Monthly Research

ロジスティック回帰分析による未知ファイル分類の有効性

株式会社ＦＦＲＩ
http://www.ffri.jp

Ver 2.00.01

静的情報によるマルウェア分類

- 実行ファイルの静的情報から、マルウェアを判別
- 情報としては、たとえば
  - セクション名
  - インポートしているDLLやAPI
  - ファイルサイズ
- 一般にマルウェアは通常ではなかなか存在しないような構造やAPIの利用をすることがあるため、これらの情報をうまく組み合わせることで、マルウェアを判別できる

問題点

- ロジスティック回帰を含め、さまざまな方法で特徴情報を加工し、分類することができるが、あるファイルセットで有効であると判定された特徴が他のファイル群でもどれくらい同様に有効なのかは分からない
- 検知率、誤検知率も同様に、学習したファイルとそうでないファイルでどれぐらいの違いがでるのか分からない

調査

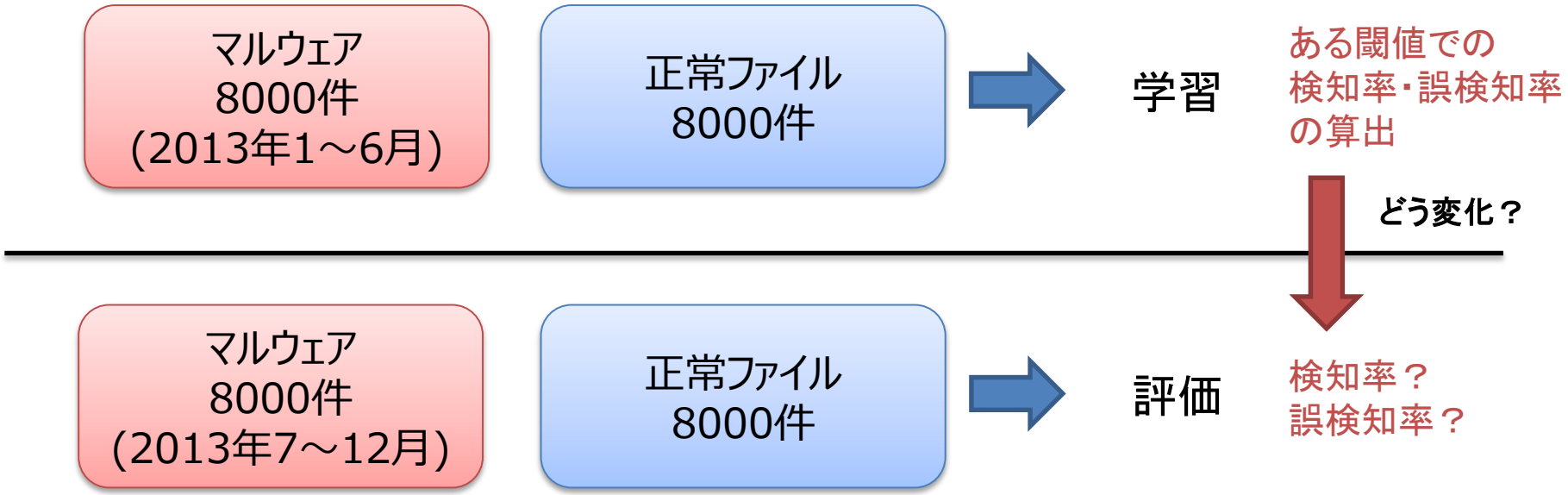
- 今回は、実行ファイルの静的情報を特徴として、ロジスティック回帰分析を行い、それぞれのどの程度の検知率、誤検知率が得られるかを調査
- 別のファイルセットで、その傾向がどの程度同じか（または異なるか）を調査
- 特に検知率については、ある期間に集められたマルウェアの特徴と、それより後の期間に得られるマルウェアの特徴がどれぐらい異なるかが重要と成る

