Towards a Flow-based Internet Traffic Classification for Bandwidth Optimization

Abuagla Babiker Mohd Bmbabuagla2@siswa.utm.my

Faculty of electrical engineering
University Technology Malaysia UTM
Johore bahru, 813100 , Malaysia

Dr. Sulaiman bin Mohd Nor sulaiman@fke.utm.my
Faculty of electrical engineering
University Technology Malaysia UTM
Johore bahru, 813100 , Malaysia

ABSTRACT

The evolution of the Internet into a large complex service-based network has posed tremendous challenges for network monitoring and control in terms of how to collect the large amount of data in addition to the accurate classification of new emerging applications such as peer to peer, video streaming and online gaming. These applications consume bandwidth and affect the performance of the network especially in a limited bandwidth networks such as university campuses causing performance deterioration of mission critical applications.
Some of these new emerging applications are designed to avoid detection by using dynamic port numbers (port hopping), port masquerading (use http port 80) and sometimes encrypted payload. Traditional identification methodologies such as port-based signature-based are not efficient for today's traffic.
In this work machine learning algorithms are used for the classification of traffic to their corresponding applications. Furthermore this paper uses our own customized made training data set collected from the campus, The effect on the amount of training data set has been considered before examining, the accuracy of various classification algorithms and selecting the best.
Our findings show that random tree, IBI, IBK, random forest respectively provide the top 4 highest accuracy in classifying flow based network traffic to their corresponding application among thirty algorithms with accuracy not less than 99.33%.

Keywords: NetFlow, machine learning, classification, accuracy, decision tree, video streaming, peer to peer.

1. INTRODUCTION

Network management is a service that employs a variety of tools, applications and devices to assist human network managers in monitoring and maintaining networks.
In recent times, ISPs are facing great challenges due to the domination of certain unregulated applications which are affecting their revenue.

Classifying traffic according to the application or application class that produces it is an essential task for network design and planning or monitoring the trends of the Internet applications. Since the early 2000s, application detection has become a difficult task because some applications try to hide their traffic from traditional detection mechanism like port based or payload based [1]
In this work, we focus on selecting the best machine learning classification algorithm for identifying flow -based traffic to their originating applications. That can be done by testing the accuracy of 30 algorithms then selecting the best 15, then doing deeper checking to reduce the best set (e.g the best 4 algorithm from an accuracy point of view).
Since our goal is to classify traffic for traffic control purpose, the extended work will focus in testing the effect of time for those four best algorithms so as to build a near real-time system for traffic control.
The remainder of this article is structured according to the following topics: Related work, methodology, results and discussion, and finally, conclusion and future work.

2. Related Work

A lot of research work has been done in the area of network traffic classification by application types and several classifiers have been suggested. However the majority of them are based on transport layer port-numbers (currently lack of efficiency because of port hopping and port tunneling), signature-base (which fails to identify encrypted payloads), heuristics or behavioral-based (which is not efficient for real time or online classification). Recently machine learning is widely used in this field. The following subsections explore these approaches in more details.

2.1 Port Number Based Classification

This method classifies the application type using the official Internet Assigned Numbers Authority (IANA) [2] list. Initially it was considered to be simple and easy to implement port-based inline in real time. However, nowadays it has lower accuracies to around about between 50% to 70% [3]. Many other studies [4, 5, 6, and 7] claimed that mapping traffic to applications based on port numbers is now ineffective. Network games, peer to peer applications, multimedia streaming uses dynamically assigned ports (port hopping) for their sub transactions, so it is difficult to classify them using this method. Also, the above mentioned applications can disguise their traffic as other known classes (such as http, and ftp).

