A Survey on Network Traffic Classification Based on Flow Accuracy

Ruhama David

Masters in Karunya University, India

Abstract

In recent years the research community is looking for classification of network application traffic to be done quickly and more accurately with increased security and QoS control compared to the past decade. This survey paper deals with the supervised Machine Learning (ML) methods based on the classification flows of different applications in the network without inspecting the packet payload. So, it is also suitable for encrypted protocols. Also accuracy of flow classification is compared with different methods such as SVM, C4.5, J48, APPR.

***Key words**: Machine Learning, Traffic Classification.*

1. Introduction

In recent years, the increase in dynamic communication has affected net communities and has attracted varied efforts at analysis. Dynamic communication protocols transfer information with dynamic parameters like unfixed TCP/UDP ports and tunnelling transferring mechanisms. Network operators ought to understand what's flowing over their networks promptly in order to react quickly in support of their numerous business goals. Traffic classification could also be a core part of machine driven intrusion detection systems. The analysis community has responded to this by work classification schemes capable of inferring application level usage patterns without deep examination of packet payloads.

The traffic classification can be done based on different levels, i.e. at social level, functional level and application level [1]. At social level, we capture the behaviour of a bunch of packets as indicated by its interactions with different hosts. At functional level, we capture the behaviour of the host in terms of its useful role within the network, namely whether or not it acts as a supplier or client of a service, or both, just in case of a cooperative application. At application level, we capture the transport layer interactions between specific hosts on specific ports with the intent of spotting the application of origin. Commonly traffic classification techniques have been based around direct inspection of every packet's contents at some point in the network.

Table 1: Network Applications for Classification

Class	Applications
Bulk	ftp
Interactive	ssh, telnet, rlogin
Mail	pop3, smtp, imap
Service	X11, dns
WWW	http, https
P2P	Kazaa, Bittorrent, Gnutalla
Multimedia	Voice, video-streaming
Game	Half-life
Attack	Worm, virus
Others	Scan, netbios, tsp, ntp

Successive IP packets having the same 5 tuples: protocol, source address, destination address, source port, destination port, port belongs to a flow whose dominant application we have a tendency to want to work out [1] [2] [3] [4] [5]. Also classification accuracy is compared with different ML methods and classification is more effective. In this paper while comparing APPR method [5] was found to have better classification flow accuracy than other methods of ML.

The rest of this paper discusses about four techniques or methods related to traffic classification based on flows using ML techniques. The last section of this survey paper is the conclusion of the comparative study of the papers.

2. ML Methods and Metrics Used for Traffic Classification

Different ML Methods are Supervised (classification), Semi-supervised and Un-supervised (clustering). Classification learning involve a ML from a set of pre-classified also called pre-labelled example, from which it build a set of classification rules to classify unseen examples. Clustering is the group of instances that have related features into clusters, without any earlier control. Most ML techniques used for network traffic classification focus on the use of supervised and unsupervised learning.

A key measure on which to make a distinction between classification techniques is how accurately the technique or model makes decisions when accessible with previously unseen data. A common way to characterize a classifier's accuracy is through metrics known as False Positives, False Negatives, True Positives and True Negatives. ML methods have two additional metrics known as Recall and Precision. Accuracy is also considered as classification metrics.

3. Comparative Study of Traffic Classification Based on Flows Using ML Techniques

3.1. BLINC: Multilevel traffic classification in the dark

In this approach [1], an application classification method supported the behaviours of the supply host at the transport layer which is divided into 3 totally different levels. The social level captures and analyses the interactions of the examined host with alternative hosts, in terms of the numbers of hosts it communicates with. The popularity of the host and of others in its community circle is considered. The role of the host, in acting as a supplier or the patron of a service, is classified at the purposeful level. Finally, transport layer information is employed, like the 4-tuple of the traffic source and destination IP addresses, and source and destination ports, flow characteristics like the transport protocol, and the average packet size.

A range of application sorts were studied during this work, including internet, p2p, knowledge transfer, network management traffic, mail, chat, and media streaming, and gaming. By analysing the social activities of the host, the authors conclude that among the host's communities, near IPs could supply identical service if they use identical service port, exact communities may indicate attacks, whereas partial communities may signify p2p or gaming applications. Additionally, most IPs acting as clients have a minimum variety of destination IPs.

Thus, specializing in the identification of that small variety of servers will facilitate client identification, resulting in the classification of an oversized amount of traffic. Classification at the functional level shows that a host is likely to be giving a facility if during a period of time it uses a small number of source ports, normally less than or equal to two for all of their flows. Client behaviour can usually be represented when the number of source ports is equal to the number of distinct flows. The steadiness of average packet size per flow across all flows at the application level is recommended to be fine characteristics for identifying definite applications, such as gaming and malware.

Completeness and accuracy are the two metrics used for the classification approach. Completeness is outlined because the magnitude relation of the quantity of flows classified by BLINC over the total range of flows, indicated by payload analysis. The results show that BLINC will classify 80% to 90% traffic flows with 95% flow accuracy. This technique has got to gather information from many flows for each host before it can choose the role of one host. Such requirements would possibly stop the utilization of this technique in real time operational networks.

3.2. Accurate Classification Based on SVM Method

In this approach [2], uses SVM (Support Vector Machine) method for traffic classification of applications in the network. This method developed during this paper for classifying seven classes of internet applications with completely different characteristics. The following steps are taken to improve the accuracy:

(i) Some features are used from network flow and from real time packet header.
(ii) Accuracy of classification for biased and unbiased training samples are compared
(iii) To obtain best combination of features classification was done by using discriminator selection algorithm.

