# Machine Learning: Network Traffic Identification

By
Alex Baker
Erik Granger
Tarek Embree

University of New Mexico
Dr. Lydia Tapia
CS 427
Fall 2016

# Introduction[1]

With millions of packets of information flowing through hundreds of thousands of nodes across the world, it would be a large and tedious job for a human to sift through all this information and make intelligent decisions about it. Because of this, there is ample motivation for some or all of this process to be automated. Automating these processes will help with security, quality of service, firewall and access control, intrusion detection, and network optimization. This automation will not only save people's time, but also save money requiring fewer people to analyze large amounts of data. Several different algorithms have been shown to help automate web traffic identification including machine learning strategies like K-means and expectation maximization. Each of these algorithms have different benefits and produce different results in identifying and classifying network traffic. Not only can machine learning be used to automatically identify network traffic, it can be used to predict it as well. Several algorithms have been used to try and help with network traffic prediction including Multi-layer Perceptron (MLP), Multilayer Perceptron with Weight Decay (MLPWD), and Support Vector Machine (SVM). These algorithms will help optimize traffic flow through a network to better utilize network resources, increasing the speed and reliability of that network.

Several papers have been published addressing these issues and discuss the different applications of these algorithms towards network traffic identification and prediction. The Network Identification paper discusses how machine learning can be used to automatically and independently identify and classify network traffic. The Network Prediction paper discuss how the MLP, MLPWD, and SVM algorithms can be used to try and predict future network traffic. The final piece is another paper on network traffic identification using machine learning. The work specifically discusses how K-means and expectation maximization can be used to classify network flow.

[1] Written by group

# Network Identification[2]

Traffic identification has multiple benefits including trend analysis, quality of service mapping, dynamic access control, and intrusion detection [1]. Traditional methods use port numbers as well as decoding protocols and packet analysis to identify network traffic. However, these approaches have several issues which make them not ideal. The issue with using port numbers to identify traffic flow is that port restrictions are not enforced, and many applications do not register which ports they use [1]. Decoding and packet analysis on the other hand are more reliable, however take a lot more processing power to analyze all possible network packets. This method also runs into problems if the packets are encrypted and can't be read or decoded [1]. This paper tries to solve this problem by applying machine learning (ML) to automatically identify and classify network traffic based on attributes such as size, duration, packet length, interarrival time distribution, and idle time. The result is a ML algorithm which will minimize processing cost and maximize the classification accuracy. This will allow the algorithm to automatically and independently identify and classify network traffic.

There are two steps in order to use machine learning to identify and classify traffic flow. The first step is to find a model of existing flow characteristics in order to learn classes. Once the classes are learned from existing network flows, new network traffic can be learned and classified [1]. In this ML algorithm, the classes are learned by using an autoclass approach, which is an unsupervised bayesian classifier capable of automatically learning classes inherent in a dataset [1]. In order to identify traffic flows based on the classes found using the autoclass approach, feature selection is used to try and achieve a high accuracy in identifying network traffic [1]. This algorithm uses sequential forward selection in order to try and identify the best attribute to use to identify the class. This is an iterative approach that continues until there is no apparent feature which would lead to a better result [1]. These features are then used to asses the quality of a class and determine intra-class homogeneity. Intra-class homogeneity is used to try and distinguish different application flows from one another, and ultimately identify the traffic flow [1].

In order to determine the accuracy of the ML algorithm, results of existing internet traffic is compared. The datasets they use are Auckland-VI, NZIX-II, and Leipzig-II traces from NLANR [1]. However, this data is anonymous so the true identification of the traffic cannot be known. This paper makes the assumption of using IANA port numbers to identify the traffic in this dataset. Since this is not ideal, the results of the ML algorithm on this dataset is considered a lower bound. The average accuracy of traffic identification of this dataset is 86.5%. The accuracy of the algorithm appears to depend on the application. The applications tested in this paper are FTP, telnet, SMTP, DNS, HTTP, AOL, Napster, and H-Life. Some applications have characteristics vastly different from other applications, while some tend to share vary close characteristics to one another. Applications with similar characteristics are harder to differentiate and identify. Analyzing a larger dataset might help to improve identification in this case because there will be more differences in characteristics available for the ML algorithm to learn.

