

Prediction Model of Word EERIE Based on BP Neural Network

Xiyu Wang*, Chenyang Li, Zekun Pang

School of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an, China

*Corresponding author: 18192319133@163.com

Abstract. This paper uses BP neural network algorithm to determine the difficulty level of the word EERIE by constructing a neural network model, and determine its relationship with the word length, word commonness, letter frequency and other attributes to predict the relevant percentage of the future date, and discusses the uncertainty of the model, the accuracy of the classification model and other characteristics of the data set. In terms of accuracy, the random forest algorithm is the best, 72%, and the k-nearest neighbor algorithm is the worst. In terms of F1 score, logistic regression algorithm performs the best conclusion.

Keywords: BP neural network algorithm, neural network model, logical regression algorithm

1. Introduction

BP neural network, as a multi-feedforward network trained according to the error back-propagation algorithm, can store and learn a large number of input and output data by simulating the function of human neurons, and does not need to describe the mapping relationship of variables. Using input and output data modeling, it has a strong simulation ability for nonlinear systems. [1] The artificial neural network has a strong ability of self-training and learning, can handle large-scale parallel processing, self-organization, nonlinear approximation, and has a strong fault-tolerant ability. It is a complex nonlinear dynamic system model that abstracts, simplifies, and simulates brain functions. It has been used in combination optimization, neural expert system, system prediction and other fields, and the application of neural network in academic evaluation has become more and more common. [2] BP neural network simulation model is widely used. In this paper, by constructing a neural network model, the difficulty level of the word EERIE is determined, and its relationship with the word length, letter frequency, word commonness and other attributes is determined. At the same time, the uncertainty of the model, the accuracy of the classification model and other characteristics of the data set are discussed in order to predict the relevant percentage of the future date. The complex nonlinear function mapping problem is solved by constructing a neural network model. BP neural network algorithm is used to convert continuous data into discrete data through simulation. The training set and verification set are randomly selected, and the expression between input and output data is obtained according to the sigmoid function. The neural network toolbox in MATLAB provides convenient conditions for solving this problem. The function of neural network toolbox is very complete, providing various MATLAB functions, including the establishment, training and simulation functions of neural network , As well as various improved training algorithm functions, users can easily carry out the design and simulation of neural networks, and can also make appropriate modifications on the basis of MATLAB source files to form their own toolkits to meet the actual needs. [3]

2. Model establishment and method

BP neural network is a feed-forward neural network that learns through error back propagation. The basic idea of BP algorithm is to carry out continuous learning and training through the forward propagation of signals and the reverse correction of errors. As an information processing unit, neural network simulates the learning, memory and induction functions of the brain to fuzzily process the expression of computation and knowledge. [4] During the forward propagation, the sample index

value is input from the input layer, and the output value is calculated through the designed weights and hidden layers of each layer. If the output value does not match the target value, the error is started to be reverse corrected from the hidden layer to the output layer. This correction is mainly achieved through the adjustment of the weights, This kind of learning and training is repeated until the error between the network output value and the target expected value reaches the error accuracy required by the system. [5]

In the training process, the network calculates the input data through the weighted sum of neurons at each layer and the activation function to get the output value, and compares it with the expected output value to get the error value. Then, the error will be transmitted from the output layer to the output layer through the back propagation algorithm, and the weight and threshold of each neuron will be adjusted according to the error to minimize the error value. This process will continue to iterate until the network output results and expected results reach a certain accuracy. The specific flow chart is shown in Figure 1.

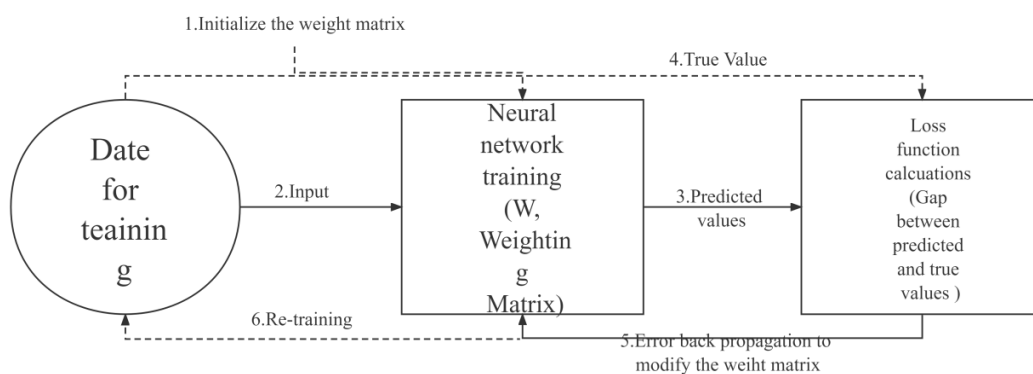


Figure 1. Network flow chart

Taking the number of attempts in the difficult mode as the input and the percentage occupied as the output, a five-layer neural network model is constructed, as shown in Figure 2, to solve the complex nonlinear function mapping problem. In this model, the BP neural network algorithm is adopted, with the number of attempts as the input and the predicted value of the proportion data as the output. Through simulation processing, the processed data is normalized to the range of (0,1), and the continuous data is converted into discrete data.

