Correlation analysis and prediction of power network loss based on mutual information and artificial neural network

Jianghong Bai¹, Mu Jiang¹, Liping Liu², Yunchao Sun², Yuxing Wang³ and Jiaan Zhang³,⁴

¹State Grid Corporation of China, 100031 Beijing, China;
²China Electric Power Research Institute, 100192 Beijing, China;
³Hebei University of Technology, 300401 Tianjin, China.
⁴Email: zhangjiaan@foxmail.com

Abstract. The research and analysis of network loss is a hot spot in power system research in recent years. This paper firstly uses the principle of mutual information to reveal the influencing factors closely related to network loss, and ranks these related factors according to the degree of influence of network loss. Secondly, according to the analysis results of mutual information, the input of neural network training is optimized, and BP neural network is used to establish corresponding prediction models for 6 different types of network losses. Finally, the actual statistical data of a certain area of Jiangsu Power Grid is used for simulation analysis. The results show that the model established by the combination of mutual information method and BP neural network has better prediction effect on network loss and less error.

1. Introduction

As an important secondary energy source, reducing the energy loss of electric energy in the transportation process has far-reaching significance for the sustainable development of the national economy [1]. Today, big data mining and artificial intelligence technologies have a wide range of applications, which makes it possible to mine potential information related to network loss from the massive data of the grid and build a network loss model.

Many researchers have done a lot of work on correlation analysis methods and network loss modeling. In the literature [2], WAMS/SCADA correlation estimation is done with the derivation of Pearson correlation coefficient. And WAMS/SCADA data fusion method based on the correlation mining of time-series data is proposed. In the literature [3], a data mining framework in support of the operating experience feedback (OEF) system is proposed. The framework combines three statistical approaches (i.e., correlation analysis, cluster analysis and association rule mining) for identifying intrinsic correlations among human factors. In the literature [4], a method for analyzing data streams based on spatio-temporal correlation using different correlation algorithms is proposed, and a discussion about co-occurrence and causality is provided. In the literature [5], the mutual information method was used to analyze the corrosion grade of the pipeline and various influencing factors. In the literature [6], the electricity imports and the influence factors of grey correlation were calculated by using grey correlation analysis method. It makes a conclusion that every influence factor's effect on imports. The above analysis method based on the correlation of big data provides a broad idea for analyzing the correlation between network loss and various factors.
At present, many scholars have conducted research on the establishment of network loss models. In the literature [7], from the perspective of a single influencing factor, aiming at the relationship between smart distribution network load and network loss, the corresponding spatio-temporal analysis is carried out, and the characteristic parameters are proposed and the spatiotemporal feature analysis model of equivalent load is established. In the literature [8], a line loss rate estimation method of transformer district based on random forest algorithm is proposed involving multi-source data. Calculative results of Shanghai Electric Power Company verify that the proposed method is feasible and has superior performance compared with the estimation results of linear regression model and regression tree model. In the literature [9], a new method of neural network power loss estimation (NN-PLE) is proposed to estimate the real-time power loss of each line in the active power distribution system. The proposed method provides high-precision fast calculation. In the literature [10], using gray comprehensive relationship, the strong relation factors correlated with line loss are extracted, then a network line loss prediction model based on portfolio of GM(1,1) and GM(1,N) is built. The above articles are all based on the perspective of a single intelligent algorithm to establish a prediction model of network loss. For complex and large grids and data, the data mining method and intelligent algorithm can be combined to optimize the establishment process of the prediction model, which can improve accuracy of model prediction, however there is little research on this aspect.

Two works was done in this paper, on the one hand, the principle and method of mutual information entropy are used to analyze the correlation between network loss and various potential related factors, and determine the order in which these factors affect the strength of network loss. On the other hand, the mutual information analysis method is used to optimize the input data of artificial neural network, which can establish a prediction model of different types of network loss.

