Multi-levels Traffic Classification Technique

Chengjie Gu, Shunyi Zhuang, Yanfei Sun, Junrong Yan
Institute of Information Networks Technology
Nanjing University of Posts and Telecommunications
Nanjing, China
jackiee.gu@gmail.com

***Abstract*—Classifying Internet network traffic flows offers substantial benefits to a number of key areas in network engineering and surveillance. However, as many newly-emerged P2P applications use dynamic port numbers and masquerading techniques, it causes the most challenging problem in network traffic classification. In this paper, we propose a novel multi-levels traffic classification technique that brings together the benefits of port mapping, signature matching and flow statistical classification techniques, motivated by variety of network activities and their requirements of traffic. Experiment results illustrate this technique can achieve high accuracy, low overheads, robustness, and real-time.**

Keywords-traffic classification; internet protocol; peer-to-peer; machine learning

I. INTRODUCTION

Traffic classification is fundamental to solve difficult network services, including traffic modeling, security surveillance, real-time quality of service and provision for future resources[1]. The ability of a network operator to accurately classify traffic into different application directly determines the success of many of the above network management tasks[2]. New application, for example P2P networks have become extremely popular for many different application like file sharing, live video streaming, IP-TV, and VoIP services.

The dynamic classification and identification of network applications responsible for network traffic flows is essential to IP network engineering. Some of the currently proposed traffic classification methods are used in traffic identification. As P2P traffic occupies large amount of bandwidth, it is more likely to incur network congestion, worsen network operating performance, thus lead to the deterioration of QoS[3]. However, in order to circumvent detection, P2P network application has started using random ports or standard port to send their traffic, even tunneled through HTTP with 80 protocol port. In despite of great effort to identify different application, it is very difficult to propose an approach to settle it completely due to the continual appearing of new applications.

The remainder of this paper is organized as follows. Section Ⅱ reviews previous background and approaches to traffic classification. Section Ⅲpresents a novel multi-levels traffic classification technique to identify all kinds of Internet traffic. We elaborate on the different key levels of the classification technique: port-based level, payload-based level and flow characteristic-based level, discuss the related concerns. Section Ⅳ gives the experimental results of classification technique. Finally, we conclude our paper in Section Ⅴ.

II. BACKGROUND AND RELATE WORK

Traditionally, ISPs have used port numbers to effectively identify and classify network traffic. This approach is extremely easy to implement and introduces very little overhead on the traffic classifier, such as port 80 is HTTP traffic, port 1214 is Kazaa P2P traffic, port 6346 is Gnutella P2P traffic and so on. A.Madhukar and Williamson show that port-base analysis is unable to identify 30-70% Internet traffic flows they investigated[4]. To avoid total reliance on the semantics of port numbers, the payload identification method was proposed. Most protocols contain a protocol special string in the payload that can be used for identification[5].

Given the shortcoming of port-based and payload-based approaches for detecting Internet traffic, newer approaches which focused on capturing and extracting commonalities in the behavior of application rely on traffic's statistical characteristics to identify the traffic[6]. Recently, machine learning methods were also used to identify Internet traffic and network application[7]. Machine learning approach relies on the premise that a set of feature for objects would be similar when objects are the same class. Machine learning techniques in traffic classification can be categorized as unsupervised and supervised[8].

III. MULTI-LEVELS TRAFFIC CLASSIFICATION TECHNIQUE

A. Multi-levels Traffic Classification

We propose a novel multi-levels traffic classification methodology which mainly consists of port-based, payload-based and flow characteristic-based levels. In port-based classification phase, traffic is classified into different categories according the port mapping, whereas in payload-based classification phase, the classifier can automatically extracts payload signatures to identify specific protocols, the traffic is categorized into distinct application types. In flow characteristic-based classification phase, we classify Internet traffic by focusing on the characteristics of the

Supported by National High-Tech Research and Development Plan of China under Grant No.2009AA01Z212 and No.2009AA01Z202).

traffic through analyzing and constructing empirical model using machine learning.