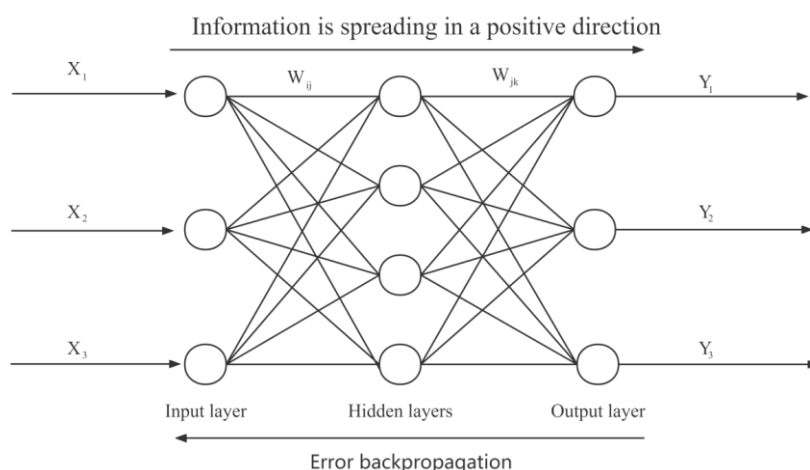


Figure 2. Five-layer neural network model diagram

The transfer function of BP neural network is divided into linear and nonlinear. The nonlinear transfer function generally uses sigmoid function. According to the different range of output values, the sigmoid function can be divided into Log-Sigmoid and Tan-Sigmoid functions. [6]

$$f(x) = \frac{1}{1+e^{-x}} \quad f(x) \in [0,1] \quad (1)$$

$$f(x) = \frac{1-e^{-x}}{1+e^{-x}} \quad f(x) \in [-1,1] \quad (2)$$

The BP neural network is trained by error back propagation. The error is calculated by the output layer, and the correction weight is transferred layer by layer through the hidden layer. The BP neural network stops training after the error is reduced to meet the system requirements through repeated iterations. The training set and test set are tested for outliers, and the outliers are removed, and normalized to obtain the data of each index. Then, 30 data are randomly selected as the training set, and the other 5 data are used as the verification set.

BP neural network is an algorithm that maps the input of neurons to the output, and its construction process includes several steps. On this basis, in order to get the relationship between input and output data, we use a function between the last hidden layer and output layer, usually sigmoid function. Get the expression between input and output data

$$y_k' = \sum_{j=1}^r v_j \cdot f[\sum_{i=1}^m w_{ij} \cdot p_i] + \theta_j \quad (3)$$

($k = 1, 2, \dots, N$), θ_j is the threshold value, y_k is the desired output value, and y_k' is the actual output value of the network.

3. Results and analysis

When designing BP neural network, it is necessary to determine the topological structure of the network (the number of hidden layers and the number of neurons in each layer), the calculation and learning rules of the initialization error of the neural network transformation function, and the normalization of the network training parameters and training samples. [7] Matlab neural network toolbox provides a large number of functions for simulation, design and training. The structure of neural network is shown in Figure 3, and the parameters of BP neural network are set as shown in Table 2. Taking the word EERIE as an example, the neural network model is used to predict the distribution of word EERIE.

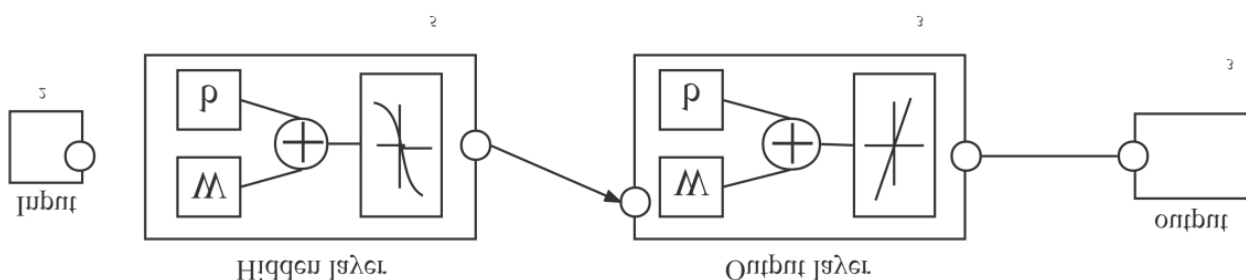


Figure 3. The neural network structure diagram

Table 1. Prediction of EERIE's proportion in different attempts

Sampling Points	Percentage /%	statistics	Precision	Accuracy
1	2.28	10.23	7.95	2.15
2	6.31	14.52	3.15	1.23
2	14.12	2.59	6.1860	7.5895
3	35.16	2.39	5.8179	5.1191
5	30.89	3.43	6.76	3.77
6	12.21	2.96	9.55	4.31
X	74.86	2.22	5.39	8.17

