Detection of Misuse Attack in NFV Networks Using Machine Learning

Ali Khalid Ali 1*, Wesam S. Bhaya 2

College of Information Technology, University of Babylon, Babylon, Iraq

Emails: kl659@yahoo.com , wesambhaya@uobabylon.edu.iq

Abstract. Network Function Virtualization (NFV) represents a virtual network whose service is provided by virtual parts of virtual machines. This type of network is easy to implement and update. In addition, NFV leading to low cost due to sharing the same resources. As is the case with other networks, NFV is not safe from attacks. Since all parts of this NFV network share the same resources, misuse attack is regarded to be the most common attack in NFVs, particularly because the attack use one or more of the resources which affect all parts of the NFV. This paper is based on using machine learning to extract rules of misuse attack detections. The tree decision C4.5 algorithm has been used to extract these rules, with nine features of network data flow. When testing the propose work with a server traffic data having more than 5 million network connections, the results show that a comparatively higher performance of the algorithm C4.5 with an accuracy of about 96%.

1. Introduction

Network Function Virtualization (NFV) can be defined as a type of network which virtualizes network services, such as routers and firewalls, in contrast to traditional networks where these parts are run on separate hardware. In NFV, these parts are packaged as virtual machines (VMs) on one hardware [1]. It can be stated that NFV is a relatively cheaper and more effective replacement of a classical network as it replaces hardware equipment with virtual software which all can be gathered in one or more physical server [2]. NFV increases the flexibility and scalability of network service providers in case of updating or adding new services, without requiring any additional hardware resources [3].

As compared to traditional network architectures, NFV has the following advantages:

i. Low cost
ii. Better efficiency of operating performance and operation
iii. Less need for configuration and resource allocation
iv. Flexible and dynamic network functions
v. Low energy consumption [4, 5].
Although this type of network only appeared during the last few years, it has been affected by different types of network attacks, especially the hypervisors and orchestrators software. One of these common attacks for this type of network is the misuse attack [6]. Misuse attack is considered a common attack to NFV networks, as all its parts share the same resources. Therefore, any attack targeting resources of one of its parts will affect the other parts of the network [7].

Machine Learning (ML) is a type of artificial intelligence (AI) application which has the capability of automatically learning and improving through experience, with no need for explicit programming. In other words, Machine Learning is an application which could access data and make use of it for self-learning [8]. Therefore, using this application with stored data traffic of network attacks can be very helpful to analyze this type of attack, as well as to extract the rules for discovering any anomaly behavior of NFV when undergoing, for example, a misuse attack, which represents the core of this research paper.

2. Related works
In [9], an approach has been proposed for identifying attack using machine learning methods, particularly the LSTM neural networks. The system uses the suggested approach onto a dataset that contains logs of interactions with an administrative interface of a login and security server. The results showed that the informed modeling can indeed capture normal behavior, in order to use it in detecting any abnormal behavior.

The study conducted in [10] captured a real time internet traffic dataset for 2 second using packet capturing engine, after which five machine learning algorithms (MLP, RBF, C4.5, Bayes Net and Naïve Bayes) were applied for network traffic classification. The results of this study showed that Bayes Net is an efficient machine learning technique for near real time traffic classification. An approach for detecting misuse attack has been proposed in [11], which depends on the building and maintenance of behavior profiles for the system users through tracking or monitoring user activity. The new activities of the network users will be put into comparison with the user's previous behavior for the detection of any potential misuse for the authorized user. The study also suggested 4 differing methods for detecting misuse in information retrieval.

Using machine learning in NFV network traffic is discussed in [12]. In this study, the implementation of autonomic management and supervised algorithms from Machine Learning has been proposed, shedding light on the analysis of the NFV network traffic during the performance of a benchmarking for the behavior of supervised machine learning algorithms for traffic classification.

The authors in [13] simulated a freeing attack using “Cloudsim” as simulation to implement resource freeing attack. The CPU's response time diagrams, as well as the obtained bandwidth have been measured. The results of the simulation indicate the significant changes in NFV network after attack within the virtualized environment. As for [14], they proposed a detection system that is based on a neural network for botnet in NFV. The system has been trained by available data sets collected in conventional networks. This study showed that their proposed detection system can detect up to 99% of the botnet attacks.

