Robust Streaming Video Traffic Classification

Jordan Ebel, December 2015
Computer Science Department, Stanford University

Background

Internet service providers use network monitoring tools to manage their networks and ensure quality of service for all subscribers. The growth of streaming video and encrypted traffic have made traditional management tools ineffective. This project aims to:

- Identify a top performing machine learning algorithm to classify streaming video traffic
- Build a real time system to classify traffic while maintaining performance, privacy, and video quality

Methods

- Datasets collected of streaming video and Internet browsing using the Wireshark packet analyzer
- Python script implemented to extract features and identify true packet classifications
- Logistic regression, Naïve Bayes, SVM, and K-Means Clustering algorithms prototyped and tuned in Matlab
- Top performing algorithm implemented in Python
- Real time traffic classification system written using trained algorithm and live feature extraction

Data

- Packets captured from the test computer for one minute, resulting in approximately 80,000 test and train packets
- Packet captures contain the timestamp and entire contents of each packet
- Figure 1 below shows an example of 6 streaming video packets from the training data set

51727	35.034829	173.194.26.109	192.168.1.252	TCP	1514	[TCP segment of a reassembled PDU]
51728	35.034829	23.246.14.169	192.168.1.252	TCP	1514	[TCP segment of a reassembled PDU]
51729	35.034830	23.246.14.169	192.168.1.252	TCP	1514	[TCP segment of a reassembled PDU]
51730	35.034832	23.246.14.169	192.168.1.252	TCP	1514	[TCP segment of a reassembled PDU]
51731	35.034832	23.246.14.169	192.168.1.252	TCP	1514	[TCP segment of a reassembled PDU]
51732	35.034833	23.246.14.169	192.168.1.252	TCP	1514	[TCP segment of a reassembled PDU]

Figure 1: Packet number, Time, Destination IP, Source IP, Protocol, Length, and Description of Streaming Video Packets

A visual observation of the training packet capture revealed interesting patterns. Streaming video traffic usually occurred in bursts, with a large number of very similar packets appearing at short intervals. Standard web traffic was more irregular and featured a wider variety in the features of the packets.

Features

To capture the patterns found in the data, features describing individual packets and groups of packets were extracted from the data. A total of 14 features were extracted, with a focus on not including features that would diminish the robustness of the system. Figure 2 below details the features and their sources.

Application Message
Unused

Transport Segment
Unused

Network Datagram
Mean and var. IP length — IP protocol
Mean and var. of TTL — IP offset
Mean and var. of protocol — IP length
IP time to live (TTL)

Link Frames
Total length — Inter-arrival time
Mean inter-arrival time
Variance of inter-arrival time

Figure 2: Packet Features

Figure 3 below shows a biplot of the training dataset and the features, projected onto the first two principle components as determined by PCA. The biplot reveals a tradeoff between length vs. time to live as the first principle component, and protocol vs. offset as the second principle component.