2.2 Payload Based Classification

In this approach packet payloads are examined to search for exact signatures of known applications. Studies show that these approaches work very well for the current Internet traffic including many of P2P traffic. So that this approach is accepted by some commercial packet shaping tools (e.g. Packeteer. Several payload-based analysis techniques have been proposed [3, 7, 8, 9].
Payload-based classification still has many disadvantages. First, these techniques only identify traffic for which signatures are available and are unable to classify any other traffic. Second, payload analysis requires substantial computationally power [1], [10] and storage capacity [11] since it analyzes the full payload. Finally, the privacy laws [10] may not allow administrators to inspect the payload and this technique will fail if payload is encrypted.
Alok Madhukar et al. [7] focus on network traffic measurement of peer to peer applications on the Internet, The paper compared three methods to classify P2P applications: port-based analysis, application-layer signature and transport layer heuristics. Their results show that classic port-based analysis is ineffective and has been so for quite some time. The proportion of "unknown" traffic increased from 10-30% in 2003 to 30-70% in 2004-2005. While application-layer signatures are accurate, it requires the examination of user-payload, which may not always be possible

2.3 Protocol Behavior or Heuristics Based Classification

Transport-layer heuristics offer a novel method that classifies traffic to their application types based on connection-level patterns or protocol behavior. This approach is based on observing and identifying patterns of host behavior at the transport layer. The main advantage of this method is that there is no need for packet payload access.

BLINK [12] proposed a novel approach to classify traffic in the dark. It attempts to capture the inherent behaviors of a host at three different levels, first, social level, which examines the popularity of the host. Second, the functional level which means whether the intended host provides or consumes the service. Finally, the application level that is intended to identify the application of the origin. The authors claimed that their approach classified approximately 80% - 90% of the total number of flows in each trace with 95% accuracy. However this method cannot be applied for inline near real time classification because of the classification speed limitation.

Fivos, et al [13] presented a new approach for P2P traffic identification that uses fundamental characteristics of P2P protocols such as a large diameter and the presence of many hosts acting both as servers and clients. The authors do not use any application-specific information and thus they expect to be able to identify both known and unknown P2P protocols in a simple and efficient way. Again, from the study done, it is anticipated that they will face a problem due to port tunneling.

2.4 Statistical Analysis Based Classification

This approach treats the problem of application classification as a statistical problem. Their discriminating criterion is based on various statistical features of the flow of packets e.g. number of packets, packet size, inter arrival time. Machine learning is used for classification. The advantage of this approach is that there is no packet payload inspection involved.

Nigel Williams et al. [14] compared five-widely used machine learning classification algorithms to classify Internet traffic. Their work is a good attempt to create discussion and inspire future research in the implementation of machine learning techniques for Internet traffic classification. The authors evaluated the classification accuracy and computational performance of C4.5, Bayes Network, Naïve Bayes and Naïve Bayes Tree algorithms using feature sets. They found that C4.5 is able to identify network flows faster than the remaining algorithms. Also they found that NBK has the slowest classification speed followed by NBTree, Bayes Net, NBD and C4.5.

Jiang, et al. [15] showed by experiments that NetFlow records can be usefully employed for application classification. The machine learning used in their study was able to provide identification accuracy of around 91%. The authors used data collected by the high performance monitor (full packet capturing system) where NetFlow record was generated by utilizing nProbe (a software implementation of Cisco NetFlow).

Wang1, et al. [16] discovered a new method based on the support vector machines (SVM) to solve the P2P traffic identification and application classification problem. Experimental results show that their method can achieve high accuracy if they carefully tune the parameters, and it also has promising identification speed.

3. Methodology

In this work, the approach used here exploits the statistical flow features of the application during the communication phase. Machine Learning (ML) is used to classify the traffic. Weka toolkit [17], MYSQL-database and other supporting programs has been used to achieve our goals. Figure 3.1 represents the flow diagram of the methodology.

This paper concentrates on an accurate classification of the bandwidth consumer applications e.g. video streaming for traffic profiling purposes. This can be used later for traffic control, capacity planning or usage based billing that can indirectly contribute on enhancing the performance of the network. The following subsections explain the flow diagram in more details.

