Vulnerability Assessment of Machine Learning Based Malware Classification Models

Godwin Raju
Concordia University of Edmonton
Edmonton, Canada
graju@concordia.ab.ca

Pavol Zavarsky
Concordia University of Edmonton
Edmonton, Canada
pavol.zavarsky@concordia.ab.ca

Adetokunbo Makanju
New York Institute of Technology
Vancouver, Canada
amakanju@nyit.edu

Yasir Malik*
New York Institute of Technology
Vancouver, Canada
ymalik@nyit.edu

ABSTRACT

The primary focus of the machine learning model is to train a system to achieve self-reliance. However, due to the absence of the inbuilt security functions the learning phase itself is not secured which allows attacker to exploit the security vulnerabilities in the machine learning model. When a malicious adversary manipulates the input data, it exploits vulnerabilities of machine learning algorithms which can compromise the entire system. In this research study, we are conducting a vulnerability assessment of the malware classification model by injecting the datasets with an adversarial example to degrade the quality of classification obtained currently by a trained model. The objective is to find the security gaps that are exploitable in the model. The vulnerability assessment is done by introducing the malware classification model to an AML environment using the Black-Box attack. The simulation provided an insight into the inputs injected into the classifiers and proves the inherent security vulnerability exists in the classification model.

CCS CONCEPTS

• **Computing methodologies** → Feature selection; **Classification and regression trees**; • **Security and privacy** → Artificial immune systems;

KEYWORDS

Machine Learning, Adversarial Machine Learning, Black-Box Attack, Poisoning Attack, Intrusion Detection System, Malware Classification.

1 INTRODUCTION

Machine learning is one of the rapidly evolving and adapting technology, that uses statistical techniques to give computers the ability to learn with data, without being explicitly programmed. Recently, there is a growing interest in applying machine learning techniques in the development of Intrusion Detection System (IDS). The machine learning model trains the IDS to predict, flag and classify malicious activities into different attack types. This would enable IDS to predict zero-day attacks and improve its detection rate while requiring minimum user maintenance due to self-learning ability of the model.

The primary focus of the machine learning model is to train a system to achieve self-reliance. Generally, machine learning based systems are trained to take actions according to the given circumstance, and the efficiency and accuracy of the model depend on the model design, learning environment, and training data. However, by observing the learning environment, the attacker can manipulate the machine learning model and exploit the security vulnerabilities in the model.

The vulnerability in the machine learning model is due to the adversely crafted inputs, i.e. when a malicious adversary manipulates the input data *a.k.a* adversarial examples. Adversarial examples are inputs to machine learning models that let the attacker exploits specific vulnerabilities of the machine learning model or algorithm to compromises the entire system. This concept is known as Adversarial Machine Learning (AML) [9][3]. AML is the study of machine learning techniques against an adversarial opponent. AML primary focus is to find vulnerabilities which are inherent to a machine learning model, which can then create opportunities for attackers to embed malicious codes during the preliminary stages to produce disruption in the results, degrade the performance of a learning model, bypass filtering rules or find new attack vectors [5][9].

In this research study, we are conducting a vulnerability assessment of the malware classification model by injecting the datasets with the adversarial example to degrade the quality of classification obtained currently by a trained model. The objective is to find the security gaps that are exploitable in the model. The vulnerability assessment is done by introducing the malware classification model to an AML environment using the Black-Box attack [10].

The rest of the paper is organized as follow, Section II presents the related research, Section III describe our methodology, section IV presents the simulation results and discussion and finally, we conclude our work in section V.

