IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 6 (Mar. - Apr. 2013), PP 57-63
www.iosrjournals.org

Categorizing Interactive IP Traffic-Skype

P.Pinky[1], S E Vinodh Ewards[2]
1 (Department of Computer Science and Engineering, Karunya University, India)
2 (Department of Computer Science and Engineering, Karunya University, India)

Abstract: *Accurate classification of traffic is basic to other numerous network activities. Traffic classification is important and it is needful in many areas like network model, network administration and safety measures of traffic. IP traffic classification comes to a part to find out the unauthorized traffic in a large network and the most essential is to grant diverse Quality of Service to the traffic from different application is a chosen feature for many IP network operators, especially for enterprise networks. So the encouraging methodology that has developed recently is traffic classification using machine learning techniques. This paper shows the overview of interactive IP traffic, analysis and classification using Machine Learning techniques.*
Keywords *- NetMate , Skype, Tcpdump, Weka,*

I. Introduction

Internet plays a major role around the world. Passing messages, video chat, video conferencing, online games and all interactive applications can be done only through internet. The government legal operations and the enterprise network operators also make use of traffic classification in internet. As a part, traffic classification plays a vital role in data communications over large network. The transmission of data in the internet must be monitored periodically to maintain network security in a crowded network, because unauthorized traffic may pass through the network. The main purpose of traffic classification is to associate the traffic to its application.

At an earlier time the classification of traffic is based on port based technique and payload based technique. The port based classification is not a valuable method because several applications are not using the standard port numbers. Some applications make use of ports other than well known ports and some may not registered with the IANA registered ports. The users can use dynamically allocated ports to obscure the application. So the port based method possesses a number of limitations. In payload system, each packet's contents are look over to detect application. Each and every application has an inimitable signature so it has to be kept up-to-date, or else several applications can be lost. The payload method is unfeasible when the packet content is encrypted. Both the port based and payload based can be stated as the traditional classification methods.

The Machine Learning technique is a promising method that has established newly over traditional classification approach to categorize the application traffic. In this paper we discuss about how the machine learning techniques are useful in Interactive IP applications.

II. Related Work

Patrick Schneider [1] explained regarding port based classification. This scheme was extensively used in classification of traffic. In port based method the budding applications are always keep away from using the accepted port number. To flee from the firewall many applications can obscure their traffic.

Patrick Haffner [2] describes the application signatures. Each and every application has a unique signature pattern and need to maintain once in a while as the signature changes over and over again. This method is impractical to detect the payload which is encrypted. Thus the machine learning technique has emerged to classify the traffic according to its application.

III. Machine Learning

In this paper we present the machine learning technique as a best method for traffic classification. Machine learning [3] is used to relate flow instances into distinct classes of network traffic. Every flow is labeled by an array of statistical features and a related feature values. The statistical features such as mean packet length, mean packet size, inter packet length etc computed over numerous packets. Each feature reveals discrete values of feature which is dependent on the class of traffic in the network to which it fit in.

The ML algorithms are divided into two categories called supervised and unsupervised. Unsupervised algorithms [4] collect flow of traffic into distinct clusters based on the identical values of feature. Those algorithm does not have a capacity of prior learning of exact class of traffic. In supervised learning [5] the traffic class must be find out in advance. The classification design that has been created by means of training instances has capable to envisage the recent hidden instances by seeing the feature values of anonymous flows. Thus

machine learning algorithms are greatly useful in IP traffic classification and in identification using statistical features. It can also make use of supervised and unsupervised learning algorithms to classify the known traffic flows and anonymous traffic flows.

IV. Supervised Learning Algorithm

There are two important steps in supervised learning:
Training: The classification model can be constructed by giving training to known traffic instances.
Testing: The model that has been made in the training phase is used to categorize the anonymous flows.
Naïve Bayes, C4.5 Decision tree, Naïve Bayes Tree and Bayesian Network are some of the supervised learning algorithm. These algorithms can be implemented in weka tool. There are two metrics used to calculate the effectiveness of the algorithm.
Recall: It represents that the percentage of class X's instances which are accurately categorized as class X.
Precision: It states that percentage of instances which have class X in the midst of accurately categorized as class X.

4.1 C4.5 Decision Tree

C4.5 Decision Tree algorithm is a tree shaped model [6]. This model has nodes which stand for features and branches symbolize values that connect to the feature. A leaf node in the tree signifies the class which concludes nodes and branches. The process will start at the root of the tree to the branches to find out the class of traffic. This continues till the leaf is encountered. C4.5 is one of the speedy classification technique and most precise classifiers.

