Classification of VPN network traffic flow using time-related features

Emmanuel Ayeleso
Faculty of Engineering
Computer Science
University of Ottawa
Email: eayel037@uottawa.ca

Alex Gagnon
School of Computer Science
Carleton University
Email: asgagnon@hotmail.com

***Abstract*—The abstract goes here.**

I. Introduction

VPN (Virtual Private Network) is a service that keeps activities over the internet private and secured. It works by routing traffic on the internet through a VPN-tunnel that encrypts data and hides user's IP address. Such VPN-tunnel makes it difficult for ISP providers, hackers, man-in-the-middle, government and regulatory bodies to snoop or view details of one's activities on the internet. This implies that VPN is an effective security and safety tool towards secured access to home, office and public internet networks that is accessed through means such as Wi-Fi, Hotspot etc. However, VPN is seen as a threat by some regulatory bodies that may want to monitor and censor activities going over the internet. For instance, some countries want to monitor activities of their citizens over the internet, tertiary institution authorities often times, want to censor what their students can access over their network in order to shun distraction.

Interestingly, VPN is gradually becoming enemy to these countries' government that want to monitor and censor their citizens activities on the internet. These governments have resorted to banning VPN technology in their countries. The problem is that such outright ban, infringes on the privacy and rights of the citizens and businesses that want to use VPN for genuinely valid and sensitive transactions that require secure tunnels. In addition, secure and safe internet-required transactions with these countries will also be adversely threaten.

We propose the use of machine learning algorithms to classify network traffic flow over the internet using time-related features as a solution to this problem. This will be a better approach to firstly classify traffic over the internet as a VPN or non-VPN traffic. Secondly, these algorithms can further be used for monitoring and censorship agenda of these regulatory authorities. Our approach is without infringement or limitation on the privacy and security of the citizens and businesses. Considering the time factor, the scope of our work shall be limited to classification of the internet traffic as either VPN of non-VPN.

We have focused on training our classification models (Decision Tree, SVM, KNN, Kmeans and Hoeffding tree) with the datasets created by Canadian Institute of Cybersecurity. In addition, procedural application of the skills learned in CSI 5155 class make our results to be unique. Our contribution shall be, training these data sets with the proposed models and using appropriate statistical tools to measure if the performance of these models are significant.

II. Description of the problem domain

Recently, research community has shifted focus towards the use of machine learning techniques to classify internet traffic effectively [2], [3], [4], [6]. This is seen as a better approach to the diminished effectiveness of port-based and the computational overheads of the deep packet inspection that are the traditional approaches[6]. The concept of machine learning techniques basically relies on statistical properties of the internet traffic which are used as features to train machine learning models for classification. Such features are: distribution of flow duration, flow idle time, packet inter-arrival time and packet lengths of various applications that have been asserted by the researchers as effective yardstick to classify internet traffic [3], [1], [7].

Two research works motivated our study [1], [7]. Gerard et. al. studied the effectiveness of flow-based time-related features to detect VPN traffic. They used Weka to implement C4.5. decision tree and KNN algorithms to classify their generated data sets using 10 folds cross validation. Sikha et. al also used the same data sets generated by Gerard et. al. study to train six different machine learning classifier models (logistic regression, SVM, Naive Bayes, kNN, Random Forest and Gradient Boosting Tree). They compare the performances of these models and recommended the best performing model for VPN and non-VPN traffic classification. As a contribution to the body of knowledge, Our study targets the use another set of classifiers in conjunction with feature engineering to achieve a better results. We shall also be measuring the performances of out trained models in terms of their statistical significance using appropriate tools.

III. Model Selection

We used six models: Decision Tree, SVM, KNN, Kmeans and Hoeffding tree to train the data sets. Our choice of these models is informed by our inquisitiveness to examine the performance of similar in operation, but different models to the ones used by Sikha et. al.

A. Decision Tree

Decision trees are supervised non-parametric estimator built by recursive partitioning. It predict the target class by making its inference from the training data set.

B. SVM

C. KNN

D. Kmeans

E. Hoeffding tree

Hoeffding tree is a variation of the decision tree that learn large data stream incrementally with the assumption that the data distribution is not changing over the time [8], [9].

IV. Experimental setup

We downloaded the data sets and converted them into Panda data frame recognised by the sklearn. Thereafter, we performed data exploration and visualization using sklearn libraries. These actions allow us to know the necessary actions required before the application of our models. We observed the following: firstly, discrepancies in a particular data set features in comparison to the others. Secondly, a lot of missing values were denoted with negative one (-1). Thirdly, need for normalization to avoid dominance of some features with huge values compared to others. Actions taken to resolve these attend to these observations are reported in subsection A -E. Thereafter, we used sklearn to implement the six algorithms using 10 fold cross validation.

A. Data sets

We used a generated real-world traffic bench-marked intrusion detection data sets created by the Canadian Institute of Cybersecurity to train the six models. The data sets have 24 attributes generated under the time frames: 15, 30, 60 and 120 seconds and the instances of 76,379, 14,670, 15,238 and 10,801 respectively. Our findings revealed that 30 seconds time frame data set had only 23 attributes and was dropped as result of its incompatibility with others.

B. Data Cleaning

Our data exploration revealed that rows concerned with negative entries as a missing value were up to forty percent of our data sets. We then wrote a script in sklearn to automatically delete these rows. Our believe is that such fictitious entries will adversely affect our models' result.

C. Merging of Data sets

?

D. Normalization

What did we used to implement this?

1) Subsubsection Heading Here: Subsubsection text here.

Evaluation criteria That is the experiment

Results and Discussion

Insights- lessons learned

V. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] Bagui, Sikha Fang, Xingang Kalaimannan, Ezhil Bagui, Subhash Sheehan, Joseph. (2017). Comparison of machine-learning algorithms for classification of VPN network traffic flow using time-related features. Journal of Cyber Security Technology. 1-19. 10.1080/23742917.2017.1321891.

[2] M. Shafiq, Xiangzhan Yu, A. A. Laghari, Lu Yao, N. K. Karn and F. Abdessamia, "Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms," 2016 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2016, pp. 2451-2455.

[3] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," in IEEE Communications Surveys Tutorials, vol. 10, no. 4, pp. 56-76, Fourth Quarter 2008.

[4] Kim, Hyun-chul Fomenkov, Marina Claffy, Kc Brownlee, Nevil Barman, Dhiman Faloutsos, Michalis. (2009). Comparison of Internet Traffic Classification Tools.

[5] Manuel Crotti, Maurizio Dusi, Francesco Gringoli, andLuca Salgarelli.Traffic classification through simple statistical fingerprinting.SIGCOMM Comput. Commun.Rev., 37(1):5–16, January 2007

[6] Jeffrey Erman, Anirban Mahanti, Martin Arlitt, IraCohen, and Carey Williamson.Offline/realtime traffic classification using semi-supervised learning.Perform.Eval., 64(9-12):1194–1213, October 2007.

[7] Habibi Lashkari, Arash Draper Gil, Gerard Mamun, Mohammad Ghorbani, Ali. (2016). Characterization of Encrypted and VPN Traffic Using Time-Related Features. 10.5220/0005740704070414.

[8] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In KDD'01, pages 97–106, San Francisco, CA, 2001. ACM Press.

[9] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In KDD, pages 97–106, San Francisco, CA, 2001. ACM Press.