% (5/7).

3) Multiple categories

The predictions of short-term failures is divided into five categories (no fault, 5 times, 6 times to10 times, 11 times to 15 times and 16 times to 20 times). The kernel function is RBF. As the fig.9 (b) shows, accuracy = 66.6667% (6/9). The multi-classification accuracy is too low to apply.

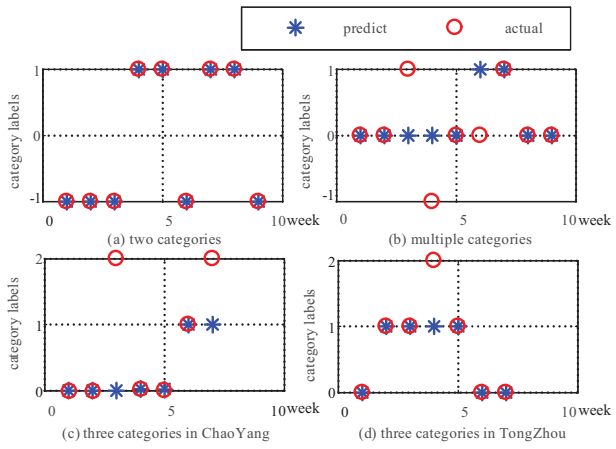

Fig. 9. The results of categories

Figure 9 (a), (c) and (b) show the results of two categories, three categories and multiple categories in Chaoyang. Based on the theory of multi-classification accuracy is lower than three categories and two categories accuracy is higher than Three categories but its practicality is very low in short-term problems prediction. The result is more useful as far as possible to use less classification when the actual requirements

can be meet. In order to verify this conclusion, this SVM classification prediction model is used in Tongzhou. As the fig.9 (d) shows, accuracy in Tongzhou is 85.7143% (6/7). It can be seen that as long as the actual needs are met, the three categories accuracy is sufficient to be applied to reality.

B. *Analysis of results*

All experiments above used RBF kernel function. In the choice of experimental kernel function, we can see that the RBF kernel function's accuracy is higher than the polynomial kernel function. It is more suitable for this kind of problem, and follows the theoretical derivation of kernel function selection in the third chapter.

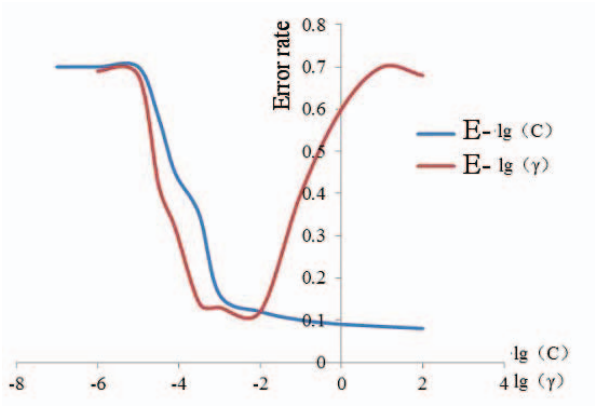

Fig. 10. The error rate varies with C and γ

Figure 10 shows the effect of the parameters γ and C on SVM. When the C is small, the estimated value of the error rate is relatively high; when C increases, the error rate is rapidly reduced, i.e., the performance is rapidly improved. After a period of time, the error rate no longer changes, that is, the change of C almost does not affect the ability of SVM promotion at this time, so the change of nuclear parameters can get optimization value in this area. Parameter γ change process of error rate is from large to small, and then from small to large, that is, there is a certain γ can get the optimal value of SVM. In the selection of model parameter, it is found that the best effect is lg(C) values between 1-2 and lg(γ) values -2.8.

V. CONCLUSION

The Bayesian model is used to analyze the correlation between factors and DN faults data. It is found that the main factors affecting DN steps are extreme weather, precipitation and other meteorological factors. After the chart analysis, the conclusion is further verified, laid the foundation For the late experimental study.

This is the use of the meteorological factors as independent variables to predict DN faults, and has high operability, adaptability and the reliability. The overall levels assessment accuracy is over 70%. The results are more credible as far as possible to use less classification when the prediction accuracy can meet the actual requirements. However, the number of DN failure factors are numerous, and many factors cannot Be quantified or data cannot be made at present, so the forecast accuracy is difficult to reach 100%. However, with the increase of data retention and data collection, the model can add more failure factors, then the prediction accuracy will be improved, which will guide DN repair work better and better.

REFERENCES

[1] L. Min, C. Li, Z. Kai and D. Qiushi, "Based on Rough Set and Support Vector Machine (SVM) in Jilin Province Power Distribution Network Transformation Project Evaluation," 2013 12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science, Kingston upon Thames, Surrey, UK, 2013, pp. 202-206.

[2] X. Peng, D. Deng, J. Wen, L. Xiong, S. Feng and B. Wang, "A very short term wind power forecasting approach based on numerical weather prediction and error correction method," 2016 China International Conference on Electricity Distribution (CICED), Xi'an, 2016, pp. 1-4.

[3] N. Shahid, S. A. Aleem, I. H. Naqvi and N. Zaffar, "Support Vector Machine based fault detection & classification in smart grids," 2012 IEEE Globecom Workshops, Anaheim, CA, 2012, pp. 1526-1531.

[4] P. Ray, D. P. Mishra and D. D. Panda, "Hybrid technique for fault location of a distribution line," 2015 Annual IEEE India Conference (INDICON), New Delhi, 2015, pp. 1-6.

[5] H. Livani, C. Y. Evrenosoglu and V. A. Centeno, "A machine learning-based faulty line identification for smart distribution network," 2013 North American Power Symposium (NAPS), Manhattan, KS, 2013, pp. 1-5.

[6] R. Bhat and A. P. Meliopoulos, "Probability of distribution network pole failures under extreme weather conditions," 2016 Clemson University Power Systems Conference (PSC), Clemson, SC, 2016, pp. 1-6.

[7] G. X. Yuan, C. H. Ho and C. J. Lin, "Recent Advances of Large-Scale Linear Classification," in Proceedings of the IEEE, vol. 100, no. 9, pp. 2584-2603, Sept. 2012.

[8] 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.

[9] Y. K. Gu and Z. X. Yang, "Reliability analysis of multi-state systems based on Bayesian network," 2013 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE), Chengdu, 2013, pp. 332-336.

[10] Shuirong Liao, Rencheng Zhang, Yijian Huang and He Xia, "Research of low-voltage arc fault classification based on support vector machine," International Conference on Automatic Control and Artificial Intelligence (ACAI 2012), Xiamen, 2012, pp. 1690-1693.

[11] O. A. S. Youssef, "An optimised fault classification technique based on Support-Vector-Machines," 2009 IEEE/PES Power Systems Conference and Exposition, Seattle, WA, 2009, pp. 1-8.

[12] H. Z. Li, S. X. Chen, T. Qian, W. H. Tang and Q. H. Wu, "Power transformer fault classification by combining genetic reduction with optimized multilayer support vector machine," 2015 IEEE Innovative Smart Grid Technologies - Asia (ISGT ASIA), Bangkok, 2015, pp. 1-5.

[13] S. Yin, Chen Jing, Jian Hou, O. Kaynak and H. Gao, "PCA and KPCA integrated Support Vector Machine for multi-fault classification," IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, 2016, pp. 7215-7220.

[14] Luger, F. George, and W. A. Stubblefield. Artificial intelligence: structures and strategies for complex problem solving. 3rd ed. DBLP, 1993.

[15] Feng Shi and Hui Wang. MATLAB Intelligent Algorithm 30 case analysis [M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 2011: 96-100.