*Corresponding Author

GECCO '19, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3326897

2 RELATED WORK

Adversarial Machine Learning techniques provide a gap analysis which could be used to improve and enhance the security aspects of the machine learning model. This section presents a review of solutions that are most relevant to this research AML provides various attack models and using multiple attack models one can corrupt a learning process. M. Barreno et. al. adopted the causative attack model to manipulate a naïve learning algorithm. The model simply yields an optimal policy for the adversary and a bound of effort is required to achieve the adversary's objective. Resulting bound is extended by using outlier's detection technique. Authors found that using hyper dimension outlier's trajectory for expansion could inject malicious points where the outlier would move next. Therefore, a dynamic machine learning model presents more opportunity for the attackers, as it is possible to twist the learning process using adversarial inputs [7]. L. Huang el. al. provides an accurate description of the adversary's control over the features by discussing domain limitations that are set upon the adversary due to the application domain itself. These limits can include how an adversary interacts with the application and what kind of data is realistically modifiable by the adversary. Contrasting feature spaces is the second limitation imposed on the adversary by the space of features used by the learning algorithm. In many learning algorithms, data is represented in a feature space in which each feature captures a relevant aspect of a data point for the learning task at hand. Another application-specific aspect of the threat discussed in the research is contrasting data distribution which not only impacts the performance of the learning algorithm but also its vulnerabilities. The data's distribution may contain properties which can conflict with the learner's assumption [6][12].

Grosse et al. studied the adversarial attack example for classification models and applied their attack model to test well known Android malware detection models. With their experiment, they were able to achieve 63% misclassification by the malware detection system. Furthermore, the authors also evaluated the defense mechanism based on DNN and suggested that defense models should be included Adversarial examples in training set to improve its robustness and combat adversaries [4]. In another work, Al-Dujaili et al. studied the methods to reduce the adversarial examples for malware detectors based on neural network. The author's suggested that the power of randomization can help in to discover malicious samples during natural training [2].

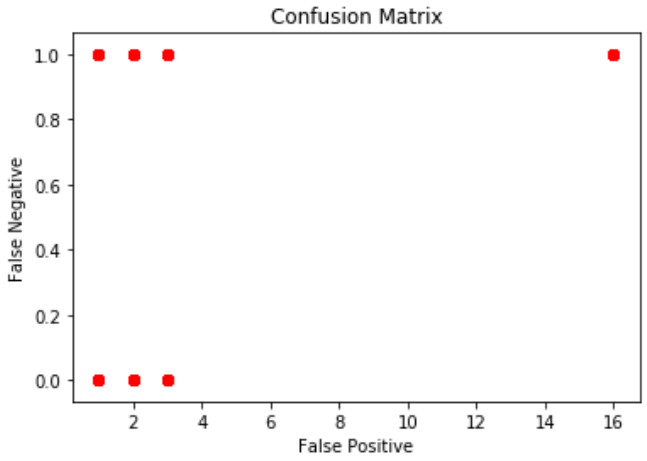
3 VULNERABILITY ASSESSMENT OF CLASSIFICATION MODEL

The objective of this research is to conduct a vulnerability assessment of the malware classification model by degrading the quality of classification obtained currently by a trained model and find the security gaps present in the model. A malware classification model is implemented which provides the attacker with complete control over the environment including classification algorithm and feature association information. The FAIL model [13] is used as a framework for an adversarial attack on the malware classification model. The FAIL model enables the attacker to successfully deploy the attack/intrusive elements into the environment with minimum resources. The vulnerability assessment is done by introducing the malware classification model to an AML environment using the Black-Box attack [10]. Using FAIL model, we were able to overcome the limitations of Black-Box attack. The four attack dimensions used for FAIL model are:

