Classification of Network Advertising Traffic

Timothy Pace, Manas Thakre, & Dania Tanzil
University of Washington Information School

Introduction

Internet traffic classification has garnered increasing industry interest by network operators due to its associated benefits in network management, security monitoring, and improving quality of service (Moore & Zuev, 2005). With the emergence of clickstream-based web-advertising models, classifying advertisement traffic is an emerging challenge — and opportunity — for mobile providers and ISP's, and may help optimize network efficiency and combat fraudulent ad traffic (e.g., spam; Chatterjee et al., 2003).

The current research aims to address the feasibility of using machine learning (ML) to classify ads on a mobile network, as well as evaluate algorithmic performance. There is an absence in existing literature on this subject. Preprocessing was completed using a novel method of labelling ad traffic, in combination with unsupervised ML methods such as principal component analysis (PCA).

Methods

Network traffic data from T-Mobile was captured through the Wireshark open source packet analysis tool from carrier provided test devices. Network traffic was collected in approximately hour-long intervals across several weeks by surfing Alexa top 500 websites.

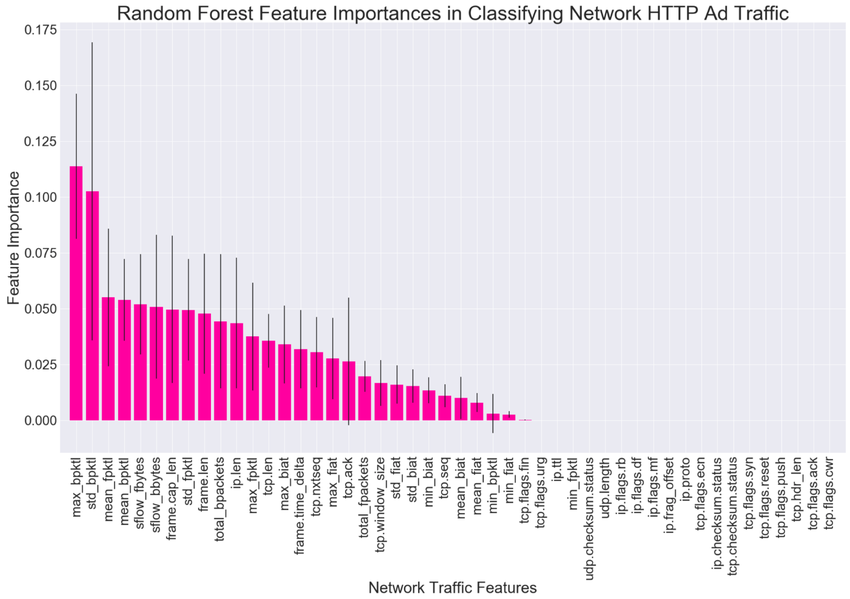
Packet header based features collected, and packet flow-based features that were calculated and employed were adapted from Alshammari and Zincir-Heywood (2011).

Approximately 4 million rows of network traffic were collected from mobile web browsing. Traffic was filtered to HTTP protocol traffic (N = 43,206). Using a novel approach, traffic from advertising domains were labeled using the latest Adblock domains (via *EasyList*). Ads represented 8% of non-encrypted browsing traffic (n = 3,548). Rows of non-ad traffic were then down-sampled.

Using Python's scikit-learn, ML classifiers and metrics were tested and adapted from Bakhshi and Ghita (2016).

