STUDY OF INTERNET TRAFFIC CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

S.Dhivya[1], P.Shanmugaraja[2]

[1] PG Scholar, Department of Information Technology, Sona College of Technology, Tamilnadu, India
[2] Associate Professor, Department of Information Technology, Sona College of Technology, Tamilnadu, India

---

Abstract - *To identify the traffic based on the application in a large network, the major part is traffic classification and it is useful to provide quality of service, lawful interception and intrusion detection. A number of limitations have been exhibited by older methods such as port-based and payload based classification. Hence Machine learning techniques are used by the research community to analyse the flow statistics for detecting network applications. The statistical based approach is used here for traffic identification and classification process. The statistical features are flow size, flow duration, TCP port, packet inter-arrival times statistics, total number of packets, mean packet length, protocol, number of bytes transferred. There are two types of machine learning algorithms such as supervised and unsupervised algorithms. These two machine learning algorithms are analysed with datasets respectively based on the set of algorithms. By evaluating the results of supervised and unsupervised algorithms respectively, best algorithm from each technique is combined together to yield better results.*

KeyWords: Machine Learning, Supervised, Unsupervised, Traffic Classification.

1. INTRODUCTION

The Internet has become the backbone of human communication. Now a day's every electronic gadget is built to communicate over the Internet. Internet traffic is heterogeneous and consists of traffic flows from a variety of applications. In the current scenario online services such as email, social networks, multimedia communication, and https traffic have become an essential need for human beings. Many new applications emerge every day and are unique and have their own requirements with the respect to network criteria's. In an Internet user insight is also important. The user may not appreciate long waiting times, whereas the application can sustain large delays. This urges the research community to contribute in the classification of Internet traffic and treat the traffic fairly for meeting the user constraints at different level of abstraction in networking devices which includes routers, application gateway etc.

Traffic classification is a method to identify different application or protocol which exist in a network. For improving the performance of network, the various actions can be performed on the traffic identified such as monitoring, control, optimisation and discovery. Once the packets are identified that it is belonging to a particular protocol or application, that packet will be marked or flagged. Older methods such as Port-based classification and Payload based classification exhibits more number of limitations so research community uses machine learning techniques for classifying the traffic and also it analyses the flow statistics to detect network applications. In the machine learning techniques, Classification is the difficult of recognizing to which of a set of groupings a new observation goes, based on a training set of data containing instances whose class membership is known. And also in machine learning terminology, classification is measured as an instance of supervised learning, where a training set of correctly classified instances is available. The equivalent unsupervised process is called as Clustering, includes grouping data into clusters based on some similarity measure. Cluster analysis is not only one particular algorithm, nevertheless the common task to be solved. It can be attained by countless algorithms that vary considerably in idea of what establishes a cluster and how to capably find them. Cluster analysis is not such a spontaneous task, hence an iterative process of knowledge discovery that includes trial and failure. It will frequently be required to change data preprocessing and model parameters till the results succeeds the preferred properties.

Real time traffic enforces delivery of real time traffic within a stipulated time period. Real time traffic flows are generated by applications like VoIP, Multimedia applications, Video Conferencing, Webinar, Online Gaming, IP-TV, Instant Messaging and interactive applications. It inflicts stringent demands to the Network. The most important task is the timely delivery of real time traffic to accumulate the original packets at the receivers' end. The efficiency of the network depends not only on bandwidth, packet loss, jitter and delay, but it also depends on the user satisfaction. But the performance of the real time

traffic is greatly affected by the delay of individual packets. They are not able to adapt to a wide range of packet delay and delay variance at the transmission over data networks. Non real time traffic flows are generated by applications like E-mail, Peer to Peer etc. They are in sensitive to delay. Many Network operators want to manage their traffic such as real time traffic or business critical traffic which is given higher priority rather than non real time flows [1].

2. RELATED WORK

Identifying network flow using port numbers was traditional in the recent past. This approach was successful in the last decade because most of the applications use port numbers assigned by Internet Assigned Numbers Authority. This approach failed when the applications failed to communicate using their standard ports Karagiannis et al (2004). Applications that belong to the current generation use ephemeral ports or random ports and also use well known port numbers such as http, ftp etc to conceal them from Firewall or any tool that classifies the application.

Techniques that rely on inspection of packet contents choi et al(2004) have been proposed to address the diminished effectiveness of port-based classification. These approaches attempt to determine whether or not a flow contains a characteristic signature of a known application. Studies show that these approaches work very well for today's Internet traffic, including P2P flows Haffner et al (2005).

