International Journal of Future Generation Communication and Networking
Vol. 8, No. 3 (2015), pp. 61-70
http://dx.doi.org/10.14257/ijfgcn.2015.8.3.06

The Abnormal Network Traffic Recognition Method Based on Optimized BP ANN Model

Xu Yabin

School of Computer, Beijing Information Science and Technology University, Beijing, China
xyb@bistu.edu.cn

Abstract

To recognize abnormal traffic in network, so as to perceive the illicit behavior in network, carry out scientific and effective management, and ensure the network security, we extracted the abnormal network traffic features and proposed an abnormal network traffic recognition method based on optimized Back Propagation Artificial Neural Networks (BP ANN). The experimental results indicate that, although the training time is longer, but the accuracy rate of BP ANN in abnormal network traffic identification is superior to other methods. And the convergence rate of optimized BP ANN model is significantly faster than traditional BP ANN model.

Keywords: *network traffic recognition; network behavior perception; BP ANN*

1. Introduction

More and more network applications appeared in front of us with the fashion of more close to our life and work, especially, the mobile Internet applications, such as Micro-bolog, Fetion and others. The convenient and cheap characteristics of Internet applications are becoming more prominent, and making the Internet application more widely and deeply. Human being more and more rely on the Internet. Internet has become one of the important survival conditions of human society. At the same time, frequent network service quality problems (such as the network congestion, the delay jitter, etc), as well as frequent network security problems (such as the Trojan virus, hacker attacks, etc.), not only caused a certain degree of distress on network managers, but also caused a certain loss and harm to network service providers and users. Thus, the network management and network security problems are becoming more and more prominent, more and more important. Therefore, it posses important meaning that research and analyze the behavior of users in network, but, it is unprecedented and challenging.

First of all, against traditional E-mail, Web, FTP and other Internet applications, the new streaming media, games, peer-to-peer (P2P) applications present a large traffic of network, complex behavior characteristics. Especially, the technology and applications based on P2P technology, such as file sharing, instant messaging and collaborative computing, streaming media transmission, spring up one after another, directly caused the changes of Internet application modes. According to statistics, the flow of P2P application is more than the sum of Web service flow and the FTP service flow, become the most application which consumed Internet link resources. It is also exacerbating the congestion extent of network in nearly double traffic. It causes a tremendous impact on network and service providers and network managers in network planning and management. The traffic identification and management technology which take P2P traffic as the main object is becoming one of the most general concerned problems in academic and engineering [1].

ISSN: 2233-7857 IJFGCN

Secondly, hackers attacking, Trojan virus intrusion, ports scanning, DDOS attacking are as flood. In order to escape law liability, escape monitoring, the malicious network behavior and new developed or improved P2P applications, instant messaging, more and more use changed ports, content encryption, transport across application protocols and other technical means. In the face of a variety of network security issues and fully anti-monitoring capability of new Internet application, recognition manners based on TCP/UDP ports and load characteristics of application layer have been difficult to fit the requirements of flow identification [2]. Efficient, accurate, intelligent, real-time Internet traffic recognition has become a highly challenging issue.

Furthermore, the real-time business of new IP-based, low retardation, little dithering are more and more, the service quality requirements on Internet are further increasing [3]. In order to meet the growing requirements of new Internet applications on QoS, it is necessary to use advanced network traffic recognition technology, and to carry out precise and differentiated network services. According to the actual needs and the network running status, the deployment of a reasonable flow management and optimized controlling measures are needed.

In view of above questions, it becomes current research keystone about how to conduct effective behavior management in the premise of not affecting the normal use, and how to conduct traffic forecasting before a large amount of data has not come and the network is not blocked. Under this situation, the perception technology of network behavior arises at the historic moment, and already became the important means for people to understand network, manage network, and better use network. At present, it is gradually independent from the network management discipline, becomes an important research branch in computer network domain. So, it has both the market prospect and the scientific research value.

