A.Jenefa, S.E Vinodh Ewards / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 2, March -April 2013, pp.1334-1339

Application Identification using Supervised Clustering Method

A.Jenefa*, S.E Vinodh Ewards**
*Department of Computer Science and Engineering, Karunya University, Coimbatore-114
** Department of Computer Science and Engineering, Karunya University, Coimbatore-114

ABSTRACT

The classification of traffic provides essential data for network management and research. Several classification approaches are developed and proposed to protect the network resources or enforce organizational policies. Whereas the port number based classification works only for some well-known applications and payload based classification is not suitable for encrypted packet payloads that make the sense to classify the traffic based on behaviors observed in networks. In this paper, a supervised clustering algorithm called *Flow Level based Classification (FLC)* is proposed to classify network flows, which comprises of flows in the same conversation. In this paper we discussed recent laurels and various research trends in supervised and unsupervised clustering algorithms. We outline the obstinately mysterious challenges in the field over the last decade and suggest strategies for tackling these challenges to promote headway in the art of Internet traffic classification.

Keywords - Clustering approaches, Machine learning approaches, Application Identification, Traffic Classification.

I. INTRODUCTION

Identifying application is essential for effective network planning and monitoring the trends of applications. The network traffic classification becomes more challenging because modern applications complicated their network behaviours. The objective of traffic classification is to understand the type of traffic carried on the Internet to protect the network resources. A number of methods have been proposed to identify and to classify the traffic into applications. There exist three methods: Payload Based classification, Port Based classification and Machine Learning approaches. However the traditional methods: Port Based [1] and Payload Based approach [11] may not work well in the modern applications. Identifying applications in networks using port based and payload based approach has been greatly diminished in recent years. But the Machine Learning approach provides a better result in application identification in which the statistical characteristics of IP flows are concerned. In this paper, we present a Machine learning approach called *Supervised Clustering approach* called *Flow Level based Classification (FLC)* is used for classifying the traffic using different statistics. Our aim is to build an efficient and an accurate classification approach using clustering techniques as the building block. Such a Clustering approach would consist of two stages: a learning phase and a classification phase. The objective of the offline learning phase is to find out the information that should be unique to or different from other applications. After the clusters are grouped by Euclidean distance, the average packet sizes for each application is calculated which will be helpful to identify the applications in the classification phase. In the classification phase, the similarity of flows is calculated and grouped by 5-tuple information and Transport Layer protocol. Once the flows are grouped, the classification phase compares the information with the offline phase for application identification.

The remainder of this paper is organized as follows. Section II explains the evaluation of both supervised and unsupervised algorithms. Each module of Flow *Level based Classification* is explained as follows. Section III, IV explains the details of learning phase and the classification phase. Section VI presents our conclusions.

II. EVALUATION OF CLASSFICATION OF TRAFFIC BY CLUSTERING APPROACHES

Normally the statistical properties are used to classify the network applications. This Section summarizes the basic concepts of clustering and outlines how the supervised and unsupervised approaches can be applied to traffic classification

1) Unsupervised Clustering Approaches

The unsupervised clustering algorithms namely), Auto Class, Simple K-Means, Expectation Maximization (EM and DBSCAN Clustering are considered in this work.

McGregor et al. [2] used Expectation maximization technique to group flows based on a set of flow statistics to classify traffic under different metrics and criteria. This algorithm is helpful only for the first step of classifying where the traffic is completely unknown, and possibly gives a hint on the group of applications that have similar traffic characteristics.

Zander et al. [3] used a probabilistic model-based clustering technique called Auto Class [4, 5] which allows for the automatic selection of clusters and the soft clustering of data. In Auto class the clusters are

labelled with the most common traffic category of the flows in it. If two or more categories are tied, then a label is chosen randomly amongst the tied category labels.

