Effective Features for Detection of Remote Access Trojans

Khin Swe Yin, May Aye Khine
University of Computer Studies, Yangon
khinsweyin2009@gmail.com,maya.khine@gmail.com

Abstract

As companies in every industry sector around the globe have lost, stolen or leaked their sensitive data to the outside world every year, the security of confidential information is increasingly important. Remote Access Trojans (RATs) are used to invade a victim's PC through targeted attacks. In the previous works features for detection of RATs were selected by the author who may be an expert in the related domain, and any feature selection method was not used. In this paper one of the feature selection methods, Information Gain is applied for evaluating and ranking features. It aims not only to reduce costs and resources for building detection system of remote Access Trojan but also to add the advantages of using feature selection method and propose new features while maintaining high accuracy. Our approach achieves 99% accuracy together with the FNR of 0.091 by Decision Trees algorithm, and this experimental result shows that our proposed features are effective to detection system of RATs.

1. Introduction

Targeted attacks are one of the biggest cyber-threats to an organization in today's Internet-connected landscape. They make the targeted company lose reputation, or cost millions in damages. Targeted attacks use Remote Access Trojans as their weapon. Traditional signature-based anti-virus scanners work by looking at the files on the computer as sequences of bytes. Malware is detected by searching for specific byte patterns that are known to occur in a given piece of malware, and if this sequence is detected, the file can be flagged as being malicious and infected with the specific piece of malware that owns that byte signature[1][2].

Although this technique has some advantages like Fast algorithms for matching and Low false positive rate, it is easily fooled by small modifications and/or obfuscation techniques. Features extracted from traffic flows are widely used to classify malicious traffic of RATs and normal application traffic in Network behavior analysis.

Features may contain irrelevant and redundant features slowing down the training and testing processes or even affect the classification performance with more mathematical complexity. However, in practice, it is worthwhile to keep the number of features as small as possible in order to reduce the cost and the complexity of building a classifier. In addition, eliminating unimportant features facilitates data visualization, improves modeling, prediction performance, and speeds up classification process. Thus, dimensionality reduction, such as feature extraction and feature selection, has been successfully applied to machine learning and data mining to solve this problem. Feature extraction techniques attempt to transfer the input features into a new feature set, while Feature Selection algorithms search for the most informative features from the original input data [3].

In this paper, we focus on feature selection and propose a scheme that selects features based on Information Gain for feature ranking. The best set of candidate features is chosen, in a filter manner. Then machine learning algorithms are used for classification. Most appropriate features can be defined by looking for the best subset that produces the highest classification accuracy.

This paper is organized as follows: Section 2 outlines the related work of this study. Section 3 presents Preliminary. Section 4 describes detailed description of our proposed method. Section 5 presents the experimental details and results. Finally, this paper is concluded in Section 6.

2. Related Work

There are many different approaches for detection systems of RATs, and the two main categories are host-based and network-based. The effectiveness of the host-based approaches is limited by its complexity and huge overhead. The network-based detection

systems generally apply the passive monitoring to machines without overhead at the end hosts. The network-based approach can be divided into two main branches- signature-based and behavior- based. It is behavior-based detection techniques that an intrusion can be detected by observing a deviation from normal behavior of the system. Advantages of behavior-based approaches are that they can detect unforeseen vulnerabilities, and discover new attacks. They are less dependent on operating system [18]. So features are extracted from network traffic to express network behavior and applied in modeling detection systems. Detection of remote access Trojans is discovering malicious traffic behaviors that are different from normal traffic.

As RATs hide their traces to be invisible and stay stealthily in the victims, their malicious traffic is less than normal traffic that does not need to be camouflaged. So the inbound traffic is larger than outbound traffic in normal applications, and this situation is reversed in the command and control traffic of RATs. The amount of packets is different between RATs and normal applications within a limit of time. Features like packet number, data size, and the duration of the session are commonly used in detection systems.

The derived features proposed by [4] include out-in-pkts, out-in-bytes, duration-versus, mean interval. The detection system of [4] could attain an accuracy of around 90%. This system extracts network features from a session that begins from a SYN packet in the TCP three-way handshake and ends with a FIN/RST packet. It takes time, and confidential information may already leaked before detection.

