ISBN 978-981-14-1455-8
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
WCSE_2019_SPRING
Yangon, Myanmar, February 27-March 1, 2019, pp. 42-47
doi:10.18178/wcse.2019.03.008

Ensemble Learning for Detecting Remote Access Trojans

Khin Swe Yin [1], May Aye Khine [2 +]

[1,2] Faculty of Computing, University of Computer Studies, Yangon, Myanmar

Abstract. Machine learning algorithms for network traffic classification has been researched for several years. They are useful for both encrypted and unencrypted network traffic classification. Nowadays malicious malware like Remote Access Trojans go through network, and they are secretly installed in a victim's computer, they stay in the victim host and communicate back to the attacker. The command and control traffic of Remote Access Trojans can be differentiated from normal traffic using machine learning based techniques. This paper compares the performance of nine supervised machine learning algorithms for detection of Remote Access Trojans. Both unbalanced and balanced dataset are applied for building model. Four ensemble learning methods give high detection rate. Among them, AdaBoost ensemble learning outperforms the competing methods, and it gets the best accuracy, least false negative rate and least false positive rate.

Keywords: machine learning algorithms, Remote Access Trojans, AdaBoost algorithm

1. Introduction

Network traffic classification is an important task for network administrators to know which applications are running on Internet. It helps to analyse different network traces and identify normal and malicious traces. Different types of applications are classified according to the organization's policy. Common applications are WWW, FTP, NTP, Telnet, DNS and P2P. Malware like Remote Access Trojans and Advanced Persistent Threats use command and control servers to control the victim, and then they access and steal sensitive information from the victim. Attacker's intrusion that uses known malware can be found by monitoring network traffic and by using tools like snort and antivirus scanners. These tools use signature database that is necessary to update regularly in order to detect new threats. Payload based technique, also called deep packet inspection (DPI) uses signatures for classification. Signatures are patterns that are associated with each application's characteristics. The classification engine compares the captured traffic against these signatures to identify the exact applications. The database needs to be updated periodically to get new signatures for new applications.

Nowadays machine learning methods are applied for both unencrypted and encrypted network traffic classification. This technique is based on features extracted from network traffic. The best features can give best accuracy and best performance result. Wireshark and tcpdump are tools that are commonly used for capturing network traffic. Different features are defined and some machine learning algorithms like K-Means, NB Tree, Naïve Bayes, Decision trees and Random Forest are used to detect Trojans [1][2][3].

There are two approaches in machine learning (1) supervised, and (2) unsupervised. The supervised learning technique trains a model with some labelled data set and then it will predict label for new data. Supervised learning is commonly used in network traffic classification and it gives good results. Decision trees and random forests are the best classifier for normal applications like WWW, FTP and for encrypted malware classification [4][5][6]. Ensemble learning algorithms are classifiers formed by a set of base

+ Corresponding author. Tel.: + 959795477763
E-mail address: khinsweyin@ucsy.edu.mm

classifiers that cooperate to get an optimal predictive model. Random Forests, Bagging, AdaBoost and Random Tree are ensemble methods and they are applied for network traffic classification.

This paper is organized as follows: Section 2 highlights related works, preliminary is explained in section 3. Section 4 describes experiment and results. This paper is concluded in section 5.

2. Related Works

Traffic classification is important for network management and network security. Researchers still find new approaches to withstand the fast changes of the Internet. There are three network traffic classification techniques- Port-based Technique, Payload Based Technique and Machine Learning (ML) techniques. Port-based technique fails when dynamic port number is assigned for applications. Payload based techniques is also called deep packet inspection (DPI) and it performs searching and comparing pattern in a payload. But it does not work in encrypted network traffic. Four machine learning methods – c4.5, SVM, Bayes Net and Naïve Bayes are used for classifying WWW, P3P, FTP, DNS and Telnet applications [1]. The best classifier C4.5's accuracy is above 78% [4].

Encrypted network traffic classification is performed with six popular machine learning algorithms [5]. The analysis is based on TLS encrypted sessions collected from malware sandbox and two geographically distinct, large enterprise networks. Linear Regression, Logistic Regression, Decision Trees, Random Forests, Support Vector Machine and Multi-layer Perceptron are applied for classifying network traffic. Random Forests is the best classifier among them.

Network traffic analysis and prediction techniques are discussed in [7]. Several classification methods and several datasets are explained for analysis. SVM, Decision Tree, Neural network and Statistical approaches are included in this analysis. The datasets are DARPA data set, NSL-KDD data set, AIDA data set, Waikato data set, Berkeley Lab data set, ACM SIGCOMM'01 data set and DARPA data set.

