Malware Detection Using Machine Learning Algorithms

Gopi Nagabhiru

*School of Computing Sciences & Computer Engineering*

*Univ. of Southern Mississippi* Hattiesburg, MS, USA [Gopi.Nagabhiru@usm.edu](mailto:Gopi.Nagabhiru@usm.edu)

Srinivasa Reddy Annapareddy

*School of Computing Sciences & Computer Engineering*

*Univ. of Southern Mississippi* Hattiesburg, MS, USA [W951458@usm.edu](mailto:W951458@usm.edu)

Nick Rahimi

*School of Computing Sciences & Computer Engineering*

*Univ. of Southern Mississippi* Hattiesburg, MS, USA [nick.rahimi@usm.edu](mailto:nick.rahimi@usm.edu)

***Abstract—*malware is the significant threat to the computer systems, networks. It can be many types and can be attacked in many ways to a network or a system. Detection of malware is a precautionary step which will help in eliminating the malware spreading. Machine learning helps in detecting the malware where we use machine learning algorithms like logistic regression, decision tree algorithm, random forest classifier, support vector machines, naïve bayes classifier.** **Machine learning-based approaches involve the use of algorithms to analyze large datasets and identify patterns indicative of malware behavior. The dataset has almost 1lakh instances where 50000 are malware and 50000 are benign. The model is trained with each algorithm and then the model is tested with the testing data where the dataset is divided into 80percent and 20percent for training and testing respectively. Cross validation is done for each model in order to avoid over-fitting problem.**

***Keywords—malware, machine learning, malware detection***

INTRODUCTION

The increasing obsolescence of hardcopy storage underscores the critical need for effective malware prevention to safeguard personal, corporate, and national security. Malware poses a significant threat by potentially extracting

sensitive information like bank details and passwords, jeopardizing security infrastructures and classified systems. Malware attacks, often involving malicious links, encompass malware downloads and password phishing, with Proofpoint's report revealing a surge in phishing incidents. To address this challenge, the paper proposes a program leveraging five algorithms, including Decision Tree and Random Forest, to accurately detect malicious messages and protect users from engaging with malware.

The subsequent sections of the paper detail the research methods, results, malware detection strategies, and key findings in the field of malware detection. This paper introduces a program designed to effectively identify malicious SMS messages, aiming to shield users from unwittingly engaging with malware. The program's development involved training datasets using five distinct algorithms, including Decision Tree, Random Forest, logistic regression, naïve bayes, support vector machines and then is validated with the testing data to give the accuracies which results with the optimistic model which can detect the malware.

While malicious attacks and their implications are a growing problem, the detection of SMS malware has not received much research.

TRAINING METHODS

In this section, we will describe the process of cleaning our data and the rules followed by each of the five algorithms we experimented with during this research.

We first tested the effectiveness of the Decision Tree (DT) algorithm on our dataset. This algorithm performs hierarchical exemplification to train data. Each layer of branches is formed through Boolean decisions based on conditional probabilities. Using supervised learning, decision trees reduce entropy by raising the influence of attributes that have a more accurate correlation with their given labels.

We then tested the effectiveness of random forest algorithm on our dataset. It is an Ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. The advantages of this algorithm is it is Robust to overfitting, handles high-dimensional data well, provides feature importance ranking.

We then tested the effective ness of the support vector machines (SVM). It is a Supervised learning algorithm that finds the hyperplane that best separates classes in a high-dimensional feature space. One side of the hyper plane determines the datapoints with malware while the other side is benign. The advantages of this algorithm is that they are Effective in high-dimensional spaces, versatile with different kernel functions, works well with small datasets. One of the application of this algorithm is this is useful for malware detection where the decision boundary between classes may be complex and non-linear.

We then tested our model with logistic regression. It is a statistical method used for binary classification tasks. It models the probability that a given instance belongs to a particular class. It is simple and easy to implement and is best suitable for binary classification tasks. Then the model is tested with naïve bayes.

TESTING AND RESULTS

We developed a Python script to measure and aggregate our benign and malicious files, then trained this new data set with each method separately in order to run our data through each algorithm. The outcomes of each algorithm will be covered in this section, along with the logic underlying their correctness or lack thereof.

Random forest algorithm gives the best accuracy so it it consider as the best algorithm even when we use cross validation where we divided the dataset into 5 folds we got the mean accuracy of 98.82%.

|  |  |
| --- | --- |
| Classifier | Average accuracy |
| DECISION TREE | 97.47% |
| RANDOM FOREST | 98.59% |
| SVM | 74.76% |
| NAÏVE BAYES | 32.39% |
| LOGISTIC REGRESSION | 25.59% |

Future improvements

In the future we can implements deep learning algorithms to

implement the same project and try to increase the accuracies

ACKNOWLEDGMENT

This study was supported in part by the University of

Southern Mississippi.

REFERENCES

[1]. Proofpoint Inc. (2022). 2022 state of the Phish Report.

https://go.proofpoint.com/en-2022-state-of-the-phish.html

[2]. Yang, F. J. (2019, December). An extended idea about

decision trees. In 2019 International Conference on

Computational Science and Computational Intelligence

(CSCI) (pp. 349-354). IEEE.

[3]. Zhang Z. (2016). Introduction to machine learning: knearest

neighbors. Annals of translational medicine, 4(11), 218.

[4]. Taud, H., & Mas, J. F. (2018). Multilayer perceptron

(MLP). In Geomatic approaches for modeling land change

scenarios (pp. 451-455). Springer, Cham.

[5]. A. Rana, A. Singh Rawat, A. Bijalwan and H. Bahuguna,

"Application of Multi Layer (Perceptron) Artificial Neural

Network in the Diagnosis System: A Systematic Review,"

2018 International Conference on Research in Intelligent and

Computing in Engineering (RICE), 2018, pp. 1-6, doi:

10.1109/RICE.2018.8509069.

[6] Rahimi, N., Maynor, J., & Gupta, B. (2020, March).

Adversarial Machine Learning: Difficulties in Applying

Machine Learning to Existing Cybersecurity Systems. In

CATA (pp. 40-47).

[7]. Thakur, S., Meenakshi, E., & Priya, A. (2017, May).

Detection of malicious URLs in big data using RIPPER

algorithm. In 2017 2nd IEEE International Conference on

Recent Trends in Electronics, Information & Communication

Technology (RTEICT) (pp. 1296-1301). IEEE.

[8]. Biau, G., & Scornet, E. (2016). A Random Forest Guided

Tour. Test, 25(2), 197-227.