From the perspective of network level, five characteristics were chosen to describe applications' network behavior[5]: (1) ratio of sent and received traffic size, (2) number of connections, (3) proportion of upload connection, (4) proportion of concurrent connection, and (5) number of distinct IP. As it uses hybridize host-based and network based techniques, it cannot avoid the limitation of host-based approach.

The behavior features were mainly distributed in the network layer, transport layer and application layer [6]. As it takes to consider many sub-connections during primary connection, standard deviation of packet interval, communications time, "heartbeat" packet to keep-alive, upload and download traffic, packet entropy and specific port in communication, features are extracted from the start of connection to the end of the whole process. So disclosure of confidential information may be occurred before detection.

Although features are extracted in the early stage of communication and it achieves over 96% accuracy, the concept of feature selection just depends on the different behavior of RATs and normal applications [7]. Any feature selection method is not applied, and the selected features cannot guarantee that they are the best ones for detecting malicious behaviors of RATs which are not included in this work.

The motivation for applying feature selection method is two-fold. First, there still remain some network traffic features which are appropriate for detection system. Second, there are many feature selection methods which can be applied in different fields to propose features. The first factor is important because there emerges many different types of RATs that uses TCP protocol nowadays, and selecting appropriate features can provide the best accuracy for detection system. Minimizing the second overhead is especially important in this study, because detection and classification are to be performed online and in real-time on production hosts.

3. Preliminary

3.1 Remote Access Trojans

Remote Access Trojan, also known as Trojans, are malware disguised as legitimate software, used to trick users into unknowingly installing malware. A RAT generally consists of two sides: A client and a server. The server-side is executed on the victim, and the client-side is executed on the server to control the victim. The attacker can control the victim's PCs remotely and secretly in order to steal confidential information, erase or overwrite data, listen the key logger or capture the system screen. They can hide themselves in process space of legitimate program and hence never appear in task manager or system monitors. RATs uses reverse connections to connect remote system and are more likely to remain undetected. So network behavioral features are extracted from the network traffic in order to detect the malicious command and control traffic of RATs. Figure 1 shows the basic functionality of RAT.

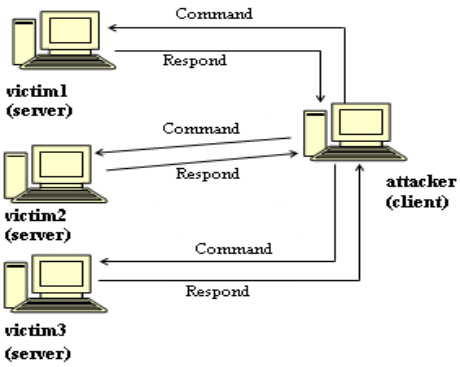

Figure 1: Basic functionality of a RAT

3.2 Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction [9].Feature selection does not only improve the quality of the model, but it also makes the process of modeling more efficient [10].The aim of feature selection is to identify the minimum number of features from the data source that are significant in building a model.

It can limit storage requirements and increase algorithm speed, save resources in the next round of data collection or during utilization, gain in predictive accuracy and gain knowledge about the process that generated the data or simply visualize the data [8].

Feature selection methods fall into three groups: Filter methods, Wrapper methods and Embedded methods [9]. Filter methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. Examples filter methods are Chi squared test, information gain and correlation coefficient scores. Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. Example of Wrapper method is the recursive feature elimination algorithm. Embedded methods learn which features best contribute to the accuracy of the model while the model is being created. The most common type of embedded feature selection methods are regularization methods. Examples of regularization algorithms are the LASSO, Elastic Net and Ridge Regression.

Among these methods we choose one of the filter methods- Information Gain for scoring features and choosing the best ones. Although Information Gain ignores the interaction with the classifier, it is independent of the classification algorithm and easily scale to very high-dimensional dataset and computationally simple and fast which is important to detect the communication of RATs early and speedily. The same features can be used in different learning algorithms for comparative analysis.

Information Gain is the expected reduction in entropy caused by partitioning the examples according to a given attribute. Information Gain, Gain(S,A) of an attribute A, relative to a collection of example S, is defined as [12]

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

3.3 Machine Learning Algorithms

Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. Machine learning algorithms are often categorized as being supervised or unsupervised. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets [13].The job of classification is to predict what class an instance of data fall into. We choose four supervised learning algorithms-Support Vector Machine (SVM), Decision Trees (DT), Naïve Bayes (NB) and Random Forest(RF) for classification.