It is discovered through research and analysis that user's behavior in network may be reflected through some kind of quota measurement [4]. In other words, we can take the change of technical data as the mapping of user behavior from the reality society to the network world. So, Network traffic measurement and analysis is the foundation of Internet behavior study. Through measurement and analysis, the basic features of Internet behavior can be mastered, and then the variation patterns of network behavior can be found, the mathematical model can be built and the network behavior can be verified. At present, the major organizations in network measurement and behavioral study aspects are MERIT, CAIDA, ITA, OC3 and IETF, and so on. The main results are focus on data collection of network equipments, from the history data of measurement to master the status of network running, in order to anticipate possible problems and reasonably arrange network load.

2. The Survey of Recognition Methods of Network Traffic

Want to sense network behavior, the key is the effective recognition of the network traffic. Using different technology to accurately distinguish each kind of flow in network according to different characteristics of flow is needed. The typical traffic recognition technologies have the following kinds [5]:

(1) The method based on port number. This method only needs to search the port--application mapping table which is established beforehand. It is simple and fast. But as the existence of hidden ports, it is unable to distinguish accurately different applications. So, it is unable to effectively sense network behavior and manage network.

(2) The method based on application layer characteristic signature. This needs to discover and withdraw the special string which each application layer load posses, and carry out in-depth examination for each packet. This method has very high accuracy, but the matching time is long, and the efficiency is low. Moreover, it can’t distinguish the

flow with encryption. And, the feature signature of application layer protocol is sometimes only in session start of some messages, so it is difficult to capture the session initiation packet.

(3) The method based on the statistical features of flows. Most network application will exchange some controlling information before the actual data transmission, and the interaction manner of different network applications posses its own features, so the different flows can be identified at the IP layer through the way of traffic statistics. This method is commonly used to identify the flow from some encrypted or non-public application protocols.

(4) The method based on machine learning. It first uses some training data (sample) to establish a classifying model, then produces a classifier based on the model and then classifies unknown data. This method recognizes based on the statistical characteristic of flows and standard sample of flows. It has the feature of intelligent discovery for new flows. Moreover, its extendibility is good, the performance is high, and may distinguish the flow with encrypted and changeable ports. But the accuracy is not very high. Relatively, the flow subdivision ability is insufficient. However this method has big potential, is the research focus, nowadays.

Thus, it can be seen that the traditional traffic recognition methods which take the port number and the features of application layer load as the foundation is difficult to deals with the anti-monitor technology, such as hidden ports, the random ports and the encryption of application layer data, and so on. The method based on machine learning does not rely on the load content of a message. The classified accuracy is mainly depends on the selection of the classify models and the training samples. The limitation of this kind of method lies in the limited application types which may be distinguished. The classified accuracy of this method is inferior to the method based on payload. Even so, the machine learning method based on statistical characteristics of flows has been widely concerned by researchers. But the accuracy of recognition, real-time processing ability, auto-learning ability and discovery ability of new application, and so on, is the challenge need to face.

Through the comparison and analysis of accuracy from several methods of machine learning, we choose the BP ANN to realize abnormal network traffic recognition. To improve the convergence speed, the optimized BP ANN abnormal network traffic identify model and algorithm is proposed.

3. The Basic Principle of BP ANN

The artificial neural network belongs to supervised machine learning method. Through learning and training for samples, it may remember the complex relations of objective things in spatial and time aspects. It especially fit for solving each kind of problems such as prediction, classification, recognition, and so on. BP ANN is a kind of counter-push learning algorithm with multi-layered network. It is one of the most widespread and successful neural networks in application at present. Its basic principle [6] is as follows:

Input signal X_i affect the output node through an intermediate node (hidden layer), by a non-linear transformation, generate an output signal Y_k. Each sample by network training includes: the input vector X and the desired output t, the deviation between network output value Y and the desired output value t. By adjusting the value of weight W_{ij} between the hidden layer node and the input node, the value of weight T_{jk} between the hidden layer node and the output node, and the threshold, make the error down along the gradient direction. After repeated learning and training, the network parameters (weights and thresholds) corresponding to the minimum error is determined, and the training will cease. At this point, the trained neural network is able to self process the entry

information of similar samples, output the information that error is minimized and non-linear conversion has been carried out.