4.2 Naïve Bayes Tree

Naïve Bayes tree [7] is the mixture of Decision tree and Naïve Bayes classifier. Naïve Bayes tree can be defined as a decision tree having nodes and branches and also can be entitled as a Naïve Bayes classifiers having leaf node. The accuracy of Naïve Bayes is more than either Decision tree or Naïve Bayes. Naïve Bayes tree acquires the benefits from decision tree and Naïve Bayes classifier. This algorithm is based on the utility of a split for every attribute.

4.3 Naïve Bayes

The root of the Naïve Bayes [8] is the Bayesian theorem and probabilistic knowledge is the foundation of this algorithm. The Naïve Bayes evaluates the probability of feature having feature values. Naïve Bayes Kernel Estimation (NBKE) and Fast Correlation-Based Filter (FCBF) are the two categories of Naïve Bayes technique used to reduce the feature and this Bayes algorithm consumes less time to construct the classification model.

4.4 Bayesian Network

The other name of Bayesian Network is Belief Networks or casual probabilistic networks [9]. A Bayesian Network is a mixture of directed acyclic graph nodes and links, and has an array of tables which contains conditional probability. The nodes in the Bayesian Network signify features and the links that denotes association between features. The conditional probability is used to find out the strength of the links. If node has a parent node then its need to maintain probability table for probability distribution of the node. If the node has several parent nodes then the probability distribution in the probability table is conditional and it is unconditional if there is absence of parent node.

V. Unsupervised Learning Algorithm

Clustering is the unsupervised algorithm which groups the objects into clusters based on its identical characteristics. This approach is unsupervised since it does not have prior learning of exact classes.

5.1 K-Means algorithm

K-Means is a partition based clustering method [10] that helps to detect the user listed clusters (k) which are expressed by centroids. Identical between flows can be calculated by Euclidean distance. The classification rule states that the Euclidean algorithm evaluates the distance between recent flow and every pre-defined cluster. If the distance between them is least then it fits into that cluster. K-Means is easy and basic analysis technique. But it is hard to classify the application if there is no majority of the clusters found.

5.2 Expectation Maximization

Expectation Maximization (EM) algorithm is otherwise called as probability clustering technique [11]. It tries to find out the maximum likelihood of the probability distribution procedure. The EM algorithm

frequently uses two methods to meet at the maximum likelihood. The initial step is expectation step which evaluates parameter that is used to rule every cluster which has distinct probability distribution. In maximization step the parameters are re-calculated by means of mean and variance till they converge to a local maximum. These two phases are repeated continually until there is a development in log-likelihood

5.3 DBSCAN algorithm

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a density based algorithms [12]. This algorithm considers density-reachability and density-connectivity views used to describe density based algorithm. To reduce the evaluation, the density based algorithm has the capability to select the best kind of clusters from the random structure of clusters. This algorithm has greater benefit than partition based method because they are not constrained to work out spherical structured clusters but they are efficient to resolve random shapes.

5.4 AutoClass algorithm

AutoClass [13] is a probability based clustering technique. It instinctively analyses usual cluster and soft clustering of data. To select the best array of parameters that governs probability distribution of every cluster. AutoClass takes more time to create model but it produces accurate clusters.

VI. Experimental Analysis

In this section, we aim to analyze the VoIP based application i.e. Skype. It is the interactive IP traffic application. The skype packets are captured using relevant traffic filter. In order to classify the skype packets using machine learning tool, features must be collected.

6.1 An Outline of Skype

Skype is a common trademarked Voice over IP (VoIP) traffic. It is peer-to peer; encrypted application shows significant challenges to legalized interception [14]. Since it is encrypted, the formation of protocol is secret. Skype allow the operators to give voice/ video calls, text messages to other skype users. From the viewpoint of protocol, skype uses suspicious resolution that is complicated to reverse engineer because of vast use of cryptography and obscure techniques. In order to defend government actions skype uses Advanced Encryption Standard (AES) which is also called as Rijndel.

In recent times government have been elucidating the Internet Service Provider (ISP) responsibilities with respect to legalized interceptions. Still there are justice and organizational issues in this approach. Legalized interception contains a law which tells that the service provider to present the data exchange between users.We can detect the skype traffic in other ways such as market investigation, by collecting the statistics of skype we can able to know how much traffic is transferred in network. Afore skype can be quickly find out by monitoring the statistical features of small packets of a flow. We can easily identify the skype traffic using machine learning techniques. Skype has its own characteristics to discover or examine it. Sinusoidal Voice Over Packet Coder (SVOPC) is the default standard codec's used by skype. It is inconsistent bit rate codec that changes its rate based on the bandwidth available and the user who is talking.