3.1 Data Collection

Figure 3.2 shows the test bed setup which is used in data collection for this work. The setup is a typical faculty environment which is generating mixed Internet traffic of different applications. UDP traffic represents a small amount of data compared to TCP traffic (74.3% of the UDP flows consist only one packet per .

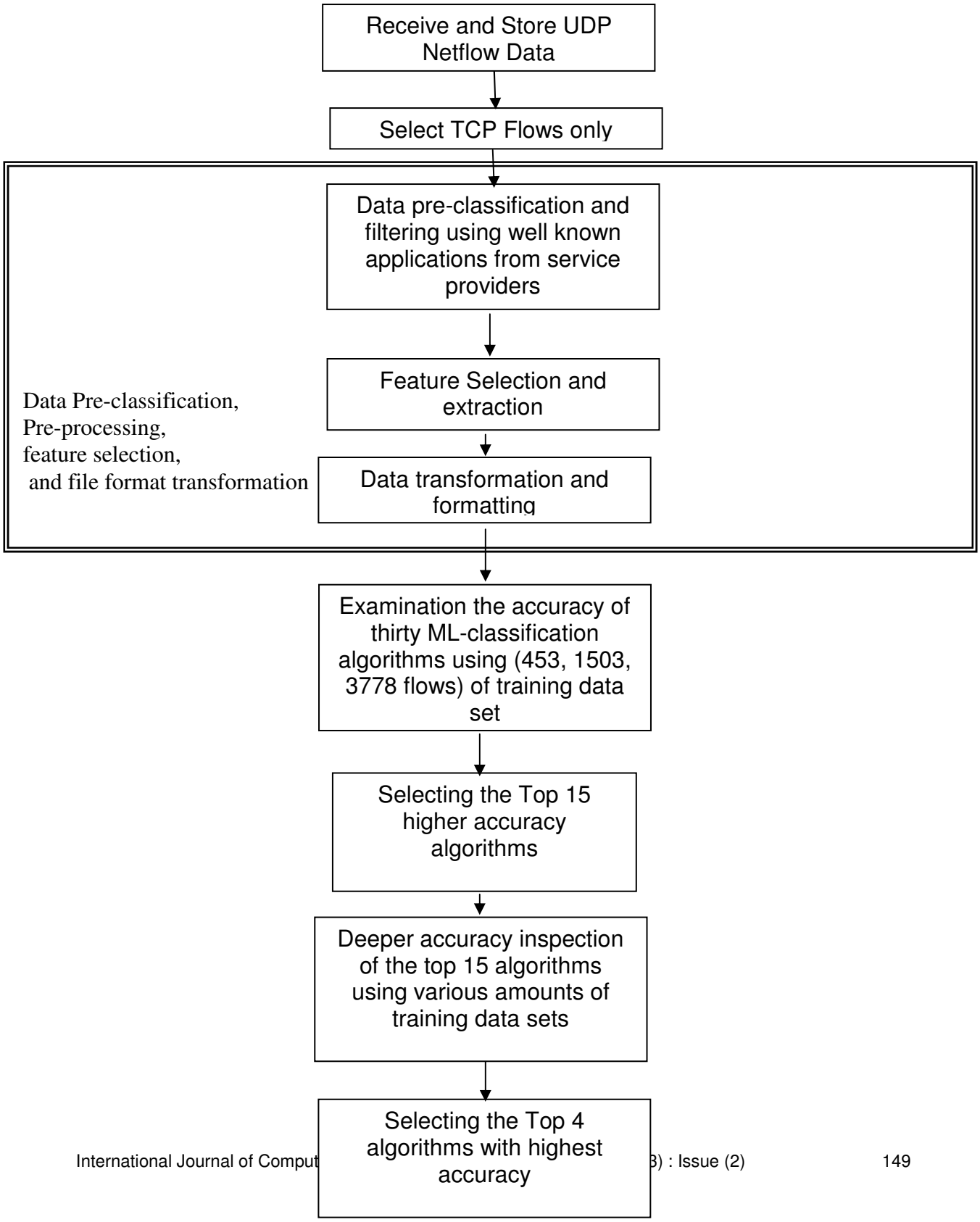

FIGURE 3-1: Methodology Flow Diagram

flow). For the purpose of evaluating algorithms in classifying the network traffic, only TCP flows have been taken into consideration.
Here, all users are connected to access switches which collapses to a distribution switch. These distribution switches in turn is connected to core routers centrally located. NetFlow data was collected from the distribution and exported to a NetFlow collector server. The collected NetFlow data were stored into Mysql database for further preprocessing, profiling and off-line classification.

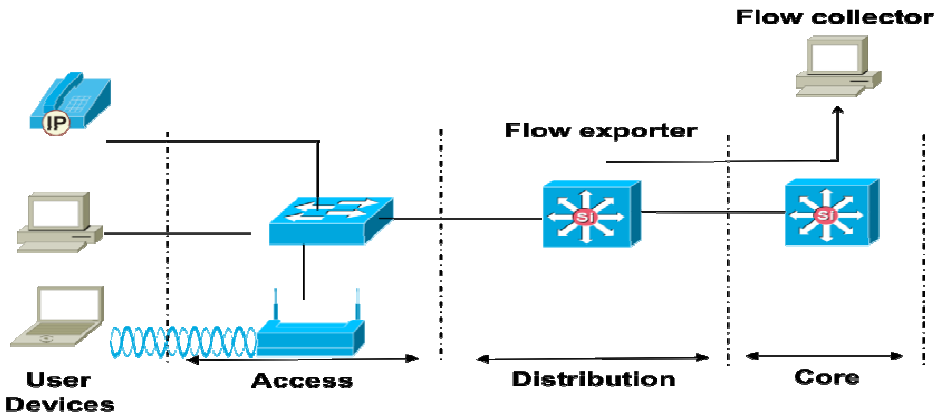

FIGURE 3-2: typical setup in a faculty with NetFlow exporter and collector

3.2 Data Pre-classification and Filtering

Using well known providers that always provide well known services, such as google.com for http, gmail and yahoo for mail service and youtube.com for video streaming, traffic pre-classification has been implemented. The pre-classified training data has been converted from table format to text file so as to train and tests Weka's classification algorithms.

3.3 Data Preparations, Preprocessing, and Accuracy Testing

From the NetFlow records, useful statistical features such as number of bytes, numbers of packets, start time and last time have been extracted. Derived features have also been produced which includes duration, mean packet size, mean inter arrival time and class.
The NetFlow data were prepared and processed in an acceptable format for further file conversion process to be compatible with Weka toolkit [18].
These pre-classified data has been used to train different classification algorithms, initially thirty machine learning classification algorithm with different amount of datasets {453, 1503, 3778 flows}. From them, the 15 best accuracy algorithms have been taken for deeper accuracy examinations using various amounts of datasets.
To accurately select the best classification algorithms that give more accuracy in our current situation, different amount of training datasets starting from (453 to 3778 flows) have been applied to each of the 15 classification algorithm so as to examine their accuracy. Also to obtain the best accuracy, according to our previous work [18] we choose flows from servers to client's direction. Finally the 4 top accuracy algorithms have been determined.

4. Results and Discussion:

In our work we neglect UDP flows because most of the applications use TCP as a transport layer protocol. Furthermore the number of UDP flows that consist of less than or equal to 2 packets per flow equals to 84.15 %

As stated in the methodology our initial testing domain started with examining the accuracy of thirty (30) machine learning classification algorithms to identify the originating application of the flow based network traffic. The 15 top (with accuracy.>70%) classification algorithms were chosen for further accuracy investigation. This has been shown in and figures 4-1 and table 4-1.