2. Algorithm principle and process
According to the “four points” principle of network loss[10], the type of network loss is first classified into the data set \( X = [x_1, x_2, \ldots, x_m] \), where \( X \) represents the matrix of the sample data of the relevant factors affecting the line loss, and \( x_i \) represents different types of network losses, such as loss of whole system, 10kV system and so on. The relevant factors affecting the network loss are composed of the data set \( Y = [y_1, y_2, \ldots, y_n] \), where \( Y \) represents the data matrix of different types of line loss samples, and \( y_j \) represents different influencing factors, such as average bus voltages in different voltage classes, sum of active load, sum of reactive load and so on. The main factors affecting the loss of different types of networks can be identified by finding the mutual information matrix of \( X \) and \( Y \), and then the neural network is used to establish a predictive model. The algorithm flow is shown in Figure 1.

![Figure 1. Algorithm flow of network loss research.](image)
2.1. Mutual information principle

In the field of information theory, mutual information is a measure which can be used to measure the amount of shared information contained in two variables, compared with the Pearson coefficient, it is not limited to a simple linear relationship, it can also describe the degree of nonlinear correlation between two variables. The formula for describing the mutual information value $I(x, y)$ between two variables is as follows:

$$I(x, y) = \int \int p(x, y) \log_2 \left( \frac{p(x, y)}{p(x)p(y)} \right) dx dy$$  \hspace{1cm} (1)

Where $p(x)$ and $p(y)$ are the edge probability density functions of $x$ and $y$; $p(x, y)$ is the joint probability density function of $x$ and $y$; however, the probability density function is difficult to obtain directly. In practice, mutual information is often obtained indirectly through information entropy. The relationship between mutual information and information entropy is as follows:

$$I(x, y) = H(x) + H(y) - H(x, y)$$  \hspace{1cm} (2)

Where $H(x)$ and $H(y)$ are the edge entropies of $x$ and $y$; $h(x, y)$ is the joint entropy of $x$ and $y$. The calculation formulas of the three are as follows:

$$H(x) = -\int p(x) \log_2 p(x) dx$$  \hspace{1cm} (3)

$$H(y) = -\int p(y) \log_2 p(y) dy$$  \hspace{1cm} (4)

$$H(x, y) = -\int \int p(x, y) \log_2 p(x, y) dx dy$$  \hspace{1cm} (5)

$I(x, y)$ can be understood as knowing the contribution of $y$ after reducing the uncertainty of $x$. The stronger the correlation between the two variables, the greater the mutual information value. If the two variables are independent of each other, the mutual information value is 0.

2.2. BP(Back Propagation) Neural Network

The full name of BP neural network is error back propagation neural network, which is a kind of artificial neural network. The BP network can learn and store a large number of input and output mode sub-mapping relationships without the need to accurately describe the mathematical equations of such mapping relationships. The BP network uses the steepest descent method to learn, and continuously adjusts the weights and thresholds of the network through backpropagation to minimize the error and sum of squares of the network, generally including the input layer, the hidden layer and the output layer.

A typical three-layer BP neural network is shown in Figure 2.

![Figure 2. Three-layer of BP neural network.](image-url)
2.3. Evaluation index ($R^2$)

This paper uses the evaluation coefficient $R^2$ to evaluate the prediction model established by neural network. The formula for $R^2$ is as follows.

$$R^2 = \frac{\sum_{i=1}^{c} o_i - \frac{\sum_{i=1}^{c} o_i \sum_{i=1}^{c} o_i}{c^2}}{\sum_{i=1}^{c} o_i^2}$$

(6)

Where $c$ is the number of samples in the test set during neural network training, $o_i$ ($i=1, 2, 3...c$) is the predicted value of the i-th network loss, and $o_i'$ ($i=1, 2, 3...c$) is the i-th true value of the network loss. The value range of $R^2$ is [0,1], and the larger the value of $R^2$, the higher the accuracy of the model prediction and the smaller the error.

