CS181 Spring 2016 Practical 2: Classifying Malicious Software — Team EXT3

Robert J. Johnson — Dinesh Malav — Matthew McKenna

March 5, 2016

Abstract

Identifying and eliminating malicious software (malware) is a modern computing task of considerable importance. Properly classifying these programs in a computationally efficient manner will lead to massive savings in time and money. Our study showed how various machine learning classifiers performed at this task when presented with malicious XML executables. A random forest classifier was shown to be the best model tested, with a categorization accuracy of .81211 on the test data.

1 Technical Approach

The training data set for our examination consisted of 3086 XML executable files. Of these, roughly half contained malware. The XML files on first glance appear quite daunting, with large hexadecimal strings and semi-random calls to various processes intermingled throughout the scripts.

2 Results

3 Discussion

All code for this project can be found at:

$$https://github.com/HarvardCS181Practical2016$$

.