

Traffic Identification Optimization of the Smallest Neighbor Method of AdaBoost-SVM

Yunhua Zhu, Xiao Cai

China College of Business Administration, Huaqiao University, Quanzhou 362021, Fujian, China

Abstract

The traditional P2P traffic identification has the shortcomings of low recognition rate and misjudgment rate is high. Considering the good classification ability of AdaBoost algorithm and generalization ability of SVM in machine learning, this paper puts forward a combination algorithm of a P2P traffic identification, which takes the SVM as the base of AdaBoost classifier and uses the smallest neighbor method to classify P2P traffic identification. Take the four kinds of simulations with P2P traffic data as the research object, the simulation results show that the combination of AdaBoost and SVM is better than that of pure AdaBoost and SVM algorithm in classification performance and classification accuracy, the average recognition rate is as high as 96.33% of combination algorithm. Key words: SUPPORT VECTOR MACHINE (SVM), THE SMALLEST NEIGHBOR METHOD, GENERALIZATION ABILITY, PEER-TO-PEER NETWORK TRAFFIC

1. Introduction

With the rapid development of peer-to-peer network technology, P2P technology has been widely applied in the streaming media transmission, file sharing, and instant messaging, etc [1]. At present, the P2P traffic has become the master of the Internet traffic, the rapid growth of P2P traffic has caused serious burden to the network bandwidth, intensified the congestion status of the network; meanwhile, a large number of P2P malicious traffic illegal connection intensifies the bandwidth consumption [2]. So the identification and control for P2P traffic becomes the key problems for network operators and managers have to solve.

At present, both at home and abroad, P2P technology has been familiar to us, the traffic identification theory research is also becoming more popular. The LASER algorithm was designed and implemented by Park, such as using the algorithm to extract the application layer content of longest common subsequence to as identification of the application. Bittorrent, and realized with LimeWire application such as feature extraction, using the extracted features to effectively

identify, each traffic makes non-response rates below 8.5% [3]. Randy et al proposed a finite state machine DPI algorithm has extensibility, solved the problem that storage space occupied large when determining the finite state machine down to identify P2P traffic [4]. Aceto G to use about the - filter for each session search depth was studied. Research shows that each session traffic generated in the probability of the first packet reached 72%, and the load of 32 bytes appeared before most of the character string, this shows that the beginning of the stream are of great help for P2P traffic identification [5]. Xu and others to a node of the ascending and descending traffic for the search. Train of thought of this method is that if we search out the characteristics of the series of the same, and then identify the P2P nodes, among them, they use the string matching algorithm for Rabin algorithm [6]. Risso et al. Research has shown that there will be millions of TCP session appears in gigabit levels of the network. This suggests that in stand-alone environment has difficulty on the P2P traffic Identification, makes the DPI (Deep Packet Identification) technology cannot be

applied to high-speed network [7]. Este A studied the stability of various features of packets in the network environment. it is concluded that in the complex network environment, the size of the package is affected by the network changes the least significant, and they found that for identifying the most effective packet handshake is often A TCP connection when the connection is established after the first packet [8].

This paper combining AdaBoost and SVM, and put forward a kind of efficient P2P traffic identification technology, make the SVM as classifier of AdaBoost, using the smallest neighbor method to classify P2P traffic identification, and make a comparison validation of combinational algorithm and P2P traffic identification of AdaBoost and SVM.

2. Support Vector Machine

SVM is a kind of machine learning method based on statistics which is proposed by Vapnik, mostly used to solve the Small sample pattern classification

$$\min \phi(W) = \frac{1}{2} \|W\|^2 = \frac{1}{2} (W \bullet W) \quad s.t \quad y_i [(W \bullet X_i) + b] - 1 \geq 0 \quad (i = 1, 2, \dots, n) \quad (3)$$

It is transformed into dualization problem further more [7]:

$$\begin{aligned} \min Q(\alpha) &= \frac{1}{2} \alpha^T A \alpha - b^T \alpha \\ s.t \quad \alpha_i &\geq 0 \quad (i = 1, 2, \dots, n) \\ y^T \alpha &= 0 \end{aligned} \quad (4)$$