Seven popular ensemble algorithms based on Decision Trees are applied for building models, it focuses on not only model accuracy but also byte accuracy and latency in order to determine that ensemble learning can be properly applied to this modeling task. Seven ensemble classifiers are OneVsOne, Error-Correcting Output-code, AdaBoost classifier (ADA), Bagging algorithm, Random Forests and Extremely Randomized Trees. Network traces are obtained from ISP traces and host traces in two different environments. The applications they classified are WWW, DNS, NTP, INT and P2P [6].

The problem of detecting malware on client computers based on HTTPS traffic analysis is studied in [8]. A malware detection method based on a neural language model and a long short-term memory (LSTM) network is derived. Malware is detected based on the host address, timestamps, and data volume information of aggregated packets that are sent and received by all the applications on the client. The previous works classifies normal applications like WWW, P2P and so on [5][6][7]. Reference [8] uses detection method based on host address, timestamps and others that are much overhead.

In this paper, the command and control traffic of remote access Trojans is differentiated from normal traffic by nine supervised algorithms. Naive Bayes, KNN, Neural network, Simple Logistic regression, Decision Trees, Bagging, Random Tree, Random Forests and AdaBoost are applied to build models for detecting Remote Access Trojans. Effective features are applied for these algorithms. Both balanced and unbalanced ratio of instances are applied for classification. Decision Trees and ensemble methods give the best accuracy, least false negative rate and least false positive rate. Among them, AdaBoost is contributed for RAT detection model and it obtains the best accuracy with least FNR and least FPR.

3. Preliminary

3.1. Remote Access Trojans

Remote Access Trojan, also known as Trojans, are malware camouflaged as legitimate application. They use drive-by-download and spear-phishing tactics in order that they can be secretly installed on endpoints [9]. The attacker applies command and control servers to control the victim's PCs remotely and secretly, and then he or she obtains opportunity to steal confidential information, erase or overwrite data,

listen the key logger or capture the system screen. As they can conceal themselves and attach to legitimate program, they rarely occur to appear in task manager or system monitors.

3.2. Wireshark

Wireshark is the world's foremost network protocol analyser [10]. It is used to capture network traffic. It is applied to defend network or to attack the victims. It includes deep inspection of hundreds of protocols, live capture and offline analysis.

3.3. Feature Extraction

Features are collected from the traffic of 10 types normal application and 10 remote access Trojans [11]. Network traffic is captured by wireshark. Features are extracted from the first twenty packets of the traffic, and seven features are used in this work. Seven features are (1) Outbyte(outbound data byte), (2) Inbyte(inbound data byte) , (3) InByteByInPac(Inbound data byte/ Inbound number of packets) , (4) OutByteByOutPac(Outbound data byte/ Outbound number of packets) , (5) Duration(time duration from the first packet to twentieth packets), (6) OutByteByInByte(Outbound data byte/ Inbound data byte), and (7) OutPacByInPac(Outbound number of packets/ Inbound number of packets).

3.4. Machine Learning Algorithms

3.4.1 Logistic Regression

Logistic regression is a statistical based model that predicts the probability of an outcome that can have two values [12]. Simple logistic regression included in WEKA is used in the experiment.

3.4.2. Decision Trees

Decision trees work like a tree structure. A dataset is broken into small subsets and an incremental tree is developed. Decision tree J48 is the implementation of algorithm ID3 developed by the WEKA project team [13] [14].

3.4.3. Random Forests

Random forests is an ensemble classifier that consists of multiple decision trees and outputs the prediction that is more accurate and stable [15].

3.4.4. KNN

K nearest neighbors is a simple algorithm that predicts new cases based on a similarity measure [16].

3.4.5. Random Tree

The Random Tree operator works like the Decision Trees but only a random subset of attributes is available for each split [13] [15].

3.4.6. AdaBoost

Adaptive Boosting, called AdaBoost, is an ensemble machine learning algorithm. AdaBoost was built based on short decision tree models, each with a single decision point called decision stumps [17].

3.4.7. Bagging

Bootstrap aggregating, also called bagging, is an ensemble learner. It is a statistical estimation technique where a statistical quantity like a mean is estimated from multiple samples of data. It works in a more robust estimate of a statistical quantity [18].

3.4.8. Neural Network

Neural networks are organized in layers. Layers are built with a number of interconnected nodes that contain an activation function. Patterns are presented to the input layer, which communicates to one or more hidden layers. The actual processing is done in this hidden layer with a system of weighted connections. Then the hidden layers link to an output layer [19].

3.4.9. Naïve Bayes

Naive Bayes is built based on Bayes theory. It is widely used for classification. Posterior class probability is obtained based on class conditional density estimation and class prior probability. The maximum posterior class probability predicts the test data [20] [21].