The models of BP ANN include: input/output model, the action function model, error calculation model and self-learning models.

(1) Input/output model

The input model of hidden node:

$$O_j=f(\sum W_{ij}\times X_i-q_j) \quad (1)$$

The output model of output node:

$$Y_k=f(\sum T_{jk}\times O_j-q_k) \quad (2)$$

Among them, f is the non-linear action function; q is the threshold of neural unit.

(2) The model of action function

The action function is a intensity function of stimulation pulse which is the reflection of underlying input to the upper node, also known as the stimulus function, generally taken to be a Sigmoid function with continuous value within (0,1):

$$f(x)=1/(1+e^{-x}) \quad (3)$$

(3) The model of error calculation

Error calculation model is the function of error between desired output and calculated output of the neural network:

$$E_p=1/2\times\sum(t_{pi}-O_{pi})^2 \quad (4)$$

Among them, t_{pi} is the desired output of node i; O_{pi} is the calculating output of node i.

(4) The model of self-learning model

The learning process of neural network is the setting and error correction process of weight matrix W_{ij}, which connect the lower layer node and upper node. Self-learning model:

$$\triangle W_{ij}(n+1)=\eta\times\Phi_i\times O_j+a\times\triangle W_{ij}(n) \quad (5)$$

Where, η is the learning factor; Φ_i is the calculation error of the output node i; O_j is the calculation output of output node j; a is the momentum factor.

4. The optimization of BP ANN model and corresponding algorithm

4.1 The design and optimization of network structure

Although the simulation results of neural network with multiple hidden layers may be better, but the solution is too complex. So, it is difficult to be applied. Theoretical study shows, a BP network with three layers has been able to map or approximate any rational function [7]. Therefore, we use BP network with a hidden layer.

After the multiple regression analysis method was used to treat the 24 parameters initially determined for network traffic recognition, we found that, in which some parameters are strongly correlated. Therefore, we deleted some input parameters which are strongly correlated. Finally, the number of effective parameters is determined as 14. Thereby, the number of input nodes is reduced.

Because we only need to determine whether the flow is normal or abnormal, so, the output layer only needs 2 output nodes.

The number of hidden layer nodes has greater impact on network performance. With too many hidden layer nodes, it will lead to online learning time too long, even can't converge. With few hidden layer nodes, the fault tolerance of network is poor. So, we can use the stepwise regression analysis method to test the significance of parameters, and dynamically delete some linear-related hidden nodes. The removal criteria of nodes is: When all weight value and thresholds value starting from the node to the next level node were down in the dead zone (usually taken ± 0.1, ± 0.05 range, etc.) , then the node can be removed[4]. The best number l of hidden layer nodes may refer to following formula:

$$l = (m+n)1/2+c \qquad (6)$$

Wherein, m is the number of input nodes; n is the number of output nodes; c is a constant between 1 and 10. Take C=5, the number of the hidden layer nodes can be calculated, it is 9.

So, the structure of neural network for abnormal network traffic recognition is 14-9-2. It is as figure1.