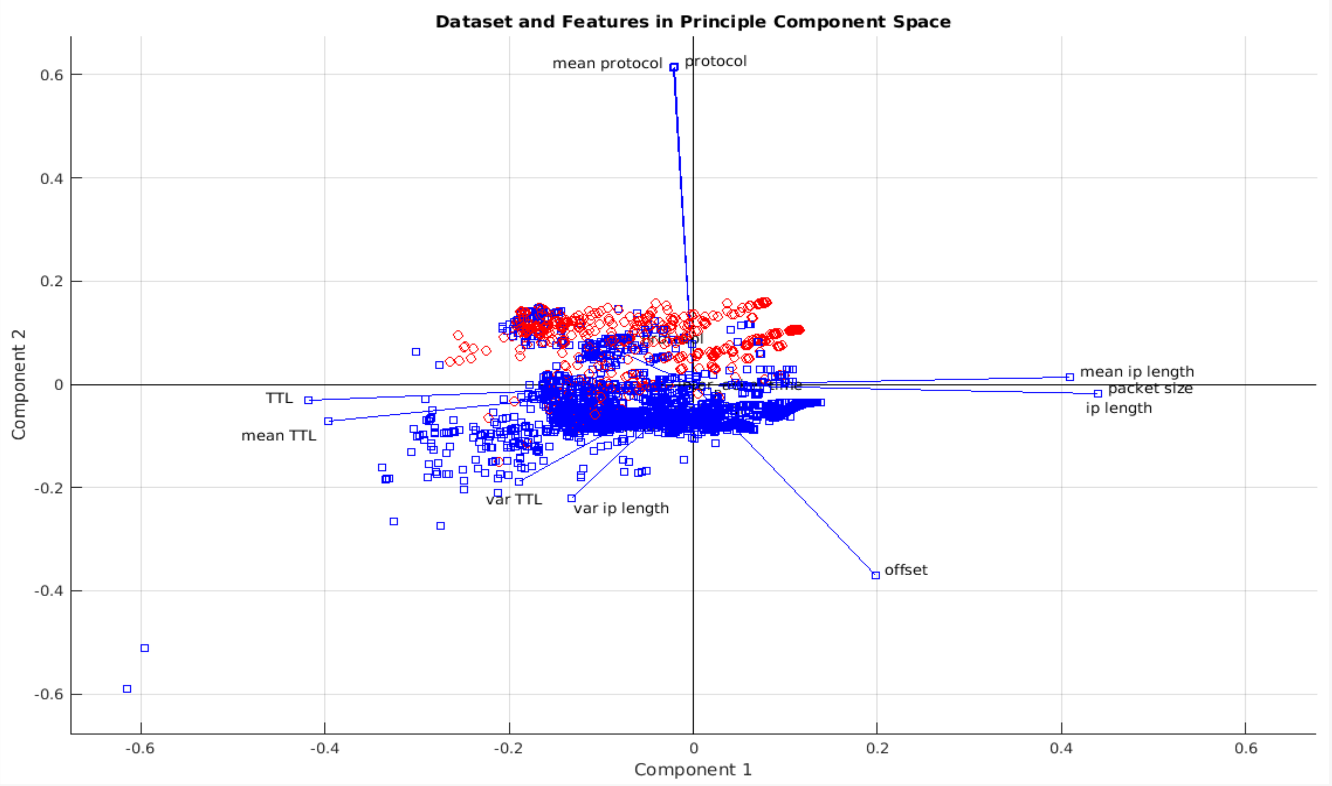

Figure 3: Biplot of Training Dataset in Principle Component Space

Supervised Learning Results

The SVM algorithm with a Gaussian kernel proved to be the top performing algorithm on the dataset. The SVM algorithm displayed test and training errors below 10% for a wide range of training set sizes, as shown in Figure 4.

