Applied Mathematics & Information Sciences
An International Journal

Research on Network Traffic Identification Based on Improved BP Neural Network

Shi Dong[1,4] *,DingDing Zhou*[2]*,Wengang Zhou*[3]*,Wei Ding*[1]*,Jian Gong*[1]

[1] School of Computer Science and Engineering, Southeast University, Nanjing 211189, CHINA
[2] Laboratory and Equipment Management Office, Zhoukou Normal University, Zhoukou 466001, CHINA
[3] School of Information and Computer Sciences, University of California, Irvine,USA
[4] School of Computer Science and Technology, Zhoukou Normal University, Zhoukou 466001, CHINA

Received: 26 Jan. 2012; Revised 25 Sep. 2012; Accepted 8 Oct. 2012
Published online: 1 Jan. 2013

Abstract: Traffic identification is a key task for any Internet Service Providers (ISP) or network administrators. Neural network is an important research method on traffic classification, this paper introduces the important methods of traffic classification,through study on Principal Component Analysis(PCA) and BP neural network. An improved BP neural network to identify traffic is proposed and $MOORE\_SET$ is used as dataset, meanwhile, building NOC_SET dataset based on CERNET(China Education and Research Network).the experiment results show that the accuracy rate of traffic classification based on the improved BP neural network model is relatively high.Finally,this paper analyzes packet sampling impact on traffic identification.

Keywords: Traffic identification,Principal Component Analysis, BP neural network, MOORESET, CERNET.

1. Introduction

Traffic identification plays an important in many fundamental network operations and maintenance activities to detect invade and malicious attacks forbid applications, bill on the content of traffics and ensure quality of service. It increasingly becomes one of the most interesting topics in network science and technology fields, especially in recent years. The current network traffic identification methods roughly are divided into five categories,(1)Port-based method; (2) Deep packet inspection(DPI) methods; (3)Network flow characteristic; (4)Host behavior [1]; (5)Machine learning methods. The machine learning methods are divided into supervised and unsupervised machine learning. These are the more classic classification method; of course, there is also individual QOS quality of service features for classification [2].Through studying on BP neural network method,we found some problems which can not achieve high identification accuracy and in order to solve some classification problems based on BP neural network. We propose a principal component analysis method (PCA) to improve BP neural network method. Experiments results show that the new BP neural network classification algorithm PCA_BP can achieve better accuracy than others. The paper is structured as follows: (1)Section 2 introduces related work of traffic identification;(2)Section 3 proposes improved BP neural network algorithms and the accuracy of traffic identification are evaluated based on the MOORE-SET;(3)At last,we list the proportion results which are classified by our identification algorithm.

2. Related work

The application identification problem has been changing due the efforts of two factors that are in a continuous competition. first, the applications, and especially those that do not want to be detected (e.g., P2P applications), in order to use the network resources without control. On the other hand, a group of network operators, investigators and even ISPs who need to know the traffic characteristics of their networks to manage the resources or even charge the users depending on their consumption.

* Corresponding author: e-mail: shdong@njnet.edu.cn

2.1. Research on traffic identification

Traffic identification has become a hot research, and it is research foundation of QOS, intrusion detection, traffic monitoring, billing and management. At the beginning of the study on port-based method, this method is used as marking and identifying the traffic type by fixed port which supplied by the IANA, the other method is aim at P2P and some certain protocols, which adopt method based on deep packet inspection, but this method has defect that can't get some encrypted information and can't get the new service type. Recently traffic identification has new method. With appearance of the new service,ML method has been applied to the traffic identification. In the fields of traffic identification, there are roughly three research directions:(1)Feature selection algorithm [3–8] (2)Classification algorithm [1,2,9–14](3)Different data sets, for example, which can be divided into full packets or sampled netflow flow [15–18] [19].Complementary information about related work in the field of traffic classification can be found in the survey of traffic classification techniques using machine learning in [20], the comparison of contemporary classification methods are presented in [17], the survey on Internet traffic identification is explored in [21] and the research review on traffic classification is shown in [22].A critical but constructive analysis of the field of Internet traffic classification is proposed in [23],and progress and suggestions are given. Although some articles have been studied on the classification algorithm, but the classification algorithm still exist some problems that need to be solved, such as the neural network classification algorithm is one point worthy of study. All previous research studies in traffic identification either use insufficient network data, usually non-public, or use very few/meaningless metrics for evaluation, making it impossible to compare results shown in different papers[21].In addition,features selection based on flow, especially the impact of the size of packet traffic is always to be concerned. Therefore, in this article we propose PCA_BP method based on improved BP neural network algorithm, while the size of different packets will influence results of classification. And we also analyze the impact of different sampling rate on classification results.