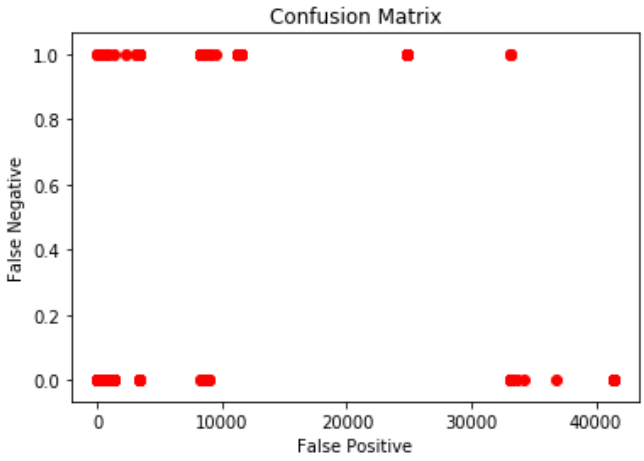
(1) ***Feature***: There are 15 features which are kept secret by the classification model, however, this subset of features is known to the adversary.
(2) ***Algorithm***: The training algorithm is not a secret in this environment. The attacker knows the design of the algorithm to craft the poison samples in the model.
(3) ***Instance***: knowledge: Labeling and classification information of the entire training set is available to the attacker.
(4) ***Leverage***: Leverage dimension is all about introducing adversarial examples into the classification features which are intentionally corrupted/polluted to introduce noise which would decrement the quality of classification done by the model. Leverage means adversary's ability to modify the subset. Here the attacker has full access to the features set which were obtained from the first dimension and is capable of modifying all the features which are selected by the classification model. The simulation record 0.35% false positive and negative before the poison attack was injected into the classifiers.

4 SIMULATION RESULTS

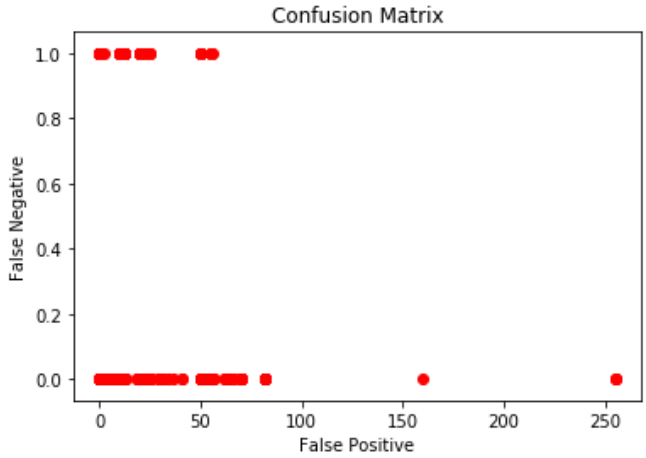
All the tests and simulations are conducted in a controlled environment. The dataset consists of the legitimate and malicious files obtained from VirusShare repository. We simulated decision tree, random forest, SVM, and linear regression classification algorithms [8][11] and [1] to inject AML examples. Parameters for each algorithm are selected based on the requirements of the algorithm (e.g. depth for the decision tree, estimators for the random forest, etc.). The adversarial inputs are generated using poison attacks for different feature sets. The injection of adversarial examples into malware classification model introduces perturbations in the results. The perturbations cause an increment of false negative and false positive in the system. The value of false positive and false negative depends upon the attack vector (i.e. classifiers), for example, `DLCharacteristics`, `ResourceMaxEntropy`, `ImageBase`, `SectionsNB` are classifying labels in our simulations. Figure 1a shows the result of poisoning the feature ''`DLcharacteristics`'' which increases the rate of false positives to 13.78% and false negatives to 11.57%. Figure 1b shows the result of poisoning the feature ''`ExportNb`'' and which increases the rate of false positives to 8.58% and false negatives to 21.53%. Figure 1c shows the result of poisoning the feature ''`ResourceMaxEntropy`'' which increases the rate of false positives to 37.57% and false negatives to 31.53%. Figure 1d shows the result of poisoning the feature ''`ImageBase`'' which increases the rate of false positives to 75.57% and false negatives to 53.53%.



