P.Pinky, S E .Vinodh Ewards / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 1, January -February 2013, pp.2099-2104

A Survey on IP Traffic Classification Using Machine Learning

P.Pinky, S E .Vinodh Ewards
*pursuing her Master's Degree, Department of Computer Science and Engineering at Karunya University, Coimbatore.
**Assistant Professor, Department of Computer Science and Engineering at Karunya University, Coimbatore.
His Areas of Interest are Networks & Network Security and Artificial Intelligence.

Abstract

Traffic classification is the major part to identify traffic based on the application in a large network. Here the traffic classification is useful to provide the Quality of Service (QOS), lawful interception and intrusion detection. The popular methods such as port and payload based techniques exhibit a number of limitations. Hence the research community uses the machine learning techniques. It analyzes the flow statistics to detect network applications. The statistical based approach is useful to assist in the traffic identification and classification process. This paper takes a review of supervised and unsupervised machine learning algorithms to identify the traffic when mixes up with other traffic.

General Terms
Machine Learning, Classification, Algorithms.

Keywords: Supervised learning, Unsupervised learning.

I. INTRODUCTION

The World is connected through Internet and it is evolving towards a vast, global infrastructure supporting data communication, digital media etc. The variety of applications is processed through Internet. In Networks all the systems are connected together and the service will be provided by means of ISP. The data transmission in the network is administered by end-to-end transmission protocols such as TCP and UDP without network monitoring, auditing and control over the traffic. Because of this, unauthorized traffic may pass through the network. Hence IP traffic classification comes to a part to detect the traffic and classify it based on the application. The classification of traffic is mainly useful for automated intrusion detection; identify patterns of denial of service attacks etc. So, real-time traffic classification has the potential to solve many network management problems. In earlier days the IP traffic classification is based on the deep packet inspection of each packet's contents. The deep packet inspection mechanisms use signature analysis to understand and verify different applications. Signatures are distinctive and they are related with every application [1, 2]. The classification engine then evaluates the traffic next to this reference to make out the exact application. Regular updates are requisite to maintain current with new applications when the payload format changes.

But the payload based inspection is unworkable when the payload is encrypted and the signature often changes for every application. So it requires more memory to store the reference database for each application's unique characteristics. Port based classification was also widely used in traffic classification. This approach is based on the association between the transport layer's port number and the corresponding application. Historically many applications use well-known port on their local host as a meeting point to which other hosts may start the communication. However, this approach has limitations. Firstly, some applications may use ports other than its well known ports and some applications may not have the IANA registered ports. Also in some cases server ports are dynamically allocated as needed. So port based and payload based techniques are fading in these days [3]. The auspicious tactic that has lately accepted is machine learning techniques.

This technique uses feature to identify the traffic. The feature may be packet size, packet length, inter arrival time etc. calculated over several packets. By using the statistical analysis (mean, standard deviation) the machine learning algorithms identify the class of traffic [4]. Once the features have been identified the machine learning classifier is trained to correlate a particular feature with a particular class of flows. Once trained the classifier is tested on unseen flows. The machine learning algorithms that have been used for IP traffic classification consists of two categories: supervised and unsupervised. In supervised learning, the class of traffic must be known before learning. A classification model is built using the training set of instances that corresponds to each class. The model that has been built in the training phase is used to categorize the new unknown instances. The input/output relationships are designed by the supervised technique which is used to map the input features to an output class. The unknown traffic can be classified by means of prior knowledge. In unsupervised clustering the objects are grouped based on its similarity. This is unsupervised because the algorithm does not have a priori knowledge of the true classes [5].

II. SUPERVISED LEARNING ALGORITHM

The machine learning techniques have two integral parts: 1. Supervised learning and 2. Unsupervised learning. With Supervised learning the class of traffic must be identified before it gets to be classified. The classification model that has been built using the training set of instances can able to predict the new instances by probing the feature values of unknown flows. The supervised learning uses weka to implement the algorithms. These algorithms' performance is calculated in terms of classification speed and the model building time.

