Kaggle 'Microsoft Malware Classification Challenge' 3rd place solution

Mikhail Trofimov, Dmitry Ulyanov, Stanislav Semenov

May 5, 2015

1 Overview

We present our solution to the Microsoft Malware Classification Challenge. In this competition we were asked to classify malware executables into nine families based on their hex dump and IDA disassembled representation. Our team extracted numerous of features from both hex dump and IDA code and fit GBDT on it. We augmented the train set using test set and used simple averaging scheme in order to improve the result.

This file briefly describes the models and software used to build our solution.

2 Features extraction

First of all, create directories and change working directories

```
bash create_dirs.sh
cd src
```

Check 'set_up.py' for all paths. Now you are able to extract all the features by running

```
bash main.sh
cd ../
```

3 Main model

Check the installation of XGBoost and run all the code in 'learning-main-model.ipynb' to get submission file. It learns single GBDT-model which solely gives private score 0.00434.

4 Semi-supervised trick

For semi-supervised trick we need to get test labels for sampling. We obtain it by mixing main model (learned at previous stage) ang 4-gramms-only-based model. To learn it, run 'learning-4gr-only.ipynb'. After, run 'semi-supervised-trick.ipynb' for getting prediction.

5 Ensembling

Originally we use sophisticated scheme but find out that it takes a lot of efforts and gives a tiny improvement, so we decide to use simple weighting.

Run all the code in 'final-submission-builder.ipynb' to get final submission which scores 0.00401.

6 Dependencies

All the code has been run on Ubuntu 14.04.1 LTS

- python 2.7.9
- ipython 3.1.0
- sklearn 0.16.1
- numpy 1.9.2
- pandas 0.16.0
- hickle 1.1.1
- pypy 2.5.1 (with installed joblib 0.8.4)
- scipy 0.15.1
- xgboost-0.3

7 Hardware

We run this code on machine with 16 cores and 120 GB RAM.

The most memory-consuming part is processing 4-gramms. All the others will require no more than 32 GB RAM.