

Toward Accurate Efficient Traffic Classification

Wei Li

Department of Computer Science, Queen Mary University of London

ABSTRACT

Online traffic classification continues to be of long term interest to the networking community. It serves as the input for practical solutions such as network monitoring, quality-of-service and intrusion detection Systems. In this paper we present a machine learning approach that accurately classifies internet traffic in near real-time, using C4.5 decision tree. Using the flow behaviour and port numbers of TCP traffic and collecting only 12 features from the start of the flow, our method can identify traffic of different types of applications with 99.8% total accuracy.

Keywords

Behaviour-based traffic classification, online classification; C4.5 decision tree.

1. INTRODUCTION

Along with the development and evolution of the applications on the Internet, an efficient application classification scheme is highly desirable to support various solutions such as advanced network monitoring, network resource management, anomaly detection, application-specific strategies and network auditing activities. Moreover, the application level knowledge of Internet is extremely useful for those who set out to model the Internet traffic or to investigate the long-term changes and requirements for the Internet.

The traffic, in principle, is the product of a complex multifactor system involving a range of networks, hosts, applications and different people closely interacting with each other. The complexity is continuously increasing as people keep producing a vast variety of network applications and application layer protocols, in many ways breaking the traditional assumptions:

- (a) [1] noted that only 50-70% of the Internet traffic was classifiable using the official International Assigned Number Authority (IANA) list. Emerging new applications and proxies often avoid the use of standard host ports.
- (b) Port-based schemes are also overly simplistic confusing applications that use common ports, such as a VoIP telephony system, chat-messenger systems such as MSN Instant Messenger, and regular web-page browsing all using the same port.
- (c) The proportion of encapsulated or encrypted traffic is increasing. Examples include proxies, VPN, tunneling, and applications using a different protocol to exchange data.

Instead of using port numbers or inspecting packet payloads, we developed a classification scheme based on packet- and flow-level behaviours which provides near real-time classification of up to 99.8% of traffic, by collecting only 12 features from in maximum 5 packets from the start of the flow.

Our methodology first divides the total traffic-mix into classifiable objects. The basic object of our network traffic classification system is the network flow: a bi-directional session between two hosts with the same 5-tuple {host-IP, clnt-IP, host-Port, clnt-Port, timestamp of the first packet}. We capture a range of traffic patterns from the beginning of a network flow before we classify the flow based on a model trained upon prior knowledge of these patterns.

2. RESULTS

Our experimental data comprises of two consecutive week-days of Internet traffic with an 8 month interval. The datasets, Day1 and Day2, are chosen from a collection detailed in [2]. We left a number of TCP traffic in the datasets unconsidered: those we haven't seen the start of the flow (typically those are with very long duration) and junk flows. The resulting data being analysed contains 31 GBytes and 42 million packets in 377 thousand TCP flows in Day1, and 28 GBytes and 35 million packets in 175 thousand TCP flows in Day2. Every flow in the two datasets was hand-classified using a content-based mechanism into one of the 10 applications classes (e.g. Web-browsing, Email, Bulk (ftp) and P2P), to serve as the ground truth data.

2.1 Offline Traffic Classification

Some potential applications such as network auditing and study of the Internet require only offline classification information. In Table.2 we compare discriminative algorithms (C4.5[3], AdaBoost [4] +C4.5, and Joint Boosting [5]) with past approaches [6, 7], based on the performance in classifying complete flows (rather than start-of-the-flow) offline. Results indicate that these methods can achieve much more satisfactory accuracy for the same dataset.

2.2 Online Traffic Classification

Provided we perform our analysis of flows offline, the practicability of such a system would be very much limited to merely analytical and auditing purposes. Therefore, our concerns are no longer solely focused upon the accuracy but also upon the latency and throughput of the system.

Since many applications would benefit from early identification of the traffic, our approach considers capturing the features upon an observation window of only a few packets at the start of the flow, instead of upon the entire flows. Further, for real-time quality and throughput, we can only capture a small amount of features. This also determines the choice of classification algorithm. We choose C4.5 not only for the high accuracy shown in empirical results, but also because of the low testing complexity, as shown in Table 3. AdaBoost+C4,5, although shows better accuracy on the complete feature set, might *overfits* more quickly with the small feature set.

Table 2 Comparison of accuracy of algorithms given complete set of features to choose from. We adopt the default algorithm implementations for C4.5 and AdaBoost in Weka [8], under a right-out-of-box condition. The accuracy values are measured by two-fold cross validation on the whole datasets unless otherwise noted. (a) Due to the time cost in training for Day1 dataset, Joint Boosting used a subset of 7340 flows (4404 flows for training and 2936 flows for testing) selected from Day1.

