A Survey on Malware Detection Using Data Mining Techniques

Jeff Malavasi
Dept. of Computing Security
Rochester Institute of Technology
Email: jm3378@rit.edu

I. Review

The authors of the paper investigate how the explosion of new evasion techniques in malware such as packing, polymorphism, and metamorphism have accelerated the need for intelligent malware detection methods. Previously, malware engines operated primarily based on signature detection, a process that creates a hash of a known malicious binary which is then stored in a database and used for classification at runtime. Unfortunately, these systems have become vulnerable to evasion and require a large amount of maintenance to stay up to date. The authors provide a comprehensive review of novel malware analysis that uses data mining (static features, dynamic features, file relations, etc.) alongside machine learning in order to automate malware classification.

In order to combat recent advancements in malware development, the authors propose a malware detection system based on data mining techniques. These systems are typically comprised of two parts: feature extraction and classification/clustering [1]. During feature extraction, specific patterns are detected from the binary of interest. The authors propose a hybrid approach that uses static and dynamic analysis to more accurately detect malicious software. Additionally, other novel features, such as file interactions, has shown to be extremely effective in identifying malware. While it is important to have a diverse feature set, the authors highlight the need for careful feature selection. Often a subset of the features can provide a more accurate model by removing redundant data. After the features are evaluated for a specific binary, the model must then classify it as benign or malicious. The authors review many potential methods of classification, but also emphasize the use of a combination of classifiers to improve accuracy (ensembles). Finally, in order to scale these systems, the authors suggest using a form of unsupervised machine learning known as clustering [1]. They review both partitioning and hierarchial clustering, which automatically group the data set based on the extracted features.

In conclusion, the authors propose an intelligent malware detection system that combines many features together to classify files. Although these systems can be extremely effective, there will always be a need to constantly update the models as new malware evolves.

References

[1] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, "A survey on malware detection using data mining techniques," in *ACM Computing SurveysCM*, 2017.