Near Real Time Online Flow-based Internet Traffic Classification Using Machine Learning (C4.5)

Abuagla Babiker Mohammed Bmbabuagla2@siswa.utm.my
Faculty of Electrical Engineering (FKE)
Deprtment of Microelectronics and
Computer Engineering MICE
Universiti Teknologi Malaysia (UTM)
Skudai, Johor, 81310 , Malaysia

Assoc.Prof. Dr. Sulaiman Mohd Nor sulaiman@fke.utm.my
Faculty of Electrical Engineering (FKE)
Deprtment of Microelectronics and
Computer Engineering MICE
Universiti Teknologi Malaysia (UTM)
Skudai, Johor, 81310 , Malaysia

Abstract

Offering reliable novel service in modern heterogeneous networks is a key challenge and an important prospective income source for many network operators and providers. Providing reliable future service in a cost effective scalable manner requires efficient use of networking and computing resources. This can be done by making the network more self enabled, i.e. making it capable of making distributed local decisions regarding the utilization of the available resources. However such decisions must be correlated in order to achieve the global overall goal (maximizing the performance and minimizing the cost)

Since network administrators are always worried about making fast decisions to monitor and regulate the Internet traffic, a novel approach for online flow-based network traffic classification is proposed. This proposal is based on Machine learning algorithm C4.5 and a custom built network traffic data set captured from a university campus environment. Furthermore the aim of this effort is to build a complete online flow based traffic classification and control system.

Validation on the proposed system is done from accuracy and time points of views. Firstly, an offline training and testing data sets are applied to Weka's C4.5 and our system. And their corresponding accuracy has been compared. Our experimental results show that the accuracy is the exactly the same. Secondly, the received UDP NetFlow packets have been send to our system and to a basic packet sniffing program and the number of NetFlow packets has been counted in each. The comparison result show that no packet overwriting due to race condition.

Keywords: NetFlow, machine learning, C4.5, online classification, accuracy, traffic control, P2P.

1. INTRODUCTION

The evolution of the current Internet into a large complex service-based network has generate a tremendous challenges and difficulties for network monitoring and control in terms of how to collect the large amount of data in the recent very fast speed wires. Furthermore how to accurately classify the Internet traffic with the exultance of new emerging applications such as peer to peer, video streaming and online gaming. These applications are considered as bandwidth hungry applications and they affect the performance of the network especially in a limited bandwidth networks such as university campuses causing performance deterioration of mission critical applications. Most of These applications use port hopping and payload encryption to avoid detection. Hence the need of online accurate detection approaches.

Traffic classification at application level is critical for protocol research, abnormity detection, accounting, network security, and network operation [1]. Internet traffic identification and classification is vital to the areas of network management and security monitoring, network planning, and QoS provision. Traditional approaches such as port-based and payload-based identification are becoming increasingly difficult with many new applications (e.g. P2P) using dynamic port numbers, masquerading techniques, and encryption to avoid detection [2].

Real-time Internet traffic classification has the potential to solve difficult network management problems for Internet service providers (ISPs) and their equipment vendors. Especially in today's high speed wires, network operators need to know what is flowing over their networks accurately so that they can react quickly in support of their various business goals [3]. Early classification is essential to allow automatic blocking, filtering, or recording of specific applications [4].

This paper proposes a novel near real time online flow based Internet traffic classification [NOFITC]. An open source code of C4.5 algorithm has been customized to work for online Internet traffic classification. Then the performance of the system has been checked from accuracy and time points of views.

Section 2 explores related work, section 3 shows the methodology, section 4 explains the experimental result, and finally section 5 concludes our work and points for future work.

2. Internet Traffic Classification – An Overview

Although a lot of respective research literatures addresses Internet traffic classification and architectural related topics, relatively little work have been done on developing solution methodologies directly related to near real time Internet traffic measurement and control.

There has been a lot of research in the area of network traffic classification by application types and several classifiers have been suggested. Although statistical based Internet traffic classification shows promising results, however relatively few work has been done related to online Internet traffic classification.

