Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 784 – 791

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

# Recent Advancement in Machine Learning based Internet Traffic Classification

Neeraj Namdev[a,*], Shikha Agrawal[a], Sanjay Silkari[a]

[a]*Department of Computer Science and Engineering, RGPV, Bhopal,462033, India*

**Abstract**

With the advancement of technology and communication system, use of internet is giving at a tremendous role. This causes an exponential growth of data and traffic over the internet. So to correctly classify this traffic is a hot research area. Internet traffic classification is a very popular tool against the information detection system. Although so many methods had been develop to efficiently classify internet traffic but among them machine learning techniques are most popular. A brief survey on various supervised and unsupervised machine learning techniques applied by various researchers to solve internet traffic classification has been discussed. This paper also present various issues related to machine learning techniques that may help interested researchers to work future in this direction.


Peer-review under responsibility of KES International

*Keywords: :* Internet; Machine learning techniques;Traffic classification

## 1. Introduction

Internet traffic defines as the density of data or information present on the Internet or in another language we can say it's a flow of data on the internet. Internet traffic classification has power to solve many network difficulties and manage different type of network problems. There are some basic function provided to government, Internet service provider (ISPs) and network administrator through Internet traffic classification. It can be used for intrusion detection system by finding patterns of denial of service (Dos) and other attacks.

It can be used for intrusion detection system by finding patterns of denial of service (Dos) and other attacks. It can also help to ISPs to monitor network traffic flow and troubleshoot the faults and other problems, it can also be used in "lawful inspection" of the payload of a packet by government to obtain users information.

There are two types of internet traffic classification techniques Port based and Payload based techniques:

### *1.1 Port Based Technique*

Port based technique most popular and common technique for traffic classification. In this technique every packet in an IP traffic carries port numbers (source port number and destination port number) which assigned by IANA[11]. The applications have well known and registered port number but this is not necessary that all applications have registered port number, some new generation applications like peer to peer (P2P), online gaming type application do not have registered port numbers, these applications uses random port numbers so due to this it is very difficult to classify such type of application using port based

* Neeraj Namdev. Tel.:+91-9907596280.
*E-mail address:*neeraj3491@gmail.com


Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.238

technique.

*1.2 Payload based Technique:*

Payload based technique overcomes the problems of port based technique. It avoids the total dependency on the semantics of port numbers. This is a deep packet inspection technique (DIP), in this technique they are matching payload of the packets with the well known signature. In this technique they can setup constrains or rules according to different application types for payload matching. This technique give very good results, it classify approx 100 % of packets correctly but only when packets are not encrypted. Payload based technique is very accurate but it have two major drawbacks. First is it cannot deal with encrypted packets because we cannot apply deep packet inspection(DPI) technique in encrypted packets and second one is it have low processing efficiency, it take too much time to classify the packets.

There are many of communication devices accessing resources and getting request to carry out their work and there is a lot of information exchanged over the internet, so accurate classification is very essential not only for QOS (Quality of service) and to maintain availability of resources but also processing of information efficiently.

## 2. Machine Learning Techniques

Looking to the importance of internet various machine learning techniques has been applied to classify internet traffic accurately and efficiently. The next subsection on introduction of ML techniques is given, which is followed by discussion on application of some of the ML techniques for solving internet traffic classification problems. There are two types of ML techniques first is supervised learning (Classification) and another one is unsupervised Learning (Clustering).

*2.1. Supervised Learning Technique*:

Supervised learning based on attributes of a class i.e. in this we choose samples on the basis of attributes collected by the whole data. The machine learning is provided with a collection of sample instances, pre-classified into classes. The output of the learning process is a classification model that is constructed by examining generalizing from providing instances. In classification approaches mainly have two phases (steps), training and testing. Learning phase that examine the provided data (called the training dataset) and constructs (builds) a classification model. And the model that has been built in the training phase is used to classify new unseen instances, in this paper we discuss the some well known supervised machine learning techniques and discuss also about issues related to different techniques.

*2.2. Unsupervised Learning Techniques*

Unsupervised learning techniques using the concept of clustering. In contrast, clustering methods, we create clusters of having same features but clustering is not provided with guidance. In clustering there is no need of the training phase.

## 3. Application of Machine learning approaches for Internet traffic classification.

*3.1. Supervised (classification) Methods*
Supervised techniques as follows:

*3.1.1. Bayes Net Method*

Bayes Net approach generally known as Belief Network. It is a Probabilistic model which uses the graph model to represent the set of random variables and their conditional dependencies. Bayes Net uses the concept of directed acyclic graph (DAG) to represent the set, in which each node represent a variable and edges among the nodes represent the relative dependencies between random variables and these relative dependencies in the graph are calculated by well known statistical and computational methods. There are two phases of bayes net approach first phase is learning of network structure, in which uses various types of search algorithm like hill climbing, tabu search etc. for identified a good network structure and second is estimate probabilistic table for each random variable. In [2013], Kuldeep singh et al. [2] uses five machine learning algorithms (MLP, RBF, C4.5, Naïve Bayes, Bayes Net) to classify real time IP traffic. In this they prepared dataset by using a packet capturing tool Wireshark and captured packets for duration of 2 second and prepared datasets and now they apply feature selection algorithms to eliminate irrelevant features for this they using correlation and consistency based feature selection algorithms for feature reduction. Correlation based FS (feature selection) algorithm is used for identifying and reducing number of features which are redundant and not defining a particular type of traffic of internet and consistency based FS algorithm first compute different number of subsets of features and after that it select the optimal subset of features which contain less number of features. Result reported in this paper show 91% of classification accuracy of Bayes net. In 2012 S. Agrawal et al. [7] uses three machine learning algorithm (C4.5, Bayes Net and RBF) to classify internet traffic classification for academic perspective. They classify the website of an educational institution into two category first is an educational website which includes website