Volume 4, No. 2, Jan-Feb 2013

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Survey on Classification of Traffic using Clustering Algorithms

A.Jenefa* and S.E.Vinodh Ewards
Department of Computer Science and Engineering
Karunya Institute of Technology and Sciences
Coimbatore, India
a.jenefa@gmail.com*, ewards@karunya.edu.in

Abstract: The traffic classification is essential for network management and it has become more challenging in the last couple of years. The research community has explored, developed and proposed several classification approaches. The continued increase of different Internet application behaviors covers up some applications to avoid filtering or blocking are among the reasons the traffic classification remains a challenge in Internet research. This survey paper looks at emerging research on both supervised and unsupervised clustering to assist in the classification process. In this article we review recent laurels and discuss various research trends in 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, Traffic Classification

I. INTRODUCTION

Network traffic classification becomes more challenging because modern applications complicated their network behaviors. Traffic classification has increased in relevance this decade, as it is now used for service differentiation, designing network, security, accounting, advertising and research purpose. Real time traffic classification has the potential to solve difficult network management problems for Internet Service Providers (ISP's). The Payload-Based classification, Port Based classification and Machine learning based approaches are the three main methods in the field of network traffic classification. In early literatures, port-based approach was widely used and it's effective for traditional applications which often use standard port assigned by IANA [1].

The proliferation of new applications that has no IANA registered ports, but instead they use ports already registered to disguise their traffic and circumvent filtering or firewalls. As application design and user behavior rendered port-based flow classification unreliable, payload based approaches emerged, which inspect packet content to identify byte strings associated with an application, or perform more complicated syntactical matching. The Payload based approach [11] identifies network traffic by searching the packet payload for signatures of known applications. Nevertheless, packet inspection approaches have several limitations. First these approaches only identify traffic for which signatures are available and are unable to classify any other traffic. Maintaining an up-to-date list of signatures is risky. Second, these techniques typically employ "deep" packet inspection because solutions such as capturing only a few payload bytes are not suffice. Third, these techniques typically require increased processing and storage processing. Finally packet inspection techniques fail if the application uses encryption. Classification of network traffic using port-based or payload-based identification approaches has been greatly diminished in recent years.

So it's better to move towards Machine Learning based method in which statistical characteristics of IP flows are concerned. The Machine Learning (ML) based approach gets some port independent statistical attributes of traffic classifier so that it can avoid the disadvantages of the above two methods.

In this paper, we explore a Machine Learning approach called *Clustering* to classify the traffic. Clustering is the grouping of instances that have similar characteristics into clusters, with or without any prior knowledge. The Clustering techniques can be divided into the categories of unsupervised and supervised methodology. Supervised Clustering requires a set of pre-classified (also called pre-labelled) examples, from which it builds a set of classification rules to classify unseen examples. The supervised approach has a higher accuracy of classifying traffic.

The supervised clustering approach offers some advantages over unsupervised learning approaches. But, the supervised approach cannot discover new applications and can only classify traffic for which it has classified training data. If the clustered data set contains encrypted P2P connections or other types of encrypted traffic. These connections would, therefore be excluded from the supervised learning approach which can only labelled training data as input. By looking at the connections in the cluster, the unsupervised cluster may be able to see the similarities between unencrypted P2P traffic and the encrypted traffic and conclude it may be P2P traffic.

Preliminary results indicate that clustering is indeed a useful technique for identifying the traffic. Our goal is to build an efficient and accurate classification tool using clustering techniques as the building block for traffic classification.

II. STATE-OF-THE-ART IN CLASSIFICATION OF TRAFFIC BY CLUSTERING APPROACHES

This Section summarizes the basic concepts of clustering and outlines how the supervised and unsupervised approaches can be applied to traffic classification

A. Unsupervised Clustering Approaches:

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

a. Clustering by Expectation Maximization:

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. The Expectation Maximization is the probability based clustering method of grouping objects. The Expectation Maximization algorithm groups the traffic flows based on similar flow features of the applications. In [2], the flow features are calculated based on the full–flow basis. First of all it groups the traffic flows into a small number of clusters based on the type of traffic and then make a classification procedure from the clusters. The procedures are helpful to find and eliminate the features which do not have a large impact on the classification of the input to the machine learning and the course of action is repetitive. The resultant estimation of performance was to select the best challenging model. The Expectation Maximization algorithm was found to separate traffic into a number of classes based on type of traffic. Nonetheless, the present results are limited in identifying applications of interest. 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.

b. Classification of Auto Class:

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. Soft Clusters permit the data objects to be marginally assigned to more than one cluster. To build the probabilistic model, the clustering algorithm determines the number of clusters and the parameters that manage the distinct probability distributions of each cluster.