2.1 Port Number Based Classification:

This approach classifies the application type using the official Internet Assigned Numbers Authority (IANA) [5] list. Initially it was considered to be simple and easy to implement port-based online in real time. However, nowadays it has lower accuracies (50% - 70%) [6]. Many other studies [7, 8, 9, and10] claimed that mapping traffic to applications based on port numbers is now ineffective.

Alok Madhukar et. Al. [9] focus on network traffic measurement of peer to peer P2P applications on the Internet. The paper compared three methods to classify P2P applications i.e. port-based analysis, application-layer signature, and transport layer heuristics. They collected their traffic trace from University Calgary Internet connection for a period of two years (2004-2005) .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, this technique requires examination of user-payload, which may not always be possible.

2.3Signature Based Payload Classification

To address the aforementioned drawbacks of port-based classification, several payload-based analysis techniques have been proposed [6, 9, 11, 12, 13, and 14]. 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. These approaches are accepted by some commercial packet shaping tools.
However, P2P applications such as BitTorrent are beginning to elude this technique by using payload encryption, variable-length padding, and/or encryption. In addition, there are some other disadvantages. First, these techniques only identify traffic for which signatures are available and are unable to classify any other traffic. Second, payload analysis consumes computational power [15, 16] because it analyzes the full payload. Third, these techniques typically require increased processing and storage capacity. [17]
Finally, the privacy laws [16, 18] may not allow administrators to inspect the payload
Liu Bin, et al. [19] presented a flexible and efficient BitTorrent measurement system using application signature analysis which has been implemented with standard hardware and Netfilter extension. They demonstrated the feasibility of this approach in a real network environment and showed that the performance is sufficient to accurately measure high volume traffic on high speed links in real-time. They claim that although the measurement system is currently geared towards BitTorrent protocols, it can be easily extended to measure other protocols running over TCP as well.

2.4 Protocol Behavior or Heuristics Based Classification

Transport-layer heuristics offer a novel method that classifies the P2P traffic based on connection-level patterns. 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.
BLINC [13] introduces a new approach for Internet traffic classification. It associates Internet hosts with applications. It looks at all flows (TCP and UDP) generated by specific hosts. BLINC is able to accurately associate hosts with the applications they provide or use (application server, web client, etc.). However BLINC has to gather information from several flows for each host before it can decide on the role of a host. These requirements prevent the use of these methods for online traffic classification. In contrast, our approach relies only on the first few packets of a TCP flow. This early classification is essential to allow automatic blocking, filtering, or recording of specific applications. It also limits the amount of memory required to store information associated with each flow.

2.5 Statistical Analysis Based Classification:

This approach treats the problem of application classification as a statistical problem. It develops its discriminating criteria based on various statistical features of the flow of packets. Machine learning is always used to build the classification model. The advantage of this approach is that there is no packet payload inspection involved.
Nigel Williams et. al. [20] compared five-widely used machine learning classification algorithms to classify Internet traffic. Their work was a good first 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 22 features and with two additional reduced feature sets. They found that the feature reduction techniques were able to greatly reduce the feature space, while only minimally impacting classification accuracy and at the same time significantly increasing computation performance. They also found that the majority of algorithms achieved similar levels of classification accuracy given their feature space and dataset. Also they discovered it was difficult making differentiation between them using standard evaluation metrics such as accuracy, recall and precision.
They found that better differentiation of algorithms can be obtained by examining computational performance metrics such as build time and classification speed. In comparing the classification speed, 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. Build time found NBTree to be slowest by a considerable margin. Our work extends this idea while providing an online Internet traffic classification by customizing the source code of C4.5 algorithm.
Jiang, et al. [21] showed by experiments, that NetFlow records can be usefully employed for application classification. The machine learning used in their study was able to provide an identification accuracy to about 91%. The authors used data collected by the high performance monitor (full packet capturing system) and then NetFlow record was generated by utilizing nPrope (a software implementation of Cisco NetFlow).
Erman, et al. [22] considered the traffic classification in the core network. The authors deployed a framework that can classify a flow using only unidirectional flow information, and they found that flow statistics from the server to client direction of TCP connection provides greater classification accuracy than flow statistics from client-to-server direction. The authors used unsupervised machine learning called clustering.

