Malware Classification into Families

Palak Agarwal(13453) and Karan Bansal(12342)

Department of Computer Science & Engineering

Abstract

The project is based on Kaggle's Microsoft Malware Challenge 2015 to classify malware into families.

- It involves training the classifier using the given database to classify the malware files(binary executables) into 9 categories of malwares.
- The main challenge we have is identifying the distinguishing features in the bytes and the asm file for classifying malware into their respective classes.

Introduction

Malware authors use automated techniques like **Polymorphism** in order to evade 'pattern matching' detection. Polymorphic malwares constantly changes itself, making it difficult for the anti-malware programs to detect it. Evolution of malicious code occurs in a variety of ways like change in filename, compression and encryption with variable keys.

Methodology

The following methods were used to complete the research:

- Random forest classifier *According to previous research random forest is the best classifier*
- N-gram based file signatures
- K-fold Cross validation

Following features were used to classify the malwares:

1. Frequency of 256 possible hex values in the bytes file corresponding to each malware
2. Occurence of instructions like mov, jmp etc. in the asm file corresponding to each malware
3. Frequency of 256 possible hex values at specific position in the asm file corresponding to each malware

Result

- After using all the proposed features, a score of 0.153125351 on the leaderboard is achieved.
- Evaluation is done using Multi-Class Logarithmic Loss.
- For each malware file in the test set, their predicted probabilities for the 9 classes was submitted.

Conclusion

The goal is to take the score as close to 0 as possible. Though, the current score is not bad but it can be further improved by identifying more **distinguishible** features for malwares belonging to different families.

Datset

Dataset in the form of .asm and .byte files

- Training & Test set - 200 GB each
- Asm file(0.4-19 millions lines)

References

[1] Daniel Bilar et al.
Statistical structures: Fingerprinting malware for classification and analysis.
Proceedings of Black Hat Federal 2006, 2006.

[2] Karthik Raman.
Selecting features to classify malware.
InfoSec Southwest, 2012.