Journal of Robotics, Networking and Artificial Life
Vol. 7(2); September (2020), pp. 103–106
DOI: https://doi.org/10.2991/jrnal.k.200528.006; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

IDS Malicious Flow Classification

I-Hsien Liu[1], Cheng-Hsiang Lo[1], Ta-Che Liu[1], Jung-Shian Li[1,*], Chuan-Gang Liu[2], Chu-Fen Li[3]

[1]*Department of Electrical Engineering/Institute of Computer and Communication Engineering, National Cheng Kung University, Tainan City 70101, Taiwan*
[2]*Department of Applied Informatics and Multimedia, Chia-Nan University of Pharmacy and Science, Tainan City 71710, Taiwan*
[3]*Department of Finance, National Formosa University, Yunlin County 632, Taiwan*

ARTICLE INFO

Article History
Received 22 October 2019
Accepted 24 April 2020

Keywords
NIDS
dynamic analysis
deep learning

ABSTRACT

We will display two different kinds of experiments, which are Network-based Intrusion Detection System (NIDS)-based and dynamic-based analysis shows how artificial intelligence helps us detecting and classify malware. On the NID, we use CICIDS2017 as a research dataset, embedding high dimensional features and find out redundant features in the raw dataset by Random Forest algorithm, reach 99.93% accuracy and 0.3% of the false alert rate. We extract the function calls in malware data by the method proposed in this paper to generate text data. The algorithm *n*-gram and Term Frequency-Inverse Document Frequency (TF-IDF) are used to process text data, converts them into numeric features, and by another feature selection methods, we reduce the training time, achieve 87.08% accuracy, and save 87.97% training time in dynamic-based analysis.

1. INTRODUCTION

In recent years, with the rapid development of Internet technology, people can find the information easily on the Internet or share information with others. Cloud technology, 5G and other emerging network technologies have also made data exchange faster and more convenient. At the same time, cyber security has become one of the biggest concerns nowadays, traditional equipment has less ability to deal with diverse and complex attack techniques. How to promote equipment and software, and improve the confidentiality, integrity and accessible of users has become an important issue. Fortunately, today's hardware is better than pass, huge amount of data can be stored and analyzed, especially in internet, leads the application of artificial intelligence technology to cyber security.

2. BACKGROUND AND RELATED WORK

2.1. Static Analysis vs. Dynamic Analysis

There are two main methods when we want to analysis the malware comes into our local devices like computers or servers, static analysis and dynamic analysis. Static analysis is a white box analysis method. As the name implies, when you analyze, it does not execute the executable file of the malicious program. Instead, it directly analyzes the internal process operation or data usage based on the binary executable file or the original code, because of its analysis way, the advantage of static analysis is that it has low infection opportunity. However, static analysis often requires reverse engineering to disassemble of the executable file. Even some malicious programs that have been protected by a shell must be unpacked by a specific tool before reverse engineering. The features commonly used in static analysis methods are the following, such as Operation Codes [1,2] and byte sequences, or extracting useful features from portable executable files. On the other hand, dynamic analysis is a black box analysis method, which means that it is necessary to start a malicious program during execution, it will be executed in a virtual environment, and record the behavior of the malicious program such as access file writing and deleting, network connection, Mutexes, Registry Keys modification and Application Programming Interface (API) function calls, etc.

2.2. NIDS vs. HIDS

Another way to detect the malware is building the Intrusion Detection System (IDS), which can be divided into two different kinds by their main function: Network-based IDS (NIDS) and Host-based IDS (HIDS). NIDS mainly detects the attack by network flow, whereas HIDS detects abnormal user behavior on local host computer. Both of them compare the log file to their database, the detection method can also separate into two ways: misuse-based and anomaly-based. Misuse-based, as known as signature-based, collect the signature of malware constantly first, and then build a malware signature database, if a network traffic flow or behavior matches the signature in malware database, it will be identified as abnormal. On the other hand, anomaly-based predefines the normal signature to detect the attack. In this study, we use CICIDS2017 [3] as experiment dataset, which collected the data from NIDS.

