Malware Classification

Catherine - Parya Jandaghi, Vibodh Fenani, I-Huei Ho

Flow

- Feature Extraction
- Random Forest Classifier
- Significant Features Combination Testing

Feature Extraction

.bytes file

1. bytes file size
2. asm file size
3. bytes and asm file size ratio
4. unigram bytes (from bytes files)
5. bigram bytes (from bytes files) – Filtered top 2000 labels from each class

Feature Extraction

.asm files

1. segment (from asm files)
    - Detected the segments in each asm file
    - Recorded the counts in each document
    - Resulted in 257 different segments.
2. 2-4 grams opcode (from asm files)
    - Detected the opcodes in each asm file
    - Selected those opcodes appeared only in 1/3 documents
    - Generated 2, 3, 4 grams opcodes by selected opcodes
    - Selected the important features by random forest classifier
    - Recorded the counts of each opcodes in each document

Feature Integration and Team Work

- Parquet files :- For large files we choose to run our parts and save every output in parquet files which we could further ease during integration features
- Maintaining both document id and hash allowed us to maintain relative order and ensure that there was no loss of data

Random Forest Classifier

- Trees Number
- Maximum Depth
- Number of Features
- Type of Features (Sparse/Dense)

Results

Bytes Size	Asm Size	Size Ratio	Unigram	Bigram	Segment	Trees	Depth	Accuracy
	v					10	5	66.00%
					v	10	5	87.10%
	v				v	10	5	90.00%
v	v	v			v	10	5	93.16%
					v	50	25	94.85%
	v		v	v	v	10	5	96.03%
	v		v	v		10	5	96.14%
v	v	v	v		v	10	8	96.32%
	v		v		v	10	5	96.58%
			v		v	10	5	96.83%
v	v	v	v		v	25	10	97.75%
			v		v	25	10	97.94%
v	v	v	v		v	50	25	98.64%
			v		v	60	30	98.75%