CSCI 8360 Data Science Practicum

Project 1: Malware Classification

Team-void

Jiahao Xu, Yang Shi, Mohammadreza Iman

Technologies

- Apache Spark on Google Cloud Platform
  - Packages: spark.ml, spark.sql
- pySpark 2.3.2
- Python 3.7.2

Feature Exploration on Bytes Data

- Unigram ends with 256 features

- Bi-grams returns 256*256 = 65,536 features

- Three-grams means 256^3 = 16,777,216 possible features

- Four-grams rise the number of possible features to 256^4 = 4,294,967,296

Features

- Unigram line of bytes (from bytes files)
  - Each line at the bytes file (ignore the address, which is the first hexadecimal token), which is a restricted 16-grams bytes
- Bigram/3gram bytes (from bytes files)
  - Selected the important features by random forest classifier
- Segment, e.g. 'data', 'idata', 'rdata' ... (from asm files)
  - The first word in each line.
- Bigrams opcodes, e.g. 'push', 'add', 'mov' (from asm files)
  - Selected the important features by random forest classifier

Classifiers

- Naive-Bayes
- Random Forest
- Xgboost

Classification Results

- Small datasets:
    - NaiveBayes:
        - bytes & unigram 0.2662
        - bytes & bigram 0.6331
        - bytes & unigram of lines 0.5799
    - Random Forest
        - bytes & unigram of lines 0.6450
        - bytes & bigram 0.8580
- Large datasets:
    - Random Forest (numtree=50, maxdepth=25)
        - only segment 0.912
        - only bigrams with spark RF feature Importance 0.965
        - only 3-grams with spark RF feature Importance 0.945
        - bytes & bigrams with segment count **0.989**
        - bytes & 3-grams with segment count 0.981
    - XGboost (beta ver.)
        - bytes & bigrams with segment count 0.120

Lessons

spark.executor.memory

spark.driver.memory