*Corresponding author. Email: jsli@mail.ncku.edu.tw

3. EXPERIMENT ARCHITECTURE

In this part, we will introduce the methods we used in this paper, for both dynamic-based analysis and NIDS-based malware classification. Our dynamic-based experiment architecture is shown in Figure 1, and NIDS-based is shown in Figure 2.

3.1. Data Pre-processing

Since the original malware data has some noise or untrainable features that makes model predict result worse, so we should take them off. The first challenge we face is that some data has no label, which makes supervised learning impossible. To deal with this problem, we proposes a method that uses a variety of anti-virus software as the basis and produces a final label by majority decision. Although the method is more complicated, it obtains a more credible label than a single anti-virus software.

The second challenge is that some non-numerical features such as strings or symbols are exist, which are untrainable features. We use different encode methods for different area to solve this problem. In dynamic-based analysis, the TF-IDF algorithm is used. It is a weighting technique that is often used in data mining and information retrieval as a statistical method. In contrast to Bag-of-Word, TF-IDF in addition to counting the frequency of occurrence of words in a single text, it is also used to assess how important a word is to each text, the length of a single text is also considered. The equation of TF-IDF are as Equations (1)–(3). The reason we use TF-IDF is that the content of raw data comes from the dynamic analysis include too many irrelevant information, and we decided to extract the API function calls of each malware. These API function calls are built on lots of words, so we take advantage of TF-IDF on processing texts, turn the API function call into numerical features.

$$tf_{t,d} = \frac{n_{t,d}}{\sum_{k=1}^{T} n_{k,d}} \tag{1}$$

$$idf_t = \log\left(\frac{D}{d_t}\right) \tag{2}$$

$$\text{score}_{t,d} = tf_{t,d} \times idf_t \tag{3}$$

On the other hand, in the NID, we use one-hot encode for low dimensional non-numerical features, and embedding for the high dimensional ones. One-hot encode can expand different m categories in the feature to m independent two-bit features, and mark the features they represent as 1 and the rest are 0. But one-hot encode will cause the dimension disaster for high-dimensional features, so we decide to use embedding, which can map high-dimensional features into low dimension properly, by optimizing the mapping matrix. For instance, the feature names "Source Port" has 52,554 different category attributes, we use embedding method to project them into two-dimensional space, as shown in Figure 3.

3.2. Feature Selection

Researchers are working hard to find a good methods to discard the redundant features in dataset, which has significant influence on model performance. The reason we use different feature selection methods for different area is that the features of dynamic-based malware analysis dataset are words, and most of NID dataset are numbers. If we use Random Forest for dynamic-based malware analysis dataset, it will take too much time to calculate the importance of features.

3.3. Application of Deep Learning Model

In the dynamic-based malware analysis, we use deep learning model such as Convolution Neural Network (CNN) [4], Multi-Layer Perceptron (MLP) [5] to classify the malware sample These model are well known in neural network study, so we do not give unnecessary details.

In the NIDS-based intrusion detection, we take model named Sequence-to-Sequence, which was proposed by Sutskever et al. [6]. The simple architecture is shown in Figure 4. The arrow between C and <EOS> is used as the boundary, left of arrow is the encoder,

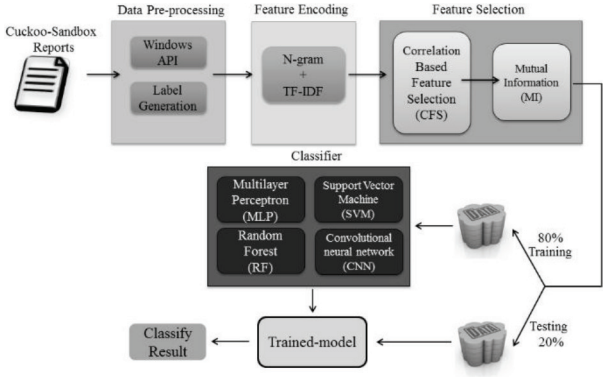

Figure 1 | Dynamic-based experiment architecture.

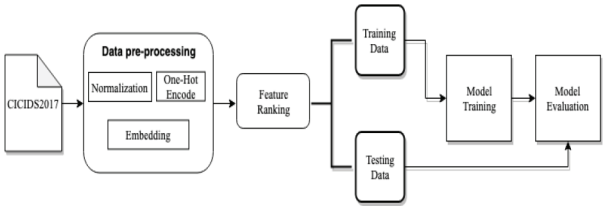

Figure 2 | NIDS-based experiment architecture.

