

Available online at <http://www.mecspress.net/ijem>

SVM Based P2P Traffic Identification Method With Multiple Properties

Yao Zhao^a, Zhixin Wei^b, Hua Zou^c

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications Beijing, China

Abstract

With the rapid development of the Internet, P2P has become the main network application in the Internet, which consumes most of the network resources. Accurately identifying and making control of the P2P traffic is of great significance. As a mature classification theory, support vector machine (SVM) algorithm is suitable for P2P traffic identification. This paper proposes a SVM based P2P flow identification method, adopting multidimensional flow properties as the input vector, which can improve the P2P flow classification accuracy. Analysis shows this method has many advantages over the other methods.

Index Terms: traffic identification; P2P; SVM.

© 2012 Published by MECS Publisher. Selection and/or peer review under responsibility of the Research Association of Modern Education and Computer Science.

1. Introduction

Investigations show that the P2P application enlarges the resource share across the whole network as well as consumes most of the network resources, which leads to congestion and other traffic problems. In order to identify the P2P traffic and to make control of it, the P2P identification has become a hot research subject.

2. Related Work

2.1. Traditional P2P Flow identification technology

The traditional traffic identification mainly uses port-based and deep packet inspection method. The P2P traffic can be identified according to the specific traffic port and special application tags of packets of the P2P traffic data[1][2]. With the rapid development of P2P, more and more P2P applications tend to use dynamical or anonymous ports as well as encrypted P2P traffic data. Under such circumstances, the above method is no longer applicable.

* Corresponding author.

E-mail address: ^azhao Yao@bupt.edu.cn; ^bwhithin@gmail.com; ^czouhua@bupt.edu.cn

Some traffic identification methods are property-based[3][4]. These methods use P2P traffic properties as a basis for judging, such as TCP and UDP traffic exist between the P2P nodes, connections P2P nodes accept and initiate at the same time, balanced upload and download traffic on P2P nodes. These methods can detect traffic with encrypted data packets and dynamic ports. However, It is only applicable to the known P2P traffic. For the unknown or new P2P traffic, it cannot work appropriately. To solve this problem, experts and scholars apply classification algorithms of the machine learning filed to the P2P traffic identification and achieve a better recognition effect.

More and more statistical decision-making, clustering pattern classification methods of the machine learning and data mining field have been applied to P2P traffic identification. Such as Bayesian classification algorithm, support vector machine (SVM), BP neural network and so on, among which SVM classification becomes a research hotspot with its superiority and significant effect in the P2P traffic classification. This paper introduces a P2P traffic identification method based on SVM classification.

2.2. Introduction to SVM

SVM is a machine learning method based on statistic theories[5]. It uses a pre-selected nonlinear transform, mapping unclassified problem of the low dimensional space to high dimensional feature space, and creates the optimal classification hyper plane in the space, which will classify the problems into two types. Therefore, it is suitable for P2P and non-P2P scene.

When using the SVM as the classification method, each sample consists of a vector and a mark. The vector is composed of one or several properties, while the mark indicates the category of the sample. In this paper, P2P flow is tagged with +1, others tagged with -1. As follows: $D_i = (x_i, y_i)$, x_i is the feature vector, y_i is the mark.

Suppose there are n d -dimension samples, the sample type mark is + 1 or - 1, can be expressed as: $(x_i, y_i), x_i \in R^d, y_i \in \{+1, -1\}, i = 1, 2, 3 \dots, n$

If there exists a hyper plane $w \cdot x + b = 0$, which makes:

$$(w \cdot x_i) + b \geq +1, y_i = +1 \quad (1)$$

$$(w \cdot x_i) + b \leq -1, y_i = -1 \quad (2)$$

At this point, to make the classification interval $\frac{2}{\|w\|}$ maximum, is equivalent to make $\|w\|$ or $\frac{\|w\|^2}{2}$ minimal. Thus, the hyper plane that meets the conditions above and make $\frac{\|w\|^2}{2}$ minimal is called optimal classification plane. The training sample points that are parallel to and nearest to the hyper plane are called a support vector.