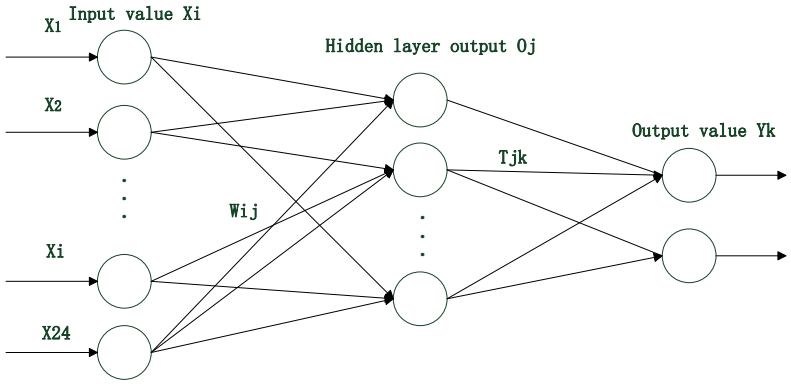


Figure 1. The Structure of BP ANN

4.2 The optimization of learning factor η

The learning factor η is called step, it is a constant in standard BP ANN. But in the actual calculation, it is difficult to give an optimum learning rate from start to finish. It can be seen from the error surface, the training number will increase when η is too small in a flat area. When, we hope the value of η is larger. But in the dramatically changed error area, if η is too large, it will stride across the narrow "pit" because of excessive adjustment, make training emerge the circumstance of shock, and increases the number of iterations on the contrary. In order to accelerate the convergence process, the best way is adaptively adjusting learning rate η. The adjusting method is as follows:

After a batch of weight adjustment, If E_p is increasing, then this adjustment is invalid, and

$$\eta = \beta \cdot \eta \qquad 0 < \beta < 1$$

If E_p is decreasing, then this adjustment is effective, and

$$\eta = \alpha \cdot \eta \qquad \alpha > 1$$

We think, adaptive adjustment method of learning factor is the simplest and most effective, and is the easiest to implement. The method can improve the learning speed of BP ANN to some extent. But the disadvantage of this method is that, α and β used to change learning factor are remained constant. If their values are selected incorrectly, the problem of low BP ANN learning efficiency is still exists. In order to overcome the disadvantage of this method, further optimized strategy of BP ANN is proposed in this paper.

The basic idea is as follows: on the basis of self-adapting adjustment method of learning factor, the concept of membership function in Fuzzy Mathematics is introduced. Depending on the error E, the parameters α and β which are used to change the learning factor have different values. As a result, this algorithm can make the learning process of network better self-adaptable and raise the learning speed.

This algorithm takes $E_p \rightarrow 0$ as a fuzzy concept. Thereby, depending on the different grade of E_p tend to 0, the parameters α and β have different values. In other words, according to the different grade of E_p tend to 0, learning factor η can be dynamically changed.

Let discourse domain is real number field R, and A expresses "the number set of $E_p \rightarrow 0$", so, $A \in F(R)$. Its membership function is:

$$A(x) = e^{-x^2}$$

The relationship between learning factor η and the total error E_p is shown in Figure 2.

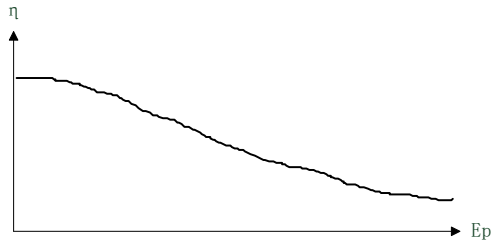


Figure 2. The Relation between Learning Rate η and Ep

It can be seen from Figure 2 that, the learning factor η of BP ANN will be decreasing with the increasing of total error E_p. Thus, the learning efficiency is effectively improved. Adopting the further optimized BP ANN, in the flat region of the error curved surface, the value of η can be automatically increased when it is too small, then, avoided the increasing of the training number by reason of the value of η is too small in the flat region. In the region of errors changes sharply, the value of η can be automatically decreased when it is too big, then avoided that the narrow "pit" is stridden, which results in training instability and increasing the number of iteration, by reason of the value of η is too big.

Corresponding algorithm is as follows:

(1) Set the initial value for weight W_{ij}. Weight value of various layers is set a smaller non-zero random number.

(2) Enter a sample $X=(x_1, x_2, \cdots, x_n)$, *and* the desired output $Y=(y_1, y_2, \cdots, y_n)$.

(3) Calculate the output of various layers based on forward propagation algorithm of BP network.

