

# The e-mail filtering system based on improved genetic algorithm

JiangHuaLi<sup>1</sup>, WangPing<sup>2</sup>

<sup>1</sup>School of Physics and Optical Electronic Information Technology  
FuJian Normal University, FuZhou, china, 350007  
Email: lilisteven@126.com

<sup>2</sup>School of Physics and Optical Electronic Information Technology  
FuJian Normal University, FuZhou, china, 350007  
Email: PWANG@fjnu.edu.cn

**Abstract**—E-mail is truly convenient for people's working and living, but the popularity of the spam requires urgent technology treatment. This article describes the basic structure of the e-mail, mainly analyzes the characteristics of the spam, then introduces an e-mail filtering system design based on improved genetic algorithm, finally makes the model experiment and tests it.

**Index Terms**—E-mail; spam; improved genetic algorithm; filtering system

## I. INTRODUCTION

With the popularity of networks, e-mails give great convenience to people's lives. Because of commercial benefits, e-mails not only give convenience, but also bring a lot of negative impacts. Spam not only waste the network bandwidth, people's time and energy, but also become a tool of the illegal criminals. The litter filtering technology at home and abroad is mainly the followings<sup>[1]</sup>:(a)The filtering technology based on the rule-assessment, it is a good pre-set rules for pattern matching on the e-mail filtering technology; (b)The blacklist filtering technology, it is based on user complaints and sampling accumulation, is a database composed by the domain name or IP;(c) The white list filtering technology, it is the opposite of the blacklist, the database is the same as the blacklist, but they are legitimate and should not be blocked; (d)The word distribution of Bayesian filtering technology, it can be classified from the database and

automatically generates filters, not needing people's ruling.

## II. E-MAIL INTRODUCTION

### A The basic principles of e-mail

E-mail system is the product of the communications and computer technology. E-mail client use WINSOCK interface to program, at present you can use two sets of interfaces: stream socket and datagram socket. Sending e-mail at least needs four different procedures: Sending e-mail client program; SMTP server program; POP3 server program; receiving e-mail client program. E-mail conversation relies on a standard set of protocol<sup>[2]</sup>: SMTP Protocol, the agreement provides a dialogue with the SMTP server about a series of orders and process standards; POP3 protocol, the protocol provides a dialogue with the POP3 server about a series of command and process standards, as well as IMAP4 protocol which is the familiar as the POP3 protocol to receive mails. SMTP (Simple Mail Transfer Protocol) is also known as Simple Mail Transfer Protocol. It is mainly for the transmission rules. POP3 (Post Office Protocol Version 3) is the e-mail system, one of the basic protocol. POP3 is used to the C / S structure of the offline model of e-mail. Server starts listening POP3 through TCP port 110. A regulated e-mail message is composed of header 1 and message body. When the

<sup>1</sup> Science and Technology Department of Fujian Province, "network based on embedded portable terminals research" (2007F5039) "

message body is divided into a number of paragraphs, each paragraph contains the header and the body, between these two parts is the blank line. E-mail header including the sender, receiver, subject, date, MIME version, the type of message and other important information. The structure of the E-mail header is complex, usually composed by a number of components, such as From, To, Date, Subject, etc. Learn the meaning of each field is crucial for the extract of the mail content. Message body is normally referred to the message content, is a series of text lines, it contains the information which send to the receiver, its type is defined by the "Content - Type", the common type are the Text / Plain (plain text), Text / Html (Hypertext) [3].

