Advances in Computer Science Research, volume 79
2nd International Conference on Sensor Network and Computer Engineering (ICSNCE 2018)

Research of Email Classification based on Deep Neural Network

Wang Yawen
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
e-mail: 475609323@qq.com

Yu Fan [a], Wei Yanxi [b]
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
e-mail: [a] 409066272@qq.com
[b] 407171251@qq.com

***Abstract*—The effective distinction between normal email and spam, so as to maximize the possible of filtering spam has become a research hotspot currently. Naive bayes algorithm is a kind of frequently-used email classification and it is a statistical-based classification algorithm. It assumes that the attributes are independent of each other when given the target value. This hypothesis is apparently impossible in the email classification, so the accuracy of email classification based on naive bayes algorithm is low. In allusion to the problem of poor accuracy of email classification based on naive bayes algorithm, scholars have proposed some new email classification algorithms. The email classification algorithm based on deep neural network is one kind of them. The deep neural network is an artificial neural network with full connection between layer and layer. The algorithm extracted the email feature from the training email samples and constructed a DNN with multiple hidden layers, the DNN classifier was generated by training samples, and finally the testing emails were classified, and they were marked whether they were spam or not. In order to verify the effect of the email classification algorithm based on DNN, in this paper we constructed a DNN with 2 hidden layers. The number of nodes in each hidden layer was 30. When the training set was trained, we set up 2000 batches, and each batch has 3 trained data. We used the famous Spam Base dataset as the data set. The experiment result showed that DNN was higher than naive Bayes in the accuracy of email classification when the proportion of the training set was 10%, 20%, 30%, 40% and 50% respectively, and DNN showed a good classification effect. With the development of science and technology, spam manifests in many forms and the damage of it is more serious, this puts forward higher requirements for the accuracy of spam recognition. The focus of next research will be combining various algorithms to further improve the effect of email classification.**

Keywords-Deep Neural Networks; Spam Email; Classification; Naive Bayes; SpamBase Data Set

I. INTRODUCTION

Email has become a major way of communication for people at present, but the problem of spam comes behind. The harm of spam is mainly manifested as the following aspects: occupying bandwidth, leading to the congestion of the email server and reducing the efficiency of the network; consuming the time of the user and affecting the work efficiency. Therefore, the effective distinction between normal email and spam, so as to maximize the possible of filtering spam has become a research hotspot currently.

Naive bayes algorithm is a kind of frequently-used email classification and it is a statistical-based classification algorithm[1-3], which has the characteristics of simple realization and fast classification. However, it assumes that the attributes are independent of each other when given the target value[4].This hypothesis is apparently impossible in the email classification, so the accuracy of email classification based on naive bayes algorithm is low. In allusion to the problem of poor accuracy of email classification based on naive bayes algorithm, scholars have proposed some new email classification algorithms. The

email classification algorithm based on deep neural network (DNN) is one kind of them.

II. THEORETICAL BASIS

The basic concept of artificial neural network is based on the hypothesis and model construction of how the human brain responds to complex problems[4-6]. The deep neural network is an artificial neural network with full connection between layer and layer, and its structure is shown in figure 1.The full connection between layer and layer means that any neuron in theith layer must be connected to any of the neurons in the $(i+1)^{th}$ layer.Although the deep neural network looks complex, it is still the same as the perceptron from a small local model.

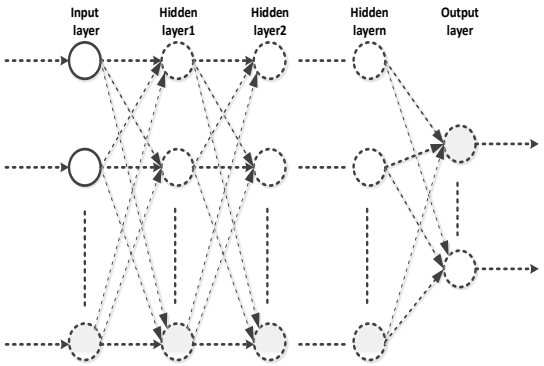

Figure 1. Structure diagram of deep neural network

We use w_{jk}^{l} to respresent the weight coefficient between the kth neuron in the$(l-1)^{th}$ layer and the jth neuron in the lth layer, b_j^l to represent the bias of the jthneuron in the lthlayer, $a_j^l$to represent the activation value of the jthneuron in the lthlayer. We can get the following relationship between the activation value of thejthneuron in the lthlayerand the activation value of all neuron sin the $(l-1)^{th}$layer:

$$a_j^l=\sigma(\sum_k w_{jk}^l a_k^{l-1}+b_j^l) \quad (1)$$

We assume that w^l is the weight coefficient matrix of all the neurons in the lthlayer, b^l is the bias matrix of the lthlayer, a^l is the activation value of the lthlayer, z^l is the weighted input of all neurons in the lthlayer, Then w_{jk}^l is the weight coefficient of row j, column k. The relationship between the activation value of the lth layer and the activation value of the$(l-1)^{th}$layer can be expressed by the following matrix relationship:

$$a^l=\sigma(z^l)=\sigma(w^l a^{l-1}+b^l) \quad (2)$$

Here σ respresents the non-linear activation function of the nodes on the hidden layers, and the traditional DNN uses sigmoid function usually, as shown in expression (3).Because the sigmoid function has properties such as monotone increasing and its inverse function has the property of monotone increasing, it is often used as a threshold function of neural networks, It maps the variables between 0 and 1.The sigmoid function curve is shown in figure 2:

$$\sigma(z)=\frac{1}{1+e^{-z}} \quad (3)$$

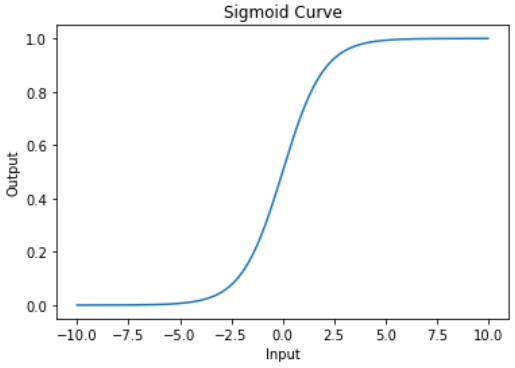

Figure 2. The sigmoid function curve

III. ALGORITHM DESCRIPTION

Implementation process of mail classification algorithm based on deep neural network was shown in Figure 3.

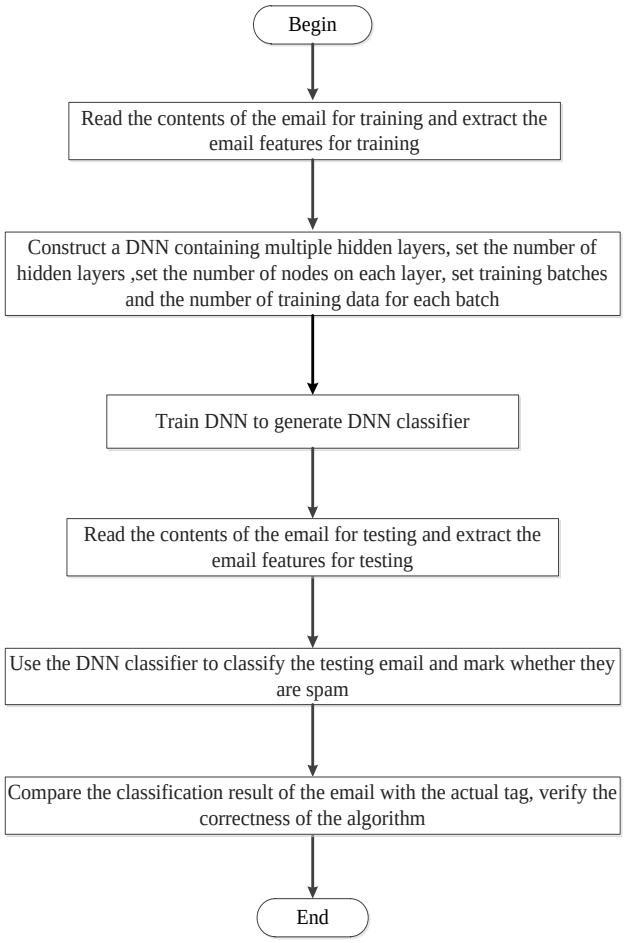

Figure 3. Algorithm execution process

Step 1: Read the contents of the email for training from the Spam Base dataset and extract the email features for training such as word_freq_, char_freq_, capital_run_length_average, capital_run_length_longest, capital_run_length_total, and so on.

Step 2: Construct a DNN containing multiple hidden layers, set the number of hidden layers(n_classes) ,set the number of nodes on each layer (hidden_units) ,set training batches(steps) and the number of training data for each batch (batch_size).

Step 3: Train DNN to generate DNN classifier.

Step 4: Read the contents of the email for testing from the SpamBase data set and extract the email features for testing such as word_freq_, char_freq_, capital_run_length_average, capital_run_length_longest, capital_run_length_total, and so on.

Step 5: Use the DNN classifier to classify the testing email and mark whether they are spam(1 or 0).

Step 6: Compare the classification result of the email (y_predict) with the actual tag(y_test), calculate the accuracy of the algorithm in the email classification(accuracy_score) and verify the correctness of the algorithm.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the effect of the email classification algorithm based on DNN, in this paper we constructeda DNN with 2 hidden layers. The number of nodes in each hidden layer was 30. When the training set was trained, we set up 2000 batches, and each batch has 3 trained data. We used the famous SpamBase dataset as the data set, which was from the UCI machine learning library at the University of California, USA. The specific situation is shown in table I.

