Feature-Selection-Based Ransomware Detection with Machine Learning of Data Analysis

Yu-Lun Wan[1]
Dept. Computer Science and Information Engineering
National Taipei University
New Taipei City, Taiwan
e-mail:
s710583117@gm.ntpu.edu.tw

Jen-Chun Chang[1]
Dept. Computer Science and Information Engineering
National Taipei University
New Taipei City, Taiwan
e-mail:
jcchang@mail.ntpu.edu.tw

Rong-Jaye Chen[2]
Dept. Computer Science
National Chiao Tung University
Hsinchu, Taiwan
e-mail:
rjchen@cs.nctu.edu.tw

Shiuh-Jeng Wang[*]
Dept. Information Management
Central Police University
Taoyuan, Taiwan
e-mail:
sjwang@mail.cpu.edu.tw
*whom correspondence

***Abstract*—Ransomwares are continuously produced in underground markets such that increasingly high-level and sophisticated ransomwares are spreading all over the world, significantly affecting individuals, businesses, governments, and countries. To prevent large-scale attacks, most companies buy intrusion detection systems to alert regarding any abnormal network behavior. However, they cannot be detected using conventional signature-based detection even though ransomwares belong to the same family. In this study, a method is provided to develop a network intrusion detection model that is based on big data technology. The system uses Argus for packet preprocessing, merging, and labeling the known malicious data. A concept of Biflow was proposed to replace the packet data. Further, we observe that the data size is reduced to 1000:1. Additionally, the characteristics of a complete traffic are obtained. Six feature selection algorithms were combined to achieve a better accuracy in terms of classification. Finally, the decision tree model of the supervised machine learning was used to enhance the performance of intrusion detection system.**

Keywords-component; ransomware; feature selection; intrusion detection system; data analysis

I. INTRODUCTION

Some shady services provide ransomware [1] as a service (RaaS). Anyone can develop a malware that includes a custom form, distribution area, way, and time zone in simple steps. This service also provides a monitoring platform for the buyer such as the distribution of malicious programs in a region, number of hosts infected, and payment of ransoms. Ransomware is a dangerous malware that can be combined with cryptosystems in which the security is evaluated based on the difficulty in solving a problem. Therefore, RaaS may transform into a serious threat since it can be easily obtained, spread, and can cause effective infection.

To solve this problem, many solutions have been developed, including infected path, network, endpoint, and intrusion detection system (IDS). Kharraz et al. [2] reported a large-scale dynamic analysis system called UNVEIL, which presented two techniques to detect ransomware: file and screen locker. File locker monitors the activity of a filesystem for every I/O request and computes its data buffer entropy. File locker further enables the comparison of each I/O data buffer before and after a particular activity. Screen locker detects the alterations in the desktop using a structural similarity index (SSIM) algorithm. Ransomware always alters the desktop to display ransom notes; therefore, screen locker can detect the ransomware immediately.

Hasan et al. [3] investigated the feature selection system of IDS that is deployed in the network environment along with the mirror traffic. To deal with a large amount of data containing various irrelevant and redundant features, feature selection plays a critical role in the preprocessing step to improve the performance of IDS. Hasan et al. further developed a random forest model for IDS based on DARPA 98 dataset by reducing the input features and processing time and by improving the false positive rate.

Ahmad et al. [4] detected the header information of the data link and network layers and deployed multipoint detection in a client–server model network. For inbound traffic, these models can be used to identify whether the incoming flows are worms or to block malicious host attacks in real time. However, several malicious servers have switched to a new domain randomly generated using a domain generation algorithm (DGA). Further, they cannot rely on the IP address or malicious domain block list to achieve complete network-level protection. Ahmad et al. combined this with the network layer 2 media access control (MAC) address, which is a unique identifier assigned to network interfaces in order to block an abnormal host. This method identified worms, such as RDP worms, and prevented ShellShock malicious attacks effectively. If the attacker deliberately avoided the feature of data link layer header information, then the cross-layer header data identification failure that is caused by this method cannot be successfully defended.

Additionally, many experts and scholars share malicious traffic analysis reports on websites or blogs such as Malicious Traffic Analysis [5] and Contagio [6]. Most of the malicious activity records enable us to understand the malicious behavior pattern. Ransomwares continue to evolve, and the available methods are not observed to be good

978-1-5386-6350-9/18/$31.00 ©2018 IEEE

enough to deal with unknown malwares using a signature-based database.