Hyper plane $w \cdot x + b = 0$ is able to correctly identify the two types of samples and expresses optimization problems with the maximum interval as:

$$y_i[(w \cdot x_i) + b] - 1 \geq 0, i = 1, 2 \dots, n \quad (3)$$

The minimum value can be obtained under the restriction of (3)

$$\Phi(w) = \frac{1}{2} \|w\|^2 = \frac{1}{2} \langle w, w \rangle \quad (4)$$

This problem can be transformed to its dual problem:

$$\sum_{i=1}^n a_i y_i = 0, a_i \geq 0, i = 1, 2 \dots, n \quad (5)$$

The maximum value can be obtained under the constraint of (5):

$$Q(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j (x_i \cdot x_j) \quad (6)$$

After computing the optimal solution of a^* , w^* and b^* , the optimal classification function is obtained as follows:

$$f(x) = \text{sgn}\{(w^* \cdot x) + b^*\} = \text{sgn}\{\sum_{i=1}^n a^* y_i (x_i \cdot x) + b^*\} \quad (7)$$

The above function is appropriate for the linear classification problems. By turning the problem into its dual problem, the optimization objective function or the classification function only involves the inner product of the high dimensional space of training samples. For the non-linear problems, it can be solved by choosing the appropriate kernel function $K(x_i, x_j)$ which can map the low dimension space to high dimension space to accumulate the inner product result. In this way, the non-linear classification can be transformed into linear classification problem. Then, the optimal objective function turns to be:

$$Q(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j K(x_i \cdot x_j) \quad (8)$$

And the related classification function turns to be:

$$f(x) = \text{sgn}\{\sum_{i=1}^n a^* y_i K(x_i \cdot x) + b^*\} \quad (9)$$

2.3. SVM based network traffic identification review

Generally speaking, the traditional statistical classification method is based on mathematical induction which get common abstraction from known data, then predict the unknown ones. This method works well only when the sample amount reaches a certain quantity. However, SVM classification method is different, it does not need to summarize the universal truth from the known, but through studying the known types, directly predict the unknown ones. At present, the development of the SVM theory is mature. Its superiority has been proved by many experiments, which shows that the application of SVM applied on P2P flow identification is worth studying.

Much research work has been done on SVM based traffic identification and made great progress. Gabriel Gómez Sena[6] use the size of the first packets on both directions of a flow as a statistical fingerprint, by comparing the centroid clustering method with the SVM clustering method prove that the SVM based method has much higher accuracy in traffic identification. Rui Wang[7] and his colleagues concern that the peer of the P2P application connect with more different address than normal nodes, based on which he proposed a sensitive feature extraction algorithm and transformed the flow data to 3-dimension feature data as the input vector of the SVM algorithm. And experiment shows this method can identify the known as well as unknown P2P traffic with an acceptable accuracy.

Both of the referenced SVM based P2P traffic identification methods work excellently on the P2P traffic identification, but they take advantage of only one or two flow statistical properties. Comparing with the single flow property as the input vector, the multiple flow properties have much better performance [8]. This paper proposes a new SVM based P2P traffic identification method which uses multiple statistical properties as the input vector of the SVM predict model

3. SVM based P2P traffic identification method with multiple properties

SVM classification method has been brought into the P2P flow identification field. Different from the existing single property input vector method, this paper adopts the multiple flow properties as the input vector of the SVM classification model. The SVM classification model is generated by the precisely pre-classified flow sample. Each time choose different properties combination as the input vector of the SVM classification model, record the classification accuracy. The flow properties combination that has highest accuracy will be finally selected as the input vector.

The key points of this method are obtaining the precisely pre-classified P2P flow samples and the selection of best combination of the flow properties. Therefore, this paper will describe both of the aspects in detail.