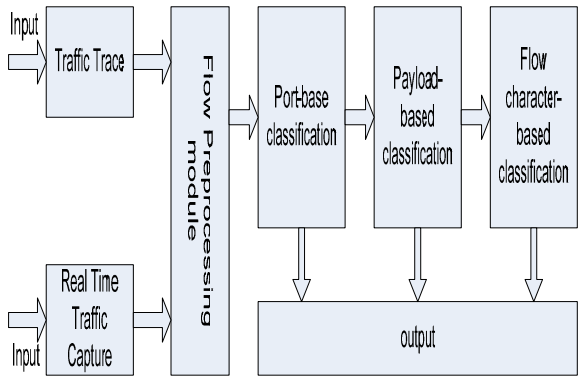

Fig.1 Multi-levels traffic classification system

In this subsection, we describe the design of the multi-levels traffic classification methodology for traffic classification. As illustrated in Fig.1, the overall architecture of multi-levels traffic classification methodology consists of three main phases, namely port-based classification level, payload-based classification level and flow characteristic-basedclassification level.

The port-based classification aims at classifying the traffic into two broad categories as known traffic and unknown traffic. It would be identified and classified by port mapping. The subsequent payload-based classification can be implemented by automatically extracting payload signatures to identify specific applications. The finally flow characteristic-basedtraffic classification employ machine learning algorithm to identify different application through constructing empirical model and analyzing flow character.

In preprocessing module, the raw packets from network link are captured and packet header information extracted from each raw packet is delivered to flow generator. A flow is defined and identifiable by the 5-tuple (source address, source port, destination address, destination port, transport protocol)[9]. In case of detecting the unknown or encrypted P2P application[10], machine learning classifiers would be used. We will elaborate the port-based traffic classifier, the payload traffic classifier and the flow characteristic-basedclassifier modules in next subsections respectively.

B. Port-based Traffic Classification

TCP and UDP provide for the multiplexing of multiple flows between common IP endpoints through the use of port numbers. Traditional port-based approach has been kept effective to traffic identification for many years since applications tended to abide by well-known port numbers, and little has changed when the first generation of P2P applications emerged as they used fixed port numbers. Thus, port-based traffic identification was reported to be unreliable as early as in 2004. But the report illustrated that port-based classification can identify 30-40% Internet traffic flows. Since non-P2P traffic has excluded flows with the dynamic and masquerade ports, which are incurred mostly by P2P applications, port-based identification is the easy and effective way to accurately classify applications according to application port matching.

C. Payload-based Ttraffic Classification

To avoid total reliance on the semantics of port numbers, many current industry products utilizes reconstruction of session and application information from each packet's content. By an exhaustive off-line search of applications and analysis of captured traffic, comprehensive application signature table is built. As illustrated in Fig.2, the payload-based traffic classifier module is comprised of the following functional blocks: payload-based classifier, signature database and signature extractor. Traffic traversing links through this signature matching component where a fast pattern matching algorithm can be classified traffic. The signature extractor can extract signatures for different applications that belong to a particular application class. The signature extractor is to automatically extract signatures for different P2P and non-P2P applications. We use the LASER algorithm to extract signatures from packet payloads. Once a signature is obtained, it is sent to store this signature in signatures database and use it to instantly classify any future flow.

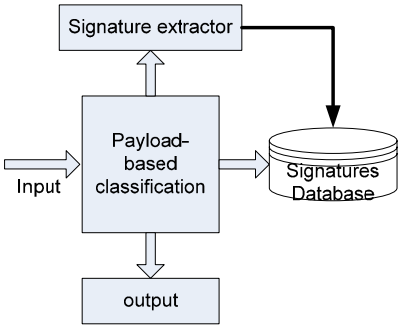

Fig.2 Payload-base traffic classification approach

D. Flow Characteristic-basedTraffic Classification

In this level, the flow characteristic-basedtraffic classifier is intended to recognize a particular class in amongst the usual mix of traffic seen on IP network. Traffic captured in real-time is used to construct flow statistics from which features are determined and then fed into the classification model. We presume that the set of features calculated from captured traffic is limited to the optimal feature set determined during training. The classifier's output indicates which flows are deemed to be members of the class of interest. Certain implementations may optionally allow the model to be updated in real-time. The optimal approach to train a supervised ML algorithm is to provide previously classified examples of every types of IP traffic: traffic matching the class of traffic that one wishes later to identify in the network, and representative traffic of entirely different applications one would expect to see in future.