### B Spam's Characteristics

At present, the major spam mail is the followings [1]: (a)The receiver dose not request or agree to receive, such as advertisements, electronic publications, various forms of promotional materials; (b) Lots of e-mail can not be rejected; (c)Hide the sender's identity, address, title and other information; (d)Contain the false information, the sender and routing information. Judging an email is whether a spam or not is mainly to recognize the characteristics of the mail, which needs extracting the feature of the required messages. We can distinguish an e-mail through the follows; (a)The sender address and receiver address is whether the same, not identical or abnormal structure is the spam; (b) E-mail copy number, forwarding address for multiple receivers, it can be suspected as the spam; (c)E-mail subject, there are always sensitive terms in the spam; (d) X-Mailer header , spam has not X-Mailer header or using a special header; (e) Received fields' frequency, usually Received lines record the transit of mail routing information, usually three or more received line in the letter can be judged as the spam; (f) Whether it contains false received field, the spam usually contains the false received line; (g) The key words of the letter, the spam has a lot of similar words or sensitive information, which requires to collect the statistical classification of words; (h)E-mail attachment types, you can check the e-mail's attachment name to determine whether the e-mail is secure; (i)The message body size, too large or too small can be recognized as spam [1].

### C e-mail classification and identification system

The classification of e-mail filtering is mainly divided into three steps: feature extraction, network training and testing. Figure 1 can show the whole process, first we extract the feature vector using the training samples, and then put all the vectors into the neural network, through the network we can get the e-mail classification, at last test it with a large number of E-mails.

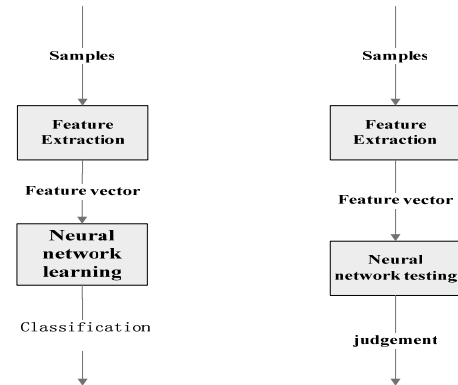


Figure 1. e-mail classification model

### D Improved Genetic Algorithm

The current means of filtering technology is mainly divided into two types, one is filtering e-mail address, and the other is filtering e-mail content. But both of these technologies are lack of intelligence and adaptability for new and emerging spam, they must be manually re-amended to adapt to the new changes. With spammers and means of diversification springing up, the traditional filter based on the old technique is difficult to adapt to the new spam, the studying of email structure according to network information, as well as transmission information and so on to identify the characteristics of the spam, automatically set up and update new features and rules of the spam, using the improved Genetic Algorithm to the design of e-mail filters are the innovations.

For the sake of brevity, this paper only consider strictly hierarchical multi-layer fully connected feed forward neural network, that is, only the neurons in the adjacent layer have a connection. For such kind of neural network, the number of neurons in the input layer and output layer are known, so the required determination is the number of hidden layer of neurons in each hidden layer. Thus the coding of the structure of neural network can use the binary code, in which the string length shows

the number of neurons in the hidden layer, and the arithmetic sum show the number of the effective neurons. This coding scheme is simple, and its decoding process is easy, and also easy to participate in genetic manipulation. But the study of the weights and thresholds of neural network is a complex problem about the continuous function optimization, if we adopt the binary code, genetic manipulation is simple, but the encoded string is more long, encoding and decoding process are more burdensome, it would increase the computing time. If we adopt the real number coding scheme, that is, the neural network weights and thresholds directly expressed with real number, which does not require encoding and decoding, it can reduce the complexity of algorithms, can improve the efficiency of the algorithm and avoid the above-mentioned shortcomings of the binary code.

In sum of the above, a coding scheme which can combine the binary with decimal coding was used<sup>[5]</sup>, that is the structure of neural network using binary encoding scheme, the thresholds and the weights using real-coded scheme. Of course it is not difficult to extend to multi-hidden layer. The feed forward neural network is shown in Figure 2,

$$X = (x_1, x_2, \dots, x_i, \dots, x_n)^T ;$$

$$O = (o_1, o_2, \dots, o_k, \dots, o_l)^T ;$$

Define respectively the input vector and output vector of neural network  $n$  and  $l$ , the input dimension and output dimension. Assuming that the number of neurons in hidden layer is up to  $m$ , the neural network encoding is as follows:  $s_1, s_2, \dots, s_j, \dots, s_m, v_{11}, v_{21}, \dots, v_{n1}, \theta_1, v_{12}, v_{22}, \dots, v_{n2}, \theta_2, \dots, v_{ij}, \dots, \theta_j, \dots, v_{nm}, \theta_m, w_{11}, w_{21}, \dots, w_{m1}, T_1, w_{12}, w_{22}, \dots, w_{m2}, T_2, \dots, w_{jk}, \dots, T_k, \dots, w_{ml}, T_l$