A. Bayesian Network

Bayesian Network [6] is one of the supervised techniques used to classify the traffic. Bayesian Network is otherwise called as Belief Networks or Causal Probabilistic Networks. It depends on a Bayesian Theorem of probability theory to generate information between nodes and it gives the relationship between nodes even if the nodes are ambiguous. It is a graphical based probabilistic model that signifies random variables and their conditional probabilities. Bayesian Network is composed of a directed acyclic graph of nodes that represent features or classes and links that represents the relationship between nodes. It also includes set of conditional probability tables which determines the strength of the links. Each node has a probability table that defines the probability distribution for the node if it has parent nodes. The probability distribution is unconditional when the node has no parents and conditional if it has one or more parents. The probability of each feature value depends on the value of the parents. Bayesian Network is a two stage process where the first stage is learning a network structure and the second being the probability tables.

Bayesian Network makes easy to study about the causal relationship between variables. In clear, the causal relationship is explained as follows: the node A and B is connected to the node C by means of links. So the node A and B is the parent node of C and C is the child node. The parent node of A and B symbolizes the causal factors of the node C. The conditional probability between the parent node and the child node is represented by the conditional probability tables. Bayesian Network is difficult to explain the conditional probability tables.

B. C4.5 Decision Tree

C4.5 Decision Tree algorithm [7] creates a tree structured model where the nodes in the tree represent features and the branches represents values which connects features. A leaf node represents the class which terminates nodes and branches. The decision tree has been built with the root as a starting point and continuous down to its leaves. To classify the object, we begin at the root of the tree then compute the test and proceeds towards the branch which yields a suitable outcome. This process prolongs until the leaf is met. If a class named by the leaf is identified then the object belongs to that class. The class instance can be determined by examining the path from nodes and branches to the terminating leaf. If all classes of an instance belong to the same class then the leaf node is labeled with that class. Otherwise the decision tree algorithm uses divide and conquer method which is used to divide the training instance set into non-trivial partitions until every leaf contain instances of only one class or until further partition is not possible. The tree will classify all instances if there is no conflicting. The prediction accuracy of unseen instances decreases by this over-fitting. To avoid this, some structures can be removed from the tree after it has formed.

The decision tree algorithm is steadfast to classify and it is understandable because the data is split into nodes and branches. C4.5 is one of the most accurate classifiers and fastest classification speed.

C. Naïve Bayes Tree

The Naïve Bayes Tree (NBTree) [8] is a combination of Decision Tree and Naïve Bayes classifier. The NBTree algorithm is labeled as a decision tree which has nodes and branches and it is also defined as a Bayes classifier on the leaf nodes. The accuracy of both Naïve Bayes and decision tree are not good enough. This paper has shown that the NBTree algorithm is more accurate than C4.5 or Naïve Bayes on certain datasets. Like most other tree based classifiers NBTree also has branches and nodes. The algorithm in [8] is mainly concerned with evaluating the utility of a split of each attribute. Utility for a particular node is calculated by making the data discrete and using the method called 5-fold cross validation for which Naïve Bayes is used to estimate the accuracy. The weighted sum of the utility of the nodes is considered as a utility of the split which is considered significant if there are at least 30 instances for the node and the relative error minimization is greater than five percent. The instances are divided based on the highest utility among all the attributes, if it is better than current node utility. If there is no such better utility a Naïve Bayes classifier is created for the current node. The NBTree gains the advantage from decision trees and Naïve Bayes classifier because each node in the decision tree is built with univariate splits and Naïve Bayes is at the leaf node.

D. Naïve Bayes

The Bayesian theorem is the basis of Naïve Bayes algorithm [9] and this method is based

on probabilistic knowledge. The Naïve Bayes classifier takes a sign from unrelated attributes to rustle up final prediction to classify the attributes. The Naïve Bayes classification uses Bayes rule to evaluate the conditional probability by examining the association between each attribute value and the class [5].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where A is class and B is fixed attribute value. To get the probability of an object which belongs to class A using these conditional probabilities multiplied together. Naive Bayes classifiers calculate the probabilities of a feature which is having a feature value. The frequency distribution cannot calculate the probability of a continuous feature if they have a large number of values. Instead it can be achieved by modeling features for the continuous probability distribution or discrete values. The classifier's performance was going down when it's considered full flow features. Instead, sub-flows are used to enhance the performance. Naïve Bayes uses two methods Naïve Bayes Kernel Estimation (NBKE) and Fast Correlation-Based Filter (FCBF) to minimize the future and improve the performance and time taken to build the classification model is less.