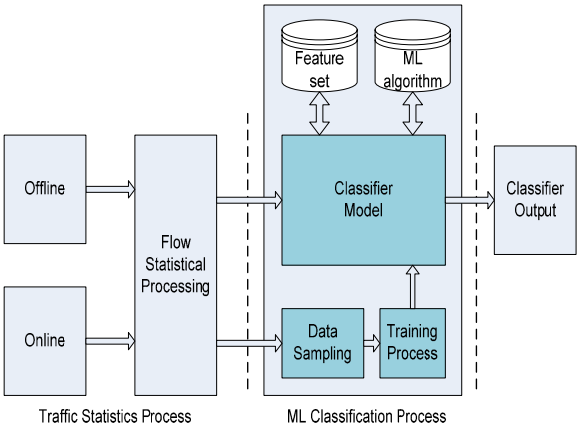

Fig. 3. Flow chart of flow characteristic-basedtraffic classification

As illustrated in Fig.3, the technique of the flow characteristic-basedclassification technique includes two steps: off-line ML modeling and on-line ML classification. Firstly, according to flow information from preprocessing module, flow statistics is computed in terms of each selected feature and stored them to the corresponding database when reaching some milestone. Secondly, training and test sets are uniformly sampled for the specific ML classification. Once the datasets is ready, the system carries out feature selection to eliminate the redundant and irrelevant features, resulting in optimal feature subset. Finally, the flow is trained using the selected ML classifier, and evaluated by performance metrics. In default, each experiment is repeated on 10 independently sampled datasets to eliminate bias. If satisfied, ML modeling is finished and ready for the traffic classification, otherwise the process is repeated. In the level of on-line ML classification, the output model produced by ML modeling is applied to classify the captured traffic. Eventually, the classification output would be applied to such network activities as QoS differentiation and network surveillance, or deliver to the fine classification module for further classification.

Prior to the ML modeling, feature selection can be executed off-line, regardless of its high complexity. We use Genetic Algorithm (GA) for optimization based on the observation that supervised ML classifiers, especially decision trees, provide the best performance with features searched by GA. Besides, in order to achieve fast and early identification, we use partial information of flows for ML modeling and classification instead of the full information when it finishes. Therefore, it is necessary to determine the appropriate number p of first packets in each flow.

Since ML approaches play an important role in multi-levels traffic classification technique, we covered the experiments of using variety of ML algorithms working on our dataset. Some effective approaches are compared in terms of computational complexity and accuracy[11][12]. We compare the relative computational complexity and accuracy through the experiments. The classification performance using different ML algorithms is demonstrated in Tab.1.

Table 1 Performance of different ML algorithms

Classifiers	Modeling time(s)	Training time(s)	Accuracy(%)
NB	0.38	18	69.76
BP	77.16	803	83.52
SMO	170.5	1472	76.68
REPTree	1.33	14	96.35
C4.5	3.47	35	96.66

From the experimental result and data, several factors prompted us to select decision tree algorithms over the other more sophisticated ML algorithms. First, decision tree algorithms could achieve better accuracy with the least amount of computational overhead. Second, it is simple and easy to implement. Third, the more complex ML algorithms requires significantly longer training time and classification time than decision tree, such as SMO. Decision tree is a simple yet successful technique for supervised classification learning. A decision tree partitions data into smaller segments called terminal nodes or leaves that are homogeneous with respect to a target variable. This partitioning continues until the subsets cannot be partitioned any further using user defined stopping criteria.