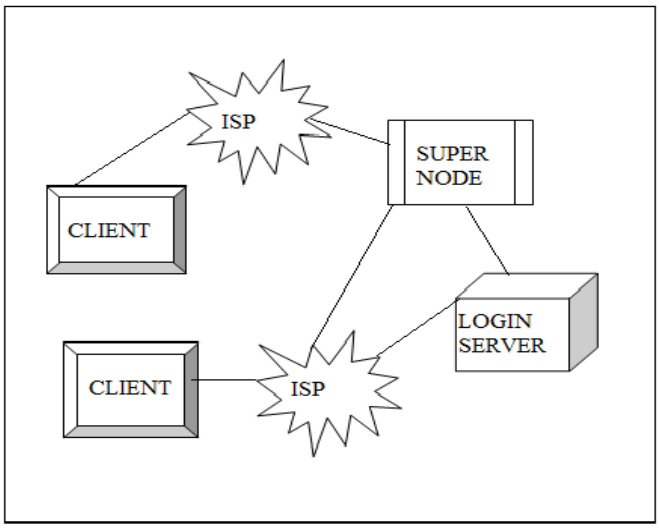

Figure1.Components of skype

Skype client, login server and super node are the three elements of skype explained in fig 1. The login server is used to store the details of the user. If the client gives username and password, the server checks the details with the saved one for authentication. If the user is authenticated then he will be allowed to initiate the communication.

Client is the user who takes part in skype. He can give his information for login, do chat, instant messages, initiate video call etc. The super node which acts as transmit for authentication details between client and the login server. Any client can turn into super node which supplies supplementary utilities to other super node and client. A super node can acts as a routing tasks which forwards requests to correct end and response the queries from other skype user or super node. The client uses UDP to transmit voice data by means of using STUN protocol to work at the back of firewall. If they are unable to send or retrieve the voice communication using UDP, then both users use super node to make a call. The skype can use TCP to transmit data than UDP if it is blocked by firewall. So machine learning techniques are used to detect skype traffic in a network.

The most common features to recognize the skype traffic is length of the packet and IAT (inter arrival time). The efficacy of every class features are evaluated in order to detect skype traffic flows. The statistical features of the traffic to be observed, gathered and combined for analysis of the skype traffic

6.2 ***Experimental flow***

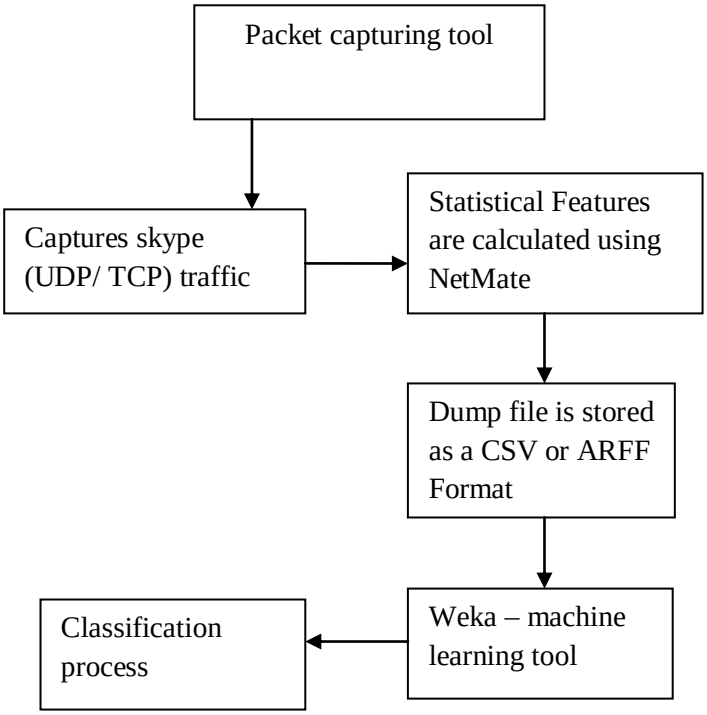

Figure2. Working flow diagram

The diagrammatic representation of our experiment is given in fig 2. The working flow is clearly explained below. Section 6.3 describes the packet capturing tool. Section 6.4 explains how statistical features are evaluated. Section 6.5 describes the machine learning tool

6.3 Packet capturing tool- Tcpdump

Tcpdump is a packet capturing tool which runs in the command line. It allows the users to capture TCP/IP packets or other packets which is being broadcast over a network in the system. In order to capture packets, tcpdump uses libpcap library to capture packets. By capturing the packets it can able to analyze the activities of the network, performance, applications that is running over the network. The importance of tcpdump is to display the communications between users in the network. Tcpdump is a simplest tool for capturing packet and store it as a dump file. Fig 3 shows the captured packet of skype.