	C4.5	AdaBoost+C4.5	Joint Boosting	Bayes Neural Network	NBs+kernel
Day1 accuracy	99.813%±0.057%	99.826%±0.055%	99.57% ^(a)	99.26%±0.4%	91.94%±0.35%
Day2 accuracy	99.942%±0.032%	99.979%±0.019%	N/A	99.84%±0.2%	92.99%±0.33%
Notes		After 10 rounds.	After 93 rounds.	As described in [7]	As described in [6]

As C4.5 will automatically select features from the given feature set, we limit the cost by reducing the number of features in the feature set.

The feature subset is selected by a correlative-based filtering algorithm. Table 4 lists these features, as well as the information and complexity properties of these features.

The maximum number of packets in observation window is tuned based on empirical results. Figure 1 below shows the result: from a total of 5 or 6 packets collected it can achieve the highest 99.842% ten-fold cross-validation accuracy in Day1 (99.834% for corresponding two-fold cross-validation). The accuracy is actually higher than the accuracy with complete flows, which mirrored the results in [9] where the authors found their highest accuracy from an observation window of 4 data packets.

3. CONCLUSIONS

In the Internet there is an ever-increasing quantity and variety of traffic motivated by a desire to identify the applications of the Internet, in this paper we present a machine learning approach for network traffic classification based on traffic behaviour. Our approach resulted in great improvements on the overall performance: the accuracy, throughput and latency. Finally, it has shown very promising feasibility for practical applications.

4. REFERENCES

- [1] A. Moore, K. Papagiannaki. Toward the Accurate Identification of Network Applications. In *Proceedings of Sixth Passive and Active Measurement Workshop (PAM 2005)*, April 2005, Boston, MA
- [2] A. W. Moore, D. Zuev, M. Crogan. Discriminators for use in flow-based classification. *Technical report*, Queen Mary University of London, 2005.
- [3] J. R. Quinlan. C4.5: Program for Machine Learning. Morgan Kaufman, San Mateo, CA. 1993
- [4] R. E. Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2001. Berkeley, CA, USA.
- [5] A. Torralba, K. P. Murphy, W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proceedings of Computer Vision and Pattern Recognition 2004 (CVPR 2004)*.
- [6] A. W. Moore and D. Zuev. Internet Traffic Classification Using Bayesian Analysis Techniques. In *Proceedings of ACM SIGMETRICS 2005*. Banff, Alberta, Canada.
- [7] T. Auld, A. W. Moore, S. F. Gull. Bayesian Neural Networks for Internet Traffic Classification. In *IEEE Transactions on Neural Networks*, Nov 2006.
- [8] I. H. Witten, E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, June 2005.
- [9] L. Bernaille, R. Teixeira, K. Salamatian. Early Application Identification. In *Proceedings of 2006 ACM conference on Emerging network experiment and technology (CoNEXT'06)*. Lisboa, Portugal.

Table 3 Algorithms for online traffic classification. The accuracies are from two-fold cross validation on Day1 dataset, using 12 features, with an observation window of 5 packets. Time values are for training and testing 325,000 objects.

	NB + kernel	C4.5	AdaBoost+ C4.5
Overall accuracy	92.38%± 0.35%	99.834%± 0.052%	99.816%± 0.057%
Complexity	O(features)+O(classes)	O(tree depth)	O(tree depth × rounds)
Testing time	1412s	1.5s	11s
Training time	<8s	133s	1505s

Table 4 Properties of a subset of features from [19]. S=start of observation; D=during observation; E=end of observation. Symmetrical Uncertainty measures the discriminative power of the individual feature.

Feature	Collection Time	Symmetrical Uncertainty	Memory Overhead	Complexity
serv port	S	0.8397	O(1)	O(1)
clnt port	S	0.0742	O(1)	O(1)
push_packets_serv	D	0.1900	O(1)	O(n)
init_win_bytes_serv	D	0.0565	O(1)	O(1)
init_win_bytes_clnt	D	0.2169	O(1)	O(1)
avg_seg_size_serv	E	0.1604	O(1)	O(n)
IP_data_bytes_med	E	0.2881	O(n)	O(n ²)
act_data_pkt_clnt	D	0.1680	O(1)	O(n)
data_bytes_var_serv	E	0.1219	O(n)	O(n)
min_seg_size_clnt	D	0.2065	O(1)	O(n)
RTT_samples_clnt	D	0.1899	O(1)	O(n)
push_packets_clnt	D	0.2123	O(1)	O(n)

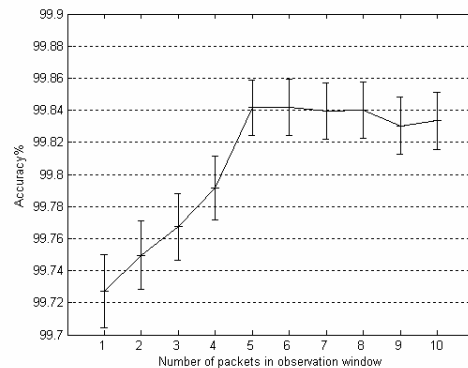


Figure 1 Accuracy with different packet number limits