評価方法

- マルウェアを16000件準備
  - 2013年1月～2013年6月までからランダムに8000件
  - 2013年7月～2013年12月までからランダムに8000件
- 正常ファイルをランダムに16000件準備
  - それらを２セットに分割（8000ずつ）

- ロジスティック解析で片方のセットで識別関数を取得。もう一つのセットに適用し、分布がどのように変化するかを確認

評価方法

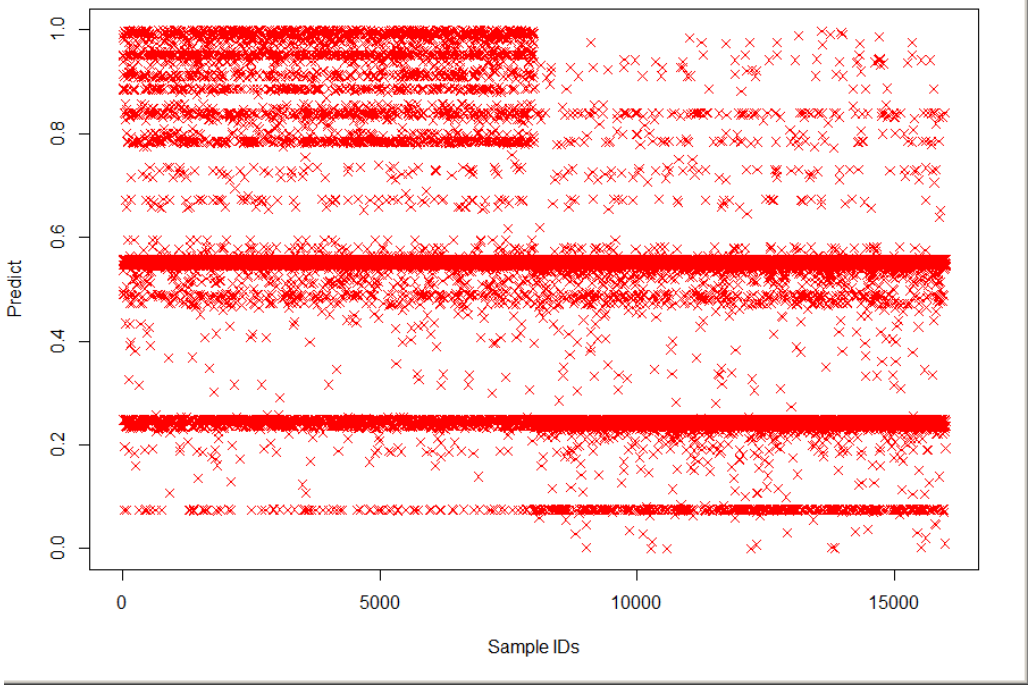

マルウェア
8000件
(2013年1～6月)
正常ファイル
8000件
学習
ある閾値での
検知率・誤検知率
の算出
どう変化？
マルウェア
8000件
(2013年7～12月)
正常ファイル
8000件
評価
検知率？
誤検知率？

特徴情報

- 以下の情報を抽出
  - ファイルサイズ
  - パックされているか (0 or 1)
  - パッカーがUPXであるか (0 or 1)
  - DLLであるか (0 or 1)
  - ドライバであるか (0 or 1)
  - VisualBasicアプリであるか (0 or 1)
  - .Netアプリであるか (0 or 1)
  - コントロールパネルアプリであるか(0 or 1)
  - GUIアプリであるか(0 or 1)
  - 不正なDOSスタブを持つか(0 or 1)
  - マルウェアが良く利用するAPIの利用数(最大 8)
  - マルウェアガ良く利用するDLLの利用数(最大 8)