2.2. BP neural network Algorithm

BP neural network, also known as back propagation neural network, which is feed-forward network composed by the nonlinear transformation unit . BP learning algorithm is the training of the basic methods of artificial neural networks.

2.2.1. Introduce BP neural network algorithms

BP neural network:
It is a common method of training artificial neural networks so as to minimize the objective function. Arthur E. Bryson and Yu-Chi Ho described it as a multi-stage dynamic system optimization method in 1969.It wasn't until 1974 and later, when applied in the context of neural networks and through the work of Paul Werbos,David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams that it gained recognition, and it led to a "renaissance" in the field of artificial neural network research.BaoJian *et al.*[24] proposed a theoretical result that an integer network can present a good performance on approximating continuous function. It is a supervised learning method, and is also a generalization of the delta rule. It requires a dataset of the desired output for many inputs, making up the training set. It is most useful for feed-forward networks (networks that have no feedback, or simply, that have no connections that loop). The term is an abbreviation for "backward propagation of errors". BP requires that the activation function used by the artificial neurons (or "nodes") be differentiable.

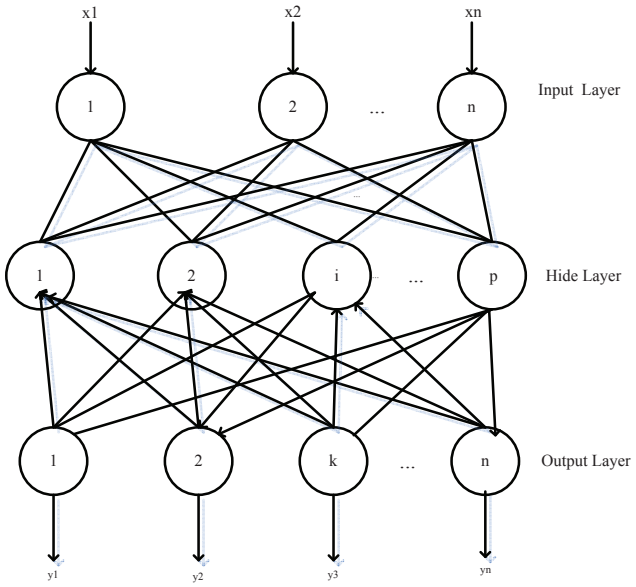


Figure 1 structure of BP neural network

3. Methodology

3.1. PCA_BP method

Suppose there are n samples, each sample has p features, then construct the n*p matrix. As follows:

$$A = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad (1)$$

When features number p of the samples are very large which enlarge dimensions of the sample,theoretically, having more features should result in more discriminating power. However, practical experience with machine learning algorithms has shown that this is not always the case. Many machine learning algorithms can be viewed as making a (biased) estimate of the probability of the class label given a set of features. This is a complex, high dimensional distribution. Thus we have to select best features to identify traffic.Moreover most algorithms will take into account the accuracy of classification results,and then analyze each feature, find out the optimal set of features.The PCA algorithm is used to deduce the input dimensions of BP neural network and predigest model complexity, and find these features to better reflect the more independent of each other for identification. These new features can be made better classification result. Assuming the original features is x_1,x_2,...,x_p,then new features will be z_1,z_2...,z_p.

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \quad \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases} \tag{2}$$

(1)z_i and z_j $(i \neq j; i, j = 1, 2, ..., m)$ has nothing to do with each other;
(2)z_1 is the biggest variance in all linear combinations x_1,x_2, ...,x_p;and z_2 is the biggest variance in all linear combinations x_1,x_2,...,x_p which is not related to z_1;z_m is the biggest variance in all linear combinations x_1,x_2,...,x_p which is not related to z_1,z_2,...,z_{m-1}. This new feature set $\{z_1$, z_2,...,$z_p\}$ is principal component of the old feature set $\{x_1$,x_2 ,...,$x_p\}$, the variance in turn decreases from the z_1 to z_m. then we can select feature of few larger variance as principal components. Introduction of principal component analysis method is described,so the steps of principal component analysis are summarized as follows:
(1)Calculate the correlation coefficient matrix

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ \vdots & & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix} \tag{3}$$

In the formula (3),r_{ij}(i,j =1,2,...,p) is correlation coefficient of x_i and x_j , which is calculated as

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x_i})^2 \sum_{k=1}^{n}(x_{ki} - \overline{x_i})^2}} \tag{4}$$

