Improved Text Classification Algorithm based on BP Neural Network

Yanhua Zhao

School of Information Technology Engineering, Tianjin University of Technology and Education, Tianjin 300222, China.

790489821@qq.com

Abstract

Natural Language Processing (NLP) is the bridge between computer and human natural language, and it is one of the research directions in the field of artificial intelligence. Text classification is one of the classical fields of natural language processing. Combined with the application of neural network in natural language processing, the efficiency of text classification can be improved. Therefore, the research goal of this paper is to use neural network to improve the speed and efficiency of text classification. The word quantization is performed after the matching method is obtained, the dimension is reduced by Principal Component Analysis (PCA), the redundant features are eliminated, and the relevant structural parameters are adjusted by back propagation Neural Network (BPNN), so as to improve the accuracy and efficiency of text classification.

Keywords

Text Categorization; BP Neural Network; Principal Component Analysis; Digital Match; Word Embedding.

1. Introduction

Natural Language Processing is an important direction in computer science and artificial intelligence, it studies how to make computers to understand and use human language to implement information interaction between human machines, and then from a lot of text Extract effective information is extracted [1]. Natural Language Processing tasks its input is a sentence or a piece of text, so there is the following characteristics: First, the input information is a one-dimensional linear word sequence; Second, the length of the input information is uncertain; In addition, there is a great relationship between the relative position of each word and sentence in the input information, and the position of the word is different and the meaning may be completely opposite. Finally, the capture of long-distance features is also very important for natural language processing. Text classification realizes the classification of emotion, emotion, attitude, theme and so on by building a model.

2. Background knowledge

Text classification refers to the process in which a text containing information is mapped by a computer to a given category or topics in advance. Its steps are preprocessing text, extracting text features, and constructing a classification model. Finally, this text document can be classified into known document categories. Chinese word segmentation is the process of reassembling a continuous word sequence into a word sequence according to a certain specification. I In English text, words use spaces as natural delimiters , while Chinese only words, sentences and paragraphs can be simply delimited by obvious delimiters, but words do not have a formal delimiter. Although English also has the problem of dividing phrases, at the word level, Chinese is much more complex and difficult than English [2].

Because of the noise words will appear after the text segmentation, and the appearance of the noise words is easy to reduce the accuracy of text classification, so it is necessary to select the features of the text after word segmentation. The text feature selection methods include mutual information method (MI), chi-square test (CHI), document frequency (DF), information gain method (IG), weighted frequency and possibility (WFO) and so on.

The commonly used method of text representation is word embedding vector method, which is an algorithm model which converts words into vector representation and uses mathematical model to deal with text corpus. The basic idea of the word vector method is to map each word into a K-dimensional real number vector through training (K is generally a super parameter in the model), and judge the semantic similarity between them by the distance between words (such as cosine similarity, Euclidean distance, etc.) [3].

3. Algorithm knowledge

Although artificial neural network (ANN) was initially inspired by biology, artificial neural network has been successfully applied in many different fields, especially for prediction and classification. The excellent feature of artificial neural network is its inherent ability of nonlinear modeling, and without any assumption about the statistical distribution followed by the observation, an appropriate model is formed adaptively based on the given data.

When using neural networks to process text, it is not the previous static mode:

1) there is no feedback in the layer (delay time window)

2) delay and feedback of layer

3) there is no feedback delay in the unit.

4) delay and feedback per unit (periodic cycle)

Therefore, there are simple feedforward neural network (BP) and radial basis function network (RBF) as nonlinear regression methods can generally be used for text recognition.

The weights between neurons and connected neurons constitute the whole neural network, and the interconnected neurons will transmit signals. The size of the weight and the connection mode (GRU, LSTM, CNN, etc.) control the information flow and intensity of the whole network, see Fig. 1.