We compared the two kinds of email filtering algorithms of DNN and naive Bayes with accuracy, which is the main evaluation standard of email filtering technology. The accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{Number of correctly identified email}}{\text{Total number of emails}} \quad (4)$$

We did five groups of experiments in this paper.The selection case of training set and testing set in each experiment is shown in table II.

TABLE I. SPAMBASE DATA SET

Index	Value	Index	Value
Total number of the email	4601	Number of attributes	57
Number of email category labels	2	Email category	Validemail， spam email
Number of the spam email	1813	The proportion of spam email	39.4%
Number of the valid mail	2788	The proportion of validemail	60.6%

TABLE II. THE SELECTION CASE OF TRAINING SET AND TESTING SET

group number	The proportion of the training set in all data	The number of email in training set	The proportion of the testing set in all data	The number of email in testing set
1	90%	4140	10%	461
2	80%	3680	20%	921
3	70%	3220	30%	1381
4	60%	2760	40%	1841
5	50%	2300	50%	2301

The experimental results were shown in Figure 4.

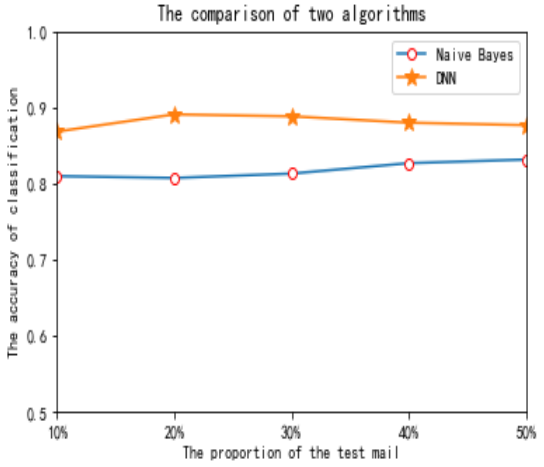

Figure 4. The comparison of accuracy of the two algorithms

The experiment result showed that DNN was higher than naive Bayes in the accuracy of email classification when the proportion of the training set was 10%, 20%, 30%, 40% and 50% respectively, and DNN showed a good classification effect.

V. CONCLUSION

The application of email classification algorithm based on deep neural network is studied in this paper. The algorithm constructed multiple hidden layers and generated DNN classifiers through training. The experiment results showed that the accuracy of the algorithm is obviously higher than the naive Bayes algorithm.

With the development of science and technology, spam manifests in many forms and the damage of it is more serious, this puts forward higher requirements for the accuracy of spam recognition. The focus of next research will becombining various algorithms to further improve the effect of email classification.

ACKNOWLEDGMENTS

New network and detection control national joint engineering laboratory fund program (GSYSJ2016017). Xi'an Technological University Principal Scientific Research Fund Project: XAGDXJJ－1315

REFERENCE

[1] Cao Cuiling, Wang Yuanyuan and Yuan Ye, “Research of a spam filter based on improved naive Bayes algorithm, ”Chinese Journal of Network and Information Security, Vol.3No.3, pp. 64-70, March 2017.

[2] Wang Zhiyong and Liu Hongmei, “DESIGN AND IMPLEMENTATION OF BAYESIAN SPAM FILTE R ING SYSTEM,” Journal of Inner Mongolia Agricultural University(Natural Science Edition), Vol.38 No.3, pp.82-86, May. 2017.

[3] Wang QingSong and Wei Ruyu, “Bayesian Chinese Spam Filtering Method Based on Phrases,” Computer Science, Vol. 43 No. 4, pp.256-259, Apr 2016.

[4] Neural Networks and Deep Learning [EB/OL]. http://neuralnetworksanddeeplearning.com.

[5] Li Kun, CHai Yumei and Zhao Hongling, “Estimation of Fetal Weight Based on Deep Neural Network,” Computer Science, Vol. 43 No. 11A, pp. 73-76, Nov 2016.

[6] Cao Meng, Li Hongyan and Zhao Rongrong, “A Pitch Detection Method Based on Deep NeuralNetwork,” Microelectronics & Computer, Vo1.33 No.6, pp. 143-146, June 2016.

[7] Ren Rongrong, Zhou Mingquan and Geng Guohua, “The multi-scale features extraction method based on deep neural network”, Journal of Northwest University(Natural Science Edition), Vol. 47No.2, pp. 215-221, Apr2017.

[8] S.L.Zhang Research on Deep Neural Networks based Models for Speech Recognition(Ph.D., University of Science and Technology of China, China 2017), p.35