SVM is a widely used technique for pattern recognition to avoid local optimization and optimal statistical classification. Here 10-fold cross validation scheme to used for the same testing and training samples. For unknown data, prediction accuracy has much influence in the performance of classification. In discriminator selection algorithm, different methods are used to get the best combination of features for classification. The main methods used are Optimum Searching, Hypo-optimum Searching methods.

Packets can be collected from 4-tuples such as source address, destination address, source port, destination port, packet length. In this approach, the best combination is obtained from 19 features in the real time packet header. For biased traffic samples 99.4% accuracy can be achieved and for biased traffic samples this method can achieve 96.9% accuracy.

3.3. Early Identifying Application Traffic with Application Characteristics

In this approach, flow behaviours are characterized based on application layer perspective is considered the negotiation behaviour of each flow attribute is considered. Also discriminators which are available at the early stage will support the real time traffic classification. The flow accuracy was tested by using several ML algorithms and accuracy of flow was compared with previous methods.

Here, the early stage identification is done after the establishment of L4-flow without inspecting the packet payload, i.e. it identifies that the particular application is in TCP/UDP. Next it will find out the particular flow for the application, so that several interaction rounds are found based on some attributes such as: transmitted size, throughput and elapsed time. L4-flow is identified by using 5-tuples: IP client and server, Port client and server, L4 protocol. Here first 20 data packets were only considered for identification, 10-fold cross validation method is used for testing and training samples of traffic traces. Then classification accuracy is calculated based on flows. Different ML schemes are used for comparing the accuracy of flows.

The result focused on 12 protocols to test the ability of attributes and compare the accuracy of ML methods. The result shows that this method can achieve high accuracy compared to previous methods i.e. J48, PART and Naive Bayes with low FP rate. This method brings 8% to 21% improvement of accuracy for flow classification. It is also suitable for encrypted protocols, without inspecting the packet content.

3.4. Early Identification of Peer-To-Peer Traffic

Here [4], only the first few bytes of the first packets of each flow are analyzed. The automatic ML algorithm is proposed for classification of flow from traffic traces. The traffic traces are from LAN, WiFi, and 3G links. Early classification algorithms are used to classify the flows of the packets. Kiss algorithm, first order Markov models are also used to compare with the performance of automatic payload-based traffic classification. Another method used is Random Forest for classification of many decision trees. The classification accuracy can improve significantly by using several decision trees and also the method has more robustness against noise.

To handle the asymmetric nature of routing the method uses only first few packets for classification of TCP/UDP. After this cross validation can be done for the training and testing datasets.

This paper showed that P2P traffic classification can be done effectively for first few bytes of the first packet of each flow. Here that no need of human expertise to design the appropriate signatures. This approach doesn't use the real traffic traces to improve the quality of labelling.

3.5. Application traffic classification at the early stage by characterizing application rounds

In this approach, they are doing the one in three is enhanced [3].Here it uses 59 protocols to test the ability of proposed classifier.

The early stage identification is done after the establishment of L4-flow without inspecting the content of packet payload. Next it finds out the particular flow for each application, so that several interaction rounds are done based on attributes such as: transmitted size, throughput, elapsed time, response time, inter arrival time and transport layer feature (protocol type, port number of client and server, flag of the first data sender). Each TCP/UDP flow is chosen to find early flows for first n rounds of the interaction of each application. Here also 5-tuples are used to identify the L4-flow: IP client and server, Port client and server, L4 protocol. After this classification, the training and testing datasets will do 10-fold cross validation by using ML algorithms and then sampling. Then classification accuracy of flow is calculated based on flows, so that 3 classes are there: TCP_{only}, UDP_{only} and

TCP+UDP. In these three classes different ML methods are applied for evaluating the accuracy.

The result focuses on 59 protocols to test the ability of attributes used and compare the accuracy of ML methods. The result shows, that this method (APPR) can achieve high accuracy than previous methods i.e. J48, PART and Naive Bayes with low FP rate. This method brings 15% to 30% improvement in accuracy for flow classification. It is also suitable for encrypted protocols, without inspecting the packet content.

4. Conclusion

This paper surveys significant work in the field of early stage application classification from the period of 2005 to 2013. The parameters used for classification of application are IP client and server, port client and server and L4 protocol. Here the classification is based on flow of packets and the accuracy is compared with different ML algorithms in supervised ML algorithm. From the analysis, the method which has used in the paper [5] can attain accuracy as nearly 99.21% when compared to C4.5, J48, RF and SVM. In APPR method they use 59 protocols and more than 25 features for early identification of the applications. In all these papers, real traffic traces are methods for accurate classification and suitable for encrypted protocols without inspecting the payload.

5. References

[1] T. Karagiannis, K. Papagiannaki, M. Faloutsos, BLINC: multilevel traffic classification in the dark, in: Proc. ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, Pennsylvania, PA, 2005, pp. 229–240.

[2] Z. Li, R. Yuan, X. Guan, Accurate classification of the internet traffic based on the SVM method, in: Proc. IEEE Int. Conference on Communications (ICC‘07), Glasgow, Scotland, 2007, pp. 1373–1378.

[3] N.F. Huang, G.Y. Jai, H.C. Chao, Early identifying application traffic with application characteristics, in: Proc. IEEE Int. Conference on Communications (ICC ‘08), Beijing, China, 2008, pp. 5788–5792

[4] B. Hullar, S. Laki, A. Gyorgy, Early identification of peer-to-peer traffic, in: Proc. IEEE Int. Conference on Communications (ICC 2011), Kyoto, Japan, 2011, pp. 1–6.

[5] Nen-Fu Huang, Gin-Yuan Jai, Han-Chieh Chao, Yih-Jou Tzang, Hong-Yi Chang, Application traffic classification at the early stage by characterizing application rounds in Information Sciences 232 (2013) 130–142