Monthly Research

Effectiveness of unknown malwaere classification by logistic regression analysis

FFRI, Inc
http://www.ffri.jp

Ver 2.00.01

Malware Classification by Static Information

- Classifies malware from static information of executables
- As examples of information it uses
  - Name of sections
  - Dlls or APIs imported
  - File size
- Since malware often has structures or APIs which are rarely used by usual executables, the combination of these information allow us to classify malware.

Problems

- These features are used in various way including logistic regression analysis and used to classify malware we still do not know if features effective to a file set is still effective to unknown file set.
- Detection rate and false positive are also suspicious if they do not differ between learning file set and other files.

Investigation

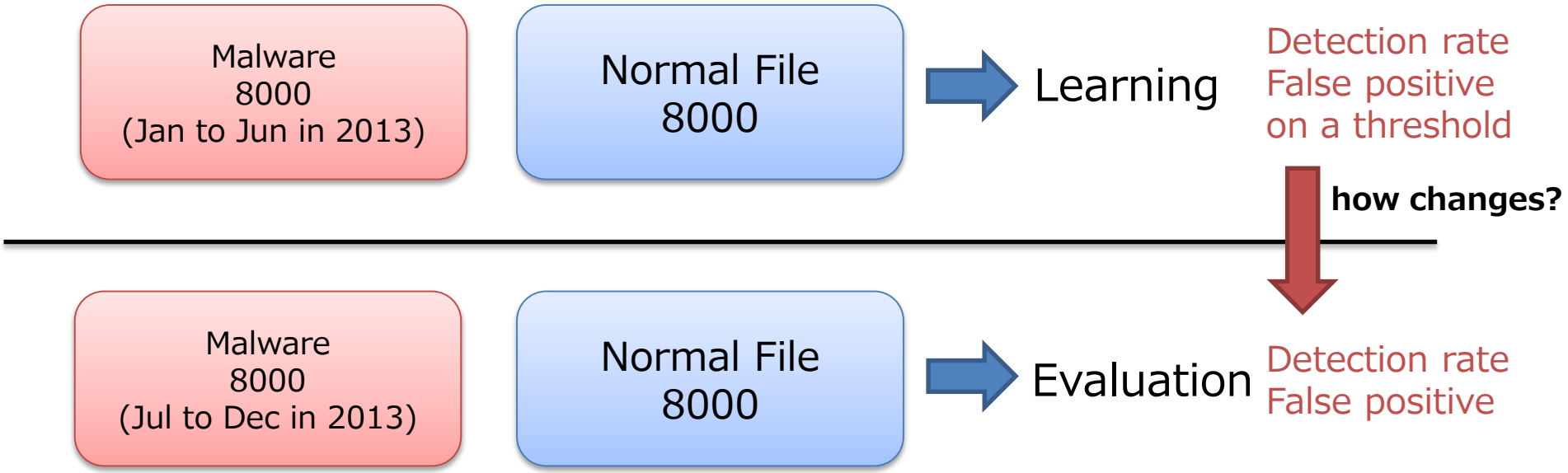
- Apply logistic regression analysis to static information of executables and find out how detection rate and false positive are.
- Investigate how the tendency of these rates differs to another file set.
- Especially for detection rate, it is important to see how the features collected from malware in a specific span and in a span after that are different.

Evaluation method

- Prepare 16000 malware
  - Randomly pick up 8000 from malware found from Jan to Jun in 2013
  - Randomly pick up 8000 from malware found from Jul to Dec in 2013
- Randomly pick up 16000 normal files
  - Divide it to two（8000 for each）

- Applying logistic regression analysis to one file set and obtain classification function. Then apply it to another file set.

