The Risk Quantification Evaluation Strategy for the Distribution Line based on EMLR

Tang Xuesong¹, Wu Bin¹, Yu Guangming¹, Fan Fulai¹, Qu Jinyu¹, Du Hongjiang¹, Wang Yunhan^{2,*}

¹State Grid Chongqing Electric Power Company Tongliang Power Supply Branch, Tongliang District, Chongqing, China

²Chongqing University of Posts and Telecommunications, Nanan District, Chongqing, China

Abstract. The electric power distribution grid directly serves the ordinary users, the stability and reliability of which are crucial to production and life. The traditional operation and maintenance of the distribution grid is based on experience and cannot scientifically predict the state of line faults. In this paper, stepwise regression is used to screen out 8 key influencing factors of line faults, and principal component analysis is performed to find the risk value of each line; so that establishes multiple linear and exponential regression equations for the line risk value. According to the existing data, the risk of the line is predicted, and the GA-BP neural network model is built so as to accurately predict the risk value of the line.

1 Introduction

In the Chinese electric power grid, the scheme overhaul mode is always adopted for the maintenance and management of power infrastructure. The overall situation is a post-maintenance system. Once the overhaul is started, it will be unable to supply energy for a long time, and will seriously affect social life. Literature [1] proposes a distribution network fault monitoring method based on big data analysis, using C-means fuzzy clustering algorithm to identify distribution line load patterns and perform load forecasting. Literature [2] introduces GIS technology to locate each tower and distribution transformer position in real time, so as to escort the normal operation of 10kV distribution network lines. Literature [3] analyzes the source and type of “dirty data” based on the characteristics of big data from distribution and consumption, so that proposes a corresponding data cleaning method. The actual calculation results verify the effectiveness and accuracy of the prediction. Literature [4] mainly analyzes and summarizes the current research status of machine learning algorithms used to process big data, including the problems faced by machine learning research institutes in the big data environment. Finally, it points out the research trend of big data machine learning.

In summary, at present, there are few studies on predicting electric power distribution grid line faults based on big data technology, so that there is a lack of universal line fault predicting models. Therefore, in this paper the data mining is analyzed to extract the key factors that may lead to line failures. In addition, a distribution grid line failure prediction model is established through standardized processing, principal component analysis and a combination of multiple linear and non-linear regression algorithms.

2 Calculate risk value of the route

According to the collected failure data on the 10kV distribution line from a power company within two years, 14 factors are obtained, which seem to lead to the fault of distribution grid line in this paper.

2.1 Failure cause factors

Figure 1. Fault Cause Factor and Explanation

Standardized stepwise regression is used to screen out the factors. After standardizing independent variables and dependent variables, stepwise regression is used to calculate the regression coefficients and their significance of each independent variable so as to eliminate variables with non-significant coefficients. The independent variable coefficient obtained after regression is called the standardized regression

* Corresponding author: longhy@cqupt.edu.cn

coefficient, and its absolute value can represent the importance of the independent variable.

With the collected data, the fourteen factors are taken as independent variables and the number of line failures are as dependent variables. Then z-score standardization is carried out for independent variables and dependent variables respectively, so that the coefficient is achieved through stepwise regression. The standardized formula is as follows:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

$$\mu = \sum_{i=1}^N x_i, \quad \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^N (x_i - \mu)^2} \quad (2)$$

It is not all of the 14 factors that has a significant influence on the occurrence of failure. Only eight factors such as the number of defects, passes the significance test. Although other factors may also have a certain influence on the occurrence of the fault, the influence of those factors is so small that will be ignored. The key factors are obtained after those factors with less influence are removed.

By sorting the key factors according to the absolute value of their standardized regression coefficients, the importance of the eight key factors can be achieved. The number of defects has the greatest influence on the number of faults and the insulation rate has the least influence on the number of faults. All of the other factors have a positive correlation with the number of line failures except that the operating years has a negative correlation with the number of line failures.

2.2 Principal component analysis method

Assuming that there are n samples and p indicators, then an $n \times p$ sample matrix x can be formed.