(4) Calculate learning error e_{ij} of various layers and the total error E_p.

(5) Determine the new size of step η, according to the degree of total error E_p approach to 0.

(6) Amend the weight W_{ij} and threshold value Φ by using standard BP algorithm.

(7) Judge whether or not meeting the requirements, according to the given smallest error, after obtaining coefficient of each layer. If the requirement is met, the algorithm is ended; otherwise, the algorithm continues to be executed from step (3).

5. Experiment Process and Experiment Results

5.1 The Data Collecting

The data source is divided into two parts, one part is taken from the campus network of Beijing Information Science and Technology University, in time for 14:00, March 6, 2013, a total of 2.9G data (2966M), 12 blocks, 4748375 data packet; The other part from the Genome Campus of University of Cambridge, a total of 0.93G, contains 5 data blocks, a total of 4956477 packets.

First, 30% of the data is extracted from the data set and taken as experimental data, then randomly divided the extracted data into two subsets, a training set, and a test set. In order to simulate real network environment, both data sets contain various types of data samples, and the approximate percentage of each type.

5.2 The Experiment Procedure

(1) Define the packages with same source IP address, same source port number, same destination IP address, same destination port number, and same transport protocol as a flow. And at the same time, extract sustained time length of a flow, the bytes of a flow, packets in a flow, forward or backward length of a packet, arriving time interval of packets (maximum, minimum, mean, standard deviation) as the network traffic characteristics, common for abnormal network traffic identification.

(2) Take the extracted feature items of a flow as the entry of BP ANN. Through the iterative calculation of characteristic values using BP ANN model, then the output term of BP ANN can be obtained.

(3) Using sniffer software produces the desired different types of experimental packages, such as: FTP, P2P, HTTP, Telnet, POP3, IMAP4, SNMP, VPN, H.323, etc. Then take these packets as standard samples to train BP ANN. And, continuously adjust the weight of each feature item based on the accuracy of recognition results, until it reaches a satisfactory accuracy. So far, the training process is completed.

(4) Apply the trained BP ANN model to the actual network environment, with identification of network traffic. Identify those flows without normal characteristics, they are the abnormal flow.

5.3 The Experiment Results

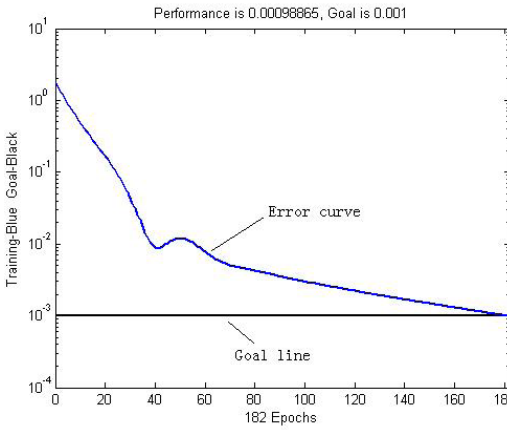
To illustrate the traffic recognition effect of the BP ANN method in real-time, we take Thunder, Bit Torrent, e-Donkey three kinds of P2P flow as the test objects. Using Bayesian, BP ANN and Support Vector Machine (SVM) three models available in Matlab tool box to test and compare. Test results are shown in table 1.

Table 1. Recognition Accuracy and Training Time of Different Machine Learning Method

Accuracy(average)			Training time(average)/s		
Bayesian	BP ANN	SVM	Bayesian	BP ANN	SVM
0.05%	95.62%	94.85%	1.242	30.262	0.924

We can see from table 1 that, although the training time of BP ANN method is longer than the SVM and Bayesian method, but the recognition accuracy is higher than SVM and Bayesian method. For real-time identification of small sample training, the accuracy of 95.62% fully meets the needs of fast, efficient and less resource-intensive requirements of real-time identification required. Thus proving that, compared with other two machine learning methods, BP neural network method has a greater advantage.