3. Methodology

In this paper, a novel online near real time flow-based Internet traffic classification [NOFITC] system has been implemented. This system is considered as a building block toward near real time Internet traffic control and bandwidth optimization.
Based on the work of [20]. An open source code of C4.5 written by the author of the algorithm [23] has been downloaded, modified, compiled, and customized to produce our novel system for an online Internet traffic classification.
The above mentioned open source code consists of two main classification module. One module works for offline classification using C4.5. The other works in an interactive mode called consult. It has the ability to receive the features from the keyboard Our [NOFITC] system is build by modifying and customizing the interactive mode module.
The customized open source code is enhanced with several new functions to achieve our goal, (e.g. online NetFlow collection, online NetFlow preprocessing and modified online user interface to adapt the classification functions to work online).
The following diagram (see figure 3-1) - represents the layering structure of the proposed system and at the same time summarizes our customized two modules (online and offline Internet traffic classification modules).

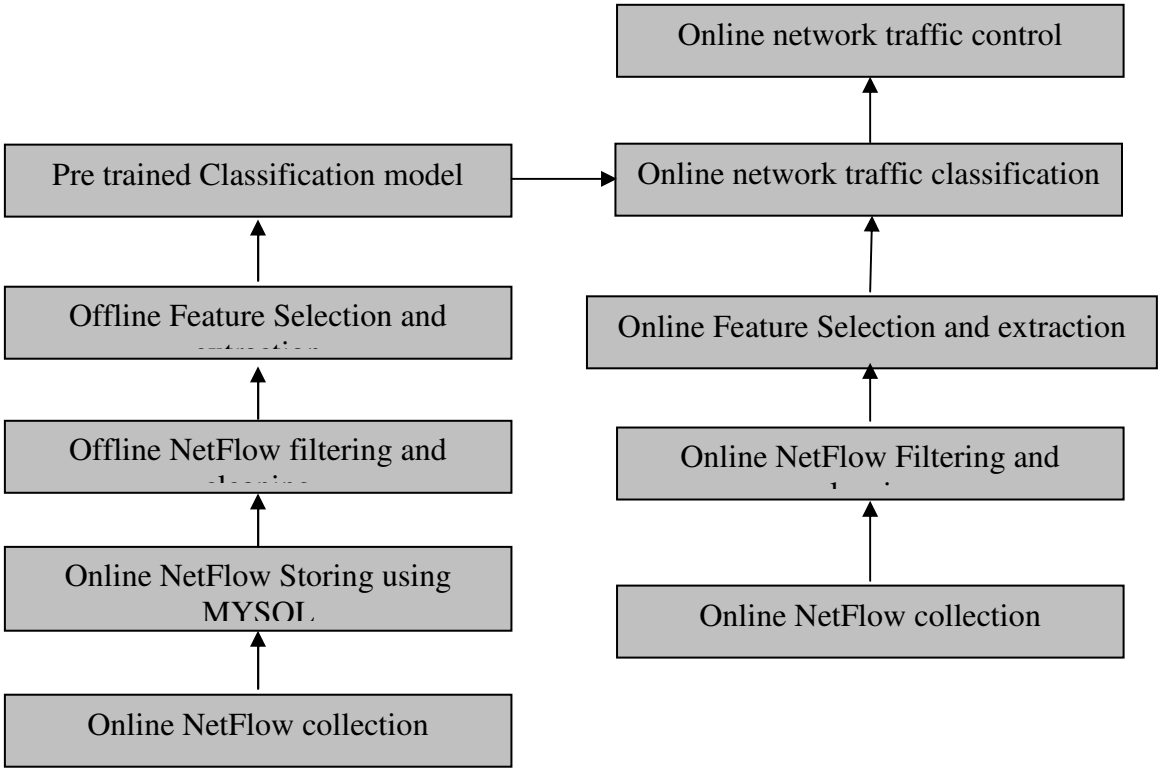