3.1. procedure description

Like most of the machine learning methods, the SVM based P2P traffic identification method includes two main phases[9], the SVM training phase and the SVM testing phase. In the SVM training phase, the pre-classified flow samples are used to produce a SVM predict model, while in the SVM testing phase, the SVM predict model will be fully tested to see its accuracy. Before the above two phase, the priority work is to get the flow samples as the input data of the classification model. The whole procedure of the method can be described in Fig 1:

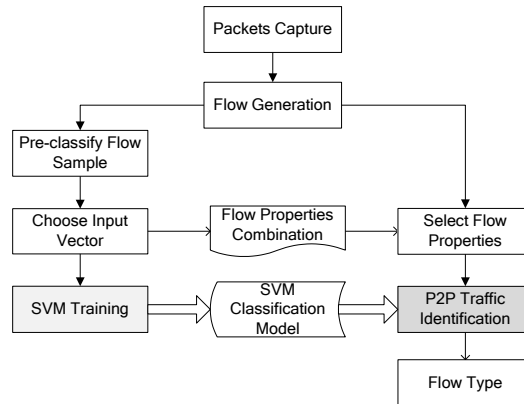


Figure 1 SVM based P2P identification procedure

Firstly, network packets are captured by capture tools. Then, flow samples are generated by a C++ program developed for this experiment. After that, flow Samples are pre-classified by marking with non-P2P traffic or P2P traffic tag. Then, the pre-classified are used by the SVM tool, combinations of different flow properties are selected as the input vectors for SVM training to generate the classification model. Select the one with the best accuracy as the final classification model, output the results of flow identification. Detailed contents will be introduced in the following content.

3.2. Packets Capture

Wire Shark is used to capture the network packets. The P2P peers use UDP to find each other and use TCP to transport the application data. To ease the analysis, before any capture, the unrelated packets like ARP, broadcast and multicast layer2 packets were all filtered, only the TCP and UDP packets were saved.

3.3. Flow Generation

A C++ program was developed to generate the flow samples. Unpack the packets, record the source IP, destination IP, source port, destination port and protocol, the packets with the same five parameters are thought to belong to the same flow, generate a flow record.

When getting a packet, the flow records of the five parameters were checked, if there is a flow record of that packet, update the flow information, and otherwise generate a new flow record. The three handshakes was thought as the beginning of a TCP flow, and the FIN/RST as the end. Besides, if no further packets come during a period of T seconds, such as 90s, the flow is thought to be over. Similarly, if there are no packets come in 90s, a UDP flow is thought to be over. Record all the flow properties, includes the total time of a flow, total packets, total bytes, the arrival time interval, the packet size, payload size etc.

3.4. Pre-classify flow sample

In order to get the accurately pre-classified flow samples, a specific LAN was established. In different time period, one or several hosts were designated to run the specific P2P application or non-P2P application. Specific application type of the flow can be identified by its IP address.

At the edge router of the LAN, mirror all the interface packets to a fixed interface of the router, capture the packets of the host directly connect to that interface for later use. Mark the type of the flow based on its IP information. The flow sample will be marked as P2P or non-P2P flow as SVM training samples.

After getting the necessary pre-classified samples, the paper will focus on the introduction of the P2P flow identification method.

3.5. SVM training process

1) Preparation work

This paper uses the open sourced LibSVM [10] tool to train and test the flow sample. LibSVM is developed by professor hih-Jen of national Taiwan University. It is an effective recognition and regression software package, including three basic tools: SVM-train, SVM-predict, SVM-scale. SVM-train is most important tool, which is used to train the sample and generate the predict model. SVM-predict is used to predict the data, output the accuracy of the test sample and predict model. SVM-scale is used to transform the vector values into $[-1, 1]$. The LibSVM source code is downloaded from the website and compiled under Linux to generate the executive file. The experiment adopted the proposed SVM classification steps:

Firstly, the flow samples are transformed into the required form, each flow sample is described as $\langle \text{label} \rangle : \langle \text{index1} \rangle : \langle \text{value1} \rangle : \langle \text{index2} \rangle : \langle \text{value2} \rangle : \langle \text{index3} \rangle : \langle \text{value3} \rangle \dots$, ended by '\n'. $\langle \text{label} \rangle$ is an integer marking the type. $\langle \text{index} \rangle : \langle \text{value} \rangle$ gives a property value. $\langle \text{index} \rangle$ starts from 1 and $\langle \text{value} \rangle$ is a real number of the property value.