Nevertheless, packet inspection approaches pose several limitations. First, these techniques only identify traffic for which signatures are available. Maintaining an up-to-date list of signatures is a daunting task. Recent work on automatic detection of application signatures partially addresses this concern Haffner et al(2005), Ma et al(2006) . Second, these techniques typically employ "deep" packet inspection because solutions such as capturing only a few payload bytes are insufficient or easily defeated. Deep packet inspection places significant processing and/or memory constraints on the bandwidth management tool. Packet inspection techniques fail if the application uses encryption. Many BitTorrent clients such as Azureus, µtorrent, and BitComet allow use of encryption.

The diminished effectiveness of the port-based and payload-based techniques motivates use of flow statistics for traffic classification Karagiannis et al (2004), Moore et al (2005). These classification techniques rely on the fact that different applications typically have distinct behaviour patterns when communicating on a network. For instance, a large file transfer using FTP would have a smaller inter arrival time between packets and larger average packet size than an instant messaging client sending short occasional messages to other clients can be distinguished from FTP data transfers because these P2P connections typically are persistent and send data bi directionally; FTP data transfer connections are non-persistent and send data only unidirectional. Although obfuscation of flow statistics is also possible, they are generally much harder to implement. There has been much work on scalable techniques for flow sampling and estimation Duffield et al(2002), Dufffield et al(2004), Estan et al(2004), Kompella et al(2005), and furthermore, the logistics for collecting flow statistics is already available in many commercial routers NetFlow(2001) solution.

3. MACHINE LEARNING TECHNIQUES

In network management, flow classification is crucial. It is used to improve the quality of service as well as network monitoring and control. From the research community, there is a lot of contribution in classifying the flow type. Flow type classification can be done using machine learning in order to build up a classifier to identify tha traffic by statistics such as the maximum packet length, minimum packet length and standard deviation.

The machine learning techniques have two integral parts: 1. Supervised learning and 2. Unsupervised learning.

3.1 Supervised Learning

Supervised algorithms are also known as classification algorithms. It is one of the machine learning task of gathering a function from the training data which are already labelled. A set of training examples includes in the training data. In this technique, each illustration is a couple comprising of an input object and a preferred output. The training data is analysed by supervised learning algorithm and produces secondary function, which are then used for mapping new examples. A best scenario will allow for the algorithm to correctly conclude the class labels for unobserved instances. There are many algorithms that can be used, the following three are used in this paper:

- Decision tree
- Naïve bayes
- Naïve bayes tree

3.1.1 Decision Tree

A flow-chart like tree is a decision tree [2], in which each internal node is a features, each branch is a values which connects features and leaf nodes is the class which terminates nodes and branches. The starting point of the tree is known as the root of the tree and continues down to the leaves. The classification of an object begins with the root of the tree, continues towards the branch till the suitable outcomes yields. And also the process continues until the leaf is met.

3.1.2 Naïve Bayes

Naïve bayes algorithm [3] is based on Bayesian theorem and probabilistic knowledge. The naïve bayes classifier takes an indication from dissimilar attributes to rush up last prediction to classify the attributes. This type of classification uses bayes rule to estimate the conditional

probability by inspecting the association between each attribute value and the class. Naive Bayes classifiers calculate the probabilities of a feature which has a feature value.

3.1.3 Naïve bayes Tree

The combination of decision tree and naïve bayes is Naïve Bayes Tree [4]. An NBTree can be considered as a dual-level classifier, root node with the decision tree classifier and numerous leaf nodes with naïve bayes classifier. Both Naïve Bayes and decision tree don't have a good accuracy. The NBT algorithm is more accurate than C4.5 or Naïve Bayes on certain datasets. Like the other tree based classifiers, NBT also has branches and nodes. The advantages of both decision tree and naïve bayes can be utilized by the NBTree algorithm and it overtakes these two classifiers.

3.2 Unsupervised Algorithms

The problem of unsupervised learning is that of trying to find hidden structure in unlabelled data. Clustering is the unsupervised learning mechanisms and it is the well-known approach used to classify the classes in the core of a group of objects. It clusters the objects based on its resemblance without any past knowledge of the true classes. The unsupervised machine learning approach depend on a classifier that has been built from clusters are found and labelled in a training set of data. The good clusters have to have intra-cluster similarity and high-inter-cluster dissimilarity. In order to classify the network traffic of unknown applications, it is a difficult problem to solve using supervised methods. It is thoroughly linked to the problem of density approximation in statistics. Still it also covers many other techniques that seek to review and explain key features of the data. Many methods in unsupervised learning are based on data mining methods used to preprocess the data. There are many algorithms that can be used, the following three are used in this paper:

- K-means
- DBSCAN
- Expectation-Maximization

3.2.1 K-means

K-means algorithm [5] is a partition based clustering technique and used to classify the traffic. it is used to classify the traffic and tries to find out user-specified number of clusters i.e., K which are denoted by using centroids. To measure the similarity between flows, Euclidean distance is used. Once the natural clusters are formed, there is a step called modelling which is used to describe a rule and that allocates a new flow to cluster. the distance measured between new flow and the cluster is called Euclidean distance. if the distance is minimum, then the new flow belongs to the cluster which is spherical in shape that is produced by the K-means algorithm. A simple and standard analysis method is a K-means clustering algorithm. The main objective is to divide n observations into K clusters, in which each observation fits to the cluster with the nearest mean. First select K initial centroids and each point is ascribed to the nearby centroids and each group of points is designated to the centroids is a cluster. Each cluster in the centroids is streamlined based on the points designated to the cluster. We repeat the update steps till the centroids keeps on same.