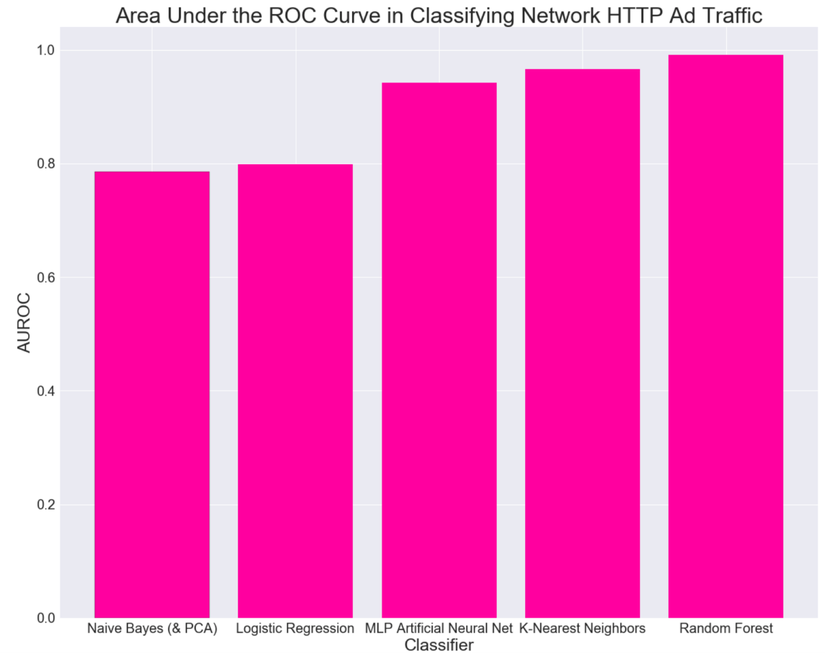
Results

Random Forest (RF) yielded the highest area under the receiver operating characteristic (AUROC) performance as a machine learning classifier of HTTP ad traffic (.99), followed by K-Nearest Neighbors (.97) and a Multilayer Perceptron Artificial Neural Network (.94). Logistic Regression and Naïve Bayes (with PCA) yielded the lowest AUROC performance (.80 and .79). Packet flow statistics, such as mean and standard deviation back and forward packet lengths, represented the majority of important RF features.



Machine Learning Classifier Algorithm	Area Under the Receiver Operating Characteristic	Accuracy Score
Random Forest	.99	97%
K-Nearest Neighbors	.97	92%
Multilayer Perceptron Artificial Neural Network	.94	89%
Logistic Regression	.80	76%
Naïve Bayes (& PCA)	.79	46%

Table 1. Machine Learning (ML) Algorithm Ad Classification AUROC & Accuracy



Discussion

The excellent (.90+) ROC AUC test and accuracy scores of three of the five ML algorithms used in the present study in classifying HTTP ads suggest that non-encrypted ads may have unique packet features differentiating them from non-ads in network traffic packets. Moreover, high ad classification performance was achieved with only modest tuning of the hyper-parameters of the different ML algorithms employed, leaving room for even greater future classification performance gains.

Conclusions

The results of the present study suggest that HTTP ads may have unique packet signatures in network traffic differentiating them from non-ads. This may enable ISP's and mobile carriers to improve quality of service for customers and mitigate spam. Future research should tune the hyper-parameters of and scale these algorithms on carrier data in real time, and model HTTPS ads. This may enable the real-time classification of and mitigate malicious advertising.

Contact
Timothy Pace (tpace211@uw.edu)
Manas Thakre (manast@uw.edu)
Dania Tanzil (daniat@uw.edu)

References

1. Alshammari, R., & Zincir-Heywood, A. N. (2011). Can encrypted traffic be identified without port numbers, IP addresses and payload inspection?. *Computer Networks*, 55(6), 1326-1350.
2. Bakhshi, T., & Ghita, B. (2016). On internet traffic classification: A two-phased machine learning approach. *Journal of Computer Networks and Communications*, 2016.
3. Chatterjee, P., Hoffman, D. L., & Novak, T. P. (2003). Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, 22(4), 520-541.
4. Moore, A. W., & Zuev, D. (2005, June). Internet traffic classification using Bayesian analysis techniques. In *ACM SIGMETRICS Performance Evaluation Review* (Vol. 33, No. 1, pp. 50-60). ACM.