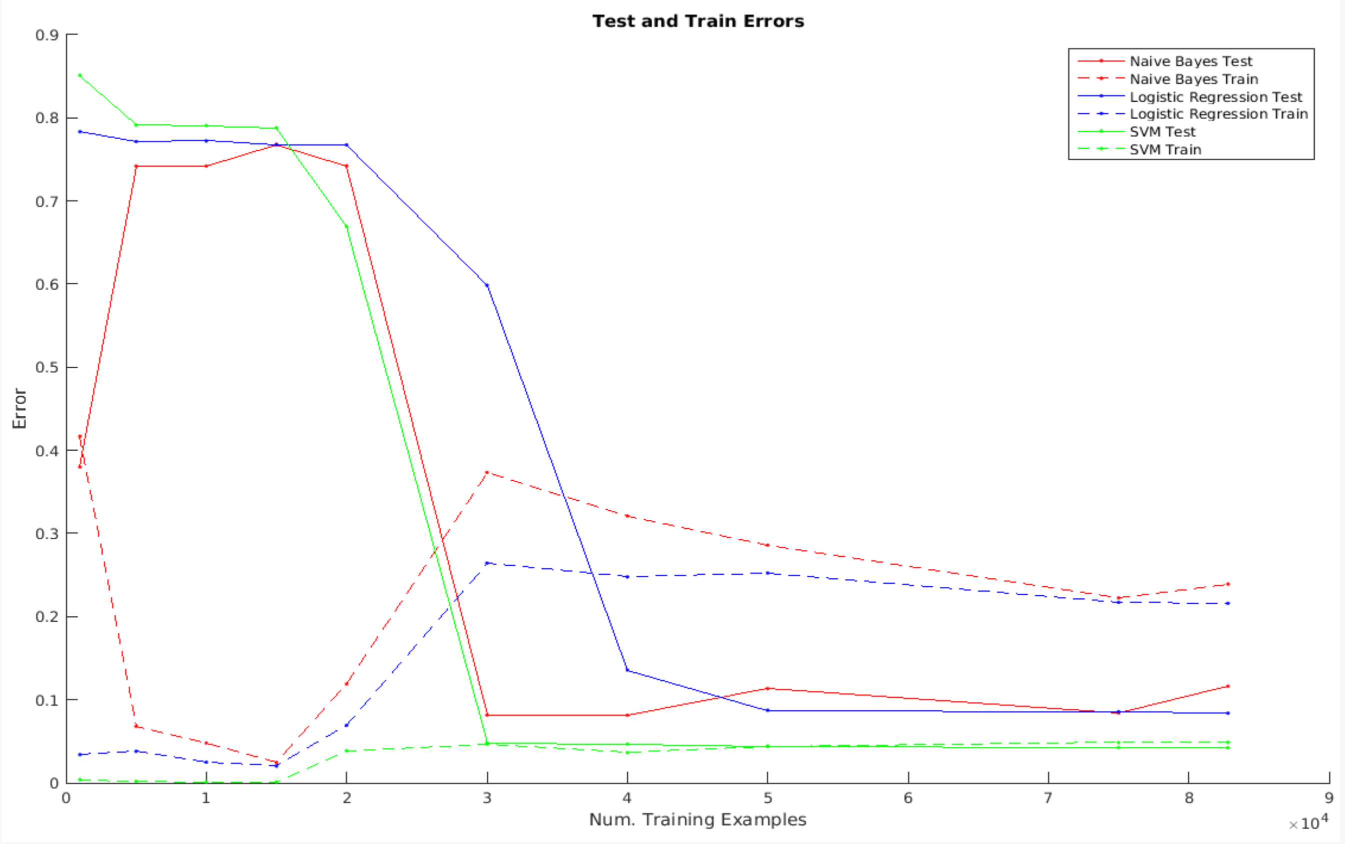

Figure 4: Test and Train Errors

The SVM algorithm was also the top performing algorithm in a Receiver Operating Characteristic (ROC) analysis, shown in Figure 5. The SVM algorithm performed very well at identifying true positives, even at a relatively low false positive rate. The SVM algorithm was also the top performer in a recall-precision analysis.

