EFFECTS OF FEATURE SELECTION ON NETWORK INTRUSION DETECTION SYSTEM ACCURACIES

Nhat Hieu Le - 101124254 Zein Hajj-Ali - 101020677

Abstract

Recently, Deep Learning (DL) and Machine Learning (ML) methods are widely targeted in cybersecurity research as a promising solution in terms of accuracy and efficiency. In this paper, we show comparisons between NIDS systems that employ different Machine Learning algorithms, Deep Learning approaches for multiclass classification to detect malicious activities on two datasets UNSW-NB15 and Bot-IoT. Furthermore, the role of numerous feature selection techniques composed of Association Rule Mining (ARM) algorithms, Deep Neural Networks and ML approaches are also discussed.

I/ Problem Statement

Heterogeneous types of malicious traffic such as zero-day attack, social engineering, denial of services (Dos) etc. has extended at an exponential rate in recent years [1]. Unsurprisingly, cybercrime is now considered a major issue for corporations in the digital era to maintain the continuous availability services to their customers to avoid reputation damage. Therefore, observing corporations' network traffic and uncovering possible intrusion and malicious software in a timely manner are ultimate goals of NIDS. Depending on the methods used, there can be tradeoffs between the accuracy of the NIDS and its speed. Due to the limitations of conventional NIDS such as high False Alarm Rate (FAR), incapable of identifying zero-day exploits etc., recent research attempts have applied artificial intelligence methods to tackle aforementioned limitations. Research has already been focused on comparing some conventional ML techniques to each other as in [2]. The aim of this project is to determine the most efficient intrusion detection system when comparing conventional machine learning techniques to some deep learning algorithms. It has also been shown that the inclusion of some irrelevant features in the decision process can increase the FAR of an anomaly detection based NIDS. Therefore, the effects of feature selection through ARM on the accuracy and speed of the algorithms will also be explored. Faster training time, reducing redundant features to avoid overfitting, increasing accuracy are the main purpose of employing feature selection methods.

II/ Work Plan

The dataset to be used in these experiments are UNSW-NB15 and Bot-IoT. Weka will be used along with CPython integration and TensorFlow to train and test the algorithms. Traditional ML algorithms like the ones tested in [2] will be compared against deep learning methods similar to what is proposed in [3]. The accuracies and efficiency of the best performing methods will be tested again after using ARM for feature selection and compared against the initial results.

III/ Expected Outcome

We assume that the training time of the deep learning models will be longer than that of the conventional methods, but the classification time might be similar. We also assume that the accuracies of the deep learning models will be competitive with the conventional model with the highest accuracy (Random Forest). Feature selection has been shown to generally reduce the false alarm rate when using the ARM method, and it is also probable that it will reduce the chances of overfitting a deep learning model on the training data.

References

1. "Malware Statistics & Trends Report | AV-TEST", Av-test.org, 2020. [Online]. Available: https://www.av-test.org/en/statistics/malware/. [Accessed: 31- Oct- 2020].
2. Das A., Ajila S.A., Lung CH. (2020) A Comprehensive Analysis of Accuracies of Machine Learning Algorithms for Network Intrusion Detection. In: Boumerdassi S., Renault É., Mühlethaler P. (eds) Machine Learning for Networking. MLN 2019. Lecture Notes in Computer Science, vol 12081. Springer, Cham, doi: 10.1007/978-3-030-45778-5_4.
3. M. Al-Zewairi, S. Almajali and A. Awajan, "Experimental Evaluation of a Multi-layer Feed-Forward Artificial Neural Network Classifier for Network Intrusion Detection System," *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, Amman, 2017, pp. 167-172, doi: 10.1109/ICTCS.2017.29.
4. N. Moustafa and J. Slay, "The Significant Features of the UNSW-NB15 and the KDD99 Data Sets for Network Intrusion Detection Systems," *2015 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*, Kyoto, 2015, pp. 25-31, doi: 10.1109/BADGERS.2015.014.