In this study, we propose to use a flow-based Biflow to replace the packet-based data. Argus was used to assemble these data on the basis of open malicious traffic datasets into binary data representing the network flows. Combining feature selection algorithms reduces the input features, thus improving the accuracy rate of the ransomware family classification. The rest of this study is organized as follows: section 2 provides the background and knowledge required for this study. Section 3 describes the experimental system architecture and procedures. Section 4 describes the evaluation of feature selection using association analysis algorithms to observe the effect on prediction accuracy. Section 5 provides the conclusions of this study.

II. Related Work

A. Argus

Argus [7] is an open-source network packet analysis tool developed by Carter Bullardn that compresses the overall data size and assembles these packets into a binary Argus packet record format. Argus consists of many scripts distributed within the "Argus-clients" package and provides various functions including checking the security status of the network according to the record file, analyzing the network usage, and network identification and visualization. Argus also supports the client–server centralized mode network and dynamic monitoring network traffic in real time. Further, it provides the basic data and data visualization chart in order to enable the users to find the problem quickly using an undefined protocol.

B. Ransomware

As the name suggests, ransomware receives payment using bitcoin, which is a digital currency platform. The initial ransomware used a symmetric-key algorithm that used the same key to perform both encryption and decryption. After completing the encryption, the key is sent to the command and control (C&C) server. It was possible to steal the key used for decryption while being transmitted across various networks. Currently, ransomwares have evolved into various sophisticated variants because they are observed to earn a large profit. The attacker eventually upgrades the ransomware to depict asymmetric encryption. The ransomware creates a random symmetric key and encodes critical data using this key. Further, the C&C server sends a public key using the Rivest, Shamir, and Adleman (RSA) algorithm to the infected target data; the public key is used to securely encrypt the symmetric key. As mentioned previously, ransomwares depict complex encryption algorithms. Therefore, you cannot reply to the encrypted file unless you obtain the decryption key.

C. Decision Tree

Decision tree is a supervised machine-learning (ML) algorithm that is commonly used in regression analysis and classification. Compared using various machines, learning models exhibit a high degree of efficiency and speed of execution. They are characterized by segmenting multiple decision points from a known set of fields and by constructing a greedy strategy from the roots to the end of the leaves to clearly understand the decision-making process.

D. Intrusion detection system

IDS is an important part of network security prevention. Most of such prevention systems are manufactured by mirroring the traffic of the network equipment or by copying the terminal equipment to the IDS analysis device or a computer equipped with the IDS software. Intrusion detection uses an anomaly feature model to identify abnormal behavior and attacks and to record and analyze the incoming traffic, generate log files, and alert immediately.

III. Our Scheme

The abnormal samples that were used to perform this study were obtained from PCAP and malware-traffic-analyses websites that provide open ransomware packet data, and the general packet is a real network packet that is collected by a web server. Further, HTTP packet samples are downloaded from the Wireshark website.

Fig. 1 depicts the research framework of this chapter. First, the collected data were randomly divided into three groups: Locky, Cerber, and normal traffic in a ratio of 1:1:2 (abnormal Locky: abnormal Cerber: general). The network data acquired in the form of PCAP files require preprocessing. Further, they were labeled.

Related research studies depicted that the classification accuracy can be increased by decreasing the number of input features. We mainly focused on the mixing of six feature selections to perform the data preprocessing.

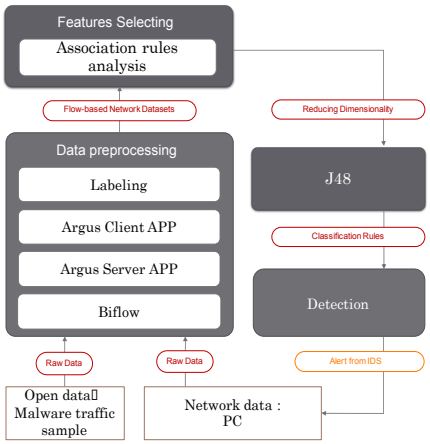

Figure 1. Flow-oriented Argus feature detection model.

A. Network packet data pre-precessing

A network packet is the smallest formatted unit of data in a physical network device. Most network sniffers obtain the location of a particular network problem by recording each packet of traffic stream. The benefits can be clearly viewed by establishing a connection. TCP uses a three-way handshake, i.e., three TCP packets are present. To perform this part, the session was divided into connection setup, data

transmission, the end of the connection, and others and were further combined into a Biflow. The data were added to the Argus server in the Argus format and were sent to Argus client with Ra output in an ASCII format.