4. Experiment and Results

A virtual environment that includes attacker and victim is set up to capture RATs traffic. 10 popular remote access trojans and 10 types of normal applications are used in this experiment. Wireshark is used to capture network traces. 10 remote access trojans are ImminentMonitor, KilerRat, NjRat, Cerberus, Xtreme, Pandora, CyberGate, SpyGate, Xena and Babylon. 10 normal applications are Dropbox, Pcloud, Skype, YahooMessenger, Facebook, Bittorrent, BitComet, Google, Firefox and Chrome. 300 normal instances and 300 RATs instances are collected to train model. Different ratios of normal and RATs instances are trained in order to determine the best method while using both unbalanced and balanced dataset.

Each RAT is run 30 times and 30 instances are collected from a RAT. Then 300 instances are obtained for 10 RATs. 30 instances are collected from a normal application and 300 instances of normal applications are obtained for 10 normal applications.

For 150 normal and 10 RATs instances, 15 instances are selected from a normal application and 150 instances are obtained for 10 normal applications. 1 instance is selected from each RAT, and 10 instances are obtained for 10 RATs.

For 300 normal and 10 RATs instances, 30 instances are selected from a normal application and 1 instance is selected from each RAT. For 300 normal and 300 RAT instances, 30 instances from a normal application and 30 instances from each RAT are selected for classification.

Weka, datamining tool is used to classify data and k-fold Cross Validation is used to validate the result of classification. 10-fold cross validation is applied in this experiment. Accuracy, False Negative Rate (FNR) and False Positive Rate (FPR) are used for evaluation. Accuracy is the correctly classified number of both normal and malicious instances on total instances. FPR is the incorrectly classified number of normal instances on the total normal instances. FNR is the incorrectly classified number of malicious RAT instances on the total RAT instances. The less FNR while maintaining high accuracy, the better the detection system is for not missing malicious sessions. How to calculate Accuracy, FPR and FNR is shown as follow:

$$Accuracy = \frac{Correctly\ Classified\ Number\ of\ normal\ and\ RAT\ Instances}{Total\ Number\ of\ normal\ and\ RAT\ Instances}$$

$$FNR = \frac{Incorrectly\ Classified\ Number\ of\ RAT\ Instances}{Total\ Number of\ RAT\ Instances}$$

$$FPR = \frac{Incorrectly\ Classified\ Number\ of\ Normal\ Instances}{Total\ Number\ of\ Normal\ Instances}$$

Table 1: Accuracy, FNR and FPR of unbalanced instances: 150 normal instances and 10 RAT instances

Performance Metrics	AdaBoost	Random forests	Random Tree	Bagging	Decision trees	Simple Logistic regression	KNN	Naïve Bayes	Neural network
Accuracy	0.988	0.988	0.988	0.981	0.981	0.981	0.969	0.963	0.95
FNR	0.1	0.1	0.1	0.2	0.2	0.2	0.3	0.6	0.7
FPR	0.007	0.007	0.007	0.007	0.007	0.007	0.013	0	0.007

The result of classifying 150 normal instances and 10 RAT instances is shown in Table 1. The FNR of Neural Network and Naïve Bayes is high. KNN, Simple Logistic regression, Decision Trees and Bagging has some FNR-0.2. Random Tree, Fandom Forests and AdaBoost get the highest accuracy and lowest FNR among the other classifiers.

Table 2: Accuracy, FNR and FPR of unbalanced instances: 300 normal instances and 10 RAT instances

Performance Metrics	AdaBoost	Random forests	Random Tree	Decision trees	Bagging	Simple Logistic regression	KNN	Naïve Bayes	Neural network
Accuracy	0.997	0.99	0.99	0.99	0.987	0.977	0.974	0.971	0.971
FNR	0.1	0.2	0.2	0.2	0.2	0.5	0.5	0.7	0.9
FPR	0	0.003	0.003	0.003	0.007	0.007	0.01	0.007	0

300 normal instances and 10 RAT instances are used for classification and the result is shown in Table 2. Neural Network, Naïve Bayes, KNN and simple logistic regression get high FNR in this unbalanced dataset. Bagging, Decision Trees, Random Tree and Random Forests obtain high accuracy and some FNR-0.2. Among them, AdaBoost's accuracy is the highest-0.997 and its FNR is the lowest-0.1.