結果

- まずロジスティック回帰分析で学習セットを分類
- 1に近いほどマルウェアらしいとする
- 今回の特徴情報でも、マルウェアと正常ファイルに違いが確認できる

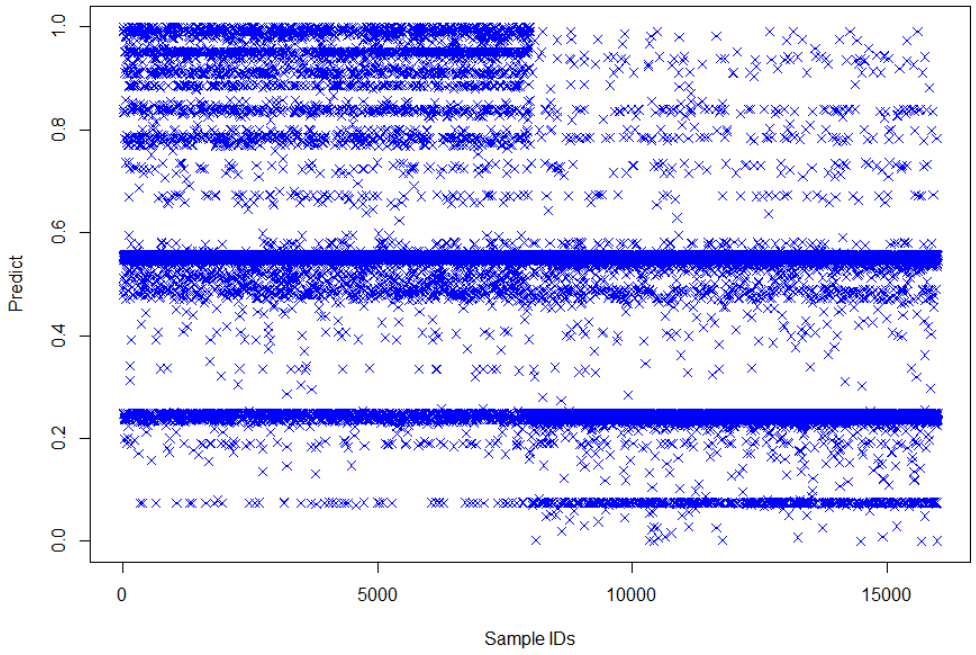


Sample ID

0～8000 = マルウェア(2013年1月～6月)
8001～16000 = 正常ファイル

結果

- 次に評価セットで同様にどのような結果になるか確認
- 同様にマルウェアと正常ファイルで違いが確認できる
- また、学習セットの結果と似た傾向が見える

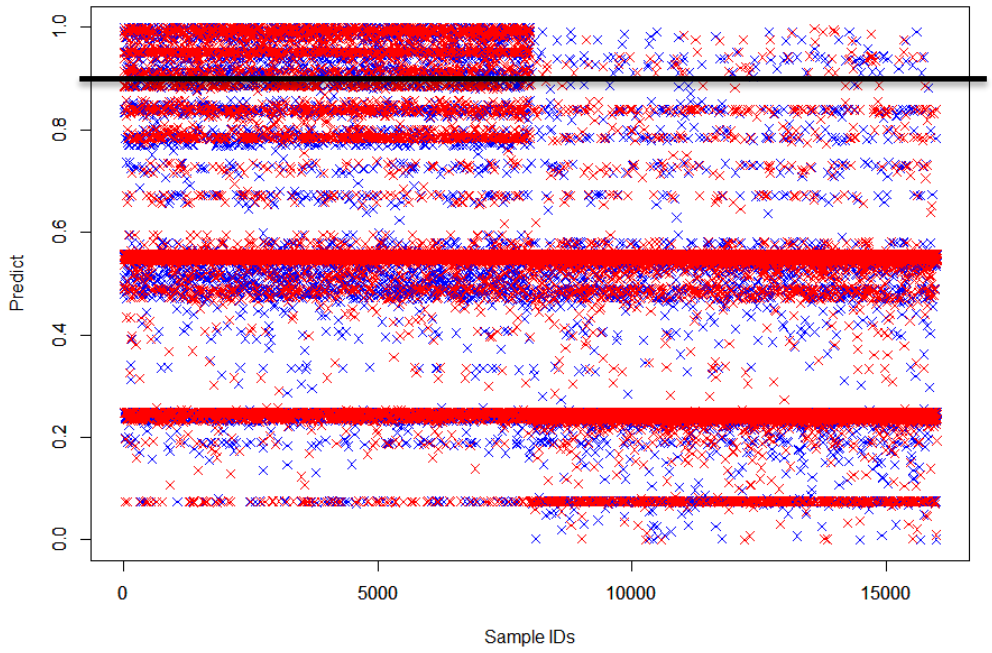


Sample ID

0～8000 = マルウェア(2013年7月～12月)
8001～16000 = 正常ファイル

結果

- 現実的な利用を考えると誤検知率は1%未満に抑えたい
- 両者を重ね合わせ、0.9以上の値を出したものをマルウェアとし、それ未満を正常ファイルと判定することとする

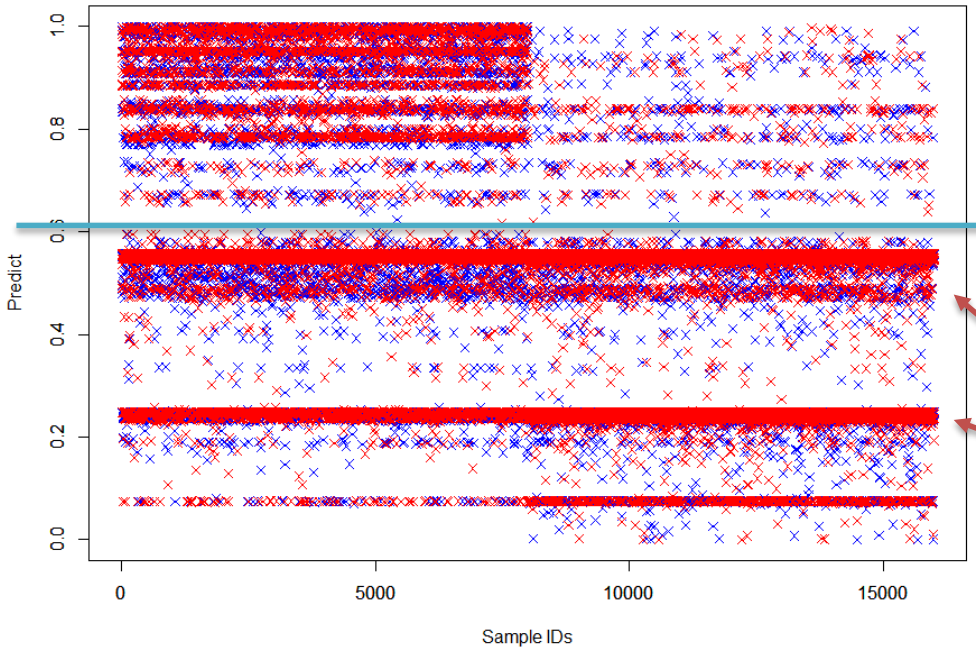


0.9を閾値とする

	マルウェア検知率	正常ファイル誤検知率
学習セット	19.2%	0.825%
評価セット	22.0%	1.13%

考察

- 評価セットにおいても大きく結果が異ならないことが分かった
- 誤検知率を許容できる場合には、より閾値を下げることで検知率を上げられる
- 一方で、今回利用した特徴からは判別できない一団が存在することも分かった



このあたりを閾値にしてもよい
（誤検知率の許容範囲と相談）

このラインにある一団は判別できない

まとめ

- 今回の手法、特徴では学習セットおよび評価セットにおいて同様の傾向が得られた
- 特にマルウェアに関しては、2013年の前半と後半で傾向が変わらないことが分かる（今回選んだ特徴情報に関して)
- 今後、特徴情報の取捨選択し、その変換方法を変えながら適用することで最適な判別方法を探る必要がある
- 特に、今回の特徴情報では判別ができないと思われるグループに属するファイルについて、それらをうまく分離する情報がないか探ることが重要になる

Contact Information

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