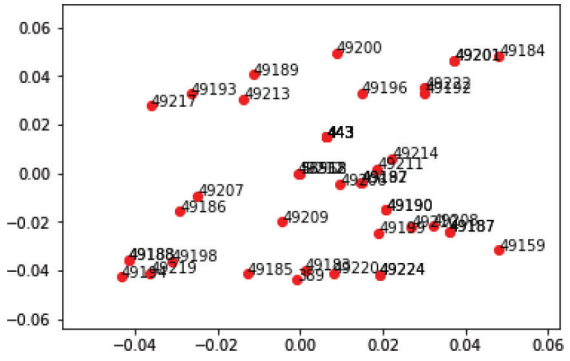

Figure 3 | The result of embedding the part of attributes in Source Port.

and the decoder for the right, both of which are composed of Long Short-Term Memory (LSTM). The main task of the encoder is to compress the input sequence vector into a content vector v (context vector) with a much smaller dimension. This content vector is also the hidden layer output of the encoder in the last layer, which represents the model's understanding of a sequence. On the decoder, there are two input sources. The first source is the content vector from the encoder, and the second source is the delay of sequence we tempt to predict. After receiving both, the decoder begins decoding and outputs the specified sequence.

4. RESULTS AND DISCUSSION

We used multiple measures to get a more persuasive result. In this part, we will introduce the evaluation metrics we used in this study.

- True positive (TP): malware sample that is correctly classified as malware.
- False positive (FP): benign sample that is incorrectly classified as malware.
- True negative (TN): benign sample that is correctly classified as benign.
- False negative (FN): malware sample that is incorrectly classified as benign.

The accuracy means the proportion of the total number of correct classifications:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{4}$$

In order to verify that the feature selection method proposed in this study does improve the model performance, we will compare four classification algorithms, both machine learning and deep learning: Random Forest (RF) [7], Support Vector Machine (SVM), MLP and CNN. To verify whether the high-dimensional feature embedding method and the Sequence-to-Sequence model has ability to improve the capability of intrusion detection, we convert high-dimensional non-numerical features of CICIDS2017 dataset: Source Port and Destination Port, into low-dimensional features by using the embedding method, and add them into training data to train the Sequence-to-Sequence model. As shown in Figure 5, we can obtain a better evaluation result based on embedding method and Sequence-to-Sequence model. The accuracy, prediction, recall and $F1$-score are 99.93%, 99.8%, 99.87% and 99.84%, respectively, which achieve an ideal result.

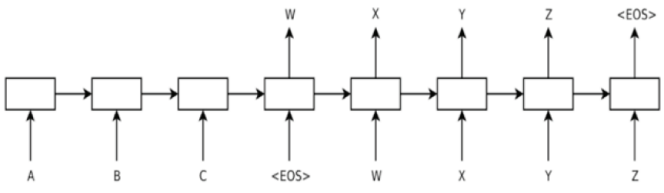

Figure 4 | Architecture of Sequence-to-Sequence [3].

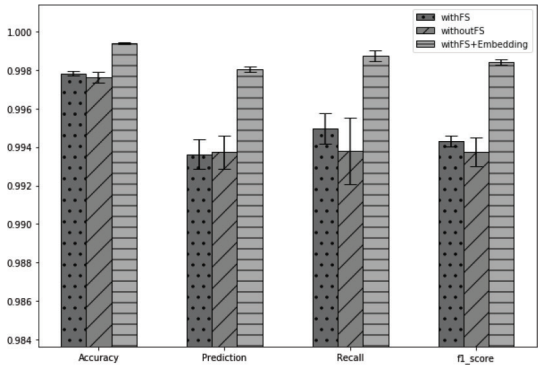

Figure 5 | Comparison of using Random Forest feature selection and embedding or not.