Evaluation methods

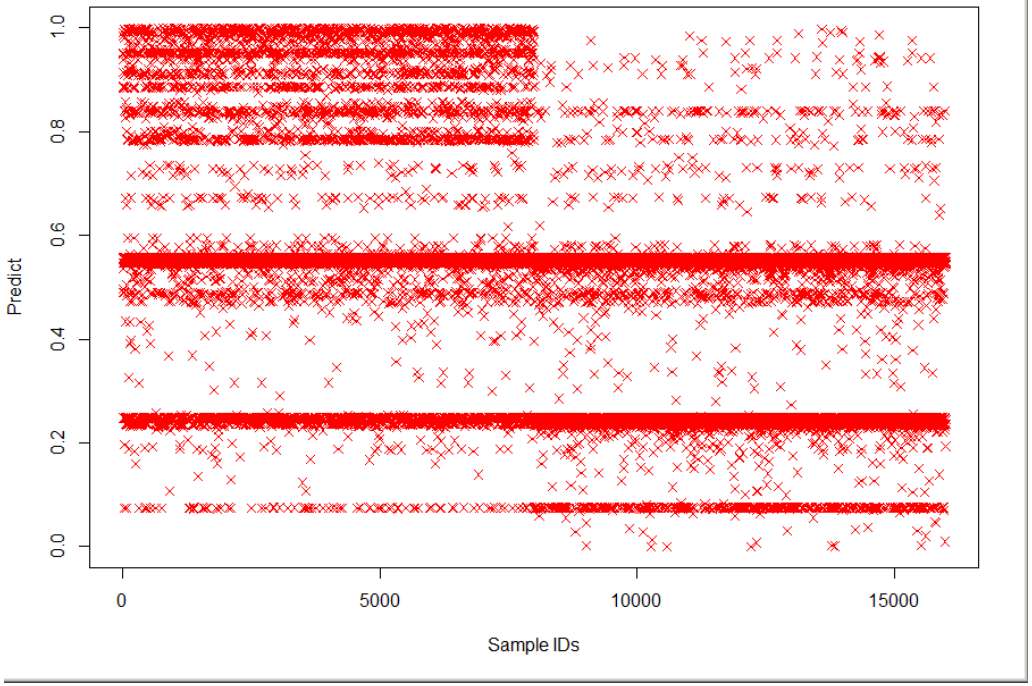


Features

- Extract features below
  - File size
  - Is packed? (0 or 1)
  - Is the packer UPX? (0 or 1)
  - Is a DLL? (0 or 1)
  - Is a driver? (0 or 1)
  - Is a VisualBasic application? (0 or 1)
  - Is a .Net application? (0 or 1)
  - Is a control panel application? (0 or 1)
  - Has GUI? (0 or 1)
  - Has invalid dos stub? (0 or 1)
  - Number of APIs often used by malware (8 at maximum)
  - Number of DLLs often used by malware (8 at maximum)

Result

- First, classify learning file set by applying logistic regression analysis
- The more it is closer to 1 the more likely it is a malware.
- The features we picked up gave us distinguishable difference between normal files and malware.

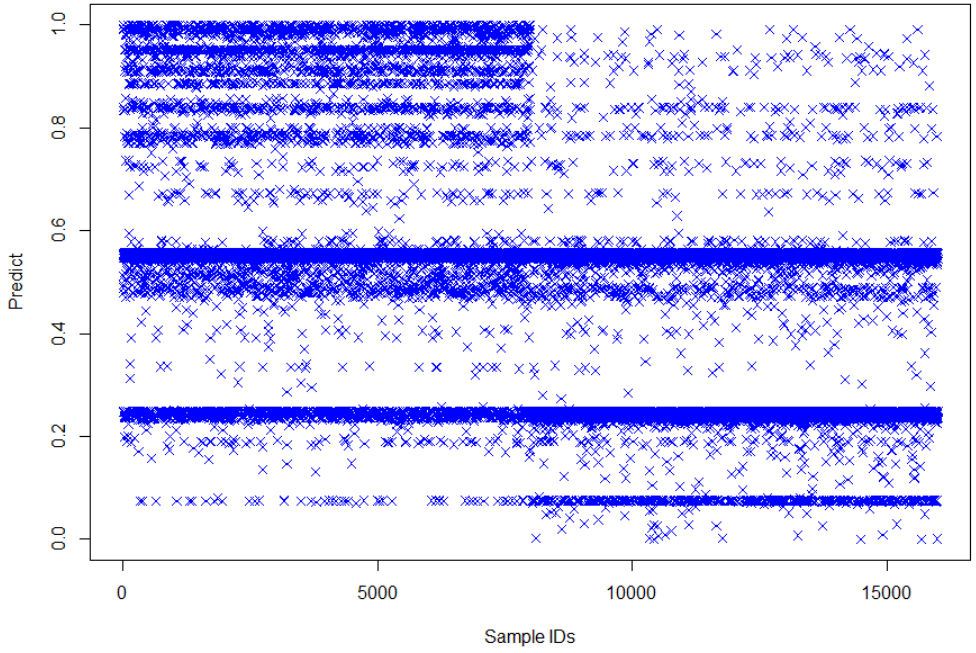


Sample ID

0 - 8000 : Malware(Jan – Jun 2013)
8001 - 16000 : Normal

Result

- Next, find out how the result for evaluation file set looks
- This also gave us the distinguishable difference between normal files and malware.
- It also has the similar result to the result of learning file set.