FIGURE (3-1) Layering Model For online traffic classification and Control System

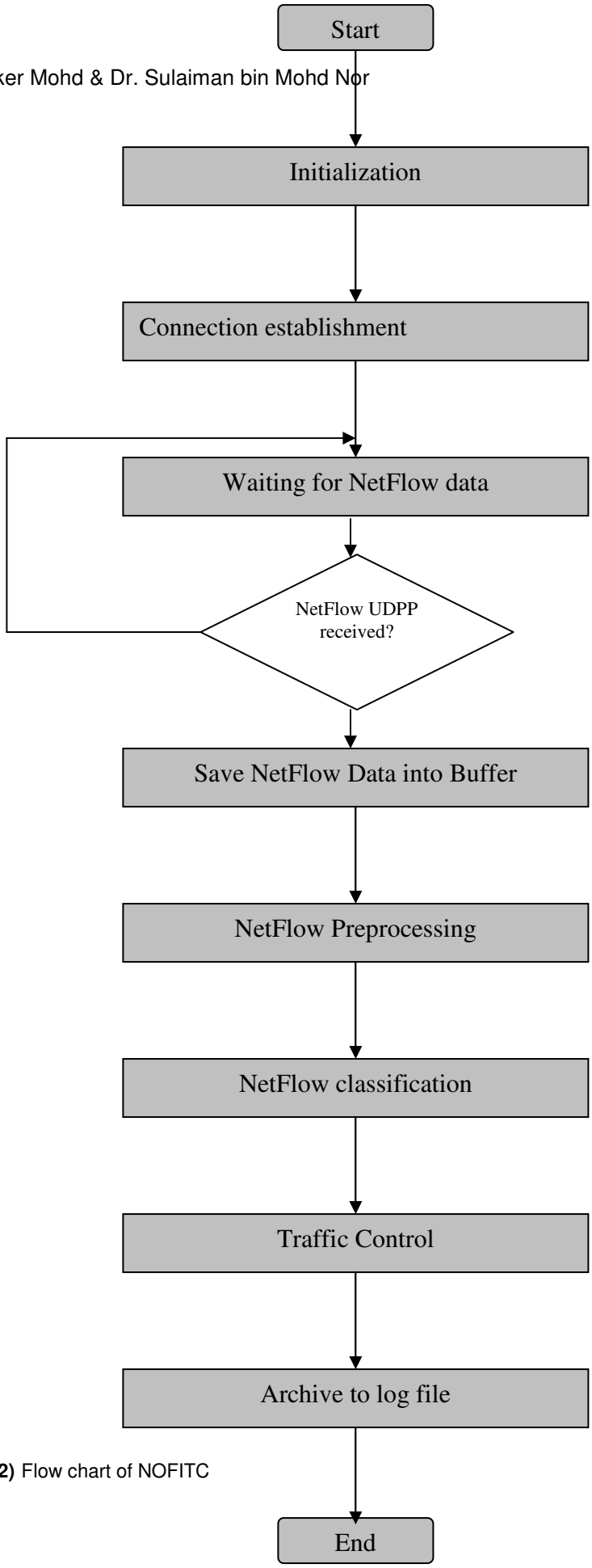

FIGURE (3-2) Flow chart of NOFITC

In this paper we will focus on the online module because the offline one has been discussed in details via our previous work [24, 25].
The following flow chart (see figure 3-2) explains the customized online traffic classification system using C4.5 algorithm.
To obtain our goal successfully and accurately, a validation process has been done according to accuracy and time points of view; firstly, offline training and testing data sets are applied to Weka's C4.5 [26] and our system [NOFITC]. The accuracy obtained by each is compared according to the training data sets.
Secondly, since our target goal is towards near real time traffic classification and control system, in this work the time factor has been considered and the performance of the proposed system examined. This was done by sending the received UDP NetFlow packets simultaneously with one copy to NOFITC and another copy to a basic packet sniffing program. Comparison between the number of received UDP NetFlow packets by the sniffer and the number of received, preprocessed and classified UDP NetFlow packets by NOFITC was done in a fixed time interval (see figure 3-3).