3.3.1 Decision Trees

In decision trees, the process is broken down into individual tests which begin at the root node and traverse the tree, depending on the result of the test in that particular node. The tree begins at the root node. From the root node the tree branches or forks out to internal nodes. The decision to split is made by impurity measures. There are 3 methods to generate decision trees: ID3, C4.5, and CART (classification and regression tree). Decision tree J48 is the implementation of algorithm ID3 developed by the WEKA project team[16][17].

3.3.2 Support Vector Machine (SVM)

Support Vector Machines (SVM) has shown great promise in binary classification. The goal of the SVM algorithm is to map the training data into a multi-dimensional feature space and then find a hyper-plane

in said space that maximizes the distances between the two categories [17].

3.3.3 Naive Bayes

Naive Bayes is a widely used classification method based on Bayes theory. Based on class conditional density estimation and class prior probability, the posterior class probability of a test data point can be derived and the test data will be assigned to the class with the maximum posterior class probability[14].Calculating the conditional probability as follows[12]:

$$P(h/D) = \frac{P(D/h)P(h)}{P(D)}$$

3.3.4 Random forest

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The method combines Breiman's "bagging" idea and the random selection of features and it improves prediction accuracy [15].

4. Our Approach

Network traffic information is collected in the early stage, and selecting features is performed using information gain, and then the selected features are learned with four supervised machine learning algorithms. We divide the whole process into three main phases: Data Preprocessing, Feature Selection, and Classification.

4.1 Data Preprocessing

The data obtained from the experiment are first processed to generate the basic features. This phase contains normalization of data. Data normalization is a process of scaling the value of each feature into a well-proportioned range. Every attribute within each record is scaled to the same range of [0-1].

4.2 Feature Selection

Even though each connection record in the dataset has 12 features, not all of these features are needed to get high accuracy. Therefore, it is important to select the most informative features of traffic data to achieve higher performance. We apply Information Gain and feature ranking to find the most important subset of features.Table1 shows the initial feature set. The features with their gain values are specified in Table 2.

Table 1: The initial set of features

No	Feature	Description
1	Pacnum	Number of packets
2	Outbyte	Outbound data byte
3	Outpac	Outbound number of packets
4	Inbyte	Inbound data byte
5	Inpac	Inbound number of packets
6	OutPacByInPac	Outbound number of packets/ inbound number of packets
7	OutByteByOutPac	Outbound data byte/outbound number of packet
8	Duration	duration of the packets from the start of the communication to the given threshold
9	Bit/s(vict_att)	bit per seconds (from victim to attacker)
10	Bit/s(att_vict)	bit per seconds (from attacker to victim)
11	SentByReceiveTrafficSize	Sent traffic size / Receive traffic size
12	OutByteByInByte	Outbound data byte / Inbound data byte

Table 2: Initial features and their gain values

No	Feature	Gain value
1	Outbyte	0.369
2	OutPac	0.366
3	OutByteByInByte	0.366
4	OutByteByOutPac	0.366
5	Inbyte	0.366
6	Pacnum	0.366
7	Bit/s(att_vict)	0.31
8	OutPacByInPac	0.278
9	Bit/s(vict_att)	0.269
10	SentByReceiveTrafficSize	0.211
11	Duration	0.204
12	InPac	0.154

4.3 Classification

Once the optimal feature subset is selected for the class, this subset is then taken into the classifier training stage. The classifier distinguishes Normal data from non-Normal.