Secondly, conduct the SVM-scale on the flow samples, the value of each property is restricted to range $[-1, 1]$ to simplify calculation complexity.

2) Kernel function and parameters selection

Appropriate kernel function and parameters can make the mapped distribution region of the sample more focused, thus strengthening the "linear divisible" degree of samples in the property space, to increase the classification precision and generalization capability.

For the same experiment data, using different kernel functions and parameters, the classification accuracy can vary widely, even for the same kernel function; classification precision also can have bigger difference with different parameters. The commonly used kernel functions [5] are:

- Linear kernel:

$$K(x_i, x_j) = (x_i \cdot x_j) \quad (10)$$

- Polynomial kernel:

$$K(x_i, x_j) = [(x_i \cdot x_j) + 1]^q \quad (11)$$

- Radical Basis function(RBF):

$$K(x_i, x_j) = \exp\left\{-\frac{\|x_i - x_j\|^2}{\sigma^2}\right\} \quad (12)$$

- Sigmoid tanh:

$$K(x_i, x_j) = \tanh(v(x_i x_j) + c) \quad (13)$$

According to the functional theory, as long as there is one kind of kernel function satisfying the Mercer condition, it corresponds to the inner product of the high dimensional space.

Polynomial kernel function can always satisfy mercer condition, but the parameters of the function is more, and the parameter selection influences classification accuracy influence; Sigmoid kernel function can satisfy Mercer condition only under specific circumstance; RBF kernel function parameter has only two parameters, at the same time, it can satisfy the Mercer conditions, so select RBF as the kernel function of the experiment.

After choosing RBF as the kernel function, the next step is to find the best parameter C and r. To simplify the experiment, use the recommended cross-validation tool “grid-search” tool provided by the LibSVM. In cross-validation, the training set is first divided into N subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining N-1 subsets. Thus, each instance of the whole training set is predicted once, the cross-validation accuracy is the percentage of data which are correctly classified. After the best(C, r) is found, the whole training set is trained again to generate the predict model.

The pre-classified training samples must be evenly distributed, so as to ensure the reliability of classification results. By testing the known classification samples, compare the results obtained to the pre-define result to see whether the predict model meets the demand.

3) Flow properties selection

In the process of the flow sample generation, all the information of the flow sample are recoded, such as the application port, the source and destination IP, the number of packets, the duration of the flow, the total bytes. Not all of the flow properties will affect the classification accuracy, only a few of the key properties affect flow classification accuracy.

A flow sample is described as:

$$F_i = (t^m, v_1, v_2, v_3, \dots, v_i, \dots, v_N)$$

$$i \in \mathbb{R}, t^m \in \{-1, +1\}, -1 \leq v_i \leq 1$$

F_i is the i th flow sample, t^m indicates the P2P flow type, +1 is P2P flow, -1 is not P2P flow;

v_i is the i th property of the flow, the value is in range of [-1, +1], after conducting the SVM-scale.

This paper adopts discriminators selection method [8] to choose the properties combination as the input vector. Firstly, choose one property as the basic property, record the classification accuracy of the SVM model, then sequentially add one property each time, record the corresponding classification accuracy, see whether the added flow property affect the classification accuracy. In this way, find out the flow properties exclusively that influence the classification result least to gain the properties combination that have the maximum classification accuracy.

The experiment proved that some of flow properties have more influence on the classification accuracy than the others, which can be selected as the input vector of the flow classification model: the flow duration, total

packets, total bytes, average maximum and minimum packet size, protocol type, average maximum and minimum payload size, average maximum and minimum interval of packet arrival time.

4. Method Analysis

SVM is a very mature theory; scientific research has proved its superiority in classification field. It also has made certain progress in its application in the flow identification. SVM method studies the known flow samples to produce a classification model to predict the unknown flow. It doesn't need to do statistical computation on the known samples which reduce the required sample number.