R is real symmetric matrix (ie $r_{ij} = r_{ji}$), so you simply only calculate the elements of upper triangular or lower triangular to achieve the goal.
(2)Calculate the eigenvalues and eigenvectors
First, the characteristic equation solutions $|\lambda i - r| = 0$ find the eigenvalues λi(i=1,2,...,p),and order them by size, that $\lambda 1 \geq \lambda 2 \geq ..., \geq \lambda p \geq 0$; and then were computed the corresponding eigenvalue λi and eigenvector $ei(i = 1, 2, ..., p)$.
(3)Calculate the principal components and the contribution rate of cumulative contribution
Contribution rate of principal components: $ri/\sum_{k=1}^{p} rk(i = 1, 2, \cdots, p)$
Contribution rate of cumulative contribution: $\sum_{k=1}^{m} rk/\sum_{k=1}^{p} rk$
Generally the rate of 85-95% of total contributions to eigenvalues $\lambda 1$, $\lambda 2$, ...,λm corresponding to the first, second... mth (mp) principal components.
(4)Calculate the principal component loading:

$$p(z_k, x_i) = \sqrt{r_k} e_{ki} \tag{5}$$

It can be further calculated principal component scores:

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1p} \\ \vdots & & \ddots & \vdots \\ Z_{p1} & Z_{p2} & \cdots & Z_{pp} \end{pmatrix} \tag{6}$$

Algorithm 1: PCA_BP algorithm

```
// Initialize the weights in the
   network (often randomly)
A = 0;
for each i ∈ [1, 249] do
    if sum(c0(1 : i)) < 90 then
        // c0 is Cumulative contribution
           rate
        A = i;
        A++;
        PCACOV(Ptrain);
        PCAprincom(Ptrain);
        generated the training set e_train;
        PCACOV(Ptest);
        PCAprincom(Ptest);
        generated the testing set e_test;
        O=neural-net-output(network, e_train);
        // forward pass
        T=output for e_train;
        Calculate error (T-O) at the output units;
        Compute Δ_wh for all weights from hidden layer
            to output layer;// backward pass
        Compute Δ_wi for all weights from input layer to
            hidden layer;// backward pass
            continued
        Update the weights in the network;
        Until all examples classified correctly or
        stopping criterion satisfied;
        Return the network;
    else
        Goto exit
```

Algorithm PCA_BP presents the training set and test set by PCA.The sequence of steps that we show in Figure.2.The procedure mainly set cumulative contribution $rate < 90$ and calculate the PCA data for training and testing data set.With these data, we re-enter as the BP network to train and test data. The process of machine learning classification is shown in Figure.2:
1.Collecting traffic: Collecting network data from network traffic and organizing the flow from traffic.
2.Selecting traffic features: Optimal selecting the known traffic features through feature selection algorithms.
3.Classified the traffic by machine learning algorithm: Using the machine learning classification algorithm to classify network traffic data.

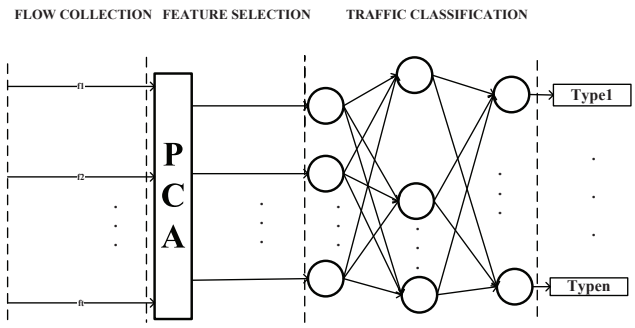


Figure 2 Process of Machine learning, traffic classification

3.2. Algorithm Evaluation

In this paper, we use the routine evaluation standard for verifying the effectiveness of our identification algorithm. The effectiveness of the current flow identification algorithm has the following three concepts evaluation criteria. And the concepts involved are as follows:
TP (true positive): The flows of application A are classified as A correctly, which is a correct result for the classification;
FP (false positive): The flows not in A are misclassified as A. For example, a non-P2P flow is misclassified as a P2P flow. FP will produce false warnings for the classification system;
FN (false negative): The flows in A are misclassified as some other category. For example, a true P2P flow is not identified as P2P. FN will result in identification accuracy loss.
The calculating methods are as follows:
Precision:
The percentage of samples classified as A that are really in class A
$Precision = \frac{TP}{TP+FP}$(6)
Recall:
The percentage of samples in class A that are correctly classified as A
$Recall = \frac{TP}{TP+FN}$(7)
Overall accuracy:
The percentage of samples that are correctly classified
Overall accuracy= $\frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i+FP_i)}$(8)

4. Experiment

4.1. dataset

4.1.1. NOC_SET dataset

In order to validate the method and analyze the impact factor,we adopt NOC_SET as dataset.as is shown in table 1.We collected network traffic data from southeast university,and use ourself $l7\_filter\_modify$ software to label the flow.$l7\_filter\_modify$ is developed based on L7filter [25].at last,we generated NOC_SET dataset.