In the formula (4): $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$,

$b = (1, 1, \dots, 1)^T$, $y = (y_1, y_2, \dots, y_n)$, $A_{ij} = y_i y_j (x_i \bullet x_j)$
The optimal classification function can be deduced through formula (4) show as formula (5):

$$f(x) = \text{sgn} \left\{ \sum_{i=1}^n \alpha_i^* y_i (X_i \bullet X) + b^* \right\} \quad (5)$$

3. AdaBoost-SVM Smallest Neighbor Method

The SVM is used as the base classifier of the AdaBoost algorithm, the Smallest Neighbor Method is used to calculate the distance of vector and the training set in order to realize the classification,

$$\min W(\alpha) = \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (X_i \bullet X_j) - \sum_{i=1}^n \alpha_i \quad \sum_{i=1}^n y_i \alpha_i = 0 \quad (6)$$

The above formula is solved and get the solution vector $\alpha^* = (\alpha_1^*, \dots, \alpha_n^*)$, if $\alpha_i^* \neq 0$, then X_i is the searching support vector. Set $t = t + 1$, the amount of the generated base classifier is carried on the iterative computation, if $t > \max \text{gen}$, then switch to Step6.

Step4:the found support vector is used to construct the positive example $X^+ = \frac{1}{C} \sum_{(x_i, y_i) \in SV^+} \alpha_i^* y_i$ and negative example

problems [9].

If linear separable sample set

$$(X_i, y_i) \quad (i = 1, 2, \dots, n; X \in R^d, y \in \{-1, 1\})$$

The discrimination function's normal pattern of $g(X) = W \bullet X + b$ dimension's linear is [3-5]:

$$g(X) = W \bullet X + b \quad (1)$$

The Classification surface equation deduced through formula (1) is show as formula (2):

$$W \bullet X + b = 0 \quad (2)$$

The formula (2) is carried on the normalization of discriminant function, factor W and b is adjust, which makes the two kinds of all samples can meet $|g(X)| \geq 1$, a this time the class interval equals to $2/\|W\|$, thus Maximum interval problem is transformed into seek the minimum of $\|W\|$.

Thus the Optimal Hyper Plane problem is transformed into the optimization problem [10]:

the training sample data set is adjusted dynamic through the searching of available training sample. The combination algorithm process is show as below:

If the original training sample set, which x_i, y_i mean the training point and type separately, n means the training sample's number. $w_i(i)$ means every $x_i \in x$ and every weight iterative returned, n_i means the size of every training subset, the process of combinational algorithm AdaBoost and SVM are show as below [7-10]:

Step1: the weight $w_i(i) = \frac{1}{n} \quad (i = 1, 2, 3, \dots, n)$ is initialized, the training time is set as $t = 0$;

Step2:according to the current distribution of weight $w_i(i)$, the number of n_i samples is choose from the original training sample, then one sample subset χ is completed, the size of subset is n_i ;

Step3: all the sample unit of training subset χ is used to search the support vector;

If

$$s.t \quad \alpha_i \geq 0 \quad (i = 1, 2, \dots, n)$$

$X^+ = \frac{1}{C} \sum_{(x_i, y_i) \in SV^+} \alpha_i^* y_i$, which $C = \sum_{y_i=1} \alpha_i^* = \sum_{y_i=-1} \alpha_i^*$. The other sample points and the distance between X^+ and X^- inside the sample set χ are separately calculated, the size of distance is used to carry on the category judgement.

Step5:all the sample data of the original training sample is carried on the classification, the weighted

error rate is calculated according to formula (7):

$$\varepsilon = \sum_{i=1}^n w(i) \cdot \varepsilon_i \quad (7)$$

In the formula (7), ε_i means wrong classified units. The adjustment rules of sample weight: if the sample is wrong classified, then the sample weight reduce; otherwise, the weight increase.

Step 6: one sample is randomly choose from the original sample data set according to the current weight distribution $W(\alpha)$, if this sample is not in the training sample and also wrong classified by the current base classifier, then this sample is added into the current training sample subset, and the minimum weight sample inside the training subset is deleted, then switch to Step3; otherwise, switch to Step 6.