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \cdots, x_p) \quad (3)$$

Constructing the correlation coefficient of the x matrix as R .

$$R = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}} \quad (4)$$

Through solving the equation $|R - \lambda I_p| = 0$, which is the characteristic equation of matrix R , it can be obtained that p eigenvalues λ_i and feature vectors a_i . For each principal component, it can be expressed as a linear combination of the original p indicators, and its expression is as follows.

$$Z_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p, (i = 1, 2, \cdots, p)$$

(5)

For each principal component, its contribution rate is $\lambda_i / \sum_{k=1}^p \lambda_k$. According to the contribution rate of each principal component, the principal components can be sorted and some of them can be selected so as to achieve the purpose of dimensionality reduction.

Taking into account the difference in the dimensions of the data of each variable and the inconsistent units, the method of normalizing each variable is adopted to eliminate the influence of the dimension. The normalization formula is (6):

$$X_{norm} = \frac{X - X_{MIN}}{X_{MAX} - X_{MIN}} \quad (6)$$

After the data is normalized, the principal component analysis is performed in order to obtain the contribution rate of each principal component.

Table 1 The contribution rate of each principal component

	CO MP. 1	CO MP. 2	CO MP. 3	CO MP. 4	CO MP. 5	CO MP. 6	CO MP. 7	CO MP. 8
STANDARD DEVIATION	1.41 55	1.12 96	1.04 15	1.02 22	0.93 65	0.86 80	0.81 05	0.55 07
PROPORTION OF VARIANCE	0.25 04	0.15 95	0.13 56	0.13 06	0.10 96	0.09 42	0.08 21	0.03 79
CUMULATIVE PROPORTION	0.25 04	0.40 99	0.54 55	0.67 62	0.78 58	0.88 00	0.96 21	1

According to the calculations, the cumulative contribution rate of the first 7 principal components has reached 96.21%, which enable to reduce the calculation dimension, so that only the first 7 principal components are selected for subsequent calculations.

After calculating the value of each principal component, the contribution rate of each principal component is set to construct a comprehensive evaluation function so as to calculate the risk score of each route and obtain the comprehensive score of the principal component of each route line as **Table 2**.

$$score = \sum_{i=1}^7 \alpha_i Z_i \quad (7)$$

Table 2 Comprehensive principal component scores of 3 partial route lines

Line unit name	score
10kV Liangba Line925 (2segment)	1.4804
10kV Liangba Line925 (1segment)	0.9681
10kV Liangban Line915	2.4074
10kV Liangping line921	2.3284
10kV Liangtai line 911	1.8476
10kV Jijiu line913	0.6900
10kV Cuiying 6 branch line of the old line	-0.3679
10kV Jijiu Line Iron and Stone 5 Branch Line	-0.4243
10kV Jijiu line and fish splashes 1 branch line	-0.1802
10kV Liangba Line Denggao 3 Branch Branch	-0.3827

Each variable in the principal component analysis is a positive indicator, that is the larger the variable value is, the greater the risk of the route. The risk value *score* of the route is calculated through the above steps, and it can also be considered that the larger the value is the risk of the route is greater.

3 Line risk model based on EMLR algorithm

3.1 EMLR algorithm

EMLR algorithm (Exponent & Multiple Linear Regression Algorithm), means: in the process of learning the training set through machine learning, adjusting each coefficient in the objective function composed of exponent function and multiple linear function. The expression of the objective functions is (8):

$$f(X) = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_i x_i + e^{\alpha_{i+1} x_{i+1}} + \dots + e^{\alpha_n x_n} \quad (8)$$

$X = (x_1, x_2, \dots, x_n)$ is a collection of all data instances. $(\alpha_1, \alpha_2, \dots, \alpha_i)$ is the coefficient that needs to be trained through machine learning, and $f(X)$ is the output value of the model. The process of machine learning is continuous iterative updates. The model's parameters are continuously updated so as to make the model output value as close as possible to the actual sample value. In general, the loss value is used to measure this closeness. The process of machine learning is of finding the $(\alpha_1, \alpha_2, \dots, \alpha_i)$ which minimizes the loss value. Through machine learning, the parameter values in the mathematical model can be calculated through the loss function defined as (9):

$$F_{cost} = \sum_{i=1}^N |f_i(X) - y_i| \quad (9)$$

Where, $f_i(X)$ is the output value calculated by the model, and y_i is the actual value of the training set. In this paper, the iterative method is adopted to minimize the value of F_{cost} .