4.1.2. $MOORE\_SET$ dataset

This $MOORE\_SET$ data was randomly sampled in several different periods from one node on the internet. This site was shared by about 1,000 researchers, technicians and management staff of three research institutions, and connected to the Internet through a full-duplex Gigabit Ethernet link. Each full-duplex traffic on this connection was captured in a full 24 hours period, so the original traffic-set contained all full duplex traffic connected the node in both link directions. Since the original traffic-set is too large, Moore divides it into ten subsets by randomly sampling method.Detail $MOORE\_SET$ dataset showed in table 2.

4.2. Feature metric

A.Moore *et al.* [26] collected 249 kinds of features of the network flow,while many features are interrelated,leading to large quantities of computation and low identification accuracy.So in this paper we choose some flow metrics as shown from table3.Metric features are composed of network behavior characteristic with label.Which is labeled by l7filter software.

4.3. Comparison of classification algorithm with Moore-set dataset:

In this paper, we adopt experimental data based on the MOORE-SET data set [2],and Use MATLAB,WEKA,the corresponding classification algorithm [27]. MOORE-SET data firstly are divided into two data which are respective

Table 1 NOC_SET dataset

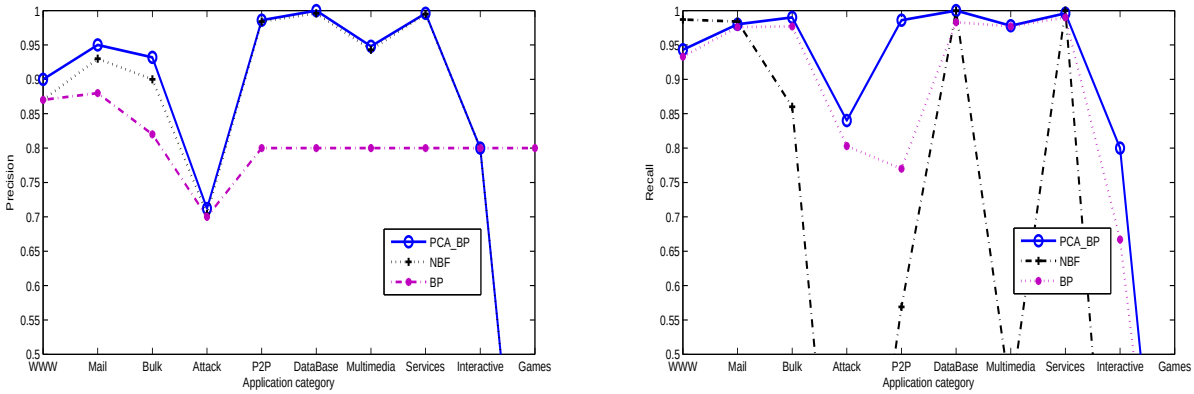
AppID	Application	Protocal	Flow number	Proportion
1	WWW	HTTP	4943	64.6
2	Bulk	FTP	39	0.5
3	Mail	IMAP,POP3,SMTP	91	1.19
4	P2P	BitTorrent,eDonkey,Gnutella,XunLei	1414	18.5
5	Service	DNS,NTP	433	5.7
6	Interactive	SSH, CVS, pcAnywhere	6	0.08
7	Multimedia	RTSP,Real	20	0.3
8	Voice	SIP,Skype	276	3.6
9	Others	games, attacks	431	5.6

Table 2 $MOORE\_SET$ dataset

AppID	Category	Application	Flow number	Proportion
1	WWW	HTTP,https	328091	86.91
2	BULK	FTP	11539	3.056
3	MAIL	pop3,Imap,Smtp	28567	7.567
4	DB	Sqlnet,Oracle	2648	0.701
5	SERV	Dns,Ntp,Ldap	2099	0.556
6	P2P	Kazaa,Bittorrent,Gnutella	2094	0.555
7	ATTACK	Worm and virus Attacks	1793	0.475
8	MULT	Media Player,Real	1152	0.305
9	INT	ssh,klogin,Telnet	110	0.029
10	GAME	Halflife	8	0.002

Table 3 bidirectional flow feature

Feature	Feature Discription
lport	low port number
hport	high port number
duration	Flow duration
Transproto	Stream transport protocol used (TCP / UDP)
TCPflags1	TCP header flag, or (OR), transport layer protocol is UDP, the feature is 0
TCPflags2	TCP header flag, or (OR), transport layer protocol is UDP, the feature is 0
pps	Packets/duration
bps	bytes/duration
Mean packets arrived time	duration/packets
Biodirection Packets ratio	Forward packets/ backward packets
Biodirection Bytes ratio	Forward bytes/ backward bytes
Biodirection Packet length ratio	Biodirection packets length ratio
Biodirection packets	Forward packets + backward packets
Biodirection bytes	Forward bytes + backward bytes
tos	Biodirection TOS OR from NETFLOW
Mean packet length	Biodirection bytes/Biodirection packets