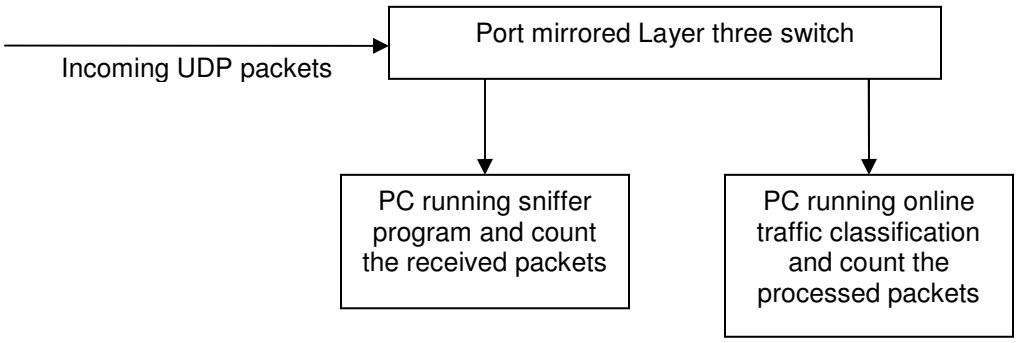

FIGURE [3-3] Performance comparison and the port mirrored switch

3.1 Online NetFlow collection, filtering preprocessing and classification:

The main difference between this work and our previous classification work [24, 25] is that the NetFlow collection filter, preprocessing and classification are done in an online manner rather than offline.
Here, to speed up the processing time, the data collection module (see figure [3-4]) has been implemented with a different approach. Furthermore the design of this module considers the time restriction so that instead of storing the NetFlow records into secondary storage device using MYSQL, the collected NetFlow records is stored into a buffer for further online processes. The collection module has the capability of receiving NetFlow UDP packet from the NetFlow exporter and deliver it to an online preprocessing, which will clean, filter, select basic features, extract derived features and calculate their corresponding values and finally it reformat the NetFlow record so as to make it ready to be classified by the online classification module.
Finally the ready NetFlow record will be send to the classification module and the classification result will be issued accordingly using the customized C4.5 source code.

4. Results

As intended in this paper, the validation of the NOFITC will be scrutinized from and accuracy and time perspectives. The following sections describe this in details.