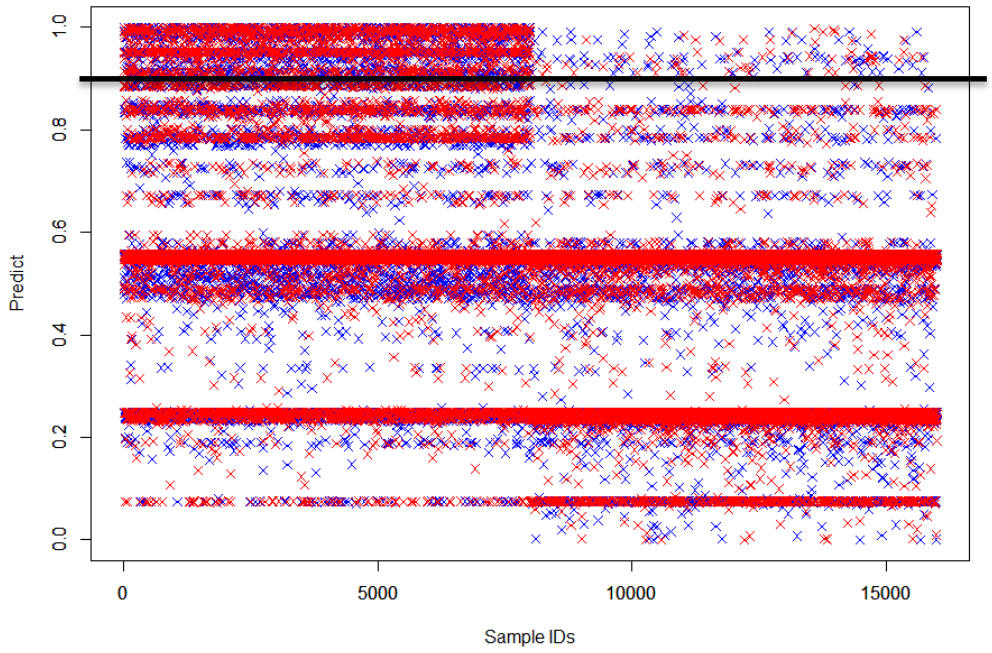


Sample ID

0 - 8000 : Malware(Jul – Dec 2013)
8001 - 16000 : Normal

Result

- From practical aspect we want to keep false positive rate less than 1.0%
- Put both results on top of each other and set the threshold to 0.9.

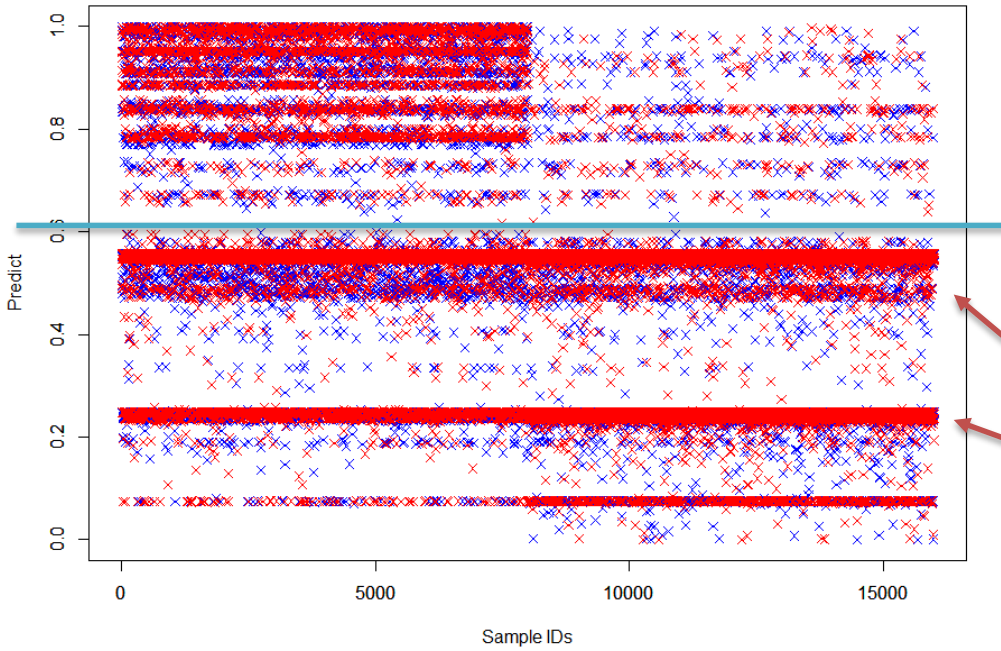


Threshold 0.9

	Detection rate	False positive rate
Learning	19.2%	0.825%
Evaluation	22.0%	1.13%

Consideration

- We can see that both results from learning file set and evaluation file set do not have big difference.
- By reducing threshold we can improve detection rate if more false positives are acceptable
- On the other hand, there are groups of files that can not be distinguishable from features we selected.



Threshold may be here.
(Needs to consult FP)

The file on this line can not be distinguishable.

Summary

- The methods and feature used this time gave us the same tendency from learning file set and evaluation file set
- Especially for malware, we found that the tendency are similar between malware from first half and latter half in 2013 (in terms of the features we selected)
- As future works, we should choose features, change the conversion of the values and find out the optimized method.
- Especially for the files than can not be classified from the features we need to investigate other features to classify them well.

Contact Information

E-Mail ：research—feedback@ffri.jp
Twitter：@FFRI_Research