Data Series →		1	2	3	4	5		Average Of Accuracy
Algorithm No ↓	Data/ algorithm	453	978	1503	2553	3078	3778	
1	Bayesnet	73.7	74.53	77.31	77.43	78.16	79.11	76.71
2	IBI	100	99.69	99.73	99.49	99.54	99.52	99.66
3	IBK	100	99.69	99.73	99.64	99.67	99.68	99.74
4	Kstar	95.36	94.06	94.67	93.57	93.3	93.19	94.03
5	BFTree	89.85	87.12	89.68	88.09	90.44	90.92	89.35
6	J 48	88.78	88.03	86.36	86.44	86.48	85.86	86.99
7	lmt	91.17	90.28	88.56	83.3	87.26	87.18	87.96
8	nbtree	75.74	80.87	82.96	83.5	81.54	83.77	81.4
9	Random forest	99.56	99.08	99.334	99.13	99.38	99.47	99.33
10	Random tree	100	99.69	99.73	99.64	99.67	99.68	99.74
11	REP tree	80.79	77.33	82.96	82.02	82.32	85.7	81.85
12	simpleCart	90.29	88.75	91.28	91.89	92.59	92.61	91.24
13	rules.jrisp	70.4	74.23	78.04	79.7	79.98	77.42	76.63
14	rules.part	86.7	85.37	83.83	85.82	80.83	84.83	84.56
15	rules.rider	80.13	81.08	81.9	81.66	82.97	83.24	81.83

Table 4-1 Accuracy matrix of algorithms (column) corresponding to the various training data sets (row)

As can be seen from table 4-1 , and figure4-1 it is clearly reported that random tree , IBI, IBK, random forest respectively provide the top 4 highest accuracy in classifying flow based network traffic to their corresponding application type among the 15 selected algorithms as an overall average accuracy. Furthermore the results also show that theses algorithms give high accuracy regardless the amount of training data sets (average accuracy of more than 99.33%).

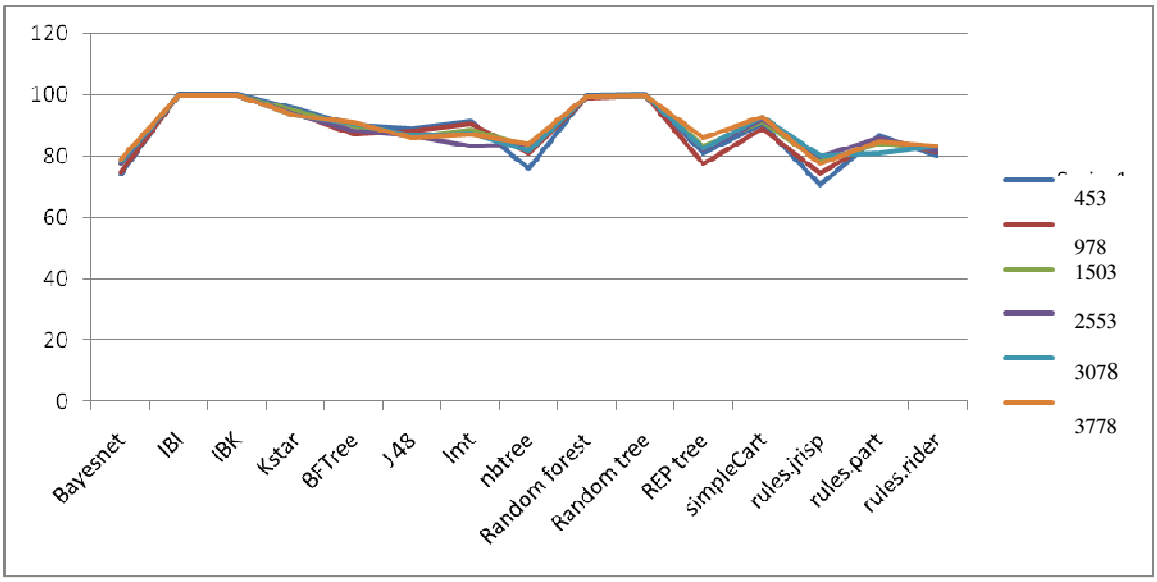

FIGURE 4-1: algorithms and their corresponding accuracy with various training data sets

5. Conclusion and Future Work

This paper evaluates the accuracy of 30 Machine learning classification algorithms as one of the important performance metrics using custom made datasets.