(a) Precision (b) Recall

Figure 3 Results of traffic classification using MOORESET

Table 4 the classification Overall accuracy rate of BP,PCA_BP,NBF(NaiveBAYES+FCBF)

Classification	Overall accuracy
BP	57.8933%
PCA_BP	99.1353%
NBF	99.6742%

20% test data and 80% training data, and then the data set collected in the same 1% proportion are analyzed by PCA method. And we compared our method that is PCA_BP with BP. In order to evaluate and analyze effectiveness of the method about PCA_BP.We study traffic classification distribution. From Figure.3,we can see that the data is divided into 10 categories(WWW,MAIL,BULK,DB,SERV, P2P,ATTACK,INTERACT,GAMES,MULT). Classification accuracy of improved BP method has been greatly increased, especially on the accuracy of database and service reflected more apparent. Classification overall accuracy of PCA_BP are shown in Table 4 based on MOORE-SET data. Table 4 indicated the improved BP neural network achieved better result than normal BP method; moreover, overall accuracy almost is same to NBF.From Figure 3(a),(b) we can see the precision and the recall of P2P and the attack have improved. The reason for high accuracy is that the proportion of P2P and ATTACK is relatively small, it will result in training BP neural network to be affected, and PCA_BP method exact high dimension features to low dimension through the principal component analysis, the impact of sample imbalance on the classification results reduce to a minimum.MOORE-SET data set has 249 features, dimensions of time and efficiency of classification is too large will have a negative impact,Lim,Y *et al.*[28] uses only 36 features which include:protocol,source and destination ports,the number of packets,transferred bytes,start time,end time,duration,average packet throughput and byte throughput,the size of the first ten packets,max/min/average standard deviation of packet sizes and inter-arrival times,the number of TCP packets with FIN, SYN, RSTS, PUSH,ACK, URG (Urgent), CWE (Congestion Window Reduced),and ECE (Explicit Congestion Notification Echo) flags set (all zero for UDP packets).This paper also build NOC_SET dataset which is constructed by bidirectional flow characteristic.

4.4. Comparison of classification algorithm with NOC-SET dataset

Experimental data for the NOC_SET data set (Table 5 as fellows)is used in experimental platform. The actual measured IP trace is analysis data [29], and the collecting site is a 10G backbone channel on Jiangsu Province border of CERNET. We adopted DPI method to mark Flow and generated NOC_SET dataset. In this section, we use the first few packets as statistical data, respectively, analyze impact of the first 1-10 packets on the results of the classification and compared PCA_BP method with NBF and BP. Traffic classification result is showed in Figure.4.

As shown in Table 5,Classification result indicates that PCA_BP could achieve better accuracy compared with BP neural and NBF.But observing from Inter and Service, classification accuracy of PCA_BP is lower than the NBF method.From Service to Inter types,Precision of NBF meth od is reduced, while the PCA_BP is in increments, so that NBF method is easily affected by the number of training samples, while the PCA_BP is not vulnerable to the impact of the training sample dataset.Bernaille *et al.*[30] mentioned packet size would affect traffic classification, showed by Figure.4,experiment results show,The statistics

Table 5 Classification performance(Precision and Recall) for NOCSET

Category	Algorithm					
	BP		PCA_BP		NBF	
	Precision	Recall	Precision	Recall	Precision	Recall
WWW	98%	100%	99%	100%	98.5%	99.2%
P2P	58%	100%	75%	100%	93.7%	91.2%
Mail	83%	91.3%	90%	99%	100%	100%
Service	58.90%	100%	70%	99%	90%	90.4%
Inter	84.5%	100%	81%	100%	84%	100%
Multimedia	100%	75%	90%	80%	60%	100%
Voice	35%	50%	45%	55%	37%	50%
Others	44%	46%	48%	77%	45%	60%
Overall Accuracy	68.83%		89.2357%		86.7764%	

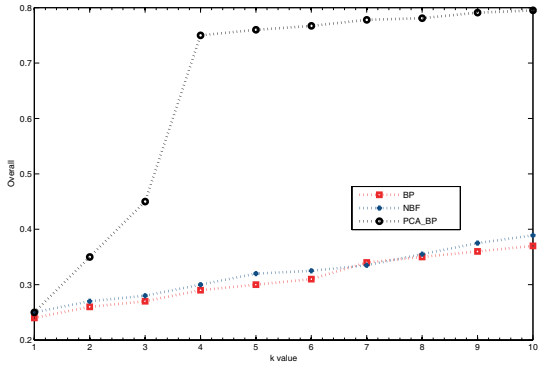


Figure 4 The impact of first several packet size on classification

characteristic of first several packets in each flow could accurately classify network traffic. Among three classification algorithm BP, PCA_BP, NBF, the overall accuracy of the PCA_BP algorithm is highest.Another noteworthy that we found the first 4 and 5 packets are the key point of traffic classification accuracy mentioned by articles [30].