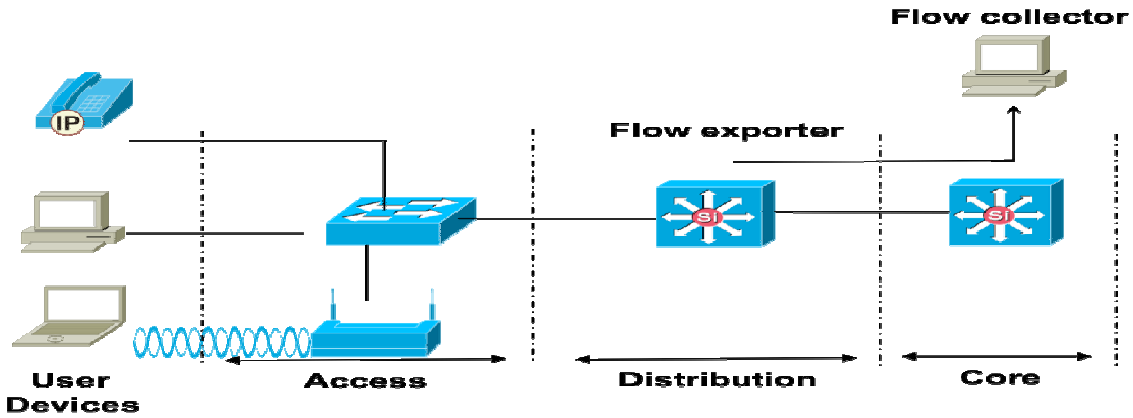

FIGURE [3-4] typical setup in a faculty with NetFlow exporter and collector

4.1 Accuracy:

Experimentally, we validated the offline classification module C4.5 with a custom build network traffic collected from the UTM campus network. Furthermore the accuracy of the open source code is compared with the accuracy of Weka's C4.5 [26]. The result of the comparison according to different training data sets is recorded in table [4-1].

As can be seen form figure [4-1] and table [4-1]The over all accuracy of the implemented system is approximately equal to the accuracy of C4.5 in Weka toolkits, and there are a little bit variation due to the differences in the pruning process which effects the tree size.

4.2 Time:

Since network administrators are always worried about making fast decisions to monitor and regulate the Internet traffic, our results show that the time for online preprocessing and classification is very small compared to the inter arrival of UDP flow packets. From performance point of view our system works perfectly with no UDP NetFlow overwriting. In other words, every UDP NetFlow packets are accounted for and analyzed with any drop in packets. To prove that, we executed the online classification system concurrently with a simple packet sniffing and filtering NOFITC program and counting simultaneously received packets and processed packets respectively from each program for a fixed time interval. More than 10,000 UDP flows were inspected and the results shows that all UDP flow packets were processed by NOFITC with any drop or over riding in packets.
This promising result is an important step in implementing our near real time online network traffic control system model.

Number of instances	Using J.Ross open source code				Using Weka's C4.5	
	Before Pruning		After Pruning			
	size	Errors %	size	Errors %	size	Errors %
76830	361	3.6	205	3.7	205	3.7069
76700	357	3.6	197	3.7	227	3.6741
76528	383	3.6	225	3.7	245	3.6549
76356	385	3.6	251	3.6	243	3.6487
76227	351	3.6	209	3.7	211	96.3176
76055	375	3.6	209	3.7	209	3.6947
75926	353	3.7	157	3.8	157	3.776
75797	359	3.7	157	3.8	157	3.7785
75668	363	3.7	215	3.7	195	3.7017
75496	359	3.6	171	3.7	189	3.7115
75367	357	3.6	175	3.7	189	3.7085
73511	345	3.6	189	3.7	201	3.658
72605	317	3.6	161	3.7	161	3.7008
71646	323	3.7	157	3.7	177	3.7099
69702	299	3.6	135	3.7	137	3.6613
67758	311	3.4	185	3.4	195	3.4372
65814	353	3.3	177	3.4	189	3.3625
63811	277	3.3	129	3.4	133	3.3803
61219	293	3.1	169	3.2	167	3.1902
58627	265	2.9	147	3	147	2.9679
56034	221	2.8	153	2.9	153	2.8572
54306	227	2.9	141	2.9	141	2.9094
49986	303	2.5	223	2.6	219	2.5967
45710	235	2.6	171	2.7	169	2.6843
41477	209	2.7	135	2.8	135	2.7582
37245	235	2.4	133	2.5	135	2.4997
32839	123	1.8	97	1.9	93	1.8575
28606	123	2.1	97	2.1	93	2.1324
24555	123	2.4	97	2.5	93	2.4842
20279	75	1.1	69	1.1	69	1.1391
16003	25	0.3	11	0.3	11	0.3249
11726	25	0.4	11	0.4	11	0.4435
7404	39	0.6	11	0.7	11	0.7023
3384	21	1.2	19	1.2	19	1.2116
2563	35	0.9	35	0.9	35	0.9364
1699	35	1.3	35	1.3	35	1.2949
396	29	4	25	4.5	25	4.5455
250	21	5.6	17	6.4	17	6.4
76	15	5.3	15	5.3	15	5.2632