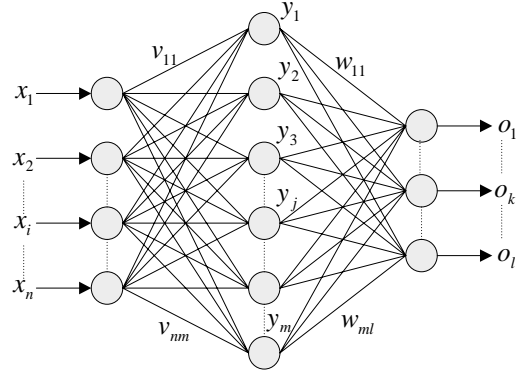


Figure2. Three-layer neural network

Neural network optimization algorithm steps are as follows:

- (a) Give a sample set of training network;
- (b) Set all the parameters, Set the group size  $N=80$ , the max time of evolution  $T=100$ , the factor for network complexity  $c = 0.001$ , the magnification factor in the fitness function  $a=100$ , the number of select strategy  $n=4$ ;
- (c) Generate randomly the initial group with the size  $N$  ;
- (d) Decode for each individual in the initial group, and obtain the actual output of network corresponding to the individual by the training sample set, and then calculate the learning error and fitness value of each individual;
- (e) Pair randomly for the individuals in the group, then cross the parent individual pairs with the probability  $P_c$ ,

$$p_c = (1 + \exp(-\frac{k_1 \lambda_1}{t}))^{-1} \quad (1)$$

Then mutate each parent individual with the probability  $P_m$ ,

$$p_m = (1 + \exp(-\frac{k_2 \lambda_2}{t}))^{-1} - 0.5 \quad (2)$$

- Then the offspring group was generated;
- (f) Decode for each individual in the offspring group, and obtain the actual output of network corresponding to the individual by the training sample set, and then estimate the learning error and fitness value of each individual;
- (g) Adopt the strategy of excellent individual protection and the competition among two, select  $N$  individuals from the offspring group to be a new generation group;

(h) Decode for the best individual, and then amend the network weights and thresholds correspond to the individual with the method of gradient descent, so can gain the new best individual. Regarding the output layer, the weight value is:

$$w_{jk} \Leftarrow w_{jk} + \frac{\eta}{p} \sum_{p=1}^p \delta_k^{p,o} y_j^p, j=1,2,\dots;m; k=1,2,\dots;l; \quad (3)$$

The threshold value is:

$$T_k \Leftarrow T_k + \frac{\eta}{p} \sum_{p=1}^p \delta_k^{p,y}, k=1,2,\dots;l; \quad (4)$$

$$\delta_k^{p,o} = (d_k^p - o_k^p) f'(net_k^p); \quad (5)$$

$f'(net_k^p)$  is the derivative of the  $k$  neuron's activation function for the output level.

Regarding the hidden layer, the weight value is:

$$v_{ij} \Leftarrow v_{ij} + \frac{\eta}{p} \sum_{p=1}^p \delta_j^{p,y} x_i^p, j=1,2,\dots;m; i=1,2,\dots;n; \quad (6)$$

The threshold value is:

$$\theta_j \Leftarrow \theta_j + \frac{\eta}{p} \sum_{p=1}^p \delta_j^{p,y}, j=1,2,\dots;m; \quad (7)$$

$$\delta_j^{p,y} = \left( \sum_{k=1}^l \delta_k^{p,o} w_{jk} \right) f'(net_j^p); \quad (8)$$

$f'(net_j^p)$  is the derivative of the  $j$  neuron's activation function for the hidden layer.