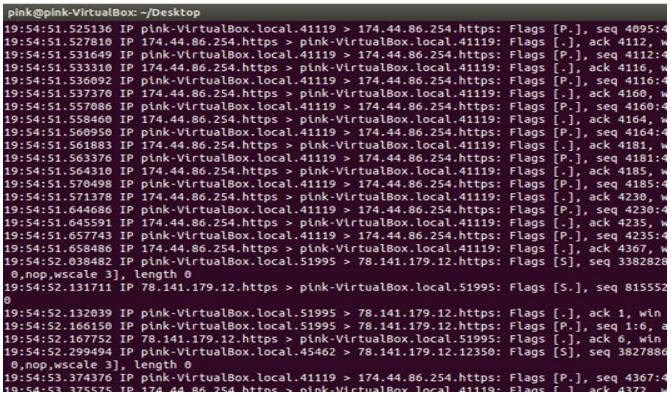

Figure3. Packets captured by tcpdump

6.4 *Feature collection by NetMate*

NetMate is Network Measurement and Accounting System which is an open source tool shared with NetAI to produce statistical features for captured packets. NetMate is used to generate 44 statistical features are shown in fig 4 from a dump file using tcpdump. It is used to change captured packets into flow statistics.

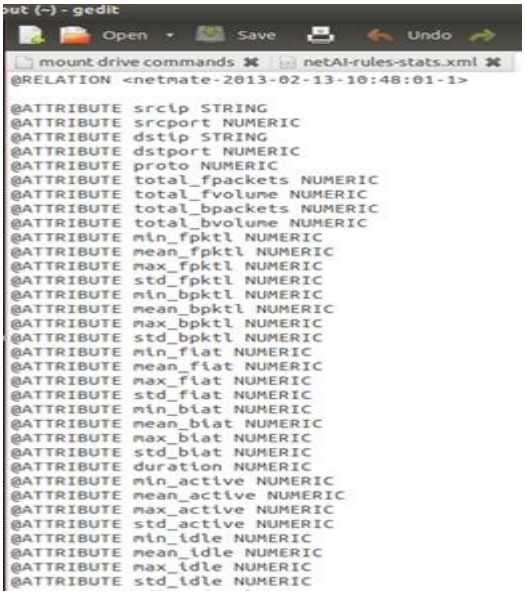

Figure4. Statistical features collected using NetMate

6.5 *Weka- The machine learning tool*

Weka (Waikato Environment for Knowledge Analysis) is the machine learning tool used to classify the network packets according to its application. The supervised and unsupervised algorithms are implemented in this machine learning tool. NetAI uses weka's abilities to produce statistics on data are attained from NetMate and categorize that data which is centered on rules and patterns.

Weka is easy to access their utilities due to graphical user interface and it is a mixture of visualization tool and machine learning algorithms for data evaluation. In order to import the statistical features into weka that is collected from NetMate, text file must be saved as a Comma Separated Value (CSV) or Attribute Relation File Format (ARFF). Fig 5 and fig 6 shows the clustering and classification