TABLE [4-1] accuracy comparison between Weka's C4.5 and the proposed system

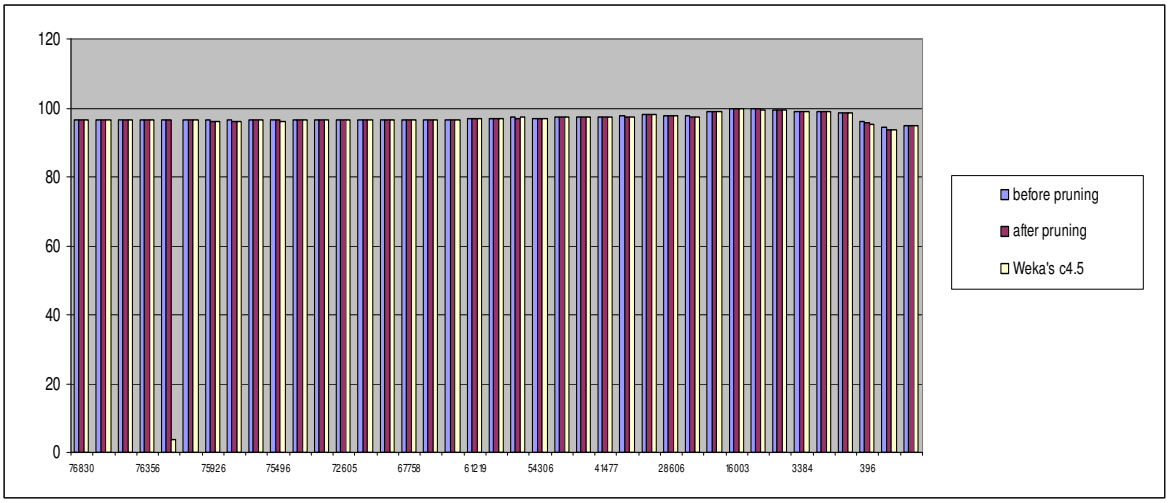

FIGURE [4.1] accuracy comparison between Weka's C4.5 and the proposed system

5. Conclusion and Future work

In this paper we customized and modified the C4.5 source code for the purpose of building a complete near real time online flow-based network traffic classification system [NOFITC].
This effort reflects three contributions. First a novel building and implementation of near real time online flow based traffic classification system [NOFITC], secondly the validation of the accuracy of the proposed system compared with Weka's C4.5. And finally the performance test that proves the system can work in real time flow-based without packet overwriting or dropping.
Although our system reported an excellent performance according to the current configuration, more testing will considered in future to check the reliability of the proposed system with different traffic rates. The proposed system is considered as a building block toward an online flow-based traffic control system, so future work will discuss online traffic control. The outcome of this effort can be directed to a policy enforcement point so as to make decision regarding bandwidth optimization by mission critical application.

References:

[1] Guangxing ZHANG, Gaogang XIE, Jianhua YANG, Yinghua MIN, Zhaomin ZHOU, Xiaodong DUAN, "Accurate Online Traffic Classification with Multi-phases Identification Methodology", Consumer Communications and Networking Conference, 2008. CCNC 2008. 5th IEEE, Page(s):141 – 146, 10-12 Jan. 2008

[2] Li Jun; Zhang Shunyi; Lu Yanqing; Zhang Zailong, "Internet Traffic Classification Using Machine Learning," Communications and Networking in China, 2007. CHINACOM '07. Second International Conference on , vol., no., pp.239-243, 22-24 Aug. 2007.

[3] Nguyen, T.T.T.; Armitage, G., "A survey of techniques for Internet traffic classification using machine learning," Communications Surveys & Tutorials, IEEE, vol.10, no.4, pp.56-76, Fourth Quarter 2008

[4] Bernaille, L., Teixeira, R., Akodkenou, I., Soule, A., and Salamatian, K. 2006. Traffic classification on the fly. SIGCOMM Comput. Commun. Rev. 36, 2 (Apr. 2006), 23-26. DOI= http://doi.acm.org/10.1145/1129582.1129589

[5] http://www.iana.org/assignments /port-numbers (last accessed July 2009)

[6] 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

[7] T. Karagiannis, A. Broido, and N. Brownlee. Is P2P Dying or Just Hiding? In GLOBECOM '04, Dallas, USA, November 2004.