3. Results and analysis

The paper takes the grid data in a certain area of Jiangsu Power Grid as the research object, and defines the network loss data set $X = [x_1, x_2, x_3, x_4, x_5, x_6]$ from the angle of partial pressure and branch line. Where $x_1...x_6$ are loss of whole system, 10kV system, 110kV system, 110kV AC line, 220kV system and 220kV AC line respectively. $Y = [y_1, y_2...y_{23}]$ is the statistical correlation factors affecting network loss, where $y_j...y_{23}$ are average bus voltages in different voltage classes, sum of active load, sum of reactive load, sum of active generation, sum of reactive generation, sum of line resistance, sum of line reactance, power supply for different voltage classes, average transformer load rate, high voltage rate and low voltage rate respectively. MATLAB is used to write a solution for mutual information values according to the principle of mutual information algorithm mentioned in 2.1, and then for each column of data in $X$, mutual information is obtained with $Y$. A matrix formed by the mutual information value $I(x, y)$ of $X$ and $Y$ is shown in Table 1.

|       | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|-------|-------|-------|-------|-------|-------|-------|
| $y_1$ | 0.44  | 0.57  | 0.35  | 0.53  | 0.54  | 0.49  |
| $y_2$ | 0.44  | 0.57  | 0.35  | 0.53  | 0.54  | 0.49  |
| $y_3$ | 0.44  | 0.57  | 0.35  | 0.53  | 0.54  | 0.49  |
| $y_4$ | 0.33  | 0.48  | 0.51  | 0.47  | 0.44  | 0.45  |
| $y_5$ | 0.33  | 0.48  | 0.51  | 0.47  | 0.44  | 0.45  |
| $y_6$ | 0.43  | 0.37  | 0.35  | 0.36  | 0.52  | 0.51  |
| $y_7$ | 0.33  | 0.48  | 0.51  | 0.47  | 0.44  | 0.45  |
| $y_8$ | 0.42  | 0.36  | 0.34  | 0.35  | 0.51  | 0.49  |
| $y_9$ | 0.54  | 0.47  | 0.40  | 0.43  | 0.58  | 0.53  |
| $y_{10}$ | 0.46  | 0.61  | 0.12  | 0.19  | 0.23  | 0.15  |
| $y_{11}$ | 0.39  | 0.15  | 0.62  | 0.56  | 0.23  | 0.23  |
| $y_{12}$ | 0.46  | 0.21  | 0.33  | 0.35  | 0.55  | 0.57  |
| $y_{13}$ | 0.46  | 0.61  | 0.12  | 0.19  | 0.23  | 0.15  |
| $y_{14}$ | 0.39  | 0.15  | 0.62  | 0.56  | 0.23  | 0.23  |
| $y_{15}$ | 0.46  | 0.21  | 0.33  | 0.35  | 0.55  | 0.57  |
| $y_{16}$ | 0.55  | 0.77  | 0.39  | 0.43  | 0.44  | 0.38  |
| $y_{17}$ | 0.22  | 0.10  | 0.07  | 0.09  | 0.22  | 0.25  |
| $y_{18}$ | 0.53  | 0.58  | 0.54  | 0.53  | 0.48  | 0.42  |
| $y_{19}$ | 0.58  | 0.48  | 0.42  | 0.49  | 0.60  | 0.53  |
| $y_{20}$ | 0.55  | 0.77  | 0.39  | 0.43  | 0.44  | 0.38  |
| $y_{21}$ | 0.53  | 0.58  | 0.54  | 0.53  | 0.48  | 0.42  |
| $y_{22}$ | 0.57  | 0.46  | 0.40  | 0.43  | 0.59  | 0.54  |
The larger the mutual information value in Tab.1, the greater the influence of the relevant indicators on the network loss, and the smaller the mutual information value, the smaller the impact of the relevant indicators on the network loss. Therefore, table 1 can help us identify the main factors affecting network loss and identify the relative strength of these factors on network loss. Taking the total loss as an example, the relevant factors affecting the total loss are arranged in descending order of numerical values as follows: $y_{20} > y_{17} > y_{12} > y_{11} > y_{10} > y_{19} > y_{22} > y_{11} > y_{13} > y_{14} > y_{16} > y_1 > y_2 > y_3 > y_6 > y_9 > y_{12} > y_{15} > y_4 > y_5 > y_7 > y_8 > y_{18}$. In this way, the relative influence of the 23 indicators on the total loss can be obtained. According to the same principle, the ranking of 23 indicators of other 5 types of network loss can be obtained separately. The order of the strength of each network loss related factor is shown in Table 2.