In [6] K-Means algorithm, an unsupervised clustering is used and it classifies different types of applications using the first few packets of the traffic flow. This algorithm is not effective if the classifier misses the first few packets of the traffic flow.

The Clustering Algorithm DBSCAN which relies on a density-based notion of clusters. Density-Based algorithms have an improvement over partition-based algorithms because it's not limited to finding spherical shaped clusters but can find clusters of random shapes. In [7] Density-Based algorithms have selected DBSCAN algorithm as a representative. This is in contrast to K-Means and Auto Class that allocates every object to a cluster.

2) Supervised Clustering Approach

Supervised Clustering requires a prior knowledge to classify the traffic flows. The phases of supervised clustering are

- *Learning Phase:* The Training phase that builds a set of classification model or rules.
- *Classification Phase:* The model that has been built in the learning phase is used to classify new unseen instances

In Karagiannis et al. [8] present a novel approach to classify traffic flows into application behaviours based on connection patterns. The connection patterns are evaluated in three different levels. 1) The social level 2) The functional 3) The application. It correlates Internet host behaviour patterns with one or more applications and filters the correlation by behaviour stratification. It is able to accurately associate hosts with the service they provide or use by inspecting all the flows generated by specific hosts. However, it cannot identify specific application sub types because it has gathered information from multiple flows for each individual host to decide the role of the host.

In Nen-Fu Huang et al. [9] present a Machine Learning technique for traffic classification. This paper addresses the problem of early identifying application traffic in protocol level. The Machine Learning involves mainly two steps. First, extensive features are defined based on statistical characteristics of application protocols such as flow duration, inter-arrival times, packet length etc. A machine learning classifier is then trained to associate set of features with known traffic classes, and apply the well-trained machine learning classifier to classify unknown traffic using previously learned rules. It's [9] also suitable to identify encrypted protocols.

In Chun-Nan Lu [10] present a Session Level Flow Classification (SLFC) algorithm to classify flows into application behaviours based on flow classification and session grouping. The flow is classified into applications by packet size distribution and then the flows are grouped as sessions by port locality. The SLFC classifies the traffic without examining the packet payloads. This method works even if the packet payloads are encrypted. Table 2 lists a summary of related works.

III. SUPERVISED LEARNING PHASE

The learning phase is operated offline to identify the application behaviours in the network. So it's helpful to spot out an individual application from a mix of applications. The objective of the offline training phase is to find out the information that should be unique to or different from other applications, to be the basis of comparison. For this reason, this learning phase first collects a set of traffic traces and tries to extract the information from the traces. A traffic filter is very much helpful to collect the traffic traces from the network because it filters out irrelevant information from the traffic collection. Our experimental research uses a one day trace collected at the edge of our university. Any packet trace analyser can be used to extract the flows. Initially, to group flows into clusters, the degree of similarity between flows is to be measured.

1) Method 1 (H1) Using Euclidean Distance

By Euclidean distance, the similarity between flows is measured; a small distance between the two flows shows a strong similarity (whereas) a large distance shows a low similarity. In between two flows, the Euclidean distance can be calculated as

$$dist(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Where n be the number of flows. If the distance between two flows is very small then the two flows would be grouped together or otherwise they are grouped as different groups. The flow having smaller value distance will be grouped together named as $S_1, S_2 \ldots\ldots S_n$ .The above step is repeated till all flows get grouped.

2) Method 2 (H2) Using per-flow Average Packet size

This method allows us to classify a flow once it has to be assigned to a cluster. The average packet sizes for each application is calculated and stored in the REF table which will be helpful to identify the application in the classification phase.

$$avg\_pkt\_size_{S_i} = \sum_{j=1}^{n} \frac{pk_j}{n}$$

Whereas pk_j be the frequent available number of packet sizes and 'n' be the number of flows in S_i .The above process is repeated for each

S_i i.e. for each application and recorded in the REF table for the classification process. At maximum all the applications shows the unique behaviour regarding patterns of transfer of packet sizes. Thus, distinguished packet sizes are helpful to discriminate certain application. Normally the packets have the same sizes across all flows but it's necessary to examine that the average packet size per flow remains constant across all flows in the network traffic.