[8] T. Karagiannis, A. Broido, M. Faloutsos, and K. cla®y. "Transport Layer Identi¯ cation of P2P Traffic. In IMC'04, Taormina, Italy, October 2004.

[9] Alok Madhukar Carey Williamson, “A Longitudinal Study of P2P Traffic Classification”, Proceedings of the 2th IEEE International Symposium on (MASCOTS '06) 2006 IEEE
[10] S. Sen, O. Spatscheck, and D. Wang."Accurate, Scalable In-Network Identi¯cation of P2P Traffic Using Application Signatures. In WWW 2004, New York, USA, May 2004.
[11] C. Dews, A. Wichmann, and A. Feldmann."An analysis of Internet chat systems". In IMC’03, Miami Beach, USA, Oct 27-29, 2003.
[12] P. Haffner, S. Sen, O. Spatscheck, and D. Wang. ACAS: "Automated Construction of Application Signatures". In SIGCOMM’05 MineNet Workshop, Philadelphia, USA, August 22-26, 2005.
[13] Karagiannis, T., Papagiannaki, K., and Faloutsos, M. 2005. BLINC: multilevel traffic classification in the dark. In Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols For Computer Communications (Philadelphia, Pennsylvania, USA, August 22 - 26, 2005). SIGCOMM '05. ACM, New York, NY, 229-240. DOI= http://doi.acm.org/10.1145/1080091.1080119
[14] S. Sen, O. Spatscheck, and D. Wang. "Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures". In WWW2005, New York, USA, May 17-22, 2004.
[15] M.S. Kim, H.J. Kang, J.W. Hong, 2003, Towards peer-to-peer traffic analysis using flows, Working paper obtained from the Distributed Processing and Network Management Laboratory. Department of Computer Science and Engineering, Pohang University of Science and Technology, Republic of Korea.
[16] Robin Sommer and Anja Feldman, Saarland University, Germany NetFlow: Information loss or win? ACM Measurement Workshop, 2002
[17] Jeffrey Erman, Martin Arlitt, Anirban Mahanti, "Traffic Classification Using Clustering Algorithms", in SIGCOMM’06 Workshops September 11-15, 2006, Pisa, Italy.
[18] M.S. Kim, H.J. Kang, J.W. Hong, 2003, Towards peer-to-peer traffic analysis using flows, Working paper obtained from the Distributed Processing and Network Management Laboratory. Department of Computer Science and Engineering, Pohang University of Science and Technology, Republic of Korea.
[19] Liu Bin, “Traffic Measurements of BitTorrent System Based on Netfilter “, C2006 IEEE
[20] Nigel Williams, Sebastian Zander, Grenville Armitrage A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification
[21] Hongbo Jiang, Andrew W.Moore, et al “Lightweight Application Classification for Network Management” ACM 2007
[22] Jeffrey Erman, Anirban Mahanti, Martin Arlitt, Carey Williamson,Identifying and Discriminating Between Web and Peer to Peer Traffic in the Network Core " August 27–31, 2007, ACM
[23] J. Ross Quainlan, “C 4.5: Programs for Machine Learning “ Morgan Kaufman Publisher, 1993
[24] Abuagla Babiker, Suliaman Mohd Nor. “Performance Evaluation of Decision Tree Algorithms for Flow-Based Network Traffic Classification IGCES2008, International Graduate Conference of Science and Engineering, UTM Johore.
[25] Abuagla Babiker, Suliaman Mohd Nor. “Towards a Flow-based Internet Traffic Classification For Bandwidth Optimization” International journal of Computer Science and Security” may 2009
[26] http://www.cs.waikato.ac.nz/ml/weka/ (last access nov 2008)