4.5. Impact of sampling rate

Nowadays,packet sampling method mainly adopt random packet sampling method which is sampling ratio 1/N.Packet sampling not only reduces number of flows, more importantly, it is possible to change distribution of flow behavior and differences degree. For example, consider the distribution of flow length as feature(unit:packet),the sampling distribution of differences before and after a long flow, the following theorem:

Theorem 1. *If the flow length in two of the original distribution of the sample space was no difference,then under any random sample ratio the distribution still was no difference*

Proof 1. *Before sampling, assume the original distribution of flow length is $p_b(len)$, packet sampling ratio is p, then after sampling,distribution of flow length is $p_b(len)$:*

$$p_a(len = l) = \sum_{i=l}^{\Theta} p_b(len = i) c_i^l p^l (1-p)^{i-l} \quad (7)$$

Assume we known that the distribution of two original flow length is no difference,then for any i, two sample space $P_b(Len = i)$ are the same; and $C_i^l, p^l, (1-p)^{i-l}$ are constants. Therefore, for any l, two sample space $P_a(Len = l)$ is the same, after sampling, distribution still is no difference. Theorem 1 shows that, in unsampled environment if flow length can't considered as behavior characteristics of protocol.Meanwhile,after any sampling ratio, χ^2 statistic should be kept at a level close to 0, it is still not regarded as feature.

Theorem 2. *If original distribution of flow length are different between in both sample space,differences of sampling distribution significant difference compared with the original distribution is smaller.*

Proof 2. *Let packet sampling ratio be p, then for each individual flow, its probability of being drawn is p_{flow}:*

$$p_{flow}(len = i) = 1 - (1-p)^i \quad (8)$$

Although the different flow length, the probability of flow being drawn is different.But when the flow length is a fixed value,that is,i is regarded as fixed value,here flow sampling rate $P_{flow}(Len = i)$ is a fixed value. Therefore
1)If the two original distribution exist the differences when Len=i,In sampling procedure the larger flow samples will be taken away more of its proportion of total flow;
2)If two original distribution is no difference when Len=i,then in the sampling procedure will be taken away two samples where the same proportion of flow.

Table 6 sampling ratio impact on precision

Application \ Sampling Ratio	p=1	p=1/10	p=1/64	p=1/128	p=1/256	p=1/1024
bulk	91.36	93.48	87.05	85.37	73.02	65.43
web	92.05	93.18	89.38	82.56	71.42	53.86
mail	99	92.77	74.59	65.60	53.39	42.53
P2P	92.94	92.81	90.42	87.59	75.09	61.29
Mm	99	64.09	31.18	20.50	20.45	82.52
serv	91.34	90.42	75.30	70.25	54.45	38.62
int	92.35	86.45	62.08	58.67	41.42	21.18
voip	92.36	93.24	83.53	80.31	66.12	43.65
others	90.30	94.13	86.72	83.26	68.27	62.74
total	89.2357	90.2137	87.35	86.59	65.5	60.43

From table 6 we can see.When the packet sampling ratio p=1/128,the overall accuracy rate of PCA_BP method is over 85%, which can be used to meet the demand of protocol identification under the environment actual sample ratio. However, when a small number of sample flow protocol is less(such as interactive application type),classification accuracy is still low. In this experiment,we compared our method with well-known NBF and BP by adopting sampling rate 1/1, 1/10, 1/256.The results are shown in the Figure.5.