Step 7: weight combination t classifier;

$$H(x) = \text{sgn}\left(\sum_{s=1}^t \ln\left(\frac{\varepsilon_s}{1-\varepsilon_s}\right) \cdot H_s(x)\right) \quad (8)$$

4. P2P Traffic Recognition Based on AdaBoost and SVM

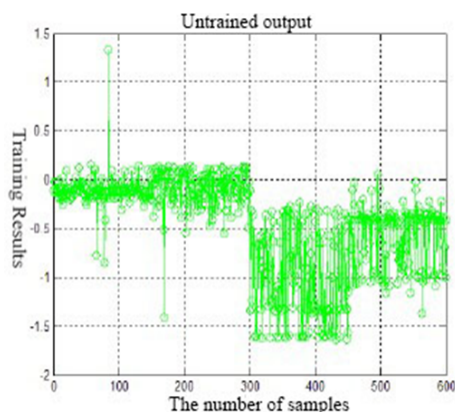
P2P Traffic Recognition process of AdaBoost and

SVM, including the data collection, data feature extraction, training sample and traffic recognition.

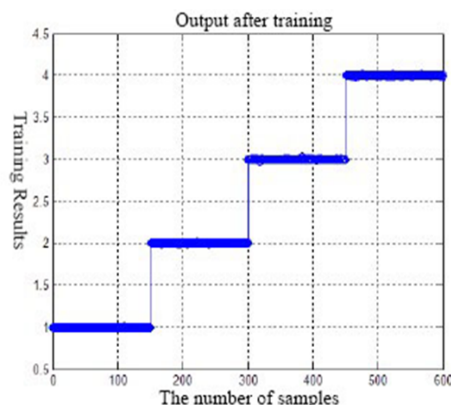
In order to verify the effect of P2P Traffic Recognition. The training time, recognized rate performances is used to calculate the recognition effect.

On the base of references literature at home and abroad, statistical data packets use 30s as a time slice. The total number of packets, upward traffic rate, average packet length, TCP traffic and The number of connections and the ratio of different IP number five traffic characteristics are choose as the input data. The Wireshark software is used to cut out each 300 number P2P traffic sample of BitTorrent, eMule, PPLive, PPStrea, 150 number sample of each kind is choose as the training subject of the combinational algorithm, others are used to test the performance of combinational algorithm. the MATLAB is used as the test platform, parameter of the SVM are: $C = 100$, $\text{Sigma} = 0.3$.

The recognition results of the P2P traffic based on AdaBoost and SVM are show as the Fig.1:



(a) before the combinational algorithm



(b) after the combinational algorithm

Figure 1. The comparison of before and after P2P traffic recognition based on AdaBoost and SVM

It can be seen from the Fig.1 that, the comparison of before and after P2P traffic recognition based on AdaBoost and SVM, the results are very obvious. Before the training, the recognition is very hard to find and disorder; after the training, the recognition is very high. In the Fig.1, 1, 2, 3, 4 separately mean the P2P traffic of BitTorrent, eMule, PPLive, PPStream. The number of sample is totally 600, 1-150 group is BitTorrent's traffic, 151-300 group is eMule's traffic, 301-450 group is PPLive's traffic, 451-600 group is PPStream's traffic.

In order to test the validity and reliability of the combination algorithm, 600 groups data are tested, the test results are show as Fig.2, the algorithm precision is really higher, but because of the similarity

between PPLive and PPStream, parts traffics may be wrong recognized.

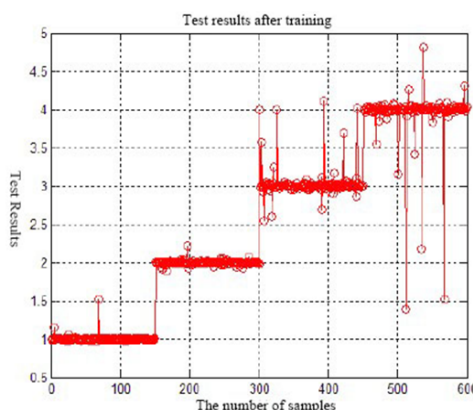
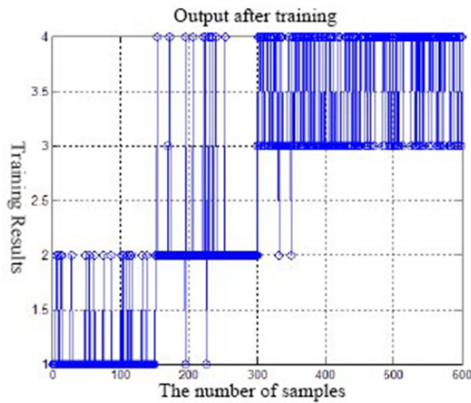