5. CONCLUSION

This study proposes a method of text processing as the main idea to extract, encode and adjust the weight of this feature of the Windows operating system application interface call, and then use the feature selection to drop redundant features step by step. In addition to the ability to reduce lots of features, the progressive feature selection method proposed in this paper can keep the information of the original features. We hope that these methods can be further used in real-time analyzes.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

ACKNOWLEDGMENTS

This work was supported by the MOST, Taiwan under contracts numbers MOST 108-2221-E-006-110-MY3 and MOST 108-2218-E-006-035-.

REFERENCES

[1] D. Bilar, Opcodes as predictor for malware, Int. J. Electron. Secur. Dig. Foren. 1 (2007), 156–168.

[2] I. Santos, F. Brezo, J. Nieves, Y.K. Penya, B. Sanz, C. Laorden, et al. Idea: opcode-sequence-based malware detection, International Symposium on Engineering Secure Software and Systems, Springer, Berlin, Heidelberg, 2010, pp. 35–43.

[3] I. Sharafaldin, A.H. Lashkari, A.A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), SciTePress – Science and Technology Publications, Funchal, Madeira, Portugal, 2018, pp. 108–116.

[4] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, International Conference on Neural Information Processing Systems (NIPS), Association for Computing Machinery, New York, NY, USA, 2012, pp. 84–90.

[5] F. Murtagh, Multilayer perceptrons for classification and regression, Neurocomputing 2 (1991), 183–197.
[6] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, USA, 2014, pp. 3104–3112.
[7] L. Breiman, Random Forests, 2001, Available from: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf (accessed February 10, 2019) [Online].

AUTHORS INTRODUCTION

Dr. I-Hsien Liu

He is a reacher fellow in the Taiwan Information Security Center @ National Cheng Kung University (TWISC@NCKU) and department of electrical engineering, National Cheng Kung University, Taiwan. He obtained his PhD in 2015 in Computer and Communication Engineering from the National Cheng Kung University. His research interests are Cloud security, Wireless Network, Group Communication and Reliable Transmission in Mobile ad hoc networks.

Mr. Cheng-Hsiang Lo

He was born in Miaoli, Taiwan, in 1994. He received the B.S. degree in Communications Engineering from Yuan Ze University (YZU), Taoyuan, Taiwan in 2016, and the M.S. degree in Computer and Communication Engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 2019.

Mr. Ta-Che Liu

He was born in New Taipei City, Taiwan, in 1995. He received the B.S. degree in Communications Engineering from Yuan Ze University (YZU), Taoyuan, Taiwan in 2017, and the M.S. degree in Computer and Communication Engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 2019. His research interest is technology.

Prof. Chuan-Gang Liu

He is an Associate Professor in the department of Applied informatics and Multimedia, Chia Nan University of Pharmacy and Science. He received the BSc degree from the Department of Electrical Engineering, Tam Kang University, in 2000. Then he graduated from the National Cheng Kung University with M.S. and PhD degrees in electrical engineering. His research interests are in the areas of Optical Networks Control, Wireless Networks, EPON, VANET, network security, cloud computing and TCP performance analysis.

Prof. Chu-Fen Li

She is an Associate Professor in the Department of Finance at the National Formosa University, Taiwan. She received her PhD in information management, finance and banking from the Europa-Universität Viadrina Frankfurt, Germany. Her current research interests include intelligence finance, e-commerce security, financial technology, IoT security management, as well as financial institutions and markets. Her papers have been published in several international refereed journals such as European Journal of Operational Research, Journal of System and Software, International Journal of Information and Management Sciences, Asia Journal of Management and Humanity Sciences, and others.

Prof. Jung-Shian Li

He is a full professor in the department of electrical engineering, National Cheng Kung University, Taiwan. He graduated from the National Taiwan University, Taiwan, with B.S. in 1990 and M.S. degrees in 1992 in electrical engineering. He obtained his PhD in 1999 in Computer Science from the Technical University of Berlin, Germany. He teaches communication courses and his research interests include wired and wireless network protocol design, network security, and network management. He is currently involved in funded research projects dealing with optical network, VANET, Cloud security and resource allocation, and IP QoS architectures. He is the deputy director general of National Center for High-performance Computing (NCHC), National Applied Research Laboratories. He serves on the editorial boards of the International Journal of Communication Systems.