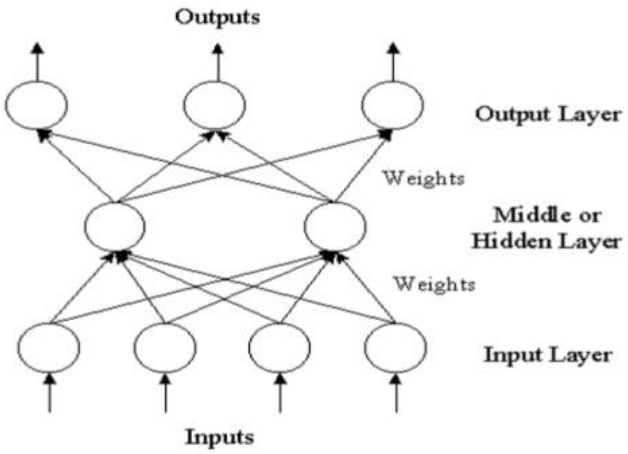

Fig. 1 Three-layer artificial neural network (ANN) model

Activation function: a neuron is activated only when its quantity reaches a certain range, and the activation function is used to estimate the energy of the neuron. Activation functions include: Sigmoid function, tanh function, maxout function, ReLU function and so on.

$$\text{sigmoid} = \frac{1}{1+e^{-x}} \tag{1}$$

The soft saturation of Sigmoid mainly lies in that the derivative value is small and the value is close to 0 at both ends. In the process of transmission from back to front, if there are many hidden layers of the neural network, the gradient will become very small and close to 0 after passing through many layers, and the gradient may disappear within 5 layers, the phenomenon of gradient disappearance will lead to insufficient training of network parameters, which is also one of the reasons why the activation function has not been developed effectively in the past 20 years.

$$\tanh(x) = 2\text{sigmoid}(2x) - 1 \quad (2)$$

Compared with the tanh network, the convergence speed is faster, and it is also soft saturation. When calculating, it is found that the average value is closer to 0. SGD will be closer to natural gradient, thus reducing the number of iterations required [4].

$$\text{ReLU} = \max(0, x) \quad (3)$$

The ReLU function is hard saturated when $x < 0$, and the corresponding weight can't be updated, because the derivative of the ReLU function is 1 when $x > 0$, so the gradient can't attenuate in the positive interval, that is, the problem of gradient disappearance is solved. However, with the advance of training, the phenomenon of "neuron death" will occur, and the output of ReLU function is not zero-centered, both of which affect its convergence.

$$\text{Maxout} = \max(x) \quad (4)$$

Maxout network can approximately fit any convex function, and when the weight and threshold w_2,b_2, …. When w_n, b_n, w is 0, it will degenerate to ReLU function.

BP neural network is a multi-layer feedforward neural network and is trained by back propagation algorithm. It is characterized by signal forward propagation and error back propagation [5]. BP algorithm solves the weight problem of inclusion layer in multi-layer model, improves the learning and memory function of neural network, and has become one of the most typical models in the application of neural network. BP neural network model is composed of input layer, hidden layer and output layer [6]. Neurons in the same layer are independent and not connected to each other. Neurons in adjacent layers are connected according to weight and interconnection structure, and then BP neural network is used for training and classification.

When a signal is input, it is first propagated to the node of the hidden layer, the signal is propagated layer by layer, and finally to the node of the output layer, and each layer has its corresponding characteristic function for transformation. Because of the signal propagates forward to the output layer, the BP neural network model is a feedforward neural network. The learning process of BP neural network is divided into forward propagation and back propagation [7]. When the input mode is given, the signal propagates from the input layer to the hidden layer, then calculates, and then transmits the result to the next layer. If the error between the actual output mode and the expected output mode is large, the error signal will be returned from the output layer through the hidden layer along the original path to the input layer, and the connection weight of each layer will be modified according to the error value to reduce the error, this process cannot stop until the result meets the condition (called back propagation). Once all the training modes are satisfied, BP network learning will be in a good state. When using BP network, we only need forward propagation, not back propagation.

4. Simulation result

The first step of text classification needs to obtain text features, the established word segmentation feature function can obtain vector data in the dictionary file, the data comes from the Chinese corpus, and the data are processed. First look at whether the length of the sentence is greater than 0, ignore if the length of the sentence is less than 0, and if the length of the sentence is greater than 0, then calculate the length of the sentence, the maximum word length and the minimum value, and then the length of the selected string cannot be greater than that length. Then the data is constantly translated and matched. The characteristic data is saved in the main function and the text is represented as a number, but such a representation is usually sparse (because the general dictionary contains tens of thousands of words, so most of the numbers are 0), so the dimension reduction method is used. that is, principal component analysis is used to eliminate redundant data and redundant features.

Through a large number of simulation training, the number of nodes in the input layer and hidden layer is determined, and the optimal network structure is selected to predict the consumption more accurately.