3.2 EMLR simulation and result calculation

Since the risk score calculated by the principal component analysis method has a negative value, the following formula (10) has to be used in so that obtains the risk value. The risk_value is used as the actual value of the dependent variable in the regression equation.

$$risk_value = [score + \text{abs}(\min(score))] \times 10 \quad (10)$$

Taking eight key factors as the independent variables of the regression equation, and solving the linear partial coefficients $(\alpha_1, \alpha_2, \dots, \alpha_5)$, the expression of the risk value is obtained as (11):

$$risk_value = 23.1669x_1 + 5.3105x_2 + 11.0058x_3 + (11) \\ 1.5687x_4 - 3.7811x_5 + e^{x_6} + e^{x_7} + e^{x_8}$$

Where, x_1 is the unit line length, x_2 is insulation rate, x_3 is the number of defects, x_4 is the degree of thunder, x_5 is the operation years, x_6 is the overload, x_7 is the surrounding trees, x_8 indicates whether there are outside objects such as colored steel around.

The most important purpose of calculating the risk value expression is to divide the risk value into different intervals, and obtain the corresponding table of the risk value and the number of line faults. So a strong correlation is required between the risk value and the number of faults.

In order to quantify the correlation between the risk value and the number of faults, Spearman coefficient is introduced. The calculation formula of the two-variable Spearman coefficient is in (12):

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (12)$$

After calculation, the Spearman coefficient of the risk value and the number of line faults is 0.4069, and the correlation is weak. In the expression of the risk value, the coefficient $(\alpha_6, \alpha_7, \alpha_8)$ of the non-linear index part defaults to 1. In order to enhance the correlation between the risk value and the number of failures, it has to keep the coefficient of the linear part unchanged, and define a new loss function $f = \frac{1}{\rho}$ to

modify the coefficient $(\alpha_6, \alpha_7, \alpha_8)$. The expression of the risk value is in (13):

$$risk_value = 23.1669x_1 + 5.3105x_2 + 11.0058x_3 + \\ 1.5687x_4 - 3.7811x_5 + e^{4.1869x_6} + e^{-10x_7} + e^{4.6524x_8} \quad (13)$$

Draw a scatter plot of the number of line failures and the risk value calculated by expressions (11) and (13). It can be found that the corresponding risk values of different failure numbers are more obvious after the coefficient is updated.

Figure 2. The distribution of risk values before

Figure 3.after update of nonlinear coefficients

4 Conclusion

In this paper, firstly stepwise regression is performed to screen out the factors that affect line faults, so that obtains 8 key fault factors and their importance ranking; then, through the principal component analysis method, the principal component scores of each line are achieved and are regarded as the risk value of the line. Finally, 8 key failure factors are used as independent variables, and the risk value normalized by risk value score is used as the dependent variable. The stochastic gradient descent method is carried out so as to train the parameters of the multiple linear and exponential regression equations, and the risk value-based Quantitative expression. Therefore, the risk value of the line can be predicted scientifically based on the value of the key failure factor, and the operation and maintenance of the distribution grid line can be guided.

Acknowledgments

This work was supported by the State Grid Chongqing Electric Power Company & State Grid Chongqing Electric Power Company Tongliang Power Supply Branch, through the scientific project No.2020-2# & contract No. SGCQTL00SJJS2000193.

References

1. Jia Yujian, Sun Shumin, Meng Yu, Yin Maolin, Wang Yue. Distribution network fault monitoring method based on big data analysis[J]. Shandong Electric Power Technology, 2017, 44(10): 1-5.
2. He Feng. 10kV distribution network line fault factor analysis and preventive measures [J]. China New Technology and New Products, 2013 (22): 85-86.
3. Ding Xiao, Sun Hong, Zheng Haiyan, Ji Cong, Xu Jinling, Zhong Chunlin, Xiong Zheng. Short-term load forecasting based on big data of power distribution[J]. Electric Power Engineering Technology, 2018, 37(03): 21- 27.
4. He Qing, Li Ning, Luo Wenjuan, Shi Zhongzhi. Overview of machine learning algorithms under big

data[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(04): 327-336.

5. Hu Yishuang, Xia Xiang, Ding Yi, Fang Jianliang. Multiple regression load forecasting based on factor and trend analysis feedback[J]. Power Demand Side Management, 2018, 20(06): 22-25+35.