Observed in Table 6, we can see it's not monotonous for

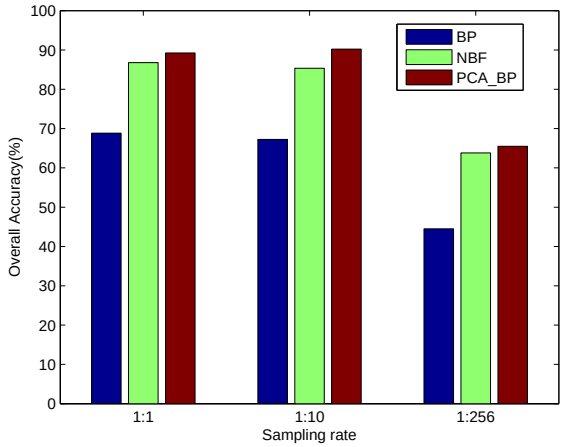


Figure 5 Impact of sampling on traffic classification results

sampling impact on application, the accuracy ratio of traffic identification is increase firstly and then decrease with the sampling ratio increased. Shown in Figure.5, when the packet sampling ratio p=1/10, the identification overall accuracy of PCA_BP and precision of most applications have been increased.Moreover when the sampling ratio is low, application behavior gets less affection, while the corresponding short flow is discarded at large probability, so that it makes application identification accuracy improved. In particular, the accuracy of OTHERS category is improved obviously due to the flow noise's lost, the accuracy ratio improves from 90% to 94%. However,with packet sampling ratio increases continually, the total identification accuracy of PCA_BP method and precision of various protocols all reduce due to the effective information and quantity of sampling flow declining, when the sampling ratio $p \leq 1/256$ and the overall accuracy less than 66%,except for WEB,BULK and P2P.Identification precision are below 70%. Further observing the table 6,with the increasing of sampling rate,we can find that the precision of flow which is reduce less become slowly decline.Formula 8 shows that the more packets appear in flow, the lower is sampling ratio of corresponding flow ,then which will be selected at larger probability.Therefore, FTP,P2P and other protocols flow in table 6, their proportion reduce slowly with the sampling ratio increases.which will show that these flows length is longer. So it can collect a larger number of flow records during sampling. At the same time, the long flow could provide the more calculable metrics and more extensive effective information for traffic identification,which make the identification accuracy of this protocol to decrease more slowly.

5. Conclusions

BP neural network is widely used in artificial intelligence fields,this paper proposed the improved BP neural network algorithm,and on this basis,the introduction of principal component analysis (PCA) model, detail study on the PCA algorithm, and adopted the improved BP neural network method (PCA_BP) to classify traffic for MOORE-SET as data set, moreover,compared with two other methods which is the BP neural network and NBF(NaiveBayes+FCBF) method,the results show that,PCA_BP neural network are greatly improved on classification accuracy.To further

prove PCA_BP method is effective,this paper collect the data in Jiangsu provincial network border and organize trace into flow record such as data sets NOC_SET,the experimental results show that PCA_BP neural network classification accuracy is high, the paper also studies that different packet size impact on classification results, the results show that the first 4 and 5 packets selected will be a key point. Moreover, we analyze the impact of sampling ratio on traffic identification and compare our method with the other two methods.Adopting our method to get much better identification results in the same sampling ratio.

6. Acknowledgments

This paper is supported by National 973 Plan Projects (2009CB320505) and National Science and Technology Plan Projects(2008BAH37B04).

References

[1] T. Karagiannis, K. Papagiannaki, M. Faloutsos, Blinc: multilevel traffic classification in the dark, in: ACM SIGCOMM Computer Communication Review, Vol. 35, ACM, 2005, pp. 229–240.

[2] A. Moore, K. Papagiannaki, Toward the accurate identification of network applications, Passive and Active Network Measurement (2005) 41–54.

[3] H. Zhang, G. Lu, M. Qassrawi, Y. Zhang, X. Yu, Feature selection for optimizing traffic classification, Computer Communications.

[4] T. Ganchev, P. Zervas, N. Fakotakis, G. Kokkinakis, Benchmarking feature selection techniques on the speaker verification task, in: Fifth International Symposium on Communication Systems, Networks And Digital Signal Processing, 2006, pp. 314–318.

[5] L. Zhen, L. Qiong, A new feature selection method for internet traffic classification using ml, Physics Procedia 33 (2012) 1338–1345.

[6] M. Sun, J. Chen, Y. Zhang, S. Shi, A new method of feature selection for flow classification, Physics Procedia 24 (2012) 1729–1736.

[7] J. ZHAO, X. HUANG, Q. SUN, Y. MA, Real-time feature selection in traffic classification, The Journal of China Universities of Posts and Telecommunications 15 (2008) 68–72.

[8] T. En-Najjary, G. Urvoy-Keller, M. Pietrzyk, J. Costeux, Application-based feature selection for internet traffic classification, in: Teletraffic Congress (ITC), 2010 22nd International, IEEE, 2010, pp. 1–8.

[9] Z. Li, R. Yuan, X. Guan, Accurate classification of the internet traffic based on the svm method, in: Communications, 2007. ICC'07. IEEE International Conference on, IEEE, 2007, pp. 1373–1378.

[10] N. Williams, S. Zander, G. Armitage, Evaluating machine learning algorithms for automated network application identification, Center for Advanced Internet Architectures, CAIA, Technical Report 060410B.

