Machine Learning

Home Work I

Android Malware Detection

Master of Engineering in Computer Science

INTRODUCTION

Malware detection has been an important topic in cyber security research. This report discusses some methods to detect a malware.

FEATURES

For each application, the Drebin dataset contains a text file. The text file describes all the properties of the application. Each property belongs to one of 8 categories (S1 to S8). Then a feature vector for this application can be the occurrence of each category in the text file. For example, one application can have a feature vector of 5-3-0–65-11-9-2. It means that S1 category occurs 5 times, S2 occurs 3 times and so on.

PROBLEMS

Malware or non-malware

The first problem can be tackled using the above feature vectors is to classify whether an application is malware or not.

A. Choice of dataset

The Drebin dataset comes with 5560 positive examples and more than 100k negative examples. This raised a problem of class imbalance. Experiment showed that a support vector machine when trained on 5560 positives + 20000 negatives had accuracy score of 0.86 and f1 score of 0.61 (because of very bad recall). But when the same model trained on 5560 positives but only 5560 negatives, it had 0.91 accuracy score but 0.90 f1 score.

B. Evaluation metrics

A good metrics for this problem is F1 score since it combines precision and recall. And in this problem, both precision and recall are important. Because people want to find all the positive examples and also not to label a negative example as positive.

C. Algorithm Used

Support vector machine, SVM was tested and performed a lot better only with default hyper parameters.

D. Performance

	Accuracy	F1 score
Support vector machine	0.9	0.90

E. Intuition explained

Since SVM was the most successful model, it was chosen to explain what are the motivations behind.

LIBRARIES

The following open source libraries were used during the experiment.

1. Numpy - scipy for reading data and building matrices.

2. Scikit-learn for machine learning algorithms and utilities.

3. Matplotlib for plotting

RESULTS

The algorithm was tested on the Malware Detection. The result for this problem performed better by using support vector machine (SVM).

CONCLUSION&SUMMARY

Machine learning algorithms are very useful in the context of malware classification. Some of them seems to perform very well in practice given the right feature vectors. They also provide interesting insights about how an application should be examined for malware detection. For example, support vector machine suggested the number of required hardware components and suspicious API calls are very important criterias.

Every individual and organization is vulnerable to the threat of malwares. Malwares have become an effective instrument to damage, destroy and incur mammoth losses not only restricted to individuals but also to highly e-secured environment of organizations. The exploitation of computer programs is being visualized as the next threat to information storing and sharing. A comprehensive research in detection, analyzing, identification, repairing, removing of malwares is required to explore this undiscovered field. Therefore, cyber crimes needs to be thoroughly and meticulously conducted similar to a murder investigation. In the good old days, digital investigators could easily explore, discover and analyze malicious code on computer systems due to the malware functionality which was easily observable; therefore little effort was required in performing in depth analysis of the code. Today, various forms of malware are proliferating, automatically spreading (worm behavior), providing remote control access (Trojan horse/backdoor behavior), and sometimes concealing their activities on the compromised host (rootkit behavior). Furthermore, malware bypass security measures & firewalls disable AntiVirus tools from within the network to external command. The increasing sophistication of malicious code & growing importance of malware analysis in digital investigation has driven advances in tools and techniques for performing autopsies and surgery on malware. The demand for formalization and supporting documentation has grown as more investigations rely on understanding malware. The results of malware analysis must be accurate and verifiable, to the point that they can be relied on as evidence in an investigation or prosecution. The above model is a very simple and helpful tool even to the least computer literate to understand and differentiate among the various types of malware.

The decision tree highlights the parameters to look out for the analysis whenever we are subjected to a cyber attack. Every antivirus is not 100% safe, malware authors are very smart and they make use of crypters & binders to bypass even Antivirus. So, instead of using so many tools, forensic investigator can make use of our model to classify the malwares. Once the investigator has searched those six parameters about the malware during investigation, it becomes easy for them to classify by our model. The identification of the binary structure of a malware helps in knowing its features, characteristics, behavior and composition. Collection of such information yields in developing its countermeasures depending upon its type (worm, rootkit). Malware analysis is like a cat & mouse game, as new malware analysis techniques are developed, malware authors respond with new techniques to thwart analysis. Antivirus use source code to detect malware i.e. too time consuming, our model is statistical & thus more successful. Forensically, Decision tree model will help in streamlining the process of investigation by concentrating only on those significant derived from the model for a forensic investigator. The advantage of this model is that if any new malware is executed or developed, Antivirus will search source code by OllyDbg, IDAPro making use of assembly language, then it will develop signature & will update in

software or countermeasure will be taken after the development of signature but when any new malware is executed in our system, then with the help of our model we can search those significant immediately, classify the category of malware & then countermeasure can be taken at that time.

Future Scope There are many ideas and methods that are yet to be implemented for the future research. 1) In future, a tool or software developed can include the significant finding obtained in our Decision tree to classify the malware. 2) Moreover the tools may be attached with the knowledge base, so that less skilled user can also use the toolkit for forensic analysis. 3) Last but not least the task of first detecting, analyzing & generating cures for unknown & malicious files is itself an individual research topic.