**Table 2.** Ranking of the relative influence of 23 indicators on different network losses.

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|-------|-------|-------|-------|-------|-------|
| $y_{20}$ | $y_{17}$ | $y_{12}$ | $y_{11}$ | $y_{20}$ | $y_{13}$ |
| $y_{23}$ | $y_{21}$ | $y_{15}$ | $y_{15}$ | $y_{23}$ | $y_{16}$ |
| $y_{17}$ | $y_{11}$ | $y_{19}$ | $y_{1}$ | $y_{10}$ | $y_{23}$ |
| $y_{21}$ | $y_{14}$ | $y_{22}$ | $y_{2}$ | $y_{13}$ | $y_{10}$ |
| $y_{10}$ | $y_{19}$ | $y_{4}$ | $y_{3}$ | $y_{16}$ | $y_{20}$ |
| $y_{19}$ | $y_{22}$ | $y_{5}$ | $y_{19}$ | $y_{1}$ | $y_{6}$ |
| $y_{22}$ | $y_{1}$ | $y_{7}$ | $y_{22}$ | $y_{2}$ | $y_{1}$ |
| $y_{11}$ | $y_{2}$ | $y_{8}$ | $y_{20}$ | $y_{3}$ | $y_{2}$ |
| $y_{13}$ | $y_{3}$ | $y_{20}$ | $y_{4}$ | $y_{6}$ | $y_{3}$ |
| $y_{14}$ | $y_{4}$ | $y_{10}$ | $y_{5}$ | $y_{9}$ | $y_{9}$ |
| $y_{16}$ | $y_{5}$ | $y_{23}$ | $y_{7}$ | $y_{19}$ | $y_{4}$ |
| $y_{1}$ | $y_{7}$ | $y_{17}$ | $y_{8}$ | $y_{22}$ | $y_{5}$ |
| $y_{2}$ | $y_{8}$ | $y_{21}$ | $y_{10}$ | $y_{4}$ | $y_{7}$ |
| $y_{3}$ | $y_{20}$ | $y_{1}$ | $y_{17}$ | $y_{3}$ | $y_{8}$ |
| $y_{6}$ | $y_{10}$ | $y_{2}$ | $y_{21}$ | $y_{7}$ | $y_{19}$ |
| $y_{9}$ | $y_{23}$ | $y_{3}$ | $y_{23}$ | $y_{8}$ | $y_{22}$ |
| $y_{12}$ | $y_{6}$ | $y_{6}$ | $y_{6}$ | $y_{17}$ | $y_{17}$ |
| $y_{15}$ | $y_{9}$ | $y_{9}$ | $y_{9}$ | $y_{21}$ | $y_{21}$ |
| $y_{4}$ | $y_{13}$ | $y_{13}$ | $y_{13}$ | $y_{11}$ | $y_{18}$ |
| $y_{5}$ | $y_{16}$ | $y_{16}$ | $y_{16}$ | $y_{12}$ | $y_{12}$ |
| $y_{7}$ | $y_{12}$ | $y_{11}$ | $y_{11}$ | $y_{14}$ | $y_{15}$ |
| $y_{8}$ | $y_{15}$ | $y_{14}$ | $y_{14}$ | $y_{15}$ | $y_{11}$ |
| $y_{18}$ | $y_{18}$ | $y_{18}$ | $y_{18}$ | $y_{18}$ | $y_{14}$ |

As can be seen from Tab. 2, the factors affecting network loss mainly include power supply, resistance, reactance, active power from the generator, bus voltage level, active load, reactive load and average load rate of the transformer. What’s more, these factors have different levels of strength for different types of network loss.

The artificial neural network model is usually used as a nonlinear function simulator for prediction. BP network can be used for line loss predictive model because of its strong self-learning ability, good robustness, high nonlinearity and fault tolerance. In view of the fact that the BP network can not automatically optimize the multivariate and the convergence of the learning process is slow, the mutual information analysis method and the BP neural network can be combined to solve the optimization problem of the input variables of the BP neural network. As is shown in Tab. 1 and Tab. 2, some indicators have less influence on the line loss. If these weak influence factors can be eliminated, it is helpful to reduce the dimensionality of the data, reduce the number of nodes input by the neural network and shorten the modeling time. Therefore, the threshold for artificially setting the relevant factors is 0.45, and 55 samples from Jiangsu Power Grid are used as the dataset of BP neural network to establish a comparative experiment on six kinds of network loss.
In the first experiment, the original 23 indicators were used as input to establish the prediction model. The structure of BP neural network is 23-9-1 (input layer node-hidden layer node-output layer node). The second experiment uses the influencing factor with the highest correlation of loss as the input to establish the prediction model. The structure of BP neural network is M-9-1 (input layer node-hidden layer node-output layer node, where M is the main type of loss number of influencing factors). The prediction model of six kinds of line loss is shown in Figure 3.