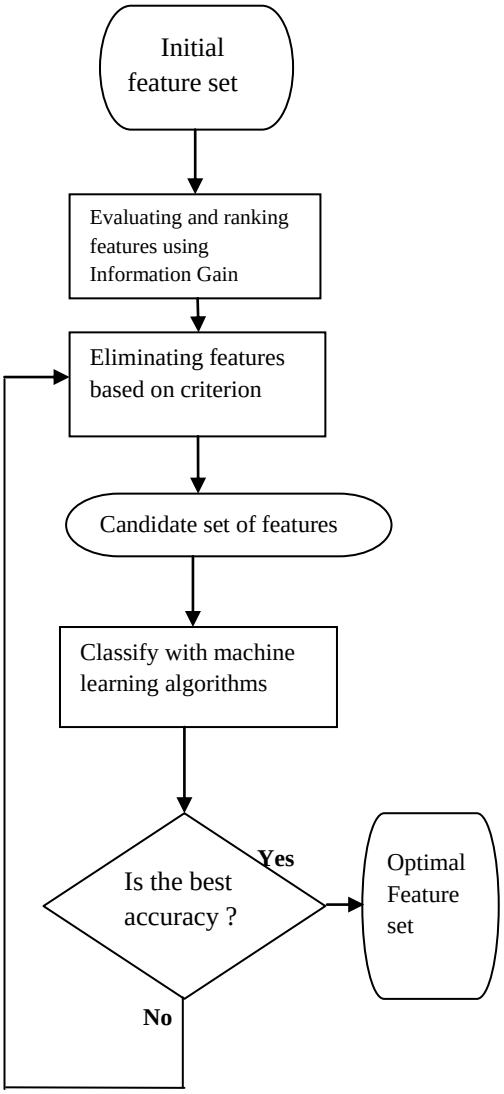

Figure 2: Proposed Feature Selection Scheme

Figure 2 shows the complete process of our approach. First, all available features are used as initial feature set. Next, the values of features are evaluated by information gain, and they are ranked. Some features are eliminated based on a criterion. Then candidate set of features are obtained and they are classified with Machine Learning algorithms. Finally, optimal feature set can be obtained after checking the accuracy they achieved. If the accuracy is not a satisfied level, a criterion can be changed. Moreover the best features can be defined depending on which features get how much accuracy and FNR.

5. Experiment and Result

A virtual environment including the attacker and the victim is set up. The attacker is a place where RAT is executed, and the victim is a place where the attacker's server.exe is executed. 5 types of Remote Access Trojans and 4 normal applications are used in the experiment. Wireshark - a network packet analyzer is run on the victim side to capture and collect all these network traffic.

As the communication characteristics of RATs in the early stage are just necessary for defining features, one sample is enough for each type of RATs. However a RAT is run many times in order to capture the various forms of behaviors that these RATs can change in the early stage. 266 sessions of RATs are collected for 5 types of RATs in our experiment. However in our experiment, the ratio 1:10 of RATs and normal sessions is used for classifying traffic. The reason is that there can be a lot of normal traffic and a smaller amount of malicious ones in the real network traces as RATs run secretly and they do not show much like normal applications. It is important to detect the secret RATs session among normal network traces. Moreover two sessions for one type of RAT is randomly chosen among the related RAT's sessions, and this way is done five times for five different RATs. Then 10 sessions for 5 RATs and 100 sessions for 4 normal applications are used for classification. When both normal and malicious RATs' traffic are changed many times and trained, the accuracy and FNR of SVM, NB and RF has a little difference. But DT is the best for most of the times. Detailed description of RATs with version number and normal application used in the experiment are shown in Table 3 and Table 4 respectively.

Table 3: Basic Information of RATs used in the experiment

RATs	Version
ImminentMonitor	4.1.0.0
KilerRat	10.0.0
NjRat	0.6.4
DarkComet	5.2-2F
Xtreme	3.8

Table 4: Normal applications used in the experiment

Normal applications	Description
Dropbox	Cloud service
Skype	Instant messenger
Facebook	Social media and service
Google	Internet-related services

Weka, datamining tool is used to load datasets, run algorithms and design and run experiments with results . Four supervised machine learning algorithms – SVM, DT, NB and RF algorithms are applied for classifying. k-fold Cross Validation is used to validate the result of classification. The data set is divided into k subsets, and the holdout method in which the data set is separated into two sets, called the training set and the testing set is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are formed a training set. We use 10 fold cross validation in the experiment. Accuracy and False Negative Rate (FNR) are used for performance measure. Accuracy gives the correctly classified number of both normal and malicious instances on total instances. FNR shows that 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.

$$Accuracy = \frac{Correctlyclassified\ number\ of\ bothinstances}{Totalnumberofinstances}$$

$$FNR = \frac{Incorrectlyclassifiednumberof\ RAT\ instances}{Totalnumberof\ RAT\ instances}$$

Firstly, 12 features – Pacnum, Outbyte, Outpac, Inbyte,Inpac,OutPacByInPac, OutByteByOutPac, Duration, Bit/s(vict_att), Bit/s(att_vict), SentByReceiveTrafficSize, OutByteByInByte, are evaluated using information gain, they are ranked. Next, they are trained in four machine learning algorithms. The accuracy of SVM, NB and DT are same, but there is a little difference in RF. Reducing features up to 6 can give RF the best accuracy- 0.991 and FNR - 0.091. Accuracy and FNR results are shown in Table 5 and Table 6 respectively.