IV. SUPERVISED CLASSIFICATION PHASE

The supervised clustering of classification phase which includes different heuristics to refine the classification of traffic. A set of experimental research has done in our traces to inspect various applications in the network. The first three methods are used to group the similar flows and the last method is used for flow classification.

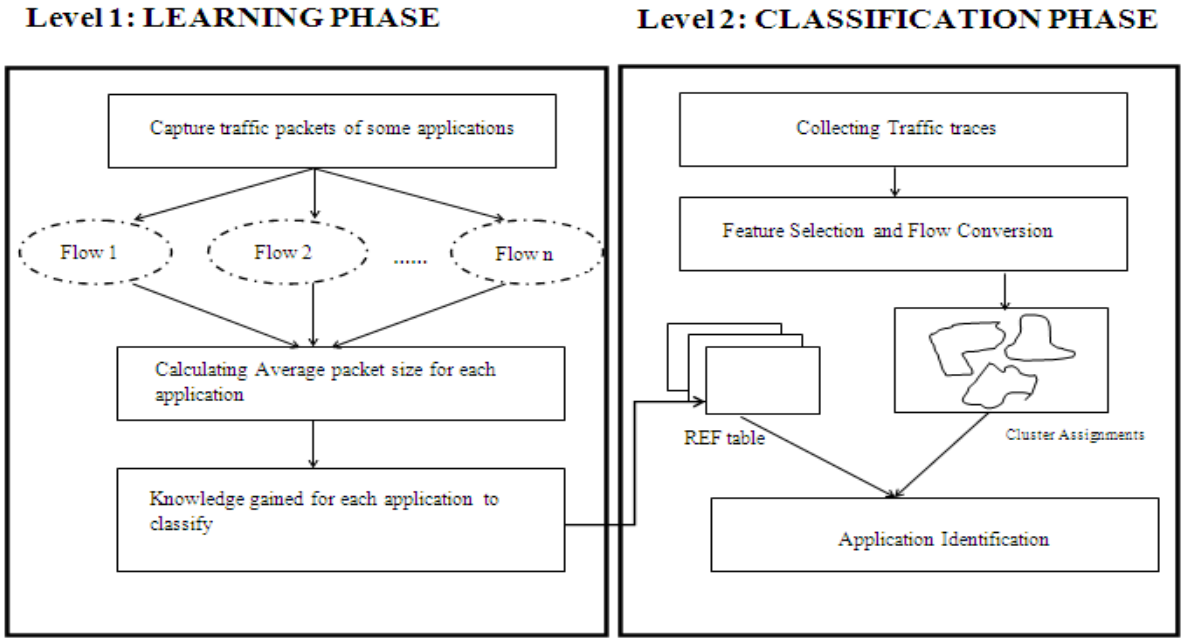

Fig 1: Overall Diagram of Application identification

TABLE 1
A SUMMARY OF RESEARCH REVIEWED

Related Work	Machine Learning Algorithms	Applications Considered	Feature Computation Overhead
McGregor et al. (2004)	Expectation Maximization (unsupervised Clustering)	HTTP,SMTP,FTP,DNS,IMAP,NTP etc.	Moderate
Zander et al. (2005)	Auto Class(Unsupervised Clustering)	HTTP,DNS,SMTP,FTP,Telnet etc.	Moderate
Bernaille et al.(2005)	Simple K-Means (unsupervised Clustering)	HTTP,FTP,NTP,HTTPS, SMTP,POP3,SSH etc.	Low
Karagiannis et al. (2005)	Supervised Clustering	All applications concerned.	High
Huang et al. (2008)	Supervised Clustering	BitTorrent,eMule,FTP,HTTP,Skype, Shoutcast, SMTP, POP3, PPLive.	Moderate
Chun-Nan Lu et al.(2009)	Supervised Clustering	BitTorrent,eMule,FTP,HTTP,Skype, Shoutcast, SMTP, POP3, PPLive.	Low
Jenefa et al. (2013) (Our Work)	Supervised Clustering	All applications concerned. (Refer Table 2)	Low

