Analytical model to detect mailicious executables

Sumedh Arani
Dept of Computer Science
PES University

Tejas Kaushik
Dept of Computer Science
PES University

Vijeth Nandan
Dept of Computer Science
PES University

Prince Kumar
Dept of Computer Science
PES University

Abstract **- A serious security threat today is malicious executables, especially new, unseen malicious executables. These new malicious executables are being created at a alarming rate and pose a serious security threat. Current anti-virus systems attempt to detect these new malicious programs with heuristics generated by hand(Signature based* and Behavior based**).**

This approach is costly and often times ineffective.In this attempt of ours, we try to build a analytical framework that would eventually be deployed to detects new, previously unseen malicious executables accurately and automatically.

We present a hybrid data mining approach to detect malicious executables. In this approach, on the basis of identified important features of the malicious and benign executables, a classifier is used to learn a classification model that can distinguish between malicious and benign executables from the patterns detected.

I. INTRODUCTION

In the current scenario, the number of devices that are getting connected to the internet is growing exponentially and as the number of users increase, it is also a growing target user base for people waiting to create havoc by the means of dubious software. One such commonly used for lancing these types of attack us popularly known as malware i.e. viruses, Trojans, and worms, which, when propagate can cause a great amount of damage to private users, commercial companies and governments.With also the rise in connected devices, we also are able to collect massive amounts of data pertaining to such scenarios.With more and more tools to gather data, we are progressing towards an era where cars are becoming driverless, voice assistants are becoming more powerful and hence we also would want that all these innovations stay safe and not be sabotaged by these malware. With the rise in such attempts to bring disruptions, there arises a need for an analytical approach to detect malicious executables on the prompt need.

II. CITATIONS AND WORK DONE BY OTHERS

A. Assumptions Made

It is a multivariate dataset with 373 instances having 513 attributes. The testing data involves 100+ non malacious examples and 250+ malicious samples. It contains labelled data. This project deals with supervised learning. Based on real code samples analysis and comparision has the author of the dataset arrived at 513 most commonly occuring features. These attributes were generated after the author of the dataset created n-gram hexdump of the executables. The hybrid features are a mix of the hexdump and the dll(dynamically linked libraries).Techniques best suited for feature extraction were made in use to find the most appropriate features. The data has it's fair proportion of missing values. So task at hand involves us to clean the data before use. Also, we assume the data true be verified and relevant to the task at hand as it is being sourced from a verifiable reputed source to collect data.

In "A Hybrid Model to Detect Malicious Executables"[1], the authors propose an efficient and scalable feature extraction technique. We extract the above mentioned features from the data and train a classifier using Support Vector Machine. This classifier achieves a very high accuracy and low false positive rate in detecting malicious executables.The dataset collected was created with ideologies from the above paper. The dataset was obtained from UCI ML Repository[2].

B.Approach Used

In the paper "Detection of Malicious Data using hybrid of Classification and Clustering Algorithms under Data Mining"[3] have discussed way to solve the problem by clubbing classification and clustering algorithms for rootkit*** detection. The methodology formulated for rootkit prediction consists of rootkit data collection, data pre-processing, and classification and performance evaluation

---

* Signature-based AV compares hashes (signatures) of files on a system to a list of known malicious files. It also looks within files to find signatures of malicious code.

** Behavior-based AV watches processes for telltale signs of Malware, which it compares to a list of known malicious behaviors.

*** Rootkits is type of software that hinders the presence and activity of malware (such as viruses, worms and Trojans) and allow attacker to capture a computer system.

phases. The job mainly is a classification oriented task and the authors have used SVM(Support Vector Machines) with KNN(K Nearest Neighbors) as their classifier. They have also used Naive Bayes, Ripper algorithms in their approaches.

C. Results And Limitations

In the case of [3], no results were reported. As in the case of [2], we've tried putting in a request to get the results.,There weren't any limitations reported.

There were no inferences made as to why clubbing of classification and clustering techniques stand to give an advantageous analysis. There are no benchmarks to which their theory holds good.This gives us an edge as not much has been discussed with regard to malicious executables.

III. PROBLEM STATEMENT

Detect malicious executables on the basis of hybrid features. It involves doing a binary classification i.e. into malicious and non-malicious executables.

IV. OUR APPROACH

We look towards cleaning the data a bit more as there is missing data. Try running PCA(Principal Component Analysis) to reduce the dimensionality of the problem. Also run for correlation tests to find how to attributes are related. Also we would be testing our cases with three different classification approaches.

1)SVM as used earlier.

2)Decision Trees.

3) Neural Networks

and analyze how each of them perform. Much research has been done with detection of toolkits rather than executables. The common approach has been to use SVM as a standard classification technique. We propose to try different techniques with suitable optimizations with changing time frame to best solve the given problem.

V. CITATIONS

[1]-A Hybrid Model to Detect Malicious Executables Mohammad M Massud, Latifur Khan, Bhavani Thuraisingham

[2] - https://archive.ics.uci.edu/ml/datasets/Detect+Malacious+

Executable(AntiVirus)

[3]-Detection of Malicious Data using hybrid of Classification and Clustering Algorithms under Data Mining Milan Jain, Bikram Pal