Table 5: Result for accuracy

no of features	Accuracy			
	SVM	NB	DT	RF
12	0.991	0.991	0.991	0.982
11	0.991	0.991	0.991	0.982
9	0.991	0.991	0.991	0.982
7	0.991	0.991	0.991	0.982
6	0.991	0.991	0.991	0.991

Table 6: Result for FNR

no of features	FNR			
	SVM	NB	DT	RF
12	0.091	0.091	0.091	0.092
11	0.091	0.091	0.091	0.092
9	0.091	0.091	0.091	0.092
7	0.091	0.091	0.091	0.092
6	0.091	0.091	0.091	0.091

6. Conclusion

In this paper, the most appropriate features are presented for classifying malicious behaviors of RATs and normal network traffic. The proposed features have been evaluated in terms of classification accuracy and low computational cost. The best accuracy and FNR helps to reduce data leakage and increase the security of confidential information. Future work will be to increase the number of RAT samples and normal applications in order to achieve comparable classification accuracy and FNR for detection system of RATs in production environments with effective feature set and no overhead.

References

[1]. Bridgwater, What is Signature Based Detection?,A, 2012. http://blogs.avg.com/business/signature-based-detection/

[2]. Aycock, J, Computer Viruses and Malware.Advances in Information Security, Springer, 2006.

[3].S. Cang and H. Yu, Mutual information based input feature selection for classification problems, Decision Support Systems, 2012.

[4].S. Li, X. Yun, Y. Zhang, J. Xiao, and Y. Wang,A General Framework of Trojan Communication Detection Based on Network Traces , IEEE

Seventh International Conference on Networking, Architecture, and Storage, pp. 49-58, 2012.

[5]. Y. Liang, G. Peng, H. Zhang, and Y. Wang, An Unknown Trojan Detection Method Based on Software Network Behavior , Wuhan University Journal of Natural Sciences, Vol. 18, No. 5, Mar, 2013,pp.369-376.

[6]. W.Jinlong, G.HaidongandX.Yixin, Closed-loop Feedback Trojan Detection TechniqueBased on Hierarchical Model,JIMET, 2015.

[7]. D. Jiang, K.Omote,An Approach to Detect Remote Access Trojan in theEarly Stage of Communication, IEEE, 2015.

[8]. Isabelle Guyon1 and Andr´eElisseeff,An Introduction to Feature Extraction,

[9]. Jason Brownlee, an Introduction to Feature Selection,2014. http://machinelearningmastery.com/an-introduction-to-feature-selection/

[10]. Feature Selection(Data Mining), 2016. https://msdn.microsoft.com/en-us/library/ms175382.aspx

[11]. Jason Brownlee, An introduction to feature selection, 2014.

[12]. TomM.Mitchell,McGrawHill,MachineLearning, 1997.

[13]. Margaret Rouse, MachineLearning, February,2016. http://whatis.techtarget.com/definition/machine-learning.

[14]. JiangtaoRen, Sau Dan Lee, Xianlu Chen, Ben Kao, ReynoldChengand David Cheung, Naïve Bayes Classification of Uncertain Data,The 2009 edition of the IEEE International Conference on Data Mining series (ICDM 2009), IEEE Computer Society, Miami, FL, USA, 6–9 December 2009, pages 944–949.

[15]. PredragRadenković 3237/10, Faculty of Electrical Engineering, University Of Belgrade. http://www.docfoc.com/predrag-radenkovi-c-323710-faculty-of-electrical-engineering-university-of.

[16]. Quinlan, J. R. (1986). Induction of decision trees, Machine Learning, 1(1), pp. 81–106.

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

[18]. Herve Debar, IBM Zurich Research Laboratory, What is behavior based Intrusion Detection?, https://www.sans.org.