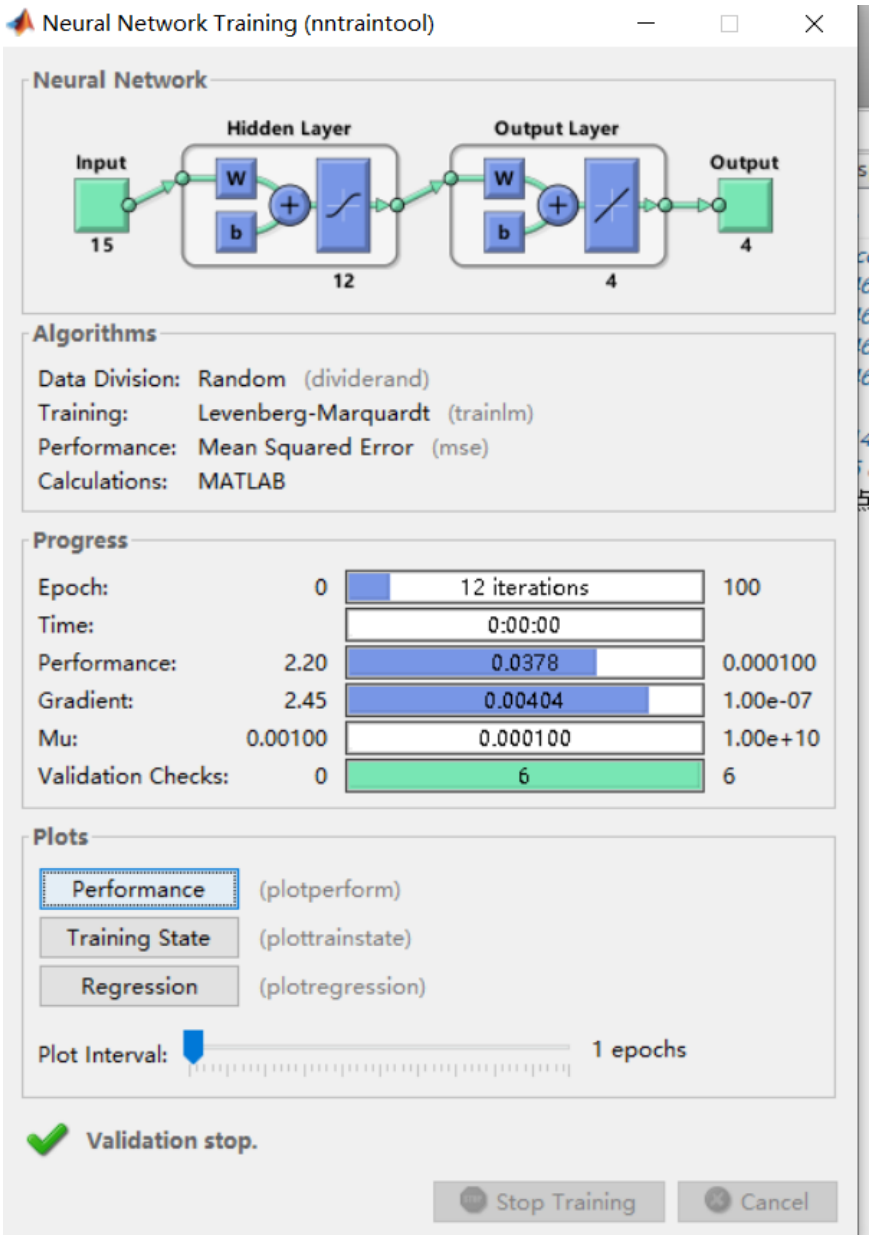

Fig. 2 Neural network toolbox operation diagram

When the number of input layer nodes is more than 6, the network learning time is generally too large. When the number is too small and less than 2, the neural network can't reflect the accuracy of the prediction. Generally speaking, the number of nodes in the hidden layer is too small, BP neural network can't establish a complex mapping relationship, so that the training of BP neural network can't effectively produce results. However, the number of nodes in the hidden layer is too large, and the error is not necessarily minimum , so the learning time of BP neural network may be too long, or the phenomenon of "over-fitting" of data may occur, which reduces the stability of BP network. Therefore, in the experiment, the number of main nodes in the hidden layer can be based on the following principles:

$$h = \sqrt{n + m} + a \text{ Or } h = \log_2 n + a \quad (5)$$

Where h is the number of nodes in the hidden layer, n is the number of nodes in the input layer, m is the number of nodes in the output layer, and a is an integer between 10 and 0.Therefore, the number of hidden layer nodes selected here is 12, the number of iterations is 100, the learning rate is 0.01, and the target error is 0.001. Some of the text data are randomly selected as input data, and the other part is used as output data to compare the test. see Fig. 3 and Fig. 4.