---

[2] Written by Alex Baker

# Network Prediction of Mobile Network Traffic[3]

As more people are using and relying on mobile networks, the demands on the resources of mobile networks are increasing. Such demands require network resources to be better managed in order to provide quality of service, congestion control, admission control, network allocation and to minimize cost. Predicting of network traffic can help address these issues. In order to better help manage mobile network resources this paper aims to compare the prediction accuracy of three machine learning algorithms: MLP - Mulit-Layer Perceptron, MLPWD - Multi-Layer Perceptron with Weight Decay and SVM - Support Vector Machine. The MLP and MLPWD algorithms are variations of Artificial Neural Networks which is a nonlinear time series model. The difference between MLP and MLPWD is the structural risk minimization approach to create the regression models [2]. The methods used to find the most accurate regression model that predicts mobile network traffic's future behavior is the Java WEKA tool implementation of the MLP, MLPWD and SVM regression models. MLP is a feed forward artificial neural network that maps input data to appropriate output by using a network of perceptrons [2]. MLPWD uses a structural risk minimization approach to create a prediction model, however, there is a compromise between complexity of the prediction model and quality of fitting the training data [2]. SVM also uses structural risk minimization to create regression models. It is a learning algorithm used for binary classification [2]. Along with the three machine learning algorithms real-life dataset from commercial trial mobile network. This dataset was composed of 1,012,959 rows which represented aggregated traffic of one specific cell in the network and 27 columns of features [2]. The data had to be prepared first by removing duplicates and selecting cells with the most data points. The features were also reduced by using the Java WEKA attribute selection tools. The resulting dataset consisted of 168 rows and 24 columns.

Two experiments were conducting to predict the accuracy of the three algorithms. Experiment I used all 24 attributes to train and test while experiment II used only one attribute along with the sliding window technique to train and test. The target class attribute used in Experiment I and the single attribute used in Experiment II is the attribute which represents "the number of active users connected to the network cell and represents the workload of the cell" [2]. It is chosen as it has a strong correlation to the other features and also represent the workload of a cell.

The results from the experiments concluded that SVM is better than MPL and MPLWD at predicting multidimensionality of real-life traffic data [2]. However, using unidimensional data MLPWD is better at predicting traffic's future behavior. In the future, conducting research to facilitate real-time data analytics for large size of data and to find a dominant feature among a set features to better predict network traffic.

[3] Written by Erik Granger

# Network Traffic Identification II[4]

Hardeep Singh, an Assistant Professor at Lovely Professional University, was able to give insight into network security by comparing two machine learning algorithms that were tasked with analyzing packets of information as they travelled through the network [3]. The algorithms are given metadata about each packet, namely: source port, destination port, source address, destination address, and protocol used. Also they have information about network traffic in general (minimum packet size, maximum packet size, variance of packet length, how long each traffic flow lasts, how long between traffic flows, and direction of packet travel.) Each algorithm was trained with known packets and then exposed to new packets and given the task of correctly classifying each one. The two algorithms chosen for this study were the K-means clustering method and the expectation maximization algorithm.

K-means is an algorithm that groups datapoints into clusters based on how close they are to each other. The distance in this case refers to the difference between the various clusters in the data collected about them. If information packets are sufficiently similar to each other this algorithm will group them into the same cluster. By clustering new packets with those that were identified during the learning portion and whose category is known the K-means algorithm makes conclusions about new, unknown packets.

The expectation maximization algorithm also sorts data into clusters. However, it does so in a more probabilistic way. It starts by taking as many clusters as it has been told to make and assigning their centers randomly. With this setup complete it refines the initial guess in an iterative way by assigning each point to a cluster with a certain probability and then it maximizes the probability of that refinement being correct. The algorithm then repeats these two steps until its output converges.

K-means was able to correctly identify packets more often over expectation maximization across five different types of file. This was only a small increase in accuracy for each file type. The largest was an approximately 3% difference in accuracy between the two methods for SMTP and ICMP. When Professor Singh adjusted the number of clusters each algorithm was allowed to use K-means once again stayed consistently ahead in accuracy. Also, the accuracy of both increased with the number of clusters they were permitted to use. So, not only did this study inform the community that K-means is a more accurate algorithm for this type of work but also suggests that clustering algorithms might be more accurate with more clusters.

[4] Written by Tarek Embree

# Conclusion[5]

Network traffic classification and prediction has proven useful for improving quality of service and network optimizations. Several machine learning algorithms were able to automate the process of analyzing network traffic to achieve these goals. While this is useful for improving network speed and quality, these algorithms can also be extended to help improve network security. By using similar methods for traffic identification, machine learning can be used to identify malicious traffic on a network, helping to increase the security and awareness of the traffic flow on a given system. Network traffic prediction can be used to help optimize quality of service protocols and other network characteristics. Being able to optimize a network for a predicted type of traffic will ultimately help to increase speed and reliability on that network. Automated network traffic identification is an exciting new application of machine learning that will continue to grow and ensure our networks are fast, safe, and secure.

[5] Written by group

# References


1. S. Zander, T. Nguyen and G. Armitage, "Automated Traffic Classification and Application Identification using Machine Learning," *The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05)l*, Sydney, NSW, 2005, pp. 250-2525
2. A. Y. Nikravesh, S. A. Ajila, C. H. Lung, and W. Ding, "Mobile Network Traffic Prediction Using MLP, MLPWD, and SVM," *2016 IEEE International Congress on Big Data (BigData Congress)*, San Francisco, CA, 2016, pp. 402-409.
3. H. Singh, "Performance Analysis of Unsupervised Machine Learning Techniques for Network Traffic Classification," *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, Haryana, 2015, pp. 401-404.