(i) Calculate the learning error and fitness value of the network corresponding to the new best individual, and then the new best individual is alternative to the original best individual.

(j) If the time of evolution reaches to the max, then goes to step k otherwise goes to step e;

(k) Amend the network weights and thresholds corresponding to the new best individual with quasi-Newton LM algorithm<sup>[6]</sup>;

(l) If the network error is less than the specified error  $E_{\min}$ , then goes to step m, otherwise goes to step k;

(m) Output the structure, weights and thresholds of the

optimized neural network.

### E Network Training

In this design, the vectors of the e-mail feature  $N=8$ , the results with 0 messages expressing the normal (category 1), with 1 expressing spam (category 2). Network structure is as follows: the input layer has eight neurons, the hidden layer has 17 neurons, and the output layer has two neurons. The activation function of the hidden layer neuron is the S-tangent; the activation function of the output layer neuron is the S-type logarithmic.

After the network is created, we need to train the network. BP training process is mainly divided into two stages: firstly set the network structure, the weights and thresholds, calculate the neuron output from the first layer, then modify the value of the thresholds and weight to calculate the total error on the impact of a two-stage. Repeat it until the value converges. The result of the Network training is shown in Figure 3.

The result of simulation experiment indicates that this optimization algorithm can not only effectively optimize the neural network structure with the weight and the threshold value, but also demonstrate that the optimized neural network has a good approaching ability and pan-ability.

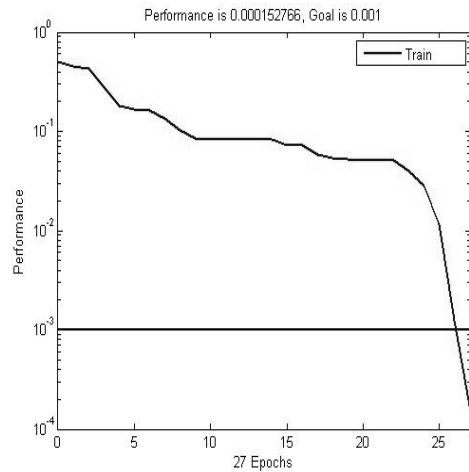


Figure 3. Network training

## III. THE EXPERIMENT OF THE E-MAIL FILTER SYSTEM

### A The design

The system is designed in Visual Studio 2005, using C# to program the e-mail feature extraction and filtering

junk e-mail procedures, the system is shown in Figure 4, and then we use 100 spams to experiment. The results show that the system can filter 95 spams; the rate is up to 95% showing that the model of e-mail classification and identification based on improved genetic algorithm can reach a relatively high accuracy.



Figure 4. E-mail filter system

## REFERENCES

- [1]. Huang Guo Yu. E-mail based on neural network classification research [D]. Chang an University master's thesis .2006:1-60.
- [2]. Zhang Yue, Shi Yang. Qinghua - SEG HDTV [N]. Computer World, 2000-2-28. C8-C9.
- [3]. Hu Yan, Teng Gui Fang, Wang Dan. Based on the structure of MIME e-mail message content extraction technology research [J].Application of reseach.2008, 5:85-88.
- [4]. Jiao Lichen. Neural Network System Theory [M]. Xi an: Xi an Electronic University Press, 1996.
- [5]. ZHEN Fang xiong, LI Yue xin. Neural network optimize based on improved genetic algorithm [J].Journal of Hubei University (Natural Science), 2006, 28(4):345-349.
- [6]. XU Jin. New Comprehensive Learning Method Based on LM-Quasi Newton Algorithm Applying for Feed-Forward Neural Network [J]. Journal of Southwest Jiao tong University, 2004, 39(5):675-678.