![Experiment 1 prediction result](image1)

![Experiment 2 prediction result](image2)

(a) Total system loss

![Experiment 1 prediction result](image3)

![Experiment 2 prediction result](image4)

(b) System loss at 10.0kV

![Experiment 1 prediction result](image5)

![Experiment 2 prediction result](image6)

(c) System loss at 110.0kV
Figure 3. The test results of six models.
The evaluation coefficient of $R^2$ as the model is in the range of $[0,1]$, the closer $R^2$ is to 1, the smaller the prediction error is, and the prediction ability of the model is better. Comparing the prediction results of each type of network loss in Figure 3, we can optimize the input of BP neural network by retaining the main influencing factors according to the threshold value, which can improve the prediction accuracy of the model and the model has achieved better prediction results. This result also proves that we artificially set the threshold value to select relevant factors for the modeling of line loss is reasonable.

4. Conclusions
In this paper, we use the principle and method of mutual information to analyze and predict the related factors of network loss, and draw the following conclusions.

In this study, the mutual information algorithm is used to identify the relevant factors affecting different types of network loss. According to the mutual information value matrix, the intensity of each related factor affecting the network loss can be analyzed, and then the influencing factors of strong correlation with network loss are retained. Therefore, the identification of network loss strong correlation factors based on data mining is realized.

In order to reduce the complexity of the network loss prediction model establishment process, we use the mutual information matrix to retain the strong correlation factors, and regard it as the input of the neural network to establish corresponding models for six different types of network loss. Finally, we verified with MATLAB that the predicted model can achieve higher precision.

Acknowledgement
Science and technology project of China EPRI: Research on identification method of main influencing factors of network loss based on current management status and big data technology.

References
[1] Lu Jiang 2014 A Study on the Line Loss of Medium and Low Voltage Based on Grid Wise Theory[J]. Telecom World 3 75-76
[2] Zhao L J, Huang L, Lv Q, et al 2017 WAMS/SCADA Data Fusion Method Study Based on Time-Series Data Correlation Mining[C]// International Conference of Artificial Intelligence, Medical Engineering, Education. Springer, Cham 120-133
[3] Zou Y, Xiao Z, Zhang L, et al 2018 A data mining framework within the Chinese NPPs operating experience feedback system for identifying intrinsic correlations among human factors[J]. Annals of Nuclear Energy 116 163-170
[4] Bermudez-Edo M, Barnaghi P, Moessner K 2018 Analysing real world data streams with spatio-temporal correlations: Entropy vs. Pearson correlation[J]. Automation in Construction 88 87-100
[5] Zhang H, Jin J, Dong S, et al 2018 A corrosion correlation analysis method based on pipeline big data[J]. Chinese Science Bulletin 8
[6] Zhang H J, Wang F, Tian Z K 2013 Grey Correlation Analysis of China's Electricity Imports and its Influence Factors[J]. Applied Mechanics & Materials 448-453 2158-2162
[7] Zhu W, Zhou Q, An H, et al 2017 Analysis of the correlation between characteristic of equivalent load and power loss in smart distribution grid[C]// 2017 2nd International Conference on Power and Renewable Energy (ICPRE). IEEE 529-533
[8] Wang S X, Zhou K, Su Y 2017 Line loss rate estimation method of transformer district based on random forest algorithm[J]. Electric Power Automation Equipment 37(11) 39-45
[9] Kashef H, Mahmoud K 2018 Abdel-Nasser M. Power Loss Estimation in Smart Grids Using a Neural Network Model[C] 2018 International Conference on Innovative Trends in Computer Engineering
[10] Weimin H E, Zou Y, Mei F, et al Electrical&Energy Management Technology