3.2.2 DBSCAN

Density Based Spatial Clustering of Applications with Noise (DBSCAN) is a density based algorithms [6]. It regard clusters as dense areas of objects that are separated by less dense areas. It has an advantage that it is not constrained to find spherical shaped clusters but it can able to find arbitrary shaped clusters when compared with partition based algorithms. This algorithm is constructed upon the outlooks of density-reachability and density-connectivity. These outlooks are based on two input parameters: one is epsilon(eps) and other is minimum number of points(minPts). The distance around an object is epsilon and that describes its eps-neighbourhood. For a given object say q, within the eps-neighbourhood when the number of objects is atleast minPts, then that q is termed as core object. The objects within the eps-neighbourhood are called directly density-reachable from q. In addition to this, an object say p, called as density-reachable if it is within the objects's eps-neighbourhood that is either directly density-reachable or density-reachable from the core object q. Both the objects p and q are termed as density-connected, if they are density-reachable from an object o exists. These density-reachable and density-connected notions are used to define the cluster. The set of objects in data set which are density-connected to a particular core object is termed as a cluster. Any object that is not a fragment of cluster is considered as a noise.

3.2.3 Expectation-Maximization

Expectation Maximization algorithm (EM) is a probabilistic clustering method [7]. It is used to find out the maximum likelihood for the parameters of the probability distribution in the model. It groups traffic based on the similar properties into distinct application types. Based on the feature, the flows are grouped into small number of clusters using EM algorithm and then develop classification rules from the clusters. The algorithm that generates clusters can be specified as either Hard or Soft clusters. In Hard clusters, assigns a given data to exactly one of several mutually exclusive groups but in soft clusters it assigns a data point to more than one group. Specify the features that don't create any effect on the classification are detached from the input to the learning phase and the process is continued. The EM algorithm first estimates the parameters of the model in each cluster and repeatedly uses two step processes in

order to converge to the maximum likelihood fit. The two step processes are expectation step and maximization step. In expectation step, the parameters are calculated that govern the different probability distribution of each cluster and in maximization step it is continually re-estimated using mean and variance until they meet to a local maximum. These local maxima are registered and the EM process is continued. These two steps are repetitive till there is improvement in log-likelihood.

4. PROPOSED SYSTEM

The proposed idea is to identify the algorithm which yields better results when comparing the results in terms of accuracy with one another in both the algorithm sets of supervised and unsupervised machine learning techniques. After identifying the best algorithm in both supervised and unsupervised algorithm sets, combine both the better algorithms together in order to yield more accurate results. Best algorithm from supervised and unsupervised techniques are NBTree and K-means respectively. Combining the best algorithms together to yield better results.

5. CONCLUSION

The dataset with flow statistical features is classified using machine learning algorithms such as supervised and unsupervised with set of algorithms respectively. By using both the machine learning techniques, identify the IP traffic and classify it. The supervised and unsupervised machine learning techniques classifies the dataset into real-time and non real-time traffic with set of algorithms respectively. The proposed system yields better results when comparing with the individual algorithm results of supervised and unsupervised algorithms respectively.

REFERENCES

[1]. Cisco NetFlow, http://www.cisco.com/warp/public/732/tech/netflow.

[2]. R. Kohavi, J. R. Quinlan, W. Klosgen and J. Zytkow, "Decision Tree Discovery," Handbook Data Mining Knowledge, 2002.

[3]. A. Moore, and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques," in SIGMETRICS'05, Banff, Canada, 2005.

[4]. R. Kohavi, "Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision –Tree Hybrid," in proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, 1996.

[5]. Carlos Bacquet, Kubra Gumus,Dogukan Tizer, "A Comparison of Unsupervised Learning Techniques for Encrypted Traffic Identification," in Journal of Information Assurance and Security, 2010.

[6]. 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.

[7]. A. McGregor, M. Hall, P. Lorier, J. Brunskill, "Flow Clustering using Machine Learning Techniques," in Proc.PAM, 2004.

BIOGRAPHIES

Dhivya S is a PG scholar in the field of Information Technology at Sona College Of Technology [autonomous] India. Holding a bachelors degree in ECE from the Sri Ramakrishna Engineering College [autonomous], Coimbatore, affiliated to Anna University Chennai, in the year 2013. My current research is undertaken on Internet Traffic Classification using Machine Learning Algorithms.