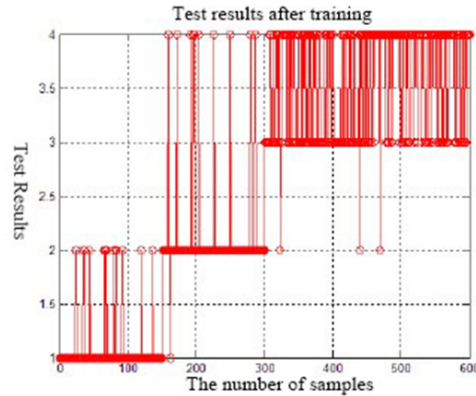
Figure 2. Test results of combination algorithm

In order to test the advantages of AdaBoost and SVM combination algorithm in doing P2P traffic recognition, it is compared with the SVM and AdaBoost

algorithm, the recognition results are show as Fig.3, the P2P traffic recognition results of AdaBoost algorithm are show as the Fig.4.

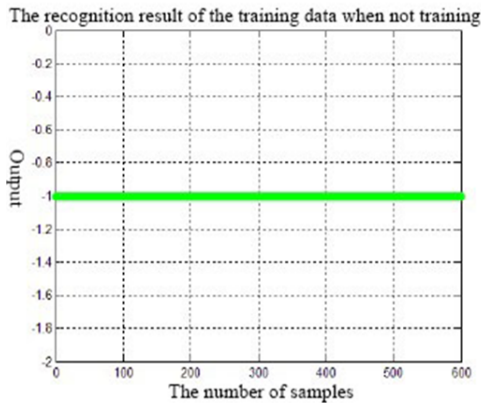


(a) training results of SVM

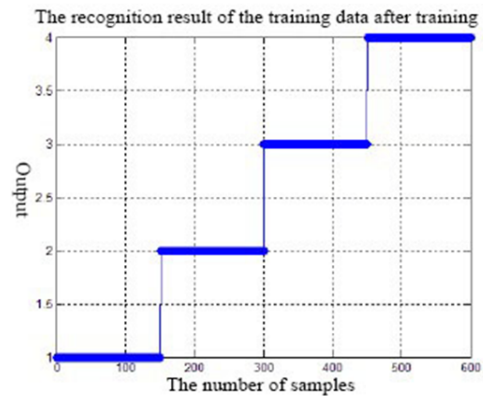


(b) test results of SVM

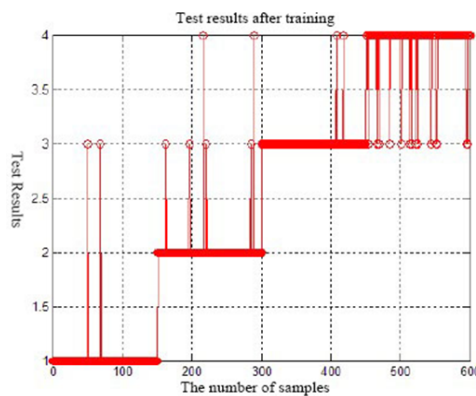
Figure 3. The P2P traffic recognition results of SVM



(a) AdaBoost before the training



(b) AdaBoost after the training



(c) test results of AdaBoost

Figure 4. The P2P traffic recognition results of AdaBoost

P2P traffic recognition results rate of AdaBoost-SVM combination algorithm, SVM and AdaBoost algorithm are show as Tab.1.

From Tab.1, Fig.5 and the P2P traffic recognition results of these three algorithms, it can be known that

the proposed combination algorithm is better than the AdaBoost, but AdaBoost is better than SVM, the combination algorithm's recognition wrong recognized rate are the opimal, thus it's superiority and reliability is verified.