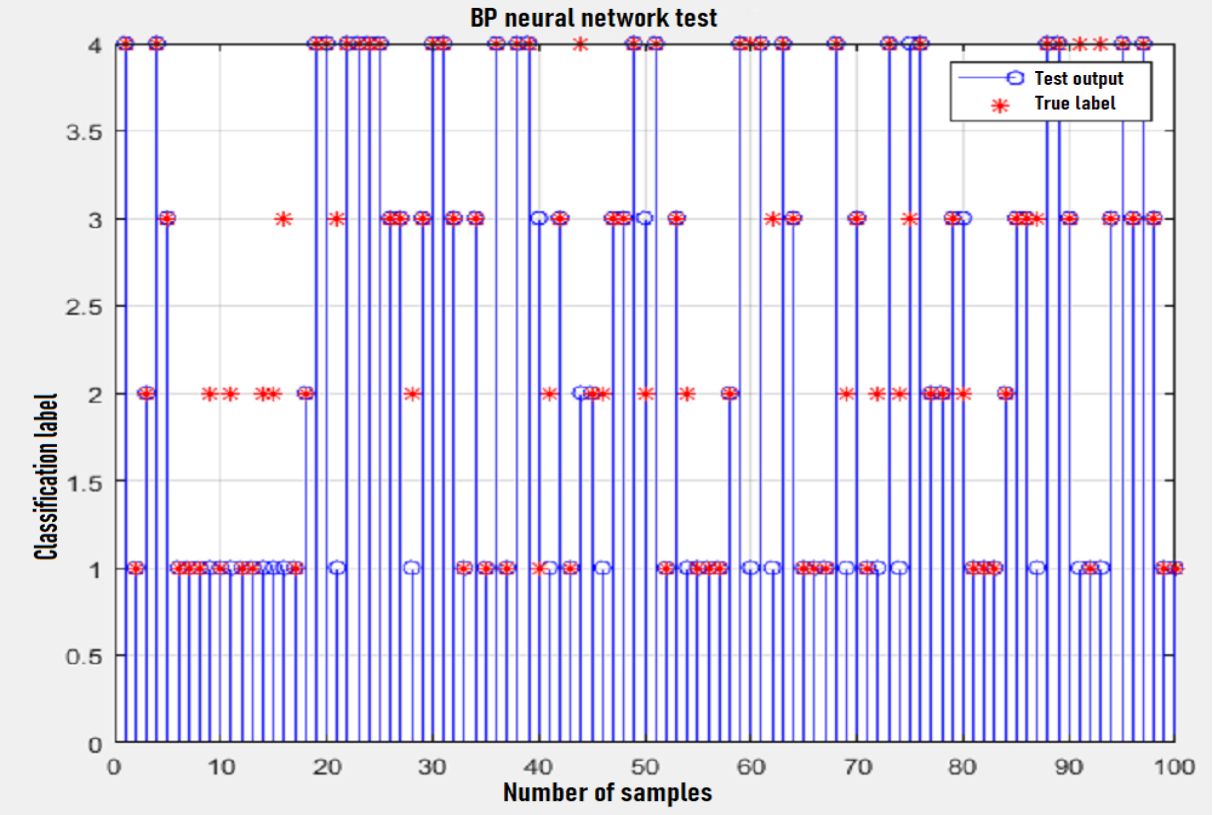

Fig. 3 Neural network training result

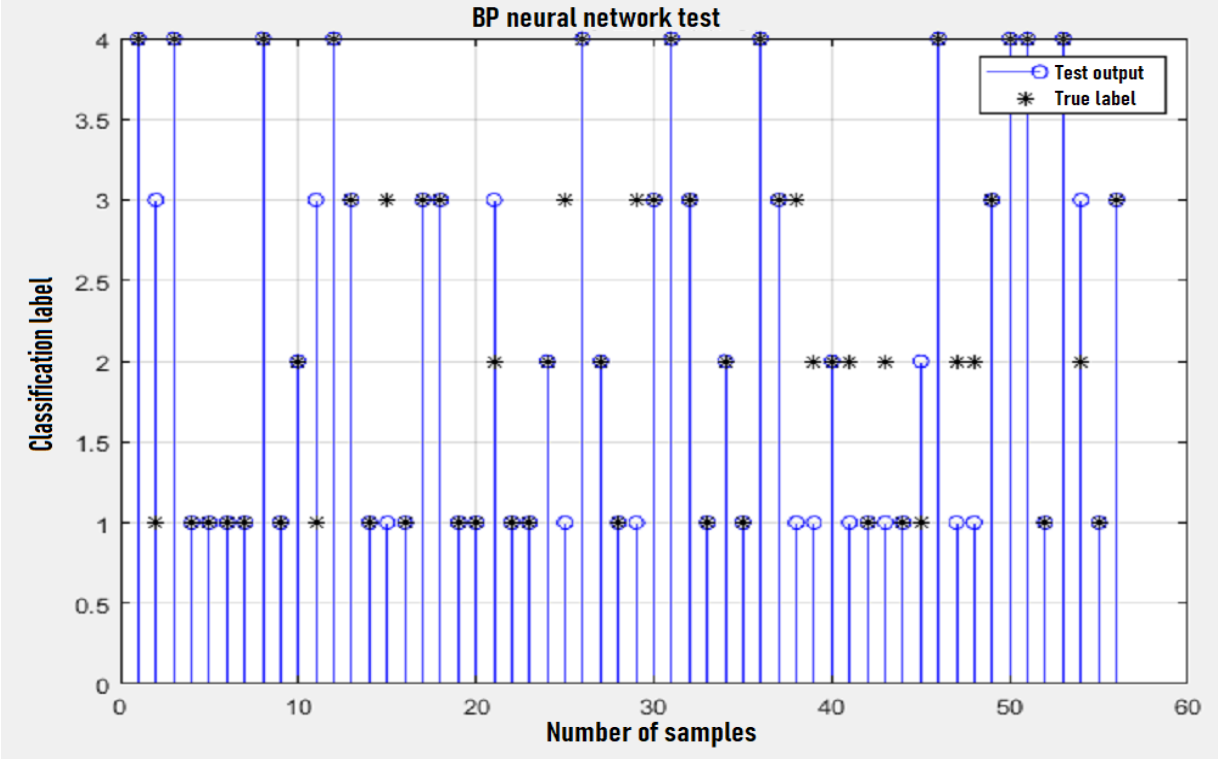

Fig. 4 Neural network test results

5. Conclusion

Based on the corpus files, this paper uses the improved text classification of BP neural network to label the text after vectorization, in which the model network is optimized for shorter texts, and principal component analysis is used to reduce the redundancy of historical information in the process

of feature extraction. The classification model in this paper has a fitting degree of more than 70% in the task of text classification using BP neural network. In a word, this paper uses the method of digital matching for text classification, uses the brief information of principal component factors, and improves the accuracy and efficiency of text classification through BP neural network.

References

[1] F.H Li, D.F Yao: A Survey of text representation and language models in Natural language processing (Network Application Branch of China computer user Association. Proceedings of the 24th annual meeting of network new technologies and applications in 2020 by the Network Application Branch of China computer user Association. Network Application Branch of China computer user Association: Beijing key Laboratory of Information Service Engineering, Beijing Union University, 2020).p.169-172.

[2] D.W Huang, S Chen, C.L Lai, D.Y Li, L.P Huang: Application of Chinese word Segmentation in Security measures Identification of Line work order, China High and New Technology, Vol. 14 (2018) No.2, p.79-81.

[3] L.L Wu.: The actual problems and contents of the implementation of continuing education policy for professional and technical personnel, Contemporary continuing Education, Vol. 38 (2020) No.2, p.9-34.

[4] R.H Li: Load identification method based on neural network (Master's degree thesis, Hangzhou University of Electronic Science and Technology, China 2018).

[5] G.X Mao, W Tan, Z.Z Chai, Y Zhao, S.J Yang: Based on BP Neural Network Breast Diameter - Tao High Model Prediction, Journal of Zhejiang Agriculture and Forestry University, Vol. 37 (2020) No.4, p.752-760.

[6] Z.G Li, Z.Y Wang, J.C Sun: Research and Design of handwritten Digital BP Neural Network based on FPGA. Computer Engineering and applications, Vol. 56 (2020) No.17, p.251-257.

[7] X.T Cui: Evaluation and countermeasure study of black and smelly water body of river based on BP neural network (Master's degree thesis, Qingdao University of Technology, China 2018).