IV. EXPERIMENT AND EVALUATION

A. Experiment Environment

As illustrated in Fig.4, experimental network were designed to evaluate our multi-levels traffic classification technique. The architecture includes some hosts running given P2P applications or non-P2P traffic and other applications and some other subnets in the information network research institute. The main applications in our experiments include HTTP, POP3, FTP, PPStream, BitTorrent, eMule, PPlive, eDonkey and Game, etc. The proposed classifier is put at the gateway of the campus network.

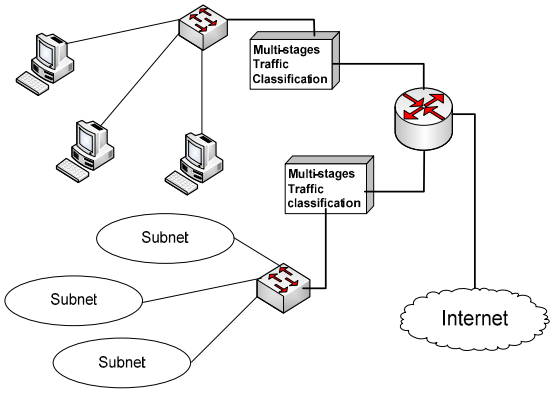

Fig. 4. The multi-levels classification prototype

The performance of classification can be measured by the accuracy and computational complexity. A good traffic

classifier aims to minimize the FN and FP. Some works make use of Accuracy as an evaluation metric. It is generally defined as the percentage of correctly classified instances among the total number of instances.

$$\text{Accuracy}=\frac{TP}{TP_i+FP_i} \quad (1)$$

The overall effectiveness of the classifiers is measured by the overall accuracy.

$$\text{Overall accuracy}=\frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m}(TP_i+FP_i)} \quad (2)$$

where m is the number of categories.

B. Advantage of Multi-levels Traffic Classification

In this section, we compare classification accuracy of multi-levels traffic classification technique with the other classification approaches solely based on port-based approach, payload-based and flow characteristic-basedusing ML. Take the dataset that we collected the link data during three days from 10:00 to 11:00 since June 19 2009 for evaluation and comparison between the different approaches.

For the multi-levels traffic classification technique, port-based approach separates known traffic from the rest of the traffic by exploiting the port protocols in the first level, the classification overall accuracy is 68.7%, 35.8% for non-P2P and P2P application respectively. In the second level, it automatically extracts payload signatures to identify specific protocols, the classification overall accuracy is 54.9%, 65.1% for non-P2P and P2P application respectively. In the third level, the relationship between the class of traffic and its observed statistical properties has been noted. We construct empirical classifier model to classify Internet traffic by focusing on the characteristics of the traffic using C4.5 algorithm, we get 97.9% and 94.6% for non-P2P and P2P application respectively using C4.5 algorithm. The experimental analysis illustrates that multi-levels traffic classification can improve the overall accuracy in different level. The overall accuracy of the multi-levels classification in different levels can be found in Tab. 2.

Table 2 Overall accuracy of the multi-levels classification

level	Traffic	TP(%)	FP(%)	Overall accuracy(%)
First level	P2P	84.37	33.02	49.87
	Non-P2P	96.58	1.65	
Second level	P2P	90.83	23.07	78.46
	Non-P2P	88.29	9.12	
Third level	P2P	93.91	4.58	95.63
	Non-P2P	97.65	3.41	

Compared with the other classification approaches, the multi-levels classification technique proposed in the paper can achieve higher overall accuracy. Moreover, the multi-levels classification technique has considerable accuracy and less overhead to satisfy some network activities. At the same time, this approach has the ability to learn online and retrains itself by the newly obtained data sets at the same time.

We calculate the relative computational time, demanded memory and accuracy through the experiments according to different approaches. The performance for different approaches is demonstrated in Tab.3. We can find that the multi-levels classification technique is able to greatly improve the accuracy, while only minimally impacting computational time and demanded memory.