To achieve this, Auto Class [3] which is an unsupervised classifier, uses the EM algorithm to determine the best clusters set from the training data. The EM algorithm has two steps: an expectation step and a maximization step. The Expectation Step estimates what the parameters are using pseudo-random numbers. In the maximization step the parameters are re-estimated continually until they converge to a local maximum. To find the global maximum, auto class repeats EM searches starting from pseudo-random points in the parameter space. The model with a parameter set having the highest probability is considered the best.

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.

c. K-Means Clustering:

K-Means [6] algorithm is a type of partition-based and an unsupervised clustering that classified different types of applications using the first few packets of the traffic flow. Flows are grouped into clusters based on the values of their first few packets. Then the flows are represented by points in a P-dimensional space with a dimension called the size of the packet. Similarity between flows is calculated by the Euclidean distance between their associated spatial representations. Once the natural clusters are formed, the procedures are defined to assign a new flow to a cluster. The Classification procedure is: Euclidean distance between the new flow and the center of each pre-defined cluster is computed, and the new flow belongs to the cluster for which the distance is minimized.

The K-Means algorithm partitions objects in a data set into a fixed number of K disjoint subsets. Initially the centers of the K clusters are chosen randomly from within the subspace. Then the objects in the data set are assigned into the closest cluster. K-Means iteratively computes the new centers of the clusters and repartitions done based on the new centers. This process is repeated until there cause no new assignments. This algorithm is not effective if the classifier misses the first few packets of the traffic flow.

d. Density Based Spatial Clustering of Applications with Noise (DBSCAN):

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.

The DBSCAN algorithm is based on the concepts of density-reachability and density-connectivity and it relies on two input parameters: Epsilon (eps) and minimum number of points (minPts).Epsilon is the distance between the objects that describes its eps-neighborhood. Mints are the minimum required points to form a core object q. All objects within its eps-neighborhood are said to be density-reachable from q.

In addition, an object p is said to be density-reachable if it is within the eps-neighborhood of an object that is directly density reachable or density-reachable from q. Moreover, objects p and q are mentioned to be density-connected if an object o exists that both p and q are density-reachable form. These notions of density-reachability and density-connectivity are used to define what the DBSCAN algorithm counts as a cluster. The Cluster is defined as the set of objects in a data set that is density-connected to a particular core object. Any object that is not a part of the cluster is considered as noise. This is in contrast to K-Means and Auto Class that allocates every object to a cluster.

The DBSCAN algorithm is defined as follows. 1) Initially all objects in the data set are assumed to be

unassigned. 2) DBSCAN selects a random unassigned object p from the data set. If DBSCAN locates p is a core object, it finds out all the density-connected objects based on eps and minPts. Then the found density-connected objects are assigned to a new cluster. 3) If DBSCAN locates p is not a core object, then p is said to be noise and DBSCAN moves onto the next unassigned object. 4) Once every object is assigned the algorithm stops.

B. Supervised Clustering Approach:

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

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

a. BLINC: Multilevel Traffic Classification in the dark:

In Karagiannis et al. [8] present a novel approach to classify traffic flows into application behaviors based on connection patterns. Connection patterns are defined by graphs, where nodes denote IP address & port pairs and edges denote flows between source and destination nodes.

The connection patterns are evaluated in three different levels. 1) The social level captures and examines the interactions of a host with other hosts. The host's popularity and that of other hosts in its community's circle are considered. 2) The functional level captures the behavior of the host in terms of its functional role such as producer and consumer of a service.