This paper adopts the multidimensional flow properties as the input vector of the P2P flow classification model, which has better accuracy than the single property. Comparison between the multidimensional flow properties and single property is showed in TABLE I, which prove that the multidimensional flow properties as the input vector of the P2P flow classification has better effect.

Table 1. Comparison between single property and multiple properties method

1	Method name	Based on the early packets size	Based on multiple properties
2	Flow generate	Save packets length < 200kbytes	Save packets with all length
3	Pre-classify sample	deep packet inspection	Specific network
4	Pre-classification accuracy	Can't classify new and unknown P2P flow	Can classify all the flow
5	Property number	Single property	Multiple property
6	Identification accuracy	About 80%	Above 90%

SVM based single flow property P2P traffic identification has its restriction. The flow type cannot be decided by one flow property. The combinations of the flow properties define the right type of the P2P flow. Although the affection of different property on the classification accuracy varies, the experiment proved that the combination of them as the study input of the SVM model has better classification effect. Besides, the pre-classify method of deep packet inspection can works well only with the existing P2P flows, for the new or unknown P2P flow, it can't correctly mark the flow type. Hence, this paper builds an experimental environment. The designated IP run the specific network applications, so the flow type can be marked accurately according to its IP, which provide completely accurate pre-classified flow samples

5. Summarize

Firstly, this paper gives a summary of P2P traffic identification, and then describes principles as well as advantages and disadvantages of each traffic classification method. After that a multi-dimensional vector-based SVM classification method is presented, which make use of the discriminatory selection to find the flow properties with the greatest impact on the classification accuracy as the input vector of the classification model, which has better classification accuracy than SVM classification based one-dimensional input vectors.

Further work will focus on the improvement of the method, enabling it to be applied to wide area networks and implementation of real-time P2P traffic identification.

Acknowledgment

This work was supported by National Key Basic Research Program of China ("973" Program) (2009CB320406), National High Technology Research and Development Program of China ("863" Program) (2008AA01A317), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (Grant No.60821001), and Beijing Municipal Commission of Education to build the project special.

References

- [1] Sen, S.,Jia Wang. Analyze P2P traffic across large network. Networking, IEEE/ACM Transactions on Volume: 12, Issue: 2.2004.
- [2] Ohzahata S ,Hagiwara Y, Terada M ,et al ,A Traffic Identification Method and Evaluations for a Pure P2P Application[M] .Lecture Notes in Computer Science ,2005.
- [3] Hongbo Jiang, Andrew W. Moore, Zihui Ge, Shudong Jin, Jia Wang. Aug. self-learning IP traffic classification based on statistical flow characteristics. Proceedings of the 2007 SIGCOMM workshop on Internet network management, 2007.
- [4] Kompella S,Wieselthier, J.E., Ephremides, A., Sherali, H.D. cross-layer peer-to-peer traffic identification and optimization based on active networking. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops, 2008.
- [5] Nello C, John S T. An introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press, 2004.
- [6] Gabriel Gómez Sena, Pablo Belzarena, Early traffic classification using Support Vector Machines, Proceeding LANC '09 Proceedings of the 5th International Latin American Networking Conference,2009
- [7] Rui Wang, Yang Liu, Yue-xiang Yang, Hai-long Wang. A new method for P2P Traffic Identification Based on Support vector Machine. AIML 06 International Conference, 13 - 15 June 2006, Sharm El Sheikh, Egypt
- [8] R. Yuan Z. Li, X. Guan, Accurate classification of the internet traffic based on the SVM method, in: Proceedings of the 42th IEEE International Conference on Communications (ICC 2007), June 2007
- [9] PAN S R,FU M ,SHI C Q. Application of the Supporting Vector Machine in P2P Traffic Identification, COMPU TER EN GINEERING & SCIENCE, Vol132 ,No12 ,2010
- [10]Chih-Chung Chang and Chi-Jen Lin. LIBSVM-A Library for support Vector Machines.<http://www.csie.ntu.tw/~cjlin/libsvm/>