The P2P traffic recognition of combination algorithm, SVM and AdaBoost algorithms are carried on 10 times, their recognition rates cooperation is show as the Fig.5. it can be seen from the Fig.5 that the

combination algorithm's recognition rate reached up to 96.33%, which is far more than SVM and AdaBoost algorithm.

Table 1. P2P traffic recognition results rate of AdaBoost- SVM combination algorithm, SVM and AdaBoost algorithm(time(s))

method	time	BitT	eMule	PPL	PPS
ombination algorithm	40.1	100.00%	99.35%	96.12%	99.56%
SVM	89.5	98.54%	90.65%	58.33%	76.74%
AdaBoost	0.81	98.67%	96.12%	98.52%	90.43%

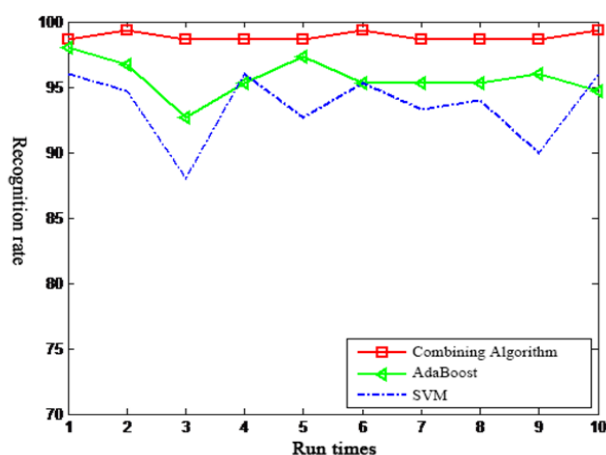


Figure 5. The recognition rate of different running times

5. Conclusion

Aiming at the low recognition rate and high wrong recognition rate of the traditional P2P traffic recognition technology, this paper proposed the P2P traffic identification optimization of the smallest neighbor method of AdaBoost-SVM, and the effectiveness of this patten is proved by experiment simulation. The mature system of P2P traffic recognition's identification, monitoring and control is not competed, such comprehensive quality system can be applied to the network supervision work, helping the network operators to monitor P2P traffic, research and development of this kind of system can be used as the next step research direction.

References

1. Chen H, Hu Z, Ye Z (2009) Research of P2P Traffic identification based on BP neural network. *Proc of the 1st Int Symp on Computer Network and Multimedia Technology*, p.p.1-4.
2. Yang A, Jiang S, Deng H, (2011)A P2P network traffic classification method using SVM. *Proc of the 9th Int Conf on Young Computer*

Scientists, p.p.398-403

3. Park B C, Won Y J, Kim M S, et al. (2012) Towards automated application signature generation for traffic identification. *Proc. of Network Operations and Management Symposium*, p.p.160-167.
4. Smith R, Estan C, Jha S, et al. (2010) Deflating the big bang: fast and scalable deep packet inspection with extended finite automata. *ACM SIGCOMM Computer Communication Review*, p.p.207-218.
5. Aceto G, Dainotti A, Donato W, et al. (2012) PortLoad: taking the best of two worlds in traffic classification. *Proc. of IEEE Conference on Computer Communications Workshops*, p.p.1-5.
6. Xu K, Zhang M, Ye M, et al. (2011) Identify P2P traffic by inspecting data transfer behavior. *Computer Communications*, 33, p.p.1141-1150.
7. Risso F, Baldi M, Morandi O, et al. (2012) Lightweight, payload-based traffic classification: An experimental evaluation. *Proc. of IEEE International Conference on Communications*, p.p.5869-5875.
8. Este A, Gringoli F, Salgarelli L. (2014) On the stability of the information carried by traffic traffic features at the packet level. *ACM SIGCOMM Computer Communication Review*, 39, p.p.13-18.
9. Fu Zhongliang, Zhao Xiang-hui, Miao Qing, Yao Yu et al. (2010) AdaBoost algorithm promotion-a set of integrated learning algorithm. *Journal of Sichuan University (Engineering Science)*, 6, p.p.91-98.
10. Zhuang Yan, Bai Zhenlin, Xu Yunfeng (2011) Research on Parameters of Support Vector Machine Based on Ant colony algorithm. *Computer Simulation*, 28(5), p.p. 216-219.