In [9], an approach has been proposed for identifying attack using machine learning methods, particularly the LSTM neural networks. The system uses the suggested approach onto a dataset that contains logs of interactions with an administrative interface of a login and security server. The results showed that the informed modeling can indeed capture normal behavior, in order to use it in detecting any abnormal behavior.
The study conducted in [10] captured a real-time internet traffic dataset for 2 seconds using packet capturing engine, after which five machine learning algorithms (MLP, RBF, C4.5, Bayes Net and Naïve Bayes) were applied for network traffic classification. The results of this study showed that Bayes Net is an efficient machine learning technique for near real-time traffic classification.

An approach for detecting misuse attack has been proposed in [11], which depends on the building and maintenance of behavior profiles for the system users through tracking or monitoring user activity. The new activities of the network users will be put into comparison with the user's previous behavior for the detection of any potential misuse for the authorized user. The study also suggested 4 differing methods for detecting misuse in information retrieval.

Using machine learning in NFV network traffic is discussed in [12]. In this study, the implementation of autonomic management and supervised algorithms from Machine Learning has been proposed, shedding light on the analysis of the NFV network traffic during the performance of a benchmarking for the behavior of supervised machine learning algorithms for traffic classification.

The authors in [13] simulated a freeing attack using “Cloudsim” as simulation to implement resource freeing attack. The CPU’s response time diagrams, as well as the obtained bandwidth have been measured. The results of the simulation indicate the significant changes in NFV network after attack within the virtualized environment. As for [14], they proposed a detection system that is based on a neural network for botnet in NFV. The system has been trained by available data sets collected in conventional networks. This study showed that their proposed detection system can detect up to 99% of the botnet attacks.

3. Algorithm and Dataset

The methodology of this research involves applying the C4.5 algorithm in machine learning onto stored data of misuse attack. The C4.5 algorithm is based on a decision tree, which will be constructed based on one or more factors. Although increasing these factors will increase the complexity of the problem, it usually leads to relatively more accurate results [15]. The dataset that is used in this research stored more than 5 million clients’ connections. Each of these connections contained 14 fields of information. About 90% of these connections were used as training data, whereas the remaining 10% was used as testing data [16-18].

Nine factors have been considered for the decision tree, these factors include:

1. Network connection protocol.
2. Minimum time connection.
3. Maximum time connection.
4. Number of zero second.
5. Number of zero size.
6. Minimum size of data.
7. Maximum size of data.
8. Number of error connection flag.
9. Number logging in connection.
These nine factors are the most effective factors than can help in differentiating between normal and misuse attack traffic. This dataset has 1,072,017 connection of misuse attacks.

4. Training Stage
As mentioned before, 90% of the data (around 4.5 million out of 5 million connections) is used as training data. The result of the training stage are shown in Table 1.

| Factors      | Number   |
|--------------|----------|
| Misuse       | 1,072,017|
| Protocol TCP | 1,072,017|
| Mini T       | 0 S      |
| Max T        | 2        |
| 0 Second     | 1,072,016|
| Mini Size    | 0        |
| Max Size     | 10,714   |
| 0 Size       | 1,072,016|
| Err Flag     | 867,446  |
| Login in     | 1        |

From Table 1, it can be noted that the connections with the misuse attack are 1,072,017, all of which used the TCP connection protocol. The time duration of all the connections, except for one, are less than one second, while the maximum time is 2 seconds. The number of connections with zero size also include all connection except one, where this single connection has a 10,714KB data connection size. The error flag was raised for 867,446 connections, and finally only one connection did login into the server.

The Entropy value is a value between zero and one:

\[ 0 \leq \text{Entropy} \leq 1 \]

Entropy is a good sign that indicates the effectivity of the factors that have been considered. The best value of entropy is zero, while the worst value is one. The value of entropy for this research is 0.000534, which is considered a very good and promising value, being so close to zero.

5. Data Testing and Results
As indicated earlier, 10% of the data size was used for testing, whereby the rules that have been extracted from the training stage are applied. Table 2 shows the values of true positive and false negative.
Table 2: True positive and false negative

| Factors        | Number  |
|----------------|---------|
| No. of Connections | 494020  |
| No. of Misuse     | 107201  |
| True Positive     | 86744   |
| True Negative     | 386558  |
| False Positive    | 262     |
| False Negative    | 20457   |

As shown in Table 2, the true positives and true negatives present higher results as compared to the false positives and false negatives. Each of the accuracy, precision, recall, specificity, and detection rate values are shown in Table 3.