B. Feature Preselections

First, 36 characteristics were selected. We intended to break through the standard pattern of traditional IP and DNS identification so that more common feature fields, such as the source IP address of the network layer, the destination IP address, and data link layer, in which the network card is able to uniquely identify the source and the destination MAC address could be ignored.

1) Optimization of the feature selection

The experiment used a total of 36 features, including Dir, and the remaining fields are strings, hexadecimal, additional treatment numeric, and nominal information.

After preprocessing, the algorithm was selected according to the six characteristic correlations: gain ratio, information gain, correlation ranking, OneR feature, ReliefF ranking, and symmetrical. Since different algorithms produce different results, all the values were limited to fall between 0 and 1. Detailed description is provided in section 4.1.

2) Data labeling and Fusion

The tagged data were trained after labeling Locky, Cerber, and normal traffic. The data in the result field are the "nominal" type, which was randomly captured in a ratio of 1:1:2 (abnormal Locky: abnormal Cerber: Normal) and was further merged into data sets for ML.

This section observed whether the "three different types of sample size" were balanced. If there is a large difference between the number of samples and others, there will be a high accuracy; however, it will also result in a high false positive.

C. Decision Tree Classification

Finally, the decision tree was classified using a J48 decision tree. Decision trees are common supervised ML models to perform classification and prediction. The tree structure presents the key parameters and conditions that eventually result in a classification path. Followed by the classification results set as IDS rules, this can be automated to cause an anomaly warning and rapid classification of the ransomware family. This is described in detail in section 4.

IV. Discussion and Analysis

A. Feature Selection Optimization

Figure 3 depicts the selection of algorithms through six feature correlations: gain ratio, information gain, correlation ranking, OneR feature, ReliefF Ranking, and symmetrical uncertainty to achieve a correlation score that is limited to fall between 0 and 1.

As depicted in Figure 3, IdleTime is the most relevant in this category. There is a correlation of more than 0.5 in all the six feature selection algorithms. Therefore, the IdleTime field becomes the first decision point during the classification, which is also known as the decision tree root. Instead, we tried to eliminate the features with a low relevance, train them with ML, and verify the accuracy of their rules.

After combining the six feature selection algorithms, the score of each feature was obtained, and a bar chart was sketched in Figure 2. In this study, five categories exist at intervals of 0.05. The original result is class 0, whereas the continually deleted less relevant features are classes 1 to 4.

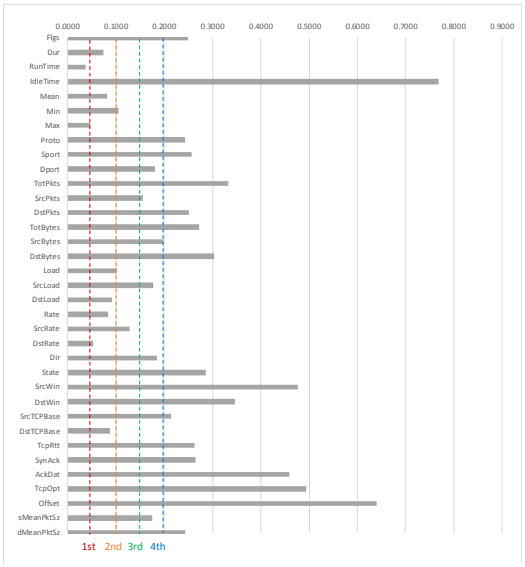

Figure 2. Four-phase progressive feature reduction.

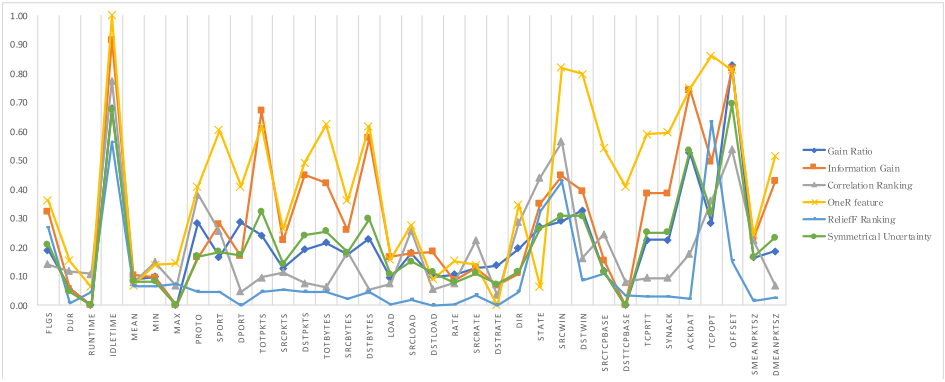

Figure 3. Scores of six feature selection algorithms.