TABLE 2
Classification of Applications using the Transport Layer Protocol

Traffic Classification	**Application Identification**	**Transport Layer Protocol Used**	**Application Layer Protocol Used**
Chat	MSN Messenger, Yahoo Messenger,AIM,IRC	TCP	chat
ftp(Data)	ftp,databases	TCP	ftp
Web	http,https	TCP	http
Mail	smtp,pop,nntp,imap,identd	TCP	mail
p2p	Bit Torrent,eDonkey,Gnutella, Pee Enabler, WinMX, OpenNap, MP2P, FastTrack, Direct Connect	TCP/UDP	p2p
Attack	Port scans,IP address scans	-----	-----
Streaming	mms(wmp),real,quicktime,shoutcast, vbrick streaming.	TCP/UDP	streaming

TABLE 3
Classification of Applications using 5-Tuple information

Source IP_ address	Destination IP_ address	Port Number used	Flow_Id
192.16.2.29	192.168.2.51	1400	x
192.16.2.29	192.168.2.51	1400	x
192.16.2.29	192.168.2.51	1400	x
192.16.2.29	192.168.2.51	1401	x+1
192.16.2.29	192.168.2.51	1401	x+1
192.16.2.29	192.168.2.51	1401	x+1
192.16.2.29	74.125.236.181	3210	y