Table 3: Accuracy, FNR and FPR of balanced instances: 300 normal instances and 300 RAT instances

Performance Metrics	AdaBoost	Simple Logistic regression	Random forests	Decision trees	Bagging	Random Tree	KNN	Neural Network	Naïve Bayes
Accuracy	0.995	0.993	0.993	0.992	0.99	0.988	0.983	0.947	0.878
FNR	0	0	0.003	0.003	0.007	0.01	0.017	0.057	0.217
FPR	0.01	0.013	0.01	0.013	0.013	0.013	0.017	0.05	0.027

In Table 3, False Negative Rate is reduced significantly while using balanced dataset. The FNR of Naïve Bayes is 0.217. KNN and Random Tree obtain about 98% accuracy and, Bagging, Decision Trees, Random Forests, Simple Logistic regression and AdaBoost get 99% accuracy. Among them, AdaBoost achieves highest accuracy 99.5%, FNR is 0 and FPR is 0.01. So AdaBoost algorithm that obtains the highest accuracy, lowest FNR and FPR outperforms the other algorithms in both unbalanced and balanced dataset.

5. Conclusion

Machine learning algorithms have been used for network traffic classification to classify applications like WWW, DNS and Telnet. Recently they are applied for encrypted network traffic classification and malware detection. In our work, five supervised machine learning methods and four ensemble methods are trained for building remote access Trojans detection model. One of the ensemble learning methods, AdaBoost, is the best model with high accuracy, least FNR and least FPR among these algorithms. Future work is to increase the number of RAT samples and normal applications in order to achieve a comparable detection model for malware in production environments with best accuracy, FNR, FPR and no overhead.

6. References

[1] S.Li,X.Yun,Y.Zhang, J.Xiao,Y.Wang. A General Framework of Trojan Communication Detection Based on Network Traces. Proc. of 2012 IEEE Seventh International Conference on Networking, Architecture, and Storage. China, 2012, pp. 49 - 58.

[2] L.Yu,P.Guojun,Z.Huanguo, W.Ying. An Unknown Trojan Detection Method Based on Software Network Behavior. Wuhan University Journal of Natural Sciences. 2013, Vol. 18, No. 5, pp. 369-376.

[3] D.Jiang, K.Omote, An Approach to Detect Remote Access Trojan in the Early Stage of Communication. Proc. of IEEE 29th International Conference on Advanced Information Networking and Applications. South Korea, 2015, pp. 706-713.

[4] M.Shafiq, X.Yu, A.Ali Laghari, L.Yao, N.Kumar Karn, F.Abdessamia. Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms. Proc. of 2nd IEEE International Conference on Computer and Communications. China, 2016, pp.2451-2455.

[5] B.Anderson, D.McGrew. Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and Non-Stationarity. Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery

and Data Mining. Canada, 2017, pp.1723-1732.

[6] S.E.Gómez, B.C.Martínez, A.J. Sánchez-Esguevillas, L.Hernández-Callejo. Computer Networks. 2017, vol-127. pp.68-80.

[7] M.Joshi, T.Aldhayni. A Review of Network Traffic Analysis and Prediction Techniques. ARXIV. July, 2015.

[8] P.Prasse, L.Machlica, T.Pevny, J.Havelka, T.Scheffer. Malware Detection by Analyzing Network Traffic with Neural Networks. Proc. of IEEE Security and Privacy Workshops. USA, 2017, pp.205-210.

[9] http://www.trusteer.com/glossary/remote-access-trojan-rat

[10] Wireshark. (Online) Available: https://www.wireshark.org.

[11] K.S.Yin, M.A.Khine. Network Behavioral Features for Detecting Remote Access Trojans in the Early Stage. Proc. of International Conference on Network, Communication and Computing. China, 2017. pp.92-96.

[12] P. Harrington. Machine Learning in Action, Logistic regression. New York, 2012, pp.83-100.

[13] T. M. Mitchell. Machine Learning, Decision Trees Learning. 1997, pp.52-76.

[14] O. Villacampa. Feature Selection and Classification Methods for Decision Making, A Comparative Analysis.2015.

[15] L. Breiman. Random Forests. Statistics Department, University of California, 2001, pp.1-32.

[16] T. M. Mitchell. Machine Learning, K-nearest Neighbor Learning;1997, pp.231-236.

[17] I.H.Witten, E.Frank, M.A. Hall. Data Mining Practical Machine Learning Tools and Techniques,3rd ed. USA, Elsevier, 2011, ch.8, pp.358- 362.

[18] I.H.Witten, E.Frank, M.A. Hall. Data Mining Practical Machine Learning Tools and Techniques,3rd ed. USA, Elsevier, 2011, ch.8, pp.352-355.

[19] T. M. Mitchell. Machine Learning, Artificial Neural Network; 1997, pp. 81-94.

[20] T. M. Mitchell. Machine Learning, Bayesian Learning; 1997, pp.154-178.

[21] J.Ren, S.D.Lee,X.Chen,B.Kao,R.Cheng,D.Cheung. Naïve Bayes Classification of Uncertain Data. In Proceedings of the IEEE International Conference on Data Mining series (ICDM 2009). USA,2009, pp.944–949.