Table 2. The parameters of the BP neural network

Maximum number of training steps	Number of steps between training results	learning rate	training target error	number of training sessions
1000	1	0.0000001	0.000001	1000

Extract the data in a specific time period from the data in Table 1. By calculating the difference between the first 30% and the last 30% of event counts in a specific time period, the negative difference means that the user plays more at the end of the observation interval, while the positive difference means that the user plays less at the end of the observation interval. Sticky vector is a detailed view of the distribution of player activities, consisting of 7 elements, each of which represents 10% of the total running time of the feature extraction cycle, and the element value is the percentage of events recorded in the corresponding time interval during the whole period. In terms of feature extraction and loss period duration, this paper uses a variety of combination and machine learning algorithms. In feature extraction, the interval of 1, 2, 3 and 7 days is considered, and the period of 1 to 7 days is predicted, and a total of 28 combinations are obtained.

We will use the obtained features to build models of random forest, logical regression and k-nearest neighbor algorithm, and use the measures described previously to evaluate these models. Random forest is an integrated machine learning method. It uses bootstrap and node random splitting technology to construct multiple decision trees, and obtains the final classification results through voting. [8] The K-nearest neighbor method usually uses the majority voting rule to determine the category of the new input instance, that is, if the majority of k instances belong to a certain class, the input instance is determined as this class. [9] If there are large differences between each feature and classification in some data sets, large errors will occur. Therefore, in terms of accuracy, the random forest algorithm is the best, at 72%, while the k-nearest neighbor algorithm is the worst, at 69%. The logistic regression model can complete the regression statistical analysis between a group of independent variables and a dependent variable. These independent variables can be continuous data or discrete data. This model is very suitable for the binary classification problem based on multivariable control, which will lead to the existence or absence of two different results. [10] Therefore, in terms of F1 score, logistic regression algorithm performs best, with the highest result of 0.78.

4. Conclusion

When solving the problem, the data is input into the network system, and the solution of the problem is obtained by calculating the function.

In this paper, BP neural network is used to predict the distribution of results, taking the number of attempts in difficult mode as the input and the percentage occupied as the output, and a five-layer neural network model is constructed to solve the complex nonlinear function mapping problem. Based on the obtained characteristics, the models of random forest, logistic regression and k-nearest neighbor algorithm are constructed, and these models are evaluated using the measures described previously. In terms of accuracy, the random forest algorithm is the best, 72%, while the k-nearest neighbor algorithm is the worst, 69%. In terms of F1 score, logistic regression algorithm performs best, with the highest result of 0.78.

Reference

- [1] Comparison and application of multiple linear regression and BP neural network prediction models Zhang Jingyang Journal of Kunming University of Science and Technology (Natural Science Edition), June 2013, 67-73
- [2] Combination evaluation of academic journals based on BP artificial neural network Yu Liping, Statistics and Decision, April 2023, 5-9
- [3] Matlab realization of Bp neural network Journal of Shiyun South Hunan University 2010-10-25 86-88
- [4] Development of roadway deformation prediction system based on BP neural network model Sheng Jianlong Mining Engineering, June 2007, 63-66
- [5] Application Research of BP Neural Network Based on Algorithm Luan Jiuyang Value Engineering June 18, 2012 298-299
- [6] Yu Meixia, Dong Gang, Xu Ruxun, Yu Guoli. Precision time base source calibration prediction model based on BP neural network China test: 1-7 [2023-02-19]
- [7] BP network realization based on MATLAB neural network toolbox
- [8] Feature selection algorithm based on random forest, Yao Dengju, Yang Jing, Zhan Xiaojuan, Journal of Jilin University (Engineering Edition), January 2014, 142-146
- [9] Comparative Study of K-Nearest Neighbor and Logistic Regression Classification Algorithms Wan Huifang, Du Yanpu Journal of Luoyang Institute of Technology (Natural Science Edition) 2016-09-25 86-89
- [10] Risk assessment and inspection of Wenchuan earthquake landslide based on logistic regression model Xu Chong, Hydrogeology and Engineering Geology, 2013, 03, 104-110