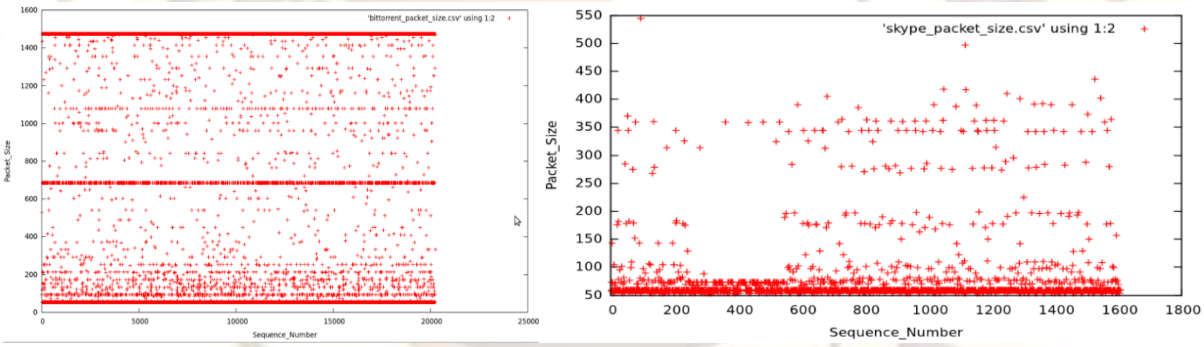

Fig 2: Distinct Packet Sizes of Bit Torrent and Skype

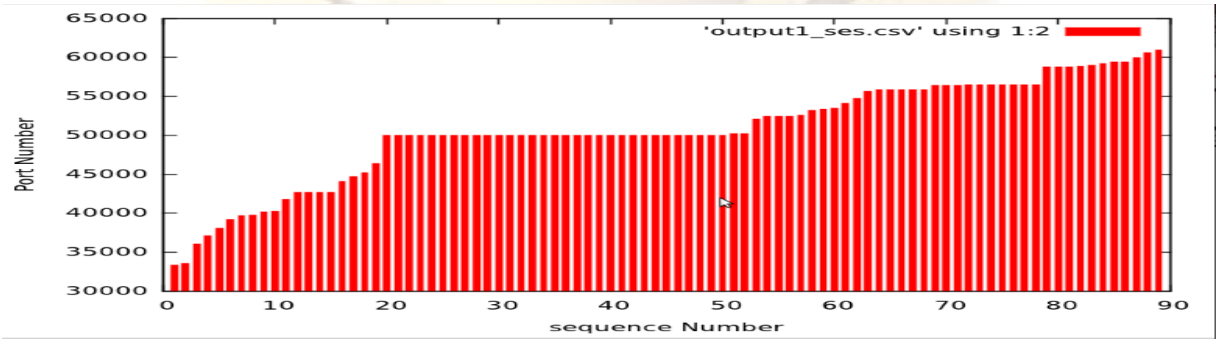

Fig 3: Adjacent Port Numbers used by flows

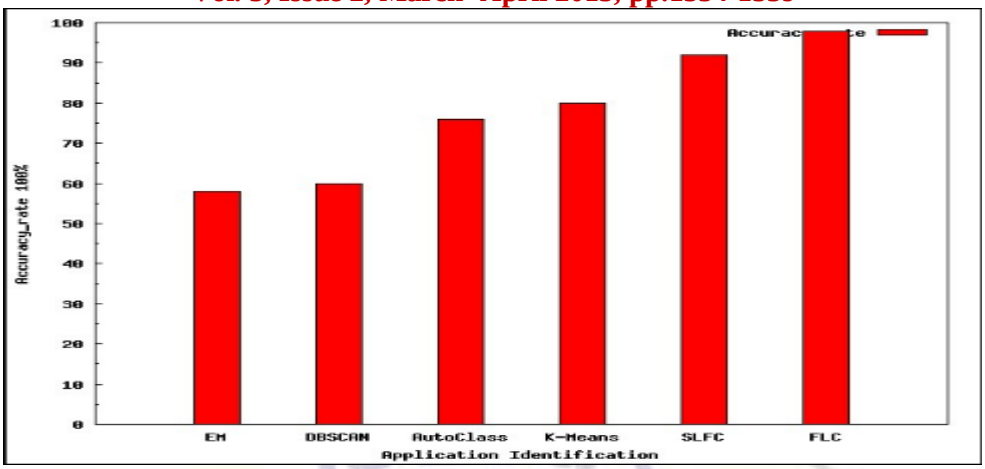

Fig 4: Accuracy rate of different Approaches

1) Method 1 (M1) The Transport Layer protocol

To distinguish the applications in the network, the protocol information is helpful to categorise into groups: i) Transport Layer Protocol (TCP) is used by FTP, mail, chat, and web. ii) Transport Layer Protocol (UDP) is normally used by games and Network Management Traffic. iii) p2p and streaming traffic uses either TCP or UDP. Thus, the method H1 helps to group the similar applications for at least to some extent. Table 2 shows the grouping of Applications using Transport Layer Protocol.

2) Method 2 (M2) The Cardinality of sets

Using ports and IPs, the behaviours of each application can be distinguished. Normally, the operating system assigns consecutive port numbers for similar flows. Figure 3 shows how the operating system assigns the consecutive port numbers for each individual flow. Here we used (Source_IP, Destination_IP, Inter-arrival Time, Port number, Flow ID) to group similar behaviours. If the Source_IP address and the Destination_IP address of different flows are same with the same port number, then the flows will be grouped together. But all flows having same port number does not belong to the same flow. So there comes, Inter-arrival time (difference of each individual arrival time of different flows) helps to group the similar flows. If the arrival time of any two flows is within a particular limit, then it will be grouped together or consider as different flow. Table 3 shows the grouping of similar flows using 5-Tuple information.

3) Method 3 (M3) The Similarity Distance

The individual similarity distance between the flows is identified by Euclidean distance. Euclidean distance helps to identify that the two different flows are identical or not. If the distance between the flows is within a specified range then it will be grouped together or else it will be grouped as different.