III. UNSUPERVISED LEARNING ALGORITHM

Clustering is the unsupervised learning mechanisms and it is the renowned approach used to classify the classes in the midst of a group of objects. It groups the objects based on its similarity without any prior knowledge of the true classes. The unsupervised machine learning approach relies on a classifier that has been built from clusters are found and labeled in a training set of data. The good clusters have to have intra-cluster similarity and high- inter-cluster dissimilarity.

The classification of traffic using the clustering algorithms can be done in two phases. The first phase consists of a set of data used to build the training model. The model which has been built is used to classify the unknown traffic in the second phase. In the first phase, the training data are used to form clusters based on its similarity. The subsequent clusters are labeled based on the class which is a majority of flows in each cluster. The classes are ascribed to the flows identified based on the label of the clusters that is more similar in each flows.

A. K-Means algorithm

K-Means algorithm is one of the unsupervised machine learning algorithms used to classify the traffic. It is a partition based clustering technique that tries to find out a user- specified number of clusters (K) which are denoted by using centroids [10]. Euclidean distance is used to measure the similarity between flows. When the natural clusters are formed, the modeling step describes a rule to allocate a new flow to a cluster. The rule is that: The Euclidean distance is measured between new flow and the cluster. If the distance is minimal, then the new flow belongs to the cluster. The clusters are spherical in shape that is produced by K-Means algorithm.

The training set contains the payload, so that the flows in each cluster are entitled with its source application. The learning output comprises of two sets: one set contains the explanation of each cluster and the other contains the structure of its application. The online flows can be classified using these sets. In the classification phase, the first P packet sizes are captured and then compared it with the new flow. When the cluster is well-defined, the flow is related to the application that is more dominant in the cluster. The K-Means algorithm faces the challenges of categorizing the application when there is dominant of any of the clusters are not found.

K-Means clustering algorithm is a simple and standard analysis method. 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.

B. Expectation Maximization algorithm

Expectation Maximization algorithm (EM) [11] is a probabilistic clustering method. 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. The EM algorithm is assured to meet the local maximum which may or may not be similar to the global maximum. In order to select the largest likelihood from the final clustering, EM has to run several times with distinct initial settings for the parameter values. It takes long time for clustering process.

C. DBSCAN algorithm

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a density based algorithms [12]. It considers clusters as dense areas of objects that are detached by less dense areas. DBSCAN algorithm uses the views of density–reachability and density-connectivity. These clustering algorithms possess gain over partition based algorithms because they are not restricted to solve spherical shaped clusters but can able to solve random shapes. DBSCAN has the ability to find out the best set of clusters from the random shapes of cluster to minimize the amount of analysis required. The DBSCAN has two input parameters: epsilon (eps) which is the distance rounds an object that describes its eps-neighborhood. The given object is represented as q, when the number of objects located within eps-neighborhood is at least minPts, then q is called as core object. Objects within its eps-neighborhood named as directly density-reachable from q. Moreover, an object p is called as density reachable if it is within the eps-neighborhood that is directly density reachable or density reachable from q. additionally, p and q are called as density-connected and the same will be stated as density-reachable if an object o exits.

These density-reachable and density-connectivity are used to explain the density based cluster algorithm. A cluster is a group of objects in a data set that are density connected to a specific core object. Any object that is not a portion of cluster is classifies as noise. DBSCAN produces lower accuracy and high precision.

D. AutoClass algorithm

AutoClass is a probabilistic model based clustering technique [13]. It automatically studies the natural cluster and soft clustering of data. Soft clusters slightly allocate the data objects to more than one cluster. To built the probabilistic model AutoClass uses Bayesian score to find out the best set of parameters that administer the probability distributions of each cluster. To achieve this, AutoClass uses EM algorithm [11] and this algorithm is assured to meet the local maximum. In order to find out the global maximum the AutoClass execute repeated EM algorithm begins from pseudo random points in the parameter space. The parameter set which has the highest probability given the present database is selected as the best. AutoClass consumes time to build the model but it yields high accurate clusters.