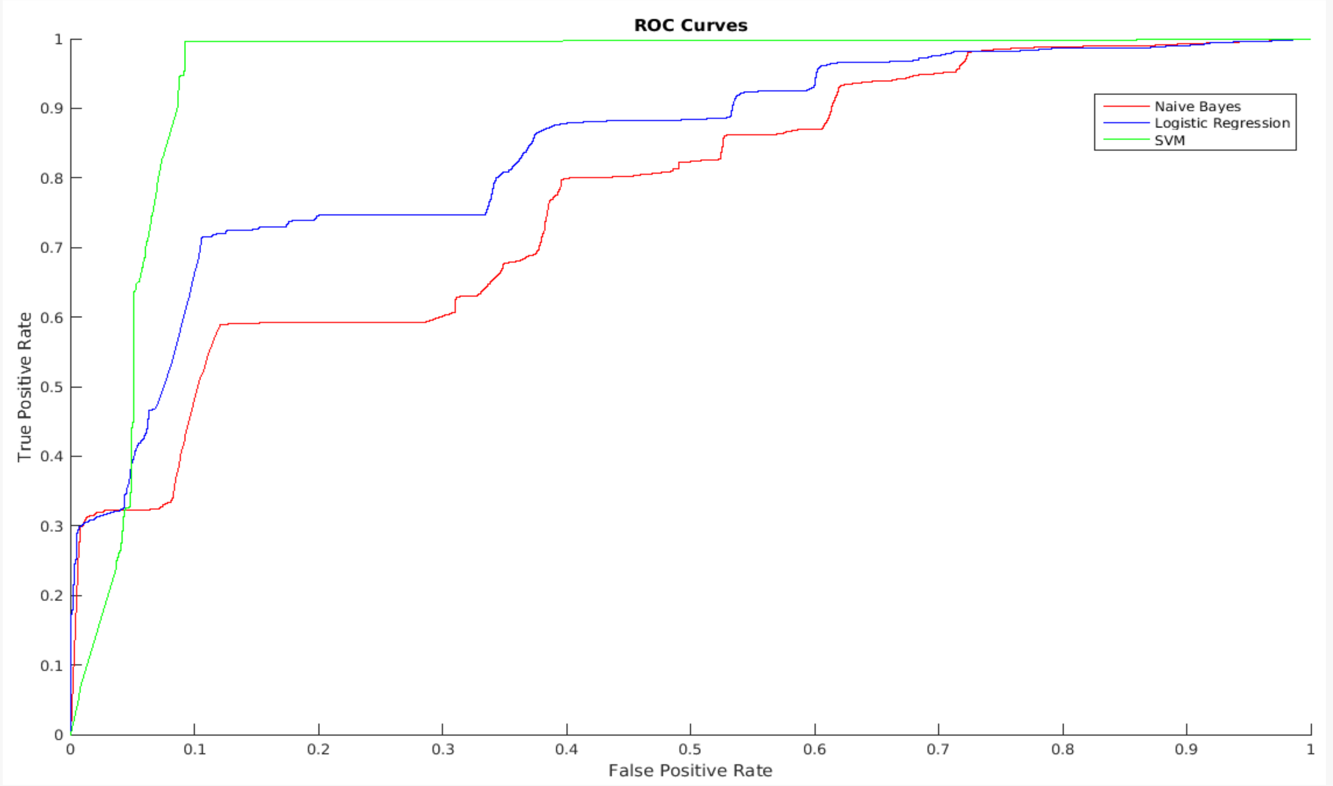

Figure 5: ROC Curves

Unsupervised Learning Results

The K-Means Clustering algorithm failed to identify either traffic group accurately. Figure 6 below shows an inaccurate grouping of clusters to labels.