4) Method 4 (M4) Flow Classification

After grouping identical flows, the average packet size of each similar flow is compared with the REF table to decide which application it should be. To classify the applications, the unknown flows are compared with the REF table one by one. Thus like SLFC [10], this method works even if the packet payloads are encrypted. These methods work very well, if the average packet size remains constant across the flows in the network traffic. Figure 2 shows that the different applications have distinct packet sizes. So it's helpful for us to effectively classify the application based on its unique behaviour of packet traces.

V. IMPLEMENTATION DETAILS

Our aim is to produce an efficient classification algorithm. Normally, the supervised and the unsupervised algorithms of machine learning are used to solve the problems in network traffic classification. Here we used, a supervised algorithm named as Flow Level based Classification to classify the network traffic. First of all, TCPDump or Wireshark is used for traffic filtration. By implementing our own supervised learning classification algorithm, effectively classify the application based on its unique behaviour of packet traces. Figure 4 shows the accuracy rate of different approaches based on our test data taken at Karunya University.

VI. CONCLUSION

Every algorithm is proposed and designed for certain purposes of improvement. We proposed Flow Level based Classification algorithm which runs in two phases: a Learning phase and a Classification phase. The Learning phase such a Supervised Clustering approach would consist of two stages: a learning phase and a classification phase. Both phases are helpful to improve the accuracy rate of traffic classification. Our proposed algorithm achieves a high accuracy rate of traffic

classification. Using the proposed algorithm, a high accuracy rate of 97.9 % is achieved.

ACKNOWLEDGEMENTS

This survey paper is made possible through the help and support from everyone. First and foremost, I would like to acknowledge and extend my heartfelt gratitude to my project guide Mr.S.E Vinodh Ewards who helped me to complete this paper possible. Second, I would like to thank my department staffs, for their vital encouragement and support. Most especially to god, who made all things possible.

REFERENCES

[1] IANA, "Internet Assigned Numbers Authority", http://www.iana.org/assignment/port-numbers.

[2] A. McGregor, M. Hall , P. Lorier , J. Brunskill , " Flow clustering using Machine Learning Techniques", in: *Proc. Passive and Active Measurement Workshop (PAM2004), Antibes Juan-Les-Pins, France*, April 2004.

[3] S. Zander, T. Nguyen, G. Armitrage, "Automated traffic classification and application identification using machine learning", *in IEEE 30^{tsh} conference on Local Computer Networks (LCN 2005),* Sydney, Australia, November 2005.

[4] P. Cheeseman and J. Strutz, "Bayesian Classification (AutoClass): Theory and Results", in *Advances in Knowledge Discovery and Data Mining*, 1996.

[5] A. Dempster, N. Laird, and D. Rubin , "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, Vol. 30, no. 1,1997

[6] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic Classification on the fly," *ACM Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review*, vol. 36, no. 2,2006.

[7] M. Ester, H. Kriegel, J. Sander, and X.Xu. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *2nd International Conference on Knowledge Discovery and Data Mining (KDD 96)*, Portland, USA, 1996.

[8] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINC: Multilevel traffic classification in the dark," in *Proc. of the Special Interest Group on Data Communication conference (SIGCOMM) 2005,*Philadelphia,PA,USA,August 2005.

[9] Nen-Fu Huang, Han-Chieh Chao, "Early Identifying Application traffic with Application Characteristics", *in Proceedings of the IEEE ICC, 2008.*

[10] Chun-Nan-Lu, Chun-Ying Huang, Ying-Dar Lin, Yuan-Cheng Lai, "Session Level Flow Classification by packet size distribution and session grouping, Journal of Network and Computer Applications(2009).

[11] S. Sen, O. Spatscheck, and D. Wang, "Accurate scalable in network identification of P2P traffic using application signatures," in WWW2004, New York, NY, USA, May 2004.