Our new findings show that random tree, IBI, IBK, random forest scored the top 4 highest accuracy classification algorithms among others for identifying flow–based network traffic to their corresponding application type.

However we started with offline classification and selected the best algorithms based on accuracy. This chosen algorithm will also be tested for processing time as a future work and the best algorithm according to time and accuracy will be used for real time inline detection.

Since our primary goal is to regulate the network traffic and to optimize the bandwidth, our future work must consider the time factor of the classification model.

6. REFERENCES

[1] Daniel Roman Koller, Application Detection and Modeling using Network Traces, master thesis “swiss federal institute of technology,2007

[2] http://www.iana.org/assignments /port-numbers

[3] A.W.Moore and D.papagiannaki, “Toward the accurate Identification of network applications”, in poc. 6th passive active measurement. Workshop (PAM), mar 2005,vol. 3431, pp 41-54

[4] Williamson, A. M. C. (2006). A Longitudinal Study of P2P Traffic Classification. Proceedings of the 2th IEEE International Symposium on (MASCOTS '06), Los Alamitos, California, IEEE. Pp 179 - 188

[5] T. Karagiannis, A. B., and N. Brownlee (2004). Is P2P Dying or Just Hiding? . GLOBECOM '04. Dallas, USA, IEEE: pp:1532 - 1538 Vol.3.

[6] Thomas, K., B. Andre, et al. (2004). Transport layer identification of P2P traffic. Proceedings of the 4th ACM SIGCOMM conference on Internet measurement. Taormina, Sicily, Italy, ACM: pp: 121 - 134

[7] Subhabrata, S., S. Oliver, et al. (2004). Accurate, scalable in-network identification of p2p traffic using application signatures. Proceedings of the 13th international conference on World Wide Web. New York, NY, USA, ACM: pp: 512 - 521

[8] Christian, D., W. Arne, et al. (2003). An analysis of Internet chat systems. Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement. Miami Beach, FL, USA, ACM: pp: 51 - 64

[9] Patrick, H., S. Subhabrata, et al. (2005). ACAS: automated construction of application signatures. Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data. Philadelphia, Pennsylvania, USA, ACM: pp: 197 - 202

[10] Feldman, R. S. a. A. "An IDS Using NetFlow Data." Retrieved march 2008.

[11] Jeffrey, E., A. Martin, et al. (2006). Traffic classification using clustering algorithms. Proceedings of the 2006 SIGCOMM workshop on Mining network data. Pisa, Italy, ACM: pp: 281 - 286

[12] Thomas, K., P. Konstantina, et al. (2005). BLINC: multilevel traffic classification in the dark. Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications. Philadelphia, Pennsylvania, USA, ACM: pp: 229 - 240

[13] Fivos Constantinou, Panayiotis Mavrommatis, “Identifying Known and Unknown Peer-to-Peer Traffic “, Fifth IEEE International Symposium on Network Computing and Applications (NCA'06) 0-7695-2640-3/06 $20.00 © 2006 IEEE

[14] Nigel, W., Z. Sebastian, et al. (2006). "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification." SIGCOMM Comput. Commun. Rev. 36(5): 5-16.

[15] Hongbo, J., W. M. Andrew, et al. (2007). Lightweight application classification for network management. Proceedings of the 2007 SIGCOMM workshop on Internet network management. Kyoto, Japan, ACM: pp: 299 - 304

[16] Rui Wang1, Y. L., Yuexiang Yang3, Xiaoyong Zhou4 (16-18 October 2006). Solving the App-Level Classification Problem of P2P Traffic via Optimized Support Vector Machines. Proceedings of the Sixth International Conference on Intelligent Systems