[11] N. Williams, S. Zander, G. Armitage, A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification, ACM SIGCOMM Computer Communication Review 36 (5) (2006) 5–16.

[12] A. Este, F. Gringoli, L. Salgarelli, Support vector machines for tcp traffic classification, Computer Networks 53 (14) (2009) 2476–2490.

[13] C. Yin, S. Li, Q. Li, Network traffic classification via hmm under the guidance of syntactic structure, Computer Networks.

[14] F. Palmieri, U. Fiore, A nonlinear, recurrence-based approach to traffic classification, Computer Networks 53 (6) (2009) 761–773.

[15] P. Teufl, U. Payer, M. Amling, M. Godec, S. Ruff, G. Scheikl, G. Walzl, Infect-network traffic classification, in: Networking, 2008. ICN 2008. Seventh International Conference on, IEEE, 2008, pp. 439–444.

[16] T. Kiziloren, E. Germen, Network traffic classification with self organizing maps, in: Computer and information sciences, 2007. iscis 2007. 22nd international symposium on, IEEE, 2007, pp. 1–5.

[17] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, K. Lee, Internet traffic classification demystified: myths, caveats, and the best practices, in: Proceedings of the 2008 ACM CoNEXT conference, ACM, 2008, p. 11.

[18] J. Erman, M. Arlitt, A. Mahanti, Traffic classification using clustering algorithms, in: Proceedings of the 2006 SIGCOMM workshop on Mining network data, ACM, 2006, pp. 281–286.

[19] V. Carela-Espanol, P. Barlet-Ros, J. Solé-Pareta, Traffic classification with sampled netflow, traffic 33 (2009) 34.

[20] T. Nguyen, G. Armitage, A survey of techniques for internet traffic classification using machine learning, Communications Surveys & Tutorials, IEEE 10 (4) (2008) 56–76.

[21] A. Callado, C. Kamienski, G. Szabó, B. Gero, J. Kelner, S. Fernandes, D. Sadok, A survey on internet traffic identification, Communications Surveys & Tutorials, IEEE 11 (3) (2009) 37–52.

[22] M. Zhang, W. John, K. Claffy, N. Brownlee, State of the art in traffic classification: A research review, in: PAM Student Workshop, 2009.

[23] A. Dainotti, A. Pescape, K. Claffy, Issues and future directions in traffic classification, Network, IEEE 26 (1) (2012) 35–40.

[24] B. Jian, Z. LiangJie, Y. Yi, Analysis on complexity of neural networks using integer weights, Appl. Math 6 (2) (2012) 317–323.

[25] J. Levandoski, E. Sommer, M. Strait, et al., Application layer packet classifier for linux (2008).

[26] A. Moore, D. Zuev, Internet traffic classification using bayesian analysis techniques, in: ACM SIGMETRICS Performance Evaluation Review, Vol. 33, ACM, 2005, pp. 50–60.

[27] T. Nguyen, G. Armitage, Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world ip networks, in: Local Computer Networks, Proceedings 2006 31st IEEE Conference on, IEEE, 2006, pp. 369–376.

[28] Y. Lim, H. Kim, J. Jeong, C. Kim, T. Kwon, Y. Choi, Internet traffic classification demystified: on the sources of the discriminative power, in: Proceedings of the 6th International COnference, ACM, 2010, p. 9.

[29] (2011). [link].
URL http://iptas.edu.cn/src/system.php
[30] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, K. Salamatian, Traffic classification on the fly, ACM SIGCOMM Computer Communication Review 36 (2) (2006) 23–26.

Dr Shi Dong was born in zhoukou, China, on NOV 05, 1980. He received the Master degree in Computer science from University of Electronic and technology of China.in 2009,nowadays pursuit for Ph.D degree in computer science from southeast university. major interests are network management,network security.

DingDing Zhou was born in Zhoukou, China, on February,1980. She received the Computer Science B.S degree in henan university.nowadays she is lecturer in zhoukou normal university. His major research interests are in high speed communications, mobility, security and QoS guarantees.

Dr Wengang Zhou received the Master degree in Computer science from University of Electronic and technology of China.in 2009,he is pursuing doctor degree in computer science from University of Electronic and technology of China.nowadays he is doctor of co-culture in computer science of University of California,Irvine major interests are network management,network security.

Prof Wei Ding received B.S degree in the Computer soft from nanjing university in 1982.respectively,She received M.S,Ph.D degree from southeast university,in 1987,1995.nowadays she is a prof in southeast university.Her major research interests are in high speed communications,management,security.

Prof Jian Gong received B.S degree in the Computer soft from nanjing university in 1982,he received Ph.D degree from southeast university in 1996.nowadays he is a prof in southeast university.his major interests are network management,network security.