Table 3: Accuracy, Precision, Recall, Specificity and Detection rate

| Factors    | Rate values       |
|------------|-------------------|
| Accuracy   | 0.95806048730722  |
| Precision  | 0.996988713422063 |
| Recall     | 0.809171556235483 |
| Specificity| 0.999322682384572 |
| Detection  | 0.809171556235483 |

From Table 3, it can be stated that there is a high percentage of accuracy (95.8%), and an even higher precision rate (99.69%). The rate values for each of the recall, specificity, and detection are (80.9%, 99.9% and 80.9%), respectively.

Using machine learning in stored data of server traffic produced results of high quality, and the C4.5 algorithm showed a high compatibility with this type of data, especially when a large number of factors has been considered. The nine factors that have been taken into consideration during the training data stage of this research lead to the high quality results that have been obtained in terms of true positive, true negative, and accuracy.

6. Conclusion
In this paper, the decision tree C4.5 algorithm of machine learning has been used to analyse a large data of server traffic for a NFV network, as well as to extract rules from this data. Nine features have been considered, and this number of features helped to produce high quality results. The true positive and true negative values have superb result as compared to the false positive and false negative. Furthermore, the results of accuracy, precision, recall, specificity and detection were highly promising as well.
References

[1] M. Pattaranantakul, R. He, Q. Song, Z. Zhang, and A. Meddahi, “NFV security survey: From use case driven threat analysis to state-of-the-art countermeasures,” IEEE Commun. Surv. Tutorials, vol. 20, no. 4, pp. 3330–3368, 2018.

[2] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, “Network function virtualization: State-of-the-art and research challenges,” IEEE Commun. Surv. tutorials, vol. 18, no. 1, pp. 236–262, 2015.

[3] Y. Park, et al., "Distributed Security Network Functions against Botnet Attacks in Software-defined Networks," in 2018 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), 2018, pp. 1-7.

[4] R. F. Moyano, et al., "A user-centric SDN management architecture for NFV-based residential networks," Computer Standards & Interfaces, vol. 54, pp. 279-292, 2017.

[5] G. Carella, et al., "Cloudified IP multimedia subsystem (IMS) for network function virtualization (NFV)-based architectures," in 2014 IEEE Symposium on Computers and Communications (ISCC), 2014, pp. 1-6.

[6] W. Yang and C. Fung, "A survey on security in network functions virtualization," in 2016 IEEE NetSoft Conference and Workshops (NetSoft), 2016, pp. 15-19.

[7] S. Lal, et al., "NFV: Security threats and best practices," IEEE Communications Magazine, vol. 55, pp. 211-217, 2017.

[8] E. Alpaydin, Introduction to machine learning: MIT press, 2020.

[9] L. Adilova, et al., "System Misuse Detection via Informed Behavior Clustering and Modeling," in 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), 2019, pp. 15-23.

[10] K. Singh, S. Agrawal, and B. S. Sohi, “A near real-time IP traffic classification using machine learning,” Int. J. Intell. Syst. Appl., vol. 5, no. 3, p. 83, 2013.

[11] R. Cathey, et al., "Misuse detection for information retrieval systems," in Proceedings of the twelfth international conference on Information and knowledge management, 2003, pp. 183-190.

[12] J. Vergara-Reyes, et al., "IP traffic classification in NFV: A benchmarking of supervised Machine Learning algorithms," in 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), 2017, pp. 1-6.

[13] S. Nayebalsadr and M. Analoui, "Simulating the Resource Freeing Attack: Using Cloudsim Simulator," Journal of Computing and Security, vol. 2, pp. 293-302, 2015.

[14] I. Letteri, et al., "Performance of Botnet Detection by Neural Networks in Software-Defined Networks," in ITASEC, 2018.

[15] S. Ruggieri, "Efficient C4. 5 [classification algorithm]," IEEE transactions on knowledge and data engineering, vol. 14, pp. 438-444, 2002.
[16] Q. Du, et al., "An approach of collecting performance anomaly dataset for NFV Infrastructure," in *International Conference on Algorithms and Architectures for Parallel Processing*, 2018, pp. 59-71.

[17] H. Kim, et al., "Design of network threat detection and classification based on machine learning on cloud computing," *Cluster Computing*, vol. 22, pp. 2341-2350, 2019.

[18] KDD. (2020). *KDD Cup*. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html