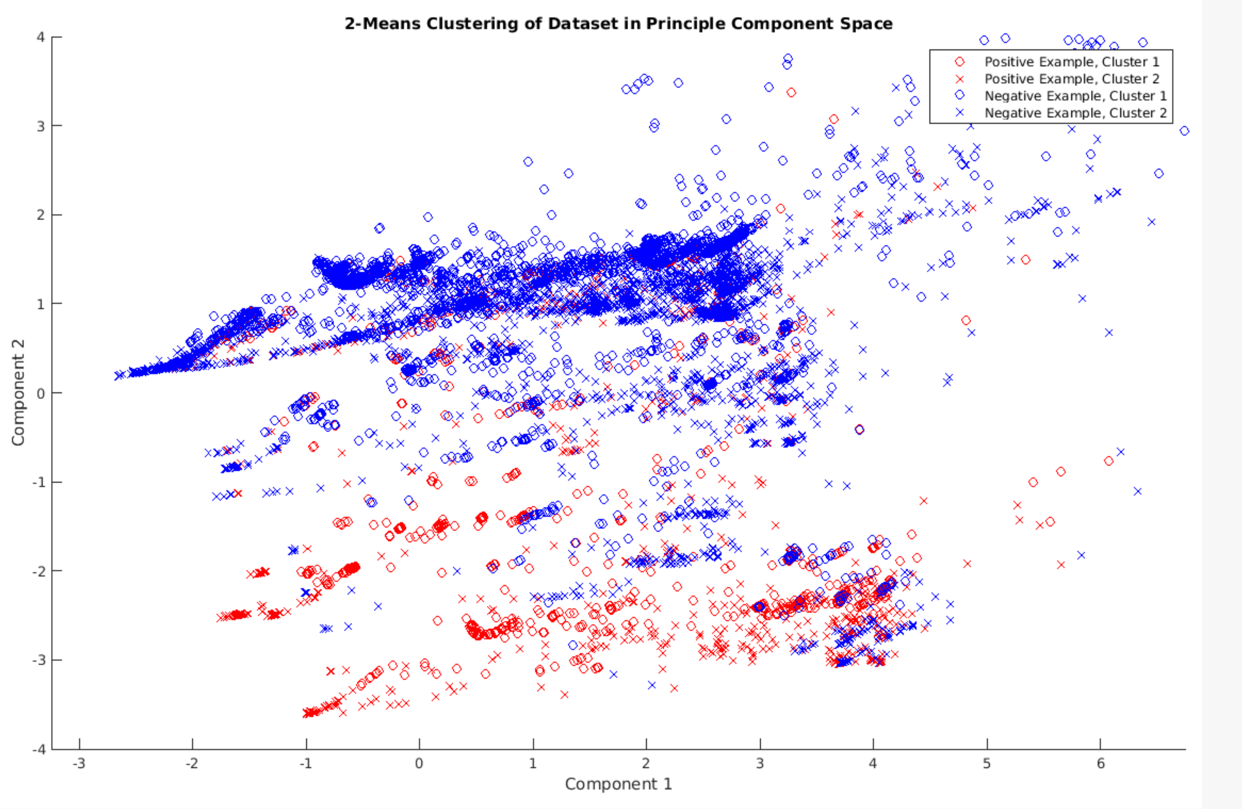

Figure 6: Clustering of Data in Principle Component Space

Real Time System

The SVM algorithm was selected as the algorithm to implement in a real time classification system, shown below in Figure 7 operating on live traffic.

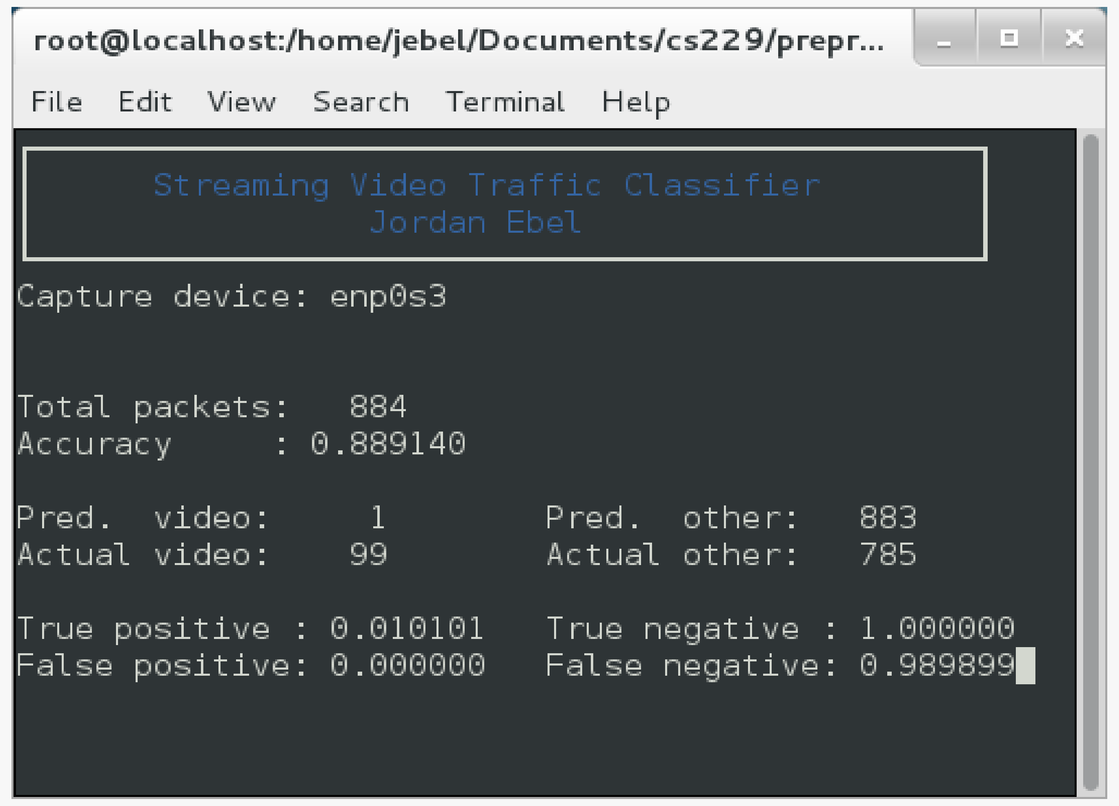

Figure 7: Real Time Classifier In Operation

The real time system did not extract features from IP addresses, port numbers, or the HTTP, TCP, or UDP headers or payloads. Therefore, the system is robust to:

- Encryption
- Web Proxies
- Port number changes
- IP masquerading
- Packet fragmentation