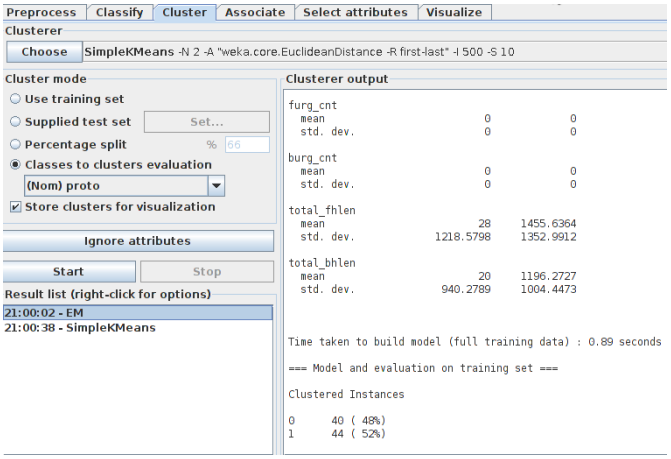

Figure5. Clustering using weka

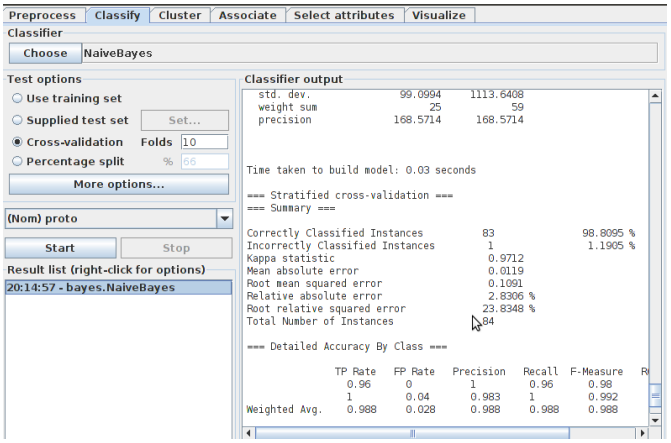

Figure6 .classification using weka

TABLE 1. BUILDING TIME OF SUPERVISED LEARNING ALGORITHM

S.No	Algorithm	Building Time
1	Naïve Bayes	0.03
2	C4.5	0.02
3	BayesNet	0.04
4	Naïve Bayes Tree	0.9

Thus the clustering techniques and classification techniques were implemented through machine learning tool and the building time for the algorithm were calculated based on our dataset that is collected in our university. The result shows that the C4.5 decision tree takes less time to build the classifier model.

VII. Conclusion

The traffic classification is the essential and it is a difficult task to categorize the network traffic based on its application. By manually it is hard to categorize the traffic in a crowded network. Thus far the traditional classification methods were used but it was not good enough to analyze and classify the encrypted application traffic. So machine learning technique gives a way to classify the traffic that is running in the network. In this paper the interactive IP traffic was taken for the analysis and it is very helpful for the network administrator to allow the authorized traffic and block the malicious traffic. The statistical features of the captured skype packets were collected and it is given to the machine learning tool for analysis. Through analysis we present the building time of algorithm for the classification model. From this model we can able to detect the unambiguous traffic.

Acknowledgement

I would like to thank my guide Mr.Vinodh Ewards and other staff members for their support and I express my gratitude to the anonymous reviewers for their comments to improve the manuscript.

References

[1] Patrick Schneider. TCP/IP Traffic Classification Based on Port Numbers. Division Of Applied Sciences, Cambridge, MA 02138.
[2] Patrick Haffner, Subhabrata sen, Oliver Spatcheck, Dongmei Wang. 2005. ACAS: Automated Construction of Application Signatures in Proc. ACM SIGCOMM MineNet.
[3] Nigel Williams, Sebastian Zander, Grenville Armitage. 2006. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP traffic Flow Classification in proc ACM SIGCOMM
[4] Erman, J., Mahanti, A and Arlitt, M. 2006. Internet Traffic Identification Using Machine Learning in proc. IEEE GLOBECOM.
[5] Li, W and Moore, A.W. 2007. A Machine Learning Approach for Efficient Traffic Classification in proc Comput.Telecommun.Syst.
[6] Kohavi, R., Quinlan, J.R., Klosgen, W and Zytkow, J. 2002. Decision Tree Discovery, Handbook Data Mining Knowledge.
[7] Kohavi, R. 1996. Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision –Tree Hybrid in proceedings of 2nd International Conference on Knowledge Discovery and Data Mining.
[8] Moore, A and Zuev, D. 2005. Internet Traffic Classification Using Bayesian Analysis Techniques in SIGMETRICS'05, Banff, Canada.
[9] Bouckaert, R. 2005. Bayesian Network Classifiers in Weka. Technical Report, Department of Computer Science, Waikato University.
[10] Carlos Bacquet, Kubra Gumus,Dogukan Tizer. 2010. A Comparison of Unsupervised Learning Techniques for Encrypted Traffic Identification in Journal of Information Assurance and Security.
[11] McGregor, A., Hall, M., Lorier, P., Brunskill, J. 2004. Flow Clustering using Machine Learning Techniques in Proc.PAM
[12] Martin Ester,Hans-Peter Kriegel, Jorg Sander, Xiawowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases in Noise in Proc of 2nd International Conference on Knowledge-Discovery and Data Mining
[13] Erman, J., Arlitt, M and Mahanti, A. 2006. Traffic Classification using Clustering Algorithms in proc of the SIGCOMM workshop on Mining network data.ACM
[14] Philip A.Branch, Amiel Heyde, Grenville J. Armitage. 2009. In proc 18th NOSSDAV.