Table 3 The performance for different approaches

Performance	Port-based	Payload-based	Character-based	Multi-levels
Time(s)	0.930	4.356	1.059	3.563
Memory(M)	5.976	18.407	5.012	10.342
Accuracy(%)	49.87	49.35	67.84	96.7

V. CONCLUSION

To meet the requirements of the network activities and take into account traffic classification challenges, we present a novel multi-levels traffic classification technique to identify all kinds of Internet traffic in this paper. A flow is identification with the protocol port-based approach, signature-based approach and flow characteristic-basedapproach using machine learning. This technique that can learn unknown and encrypted traffic with minimum manual intervention can be used for classifying Internet traffic. Experiment results clearly illustrate that the multi-levels traffic classification technique can be competent for classifying internet traffic from gigabit network stream online. It also shows good extension ability to learn new traffic features and indentify new applications.

ACKNOWLEDGMENT

This research is funded by National High-Tech Research and Development Plan (863) of China (No.2009AA01Z212, No.2009AA01Z202), Natural Science Foundation of Jiangsu Province (No.BK2007603); High-Tech Research Plan of Jiangsu Province (No.BG2007045); Research Climbing Project of NJUPT(No.NY2007044). We would like to thank the reviewers for their very helpful comments and feedbacks to improve the manuscript.

REFERENCES

[1] B.Choi, S.Moon, Z.Zhang, K.Papagiannaki, Analysis of point-to-point packet delay in an operational network, Computer Networks, Vol. 51, No.13, 2007, pp. 33–37.

[2] Suresh K. Naira, and David C. Novakb, A traffic shaping model for optimizing network operations, European Journal of Operational Research, Vol.180, No.3, 2007, pp.1358-1380.

[3] B.Marco, Mellia, Antonio Pescapè and LucaSalgarelli, Traffic classification and its applications to modern networks, Computer Networks, Vol. 53, No 6, 2009, pp. 759-76.

[4] A.Madhukar. C.Williamson, A longitudinal study of P2P traffic classification, Proceedings of the 14th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Monterey,CA, September, 2006, pp.179–188.

[5] L.Bernaille, R.Teuxeira, I.Akodkenous, Traffic classification on the fly.ACM SIGCOMM Computer Communication Review,Vol.36, No.2,2006,pp23–26.

[6] S.Sarvotham, R.Riedi, and R.Baraniuk, Connection-level analysis and modeling of network traffic, Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement. San Francisco, California, USA: ACM, 2001, pp. 99–103.

[7] A. W. Moore and D. Zuev, Internet traffic classification using Bayesian analysis techniques, Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, Banff, Alberta, Canada, June, 2005, pp.50–60.

[8] S.Zander, T.Nguyen, and G.Armitage, Automated traf c classi cation and application identi cation using machine learning, IEEE 30th Conference on Local Computer Networks (LCN 2005), Sydney, Australia, November 2005, pp. 250–257.

[9] T.Karagiannis, A.Roido, M.Aloutsos, Transport layer identification of P2P traffic. Proceedings of the 2004 ACM SIGCOMM Internet Measurement Conference (IMC'2004), Italy, October, 2004, pp.121–134.

[10] A.W. Moore and K.Papagiannaki, Toward the accurate identification of network applications, Proceedings of the Passive and Active Measurement Workshop (PAM'2005), Boston, MA, USA, March 31–April 1, 2005, pp.41–54.

[11] Tom Auld, Andrew W. Moore, Stephen F. Gull, Bayesian neural networks for internet traffic classification. IEEE Transactions on Neural Networks, Vol.18, No.1, 2006, pp. 223–239.

[12] S.Zander, T.Nguyen, and G.Armitage. Automated traffic classification and application identification using Machine Learning, Proceedings of the IEEE Conference on Local Computer Networks 30th Anniversary, Sydney, Australia, November,2005,pp.250–257