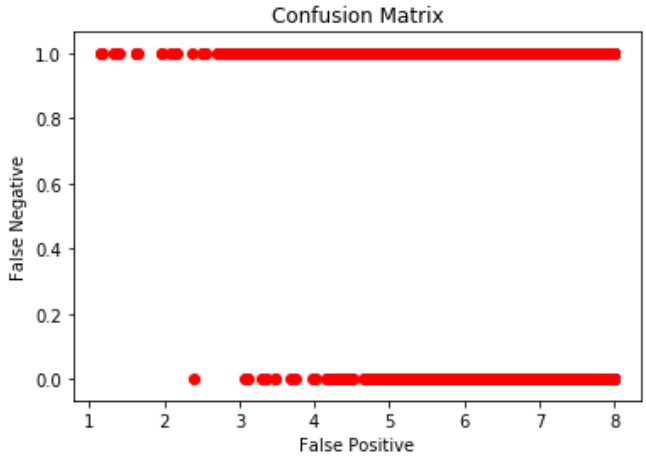
(a) DLcharteristics



(b) ExportNb



(c) ResourceMaxEntropy



(d) ImageBase

Figure 1: The False Positive and Negative Ratios of Classifier Features After Poison Attack.

5 CONCLUSION

This research work investigates the possibilities of injecting adversarial examples into machine learning based classifiers to exploits the security vulnerability. The vulnerability of machine learning systems against poisoning attack is incorporated using AML techniques, that leads to an arms race, where system defense becomes adequate and classifier results are misleading. The simulation provided an insight into the inputs injected into the classifiers and proves the inherent security vulnerability exists in the classifier model, which allow the attacker to craft similar attack for other models using the transferability property. We notice, when adversarial inputs are not moderated and injected on a large-scale, it would introduce high-levels of noise into the model which results in more misclassification of a data. We have also studied mitigation techniques to combat AML attacks. There are mainly two security approaches reported in the literature for machine learning models, i.e. reactive, and proactive security. The reactive security model would ensure countermeasure after the system detects intrusion, whereas proactive model would devise methods to simulate the attack on the model and comprehend the impact and effects of the adversarial attack to develop a mitigation strategy. This paper can be viewed as an effort to developing a proactive approach, where we tested different classification model to uncover their vulnerabilities and improve its defense. In the next step, we are working on developing a framework to analyze a continuous adversarial attack and its impact analysis model to devise effective mitigation strategies against adversaries.

REFERENCES

[1] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2, no. 3 (2002): 18-22

[2] Al-Dujaili, Abdullah, Alex Huang, Erik Hemberg, and Una-May O'Reilly. "Adversarial deep learning for robust detection of binary encoded malware", In 2018 IEEE Security and Privacy Workshops (SPW), pp. 76-82. IEEE, 2018.

[3] Duddu, Vasisht. "A Survey of Adversarial Machine Learning in Cyber Warfare". Defence Science Journal 68, no. 4 (2018): 356-366.

[4] Grosse, Kathrin, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. "Adversarial examples for malware detection". In European Symposium on Research in Computer Security, pp. 62-79. Springer, Cham, 2017.

[5] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". arXiv preprint arXiv:1412.6572 (2014).

[6] Huang, Ling, Anthony D. Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J. D. Tygar. "Adversarial machine learning", In Proceedings of the 4th ACM workshop on Security and artificial intelligence, pp. 43-58. ACM, 2011.

[7] Barreno, Marco, Blaine Nelson, Anthony D. Joseph, and J. Doug Tygar. "The security of machine learning". Machine Learning 81, no. 2 (2010): 121-148.

[8] Adankon, Mathias M., and Mohamed Cheriet. ""Support vector machine". Encyclopedia of biometrics (2015): 1504-1511.

[9] Papernot, Nicolas, Ian Goodfellow, Ryan Sheatsley, Reuben Feinman, and Patrick McDaniel. ""Cleverhans v1. 0.0: an adversarial machine learning library". arXiv preprint arXiv:1610.00768 10 (2016).

[10] Papernot, Nicolas, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. "Practical black-box attacks against machine learning". In Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp. 506-519. ACM, 2017

[11] Suthaharan, Shan. "Machine learning models and algorithms for big data classification". Integr. Ser. Inf. Syst 36 (2016): 1-12

[12] Chen, Sen, Minhui Xue, Lingling Fan, Shuang Hao, Lihua Xu, Haojin Zhu, and Bo Li. "Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach". Computers & Security (2018):326-344.

[13] Suciu, Octavian, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. "When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks". In 27th USENIX Security Symposium (USENIX Security 18), pp. 1299-1316. 2018