Classifying Malware into Families based on File Content and Characteristics

Karan Bansal(12342), Palak Agarwal(13453) and Tarun Singh(13111068) *

*Officially not participating in the course

Submitted to Prof. Amitabha Mukherjee for partial fulfilment of the course requirements for CS365A

Computer users often download malware to their computer by unknowingly visiting a malicious web- page hosting a drive-by download attack, by clicking on a malicious link included in email, or by inserting a USB thumb drive containing malware into their computer. The amount of new malware is growing at a staggering rate. Analysts are able to manually investigate a small number of unknown files, but the best large-scale defense for detecting malware is automated malware classification.

Polymorhpic Malware | Bloom Filters | SVM | Disassemblers

Problem Statement

We aim at providing a machine learning approach towards automated classification of malwares into various families. We plan to use classification methods based on various algorithms and then choose the one that gives the best result. Later, we will apply the grouping criteria to new files encountered on computers in order to detect them as malicious and of a certain family.

Motivation

In order to evade detection, Malware authors use automated methods such as Polymorphism, where a program generates a unique, new instance of a malware family for each victim, to create new malware. To combat this threat, anti-virus companies must utilize machine learning methods to automatically detect new instances of malware. In order to be effective in analyzing this large amount of data, we need to be able to group them into groups and identify their respective families.

Dataset

Through its open Malware Classification Challenge on kaggle, Microsoft has provided us with an unprecedented malware dataset to help us in formulating effective techniques for grouping variants of malware files into their respective families.

Challenge

We would like to tackle the challenge of discovering malware signature by analysing and observing the patterns in the assembly level instructions of various x86 binary malware executables captured by the IDA disassembler. The goal of this challenge is to classify the various given potential malware binary executables into various classes with some probability using the various function calls i.e., signatures as features for classification in the machine learning algorithms.

Methodology

Support Vector Machine (SVM)-The SVM is a widely used supervised learning model with associated learning algorithms to analyze high dimensional and sparse data and recognize patterns.The concept of SVM is to classify each data sample into one of two categories: positive class denoted by +1 and negative class denoted by 1. It boils down to find a decision boundary a plane (a hyperplane for n ¿ 3), which divides data into two sets, one for each class.

Naive Bayes (NB) -A NB classifier employs Bayes theorem with strong independence assumptions. In this technique, it is assumed that the features of the files are independent of each other. The learning algorithm based on Bayes classifier allows to combine prior knowledge and current measurements. Parameter estimation for naive Bayes models uses the method of maximum likelihood.

k-Nearest Neighbors (kNN)-The general idea of the kNN classifier is to classify a given query sample based on the class of its nearest neighbors in the dataset. The classification process is performed in two phases. In the first phase the nearest neighbors are determined. The neighbors of a query sample are selected based on the measured distances. In the second phase, the class is determined for a query sample based on the outcomes of the k selected neighbors.

N-Grams Extraction - An n-gram is an n-character slice of a longer string. To extract n-grams, we use the disassembled file that contains the contents of a file into a long string of hexadecimals. The string is then processed into a set of overlapping n-grams. In our study, we explore n-grams of several different lengths.

References

The following are the online sources we referred to for our project :

[1] Siddiqui, Muazzam, Morgan C. Wang, and Joohan Lee. ”Detecting internet worms using data mining techniques.” Journal of Systemics, Cybernetics and Informatics 6.6 (2008): 48-53.

[2] Zhou, Xin, et al. ”MRSI: A fast pattern matching algorithm for anti-virus applications.” Networking, 2008. ICN 2008. Seventh International Conference on. IEEE, 2008.

[3] Lum, P. Y., et al. ”Extracting insights from the shape of complex data using topology.” Scientific reports 3 (2013).

[4] Chaumette, Serge, Olivier Ly, and Renaud Tabary. ”Automated extraction of polymorphic virus signatures using abstract interpretation.” Network and System Security (NSS), 2011 5th International Conference on. IEEE, 2011.

[5] Dahl, George E., et al. ”Large-scale malware classification using random projections and neural networks.” Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[6] Kruczkowski, Micha, and Ewa Niewiadomska-Szynkiewicz. ”Comparative Study of Supervised Learning Methods for Malware Analysis.” Journal of Telecommunications Information Technology 2014.4 (2014).

[7] Swathigavaishnave, D., and R. Sarala. ”Detection of Malicious Code-Injection Attack Using Two Phase Analysis Technique.” International Journal of Computer Applications 45 (2012).

[8] Schultz, Matthew G., et al. ”Data mining methods for detection of new malicious executables.” Security and Privacy,

2001. SP 2001. Proceedings. 2001 IEEE Symposium on. IEEE, 2001.

[9] Bilar, Daniel. ”Statistical structures: Fingerprinting malware for classification and analysis.” Proceedings of Black Hat Federal 2006 (2006).

[10] Kinable, Joris, and Orestis Kostakis. ”Malware classification based on call graph clustering.” Journal in computer virology 7.4 (2011): 233-245.

[11] Briones, Ismael, and Aitor Gomez. ”Graphs, entropy and grid computing: Automatic comparison of malware.” In Proceedings of the 2004 Virus Bulletin Conference. 2004.

[12] Griffin, Kent, et al. ”Automatic generation of string signatures for malware detection.” Recent Advances in Intrusion Detection. Springer Berlin Heidelberg, 2009.

[13] Santos, Igor, et al. ”N-grams-based File Signatures for Malware Detection.”ICEIS (2) 9 (2009): 317-320.

[14] Siddiqui, Muazzam, Morgan C. Wang, and Joohan Lee. ”A survey of data mining techniques for malware detection using file features.” Proceedings of the 46th Annual Southeast Regional Conference on XX. ACM, 2008.

[15] Raman, Karthik. ”Selecting features to classify malware.” InfoSec Southwest(2012).