TABLE I
SUMMARIZED RESULTS OF SUPERVISED LEARNING ALGORITHMS

Algorithm	Feature	Model Building Time
Bayesian Network [6]	• Packet length (min, mean, max, std deviation) • Inter-Arrival time (min, max, mean, std deviation)	less
C4.5 Decision Tree [7]	• Statistical features such as Packet Length, Inter-Packet Length and InterPacket Arrival time.	less
Naïve Bayes Tree [8]	• Inter-Packet Arrival time for both the direction • Packet length for both the direction • calculate statistics for all features	more
Naïve Bayes [9]	• Inter-Packet Arrival time for both the direction • Inter packet length variation for both the direction • IP packet length • calculate statistics for all features	less

TABLE II
SUMMARIZED RESULTS OF UNSUPERVISED LEARNING ALGORITHMS

Algorithm	Merits	Feature	Feature evaluation overhead
K-Means [10]	simple to classify and effectively classify large dataset	consider first few packets of bidirectional flows	low
EM [11]	easy to find out the maximum likelihood	consider full flows	medium
DBSCAN[12]	ability to detect noise and good efficiency on large database	DBSCAN uses two parameters to form cluster	high
AutoClass	Automatically determines the number of clusters	full flows	medium

IV. CONCLUSION

This paper surveys the techniques of machine learning based classification for IP traffic. In earlier days the payload and port based techniques are used to classify the traffic. But it was not efficient to classify accurately. So the machine learning techniques has emerged to classify the real time traffic based on its application. In this survey paper, the supervised and unsupervised learning algorithms are reviewed. The supervised ML algorithms such as Naïve Bayes, NBTree, C4.5 Decision tree, Bayesian Network are evaluated in terms of classification speed for calculating the computational performance and the classification model building time. In order to classify the packets, the algorithms were suggested to consider only the small number of packets not the full flows. The benefit of small number of flows is timely classifying the packets and minimizes the buffer space to store the packets' information. By this, the performance is enhanced. The metrics used to evaluate the effectiveness are Precision and recall. The classification algorithm such as Naïve Bayes is used along with the clustering algorithm AutoClass to evaluate effectiveness. AutoClass provides better Precision and Recall than Naïve Bayes when it considers individual traffic classes. But the model building time is greater than Naïve Bayes. The model building time of supervised learning is ranked in descending order: NB Tree, C4.5, Bayesian Network, Naïve Bayes Density and Naïve Bayes Kernel. C4.5 is the fastest algorithm when using any feature set. Its classification speed is high compared to other algorithms. In unsupervised algorithm, the clustering process uses Accuracy as a metric. The accuracy finds out the algorithm which is able to create clusters that have only a single category of traffic. AutoClass algorithm produces best accuracy. For K-Means, the overall accuracy increases when the number of clusters increases. DBSCAN produces very low overall accuracy but it has the ability to detect noise and works efficiently on large database. Our work is in progress to enhance the effectiveness of Machine learning algorithm.

ACKNOWLEDGEMENT

I'm very grateful to my guide Mr. Vinodh Ewards and other staff members for their thought-provoking discussions and support and special thanks to anonymous reviewers for their valuable feedback.

REFERENCES

[1] Patrick Haffner, Subhabrata sen, Oliver Spatcheck, Dongmei Wang. 2005. ACAS: Automated Construction of Application Signatures in Proc. ACM SIGCOMM MineNet.
[2] Sen, S., Spatscheck, and Wang, D. 2004. Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures in WWW2004.
[3] Patrick Schneider. TCP/IP Traffic Classification Based on Port Numbers. Division Of Applied Sciences, Cambridge, MA 02138.
[4] Li, W and Moore, A.W. 2007. A Machine Learning Approach for Efficient Traffic Classification in proc Comput.Telecommun.Syst.
[5] Erman, J., Mahanti, A and Arlitt, M. 2006. Internet Traffic Identification Using Machine Learning in proc. IEEE GLOBECOM.
[6] Bouckaert, R. 2005. Bayesian Network Classifiers in Weka. Technical Report, Department of Computer Science, Waikato University.

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

[8] 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.

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

[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