To illustrate the superiority of the optimized BP ANN algorithm, we give the error curves of the standard BP algorithm and the optimized BP algorithm, shown in Figure 3 and Figure 4 by each.



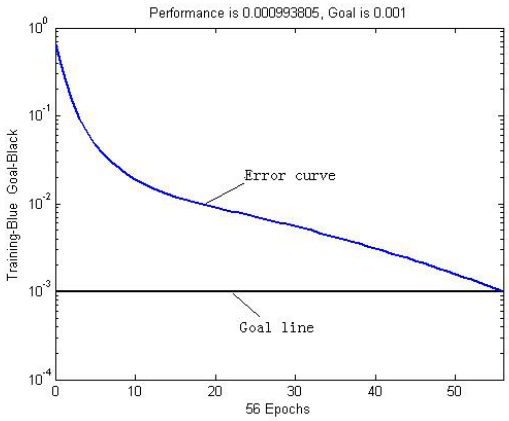


Figure3. The Standard BP Algorithm **Figure 4. Improved BP Algorithm**

From the figures above we will see that the error precision between training and goal can be implemented in 182 steps by standard BP ANN algorithm, but in 56 steps by optimized BP ANN algorithm. So, the optimized BP algorithm is reasonable, effective and feasible.

6. Conclusion

On the basis of extracting recognizable features of abnormal network traffic, we select the models of Bayesian, BP ANN and SVM to carry out identifies studies, and obtained important experiment data. The result of data analysis showed that, although the training time of BP ANN is significantly longer than the Bayesian and SVM, but its recognition accuracy is the highest. In order to improve the convergence speed, we determine the parameters α and β which change the learning factor η according to the membership function. So, the learning process of BP ANN has better adaptability. Thus, the learning speed of BP ANN is improved, and can achieve rapid identification of abnormal network traffic. Comparative experimental results show that, the convergence rate of optimized BP ANN model is significantly improved.

Through applying the model of improved BP ANN to identify network traffic, the abnormal flow can be recognized effectively. Furthermore, help us to sense the behavior of the illicit flow. And, depending on special network management strategy, the recognized illicit flow can be suppressed or limited in accordance with the occupation ratio of whole network bandwidth. The network bandwidth can be managed efficiently in this way and we can even make different charging strategy according to different flows.

Acknowledgement

This paper is supported by:

1. The National Natural Science Foundation of China under Grant No. 61370139;
2. The Beijing Key Laboratory of Internet Culture and Digital Dissemination Research under Grant No. ICDD201309;
3. The Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges Under Beijing Municipality No. IDHT20130519)

References

[1] L. Xizi and X. Yabin, "Design of Peer-to-peer Traffic Classification System Model Based on Cloud Computing", Applied Mechanics and Materials, vol.1, no.182-183, **(2012)**, pp.1347-1351.
[2] Q. Chenxi and X. Yabin. "A Fast Identification Approach to Social Network Trafic Based on Unsupervised Learning, Mathmatics In Practice and Theory, vol.3, no.44, **(2014)**, pp.100-107.
[3] Fowler and Scott, Impact of Denial of Service Solutions on Network Quality of Service, Security and Communication Networks, vol.4, no.10, **(2011)**, pp.1089-1103.
[4] He Weisong, "Network Traffic Sensing Method Study of Abnormal Behavior in Backbone", University of Electronic Science and Technology [D], **(2011)**.
[5] J. Zheng and Y. Xu, "Identification of Network Traffic Based on Support Vector Machine", Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering, **(2010),** pp.3286-3290.
[6] byxdaz, "The Basic Principle of BP Neural Network", http://blog.csdn.net/byxdaz/article/details/ 534855, **(2005)**.
[7] C. Dongwen, "Application of Hidden Multilayer BP Neural Network Model", Journal of China Hydrology, vol.1, no.33, **(2013)**, pp.68-73.

Author

Xu Yabin, (1962-), Professor, the main research areas are network measurement, Cloud computing, data mining and knowledge discovery