Table I. A summary of research reviewed

Sr. No.	*Related Work*	*Features*	*Algorithms*	*Feature Computation Overhead*
1	McGregor et al. (2004)	▪ Inter-arrival and Packet Length statistics. ▪ Byte Counts ▪ Connection Duration ▪ No of Transitions between transaction mode and bulk transfer	Expectation Maximization (unsupervised Clustering)	Moderate
2	Zander et al. (2005)	▪ Flow size and Duration ▪ Packet length statistics ▪ Inter-Arrival time statistics	Auto Class(Unsupervised Clustering)	Moderate
3	Bernaille et al.(2005)	Packet Size of TCP flow (First few packets)	Simple K-Means (unsupervised Clustering)	Low
4	Chun-Nan Lu et al.(2009)	▪ Packet size Distribution ▪ Packet size ▪ Inter-arrival Time	Supervised Clustering	Low
5	Karagiannis et al. (2005)	▪ Port Relationship and Host Negotiation ▪ Host Relationship ▪ Social level, network level, and application level are used to classify the application behavior.	Supervised Clustering	High
6	Huang et al. (2008)	• Flow statistics • Elapsed Time • Transmitted Time • Throughput • Response Time • Inter-Arrival Time	Supervised Clustering	Moderate

3) The application level captures the transport layer interactions between hosts on specific ports with the intent to identify the origin of the application. The classification is done by using four tuples (Source IP address, destination IP Address, source port and destination port) and average packet size. In BLINC, traffic classification is based on the analysis of host behavior. It correlates Internet host behavior patterns with one or more applications and filters the correlation by behavior 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.

b. Early identifying Application traffic with Application characteristics:

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. Both TCP and UDP flows are considered. Flows are categorized in L7 perspective to reflect L7 interaction behaviors. 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.

c. Session Level Flow Classification (SLFC) by Packet Size distribution and Session Grouping:

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 training phase of session level flow classification algorithm finds out the representative of each pre-selected application and later on in the classification phase the representation of an application is compared with the traffic flows for classification. Moreover, the flows will be clustered as sessions by checking the port locality because os often uses successive port numbers for an application to setup with remote hosts. If a flow of a session is classified as different applications, an arbitration algorithm is used to make the correction. The SLFC classifies the traffic without examining the packet payloads. This method works even if the packet payloads are encrypted. But the accuracy rate of classifying the traffic is not good enough for the applications having the same packet size.

III. SURVEY ANALYSIS

Our analysis is aimed to produce an efficient classification algorithm. Both supervised and unsupervised algorithms are used to solve the problems in network traffic classification. The three different unsupervised clustering algorithms namely K-means, DBSCAN, and Auto Class are evaluated. DBSCAN clusters the connection into a small subset of clusters. This is much more helpful to classify the single category of traffic. Any connection that is not assigned will be considered as noise. The connections that are considered as noise decrease the accuracy of the DBSCAN algorithm because they are considered as misclassified. The K-Means algorithm produces clusters that are spherical in shape whereas the DBSCAN produce random shaped clusters. But K-Means is much faster in clustering the data objects. With random shaped cluster of objects the DBSCAN can be able to find out the best set of clusters which reduces the amount of analysis required. The overall accuracy of K-Means algorithm steadily increases as the number of clusters increases.

The Auto Class algorithm which automatically finds out the number of clusters. The Auto class algorithm has a higher accuracy rate when it is compared with the other two. The supervised clustering approach offers some advantages over unsupervised learning approaches. It builds a set of classification rules to classify unseen network traffic flows. But, the supervised clustering approach cannot discover new applications and can only classify traffic for which it has classified training data. BLINC that 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 [10] it is assumed that the different application uses different packet size and by using that the traffic has been classified. In which it has a higher accuracy rate than the K-Means clustering algorithm but it provides a poor classification result for the applications having the same packet size. So it is better to move onto the approach by combining the both.

IV. CONCLUSION

Every clustering algorithm has some disadvantage and designed for certain purposes of improvement. Our work is in progress of producing efficient classification tool for network traffic flows by combining both supervised and unsupervised clustering algorithms of applications. Thus it is proposed from the above survey that both labelled and unlabelled data sets are used to classify the network traffic using a suitable algorithm like SLFC using a certain proposed framework to classify the network traffic flows.

V. ACKNOWLEDGEMENT

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 have helped me to complete this survey 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.

VI. REFERENCES

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

[2]. 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]. 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.