B. *Effect of Feature Optimization on the Accuracy of Impact*

A true positive indicates that the actual category matches the predicted category. If the actual category label is Cerber and the predicted result is Cerber, then it can be classified as TP. A false positive indicates that the actual classification of the category does not match with the predicted result; i.e., an inaccurate classification into the second category. We further assume that the actual label for the Cerber category can be classified as Locky or N, i.e., belonging to FP and FN. The confusion matrix that was predicted by the ransomware in this experiment is depicted in Table I.

TABLE I. CONFUSION MATRIX FOR RASTER SOFTWARE PREDICTION

Classification		Predicted Class		
		Cerber	Locky	N（Normal）
True Class	Cerber	TP	FP	FN
	Locky	FP	TP	FN
	N	FP	FP	TN

This section focuses on the accuracy of the classification rule. The precision rate can further be defined as follows:

$$Precision\ Rate = \frac{TP}{TP+FP} \quad (1)$$

TABLE II. TRAINING RESULTS AFTER DIMENSIONAL REDUCTION

class	Attributes	NoL	SoT	Precision
4（0-0.2 ）	19	16	25	90.1919
3（0-0.15）	25	19	31	90.6183
2（0-0.1 ）	28	19	31	89.339
1（0-0.05）	32	19	31	89.1258
0	36	19	31	89.1258

The training process is depicted in Table II. Class 0 is the first choice from among all the features. The results of J48 were expressed as NoL (Number of Leaves), SoT (Size of Tree), and precision after performing cross-validation. Each class was rated to be 0.05 for a level, while the less relevant features were eliminated. By summing up all the classes, the accuracy is observed to be the true positive rate of the total classification result. If this classification method would have been used to obtain a higher accuracy, it may indicate that the feature selection algorithm is more sensitive so as to perform accurate classification. It can further achieve a higher accuracy for detection in an unknown network traffic.

Classes 1 to 4 depicted that not only has the accuracy been improved but the size of the decision tree has also been significantly reduced. This indicates that eliminating less relevant features can reduce the complexity of decisions and enhance the classification accuracy and accelerate IDS for identification in an unknown traffic.

V. CONCLUSION

The trend of using ransomware is not new. However, its distribution is observed to develop continuously. Further, customized ransomware services can be obtained by paying a small amount. Although the development of ransomware is very difficult, anyone can launch the corresponding malicious code. In this study, we designed a system containing an Argus server and client applications and proposed a flow-oriented method as Biflow to create a ransomware detection module which based on open malicious network traffic datasets. These datasets were reducing the data size to 1000:1. This study focuses on whether the feature selection affected the classification precision rate by combining six feature selection algorithms to analyze the preselected column and its relevance for classification. The input features were observed to gradually reduce. Not only does the accuracy depict an increase, but the number of nodes in the decision tree also decrease. Finally, lesser number of rules improve the performance of intrusion detection, decrease the time required, and result in an early detection of the abnormal traffic.

ACKNOWLEDGMENT

This research was partially supported by the Ministry of Science and Technology of the Republic of China under the Grant MOST 106-2221-E-305-005- and MOST 106-2221-E-015-001-.

REFERENCES

[1] TrendMicro. Ransomware: 10 Years of Bullying, Fear-mongering and Extortion.[Online]. Available: http://www.trendmicro.com.sg/, accessed Oct. 2017.

[2] A. Kharraz, S. Arshad, C. Mulliner, W. Robertson, and E. Kirda, “UNVEIL: A Large-Scale, Automated Approach to Detecting Ransomware,” In 25th USENIX Security Symposium, 2016.

[3] M. A. M. Hasan, M. Nasser, S. Ahmad and K. I. Molla, “Feature Selection for Intrusion Detection Using Random Forest,” Journal of Information Security, vol. 7, no. 3, pp. 129–140, 2016.

[4] M. A. Ahmad, S. Woodhead, and D. Gan, “Early Containment of Fast Network Worm Malware,” In Information and Computer Science National Foundation for Science and Technology Development Conference on IEEE, 2016.

[5] Malware-traffic-analysis. A Source for Pcap Files and Malware Samples.[Online]. Available: http://www.malware-traffic-analysis.net/, accessed Oct. 2017.

[6] Contagio. Malware Analysis and Malware Samples.[Online]. Available: http://contagiodump.blogspot.tw/, accessed Sep. 2017.

[7] QoSient. Argus: Network Audit Record Generation and Utilization System.[Online]. Available: http://www.qosient.com/Argus/, accessed Sep. 2017.