Monthly Research

静的情報に基づいたマルウェア判定指標の検討

株式会社ＦＦＲＩ
http://www.ffri.jp

Ver2.00.01

Agenda

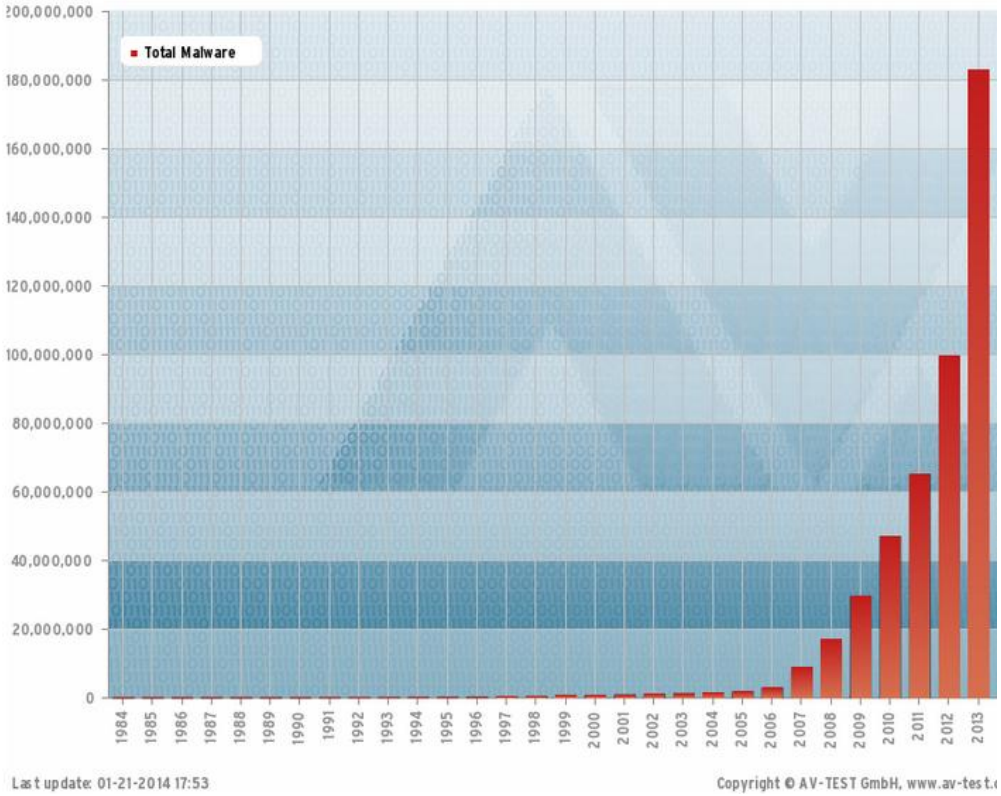
- 背景と目的
- 実験概要
- 実験結果
- 評価及び考察
- まとめ

背景と目的

- マルウェアの急増に伴い、従来のパターンマッチングによる検知が困難になっている
- そのため、事前情報がない状態でも有効に機能する検知技術が必要となっている
- その一手法として静的ヒューリスティック検知が提案、実用化されている
- 多くの場合、この設計・実装はマルウェア解析者等の専門家の知見に基づいて検知ロジックの研究開発が行われている
- 上記について定量的な指標に基づいて研究開発を行う方法として回帰分析を用いた手法について検討する
- 本書では上記手法の概要、手段、評価の観点等について整理を行う

背景と目的(補足)

- 近年マルウェアは指数関数的に増加している(2014/1時点)



http://www.av-test.org/en/statistics/malware/

背景と目的(補足)

- なぜ回帰分析か?
  - マルウェア判定に適用可能な手法（分類タスク）は他にも存在する
    - 決定木、ランダムフォレスト、ニューラルネットワーク、SVM等
  - マルウェア検知（情報セキュリティ）は、比較的誤りが許されない適用領域
  - 未知データによる誤り（誤検出）の危険性
  - 誤り発生時の原因究明、説明責任、継続的な改善可能性が必要
    - ブラックボックスの手法では対応できない場合が多い
  - この観点において回帰分析は有力な一手法となり得る
    （検知ロジック自体よりかはその研究開発に利用）

実験概要

- 分析の目的は下記の通り
  - どの特徴がどの程度マルウェア、正常系の判別に貢献するのか把握する
  - どの特徴の組み合わせが最適なのか把握する
- FFRI所有のファイルセットよりマルウェア、正常系ファイルをそれぞれ5,000件ずつ無作為に抽出
- 上記に対して「Attributes of Malicious Files」にてレポートされている特徴を材料に分析を実施（同レポートの概要は後述）
  - (SANS Institute InfoSec Reading Room) https://www.sans.org/reading-room/whitepapers/malicious/attributes-malicious-files-33979
- ロジスティック回帰を利用して上記の分析を実施
- 実験には下記のツールを利用
  - R 3.0.2, python, pefile-1.2.10-139 (http://code.google.com/p/pefile/)

「Attributes of Malicious Files」の概要(1/2)

- マルウェアを2,500,000件、正常系ファイルを65,000件用意
- これらに対して主にPEヘッダー中の様々なフィールド値の出現傾向を調査し、下記を分析、レポート
  - マルウェアに多く見られるフィールド値の傾向
  - 上記傾向に基づいた検知ルール
  - 検知ルールを上記ファイル群に適用した場合の検出率・誤検出率
- 例)
  - マルウェアのファイル群には、PEヘッダ中のFILE_HEADER.TimeDateStamp値を意図的に改変し、「1992年以前また未来の日時(※1)」に設定したものが存在する
  - これに基づいて下記検知ルールを作成し、評価した結果は下記の通り(※2)

年	該当正常ファイル	該当マルウェア	差分
< 1992	0.01%	11.72%	11.71%
1992-2012	99.98%	87.93%	-
>2012	0.00%	0.35%	0.35%

※1 当該レポートは2012年公開
※2 元のレポート内容に基づいて作成

「Attributes of Malicious Files」の概要(2/2)

- 総括にて右記の検知ルールを提案
（計28項目）

- 独立した検知ルール及びその評価結果であり、それぞれを組み合わせた結果については言及されていない

	Detection Rule	Detection Rate	False Positive
FILE HEADER	*Year* < 1992 or *Year* > 2012	12.05%	0.35%
	NumberOfSections < 1 or *NumberOfSections* >9	3.64%	0.87%
	PtrToSymTable > 0	1.20%	0.17%
	Characteristics (BYTE_RESERVED_LO=1)	14.99%	0.29%
	Characteristics (BYTE_RESERVED_HI=1)	14.98%	0.26%
	Characteristics (RELOCS_STRIPPED=1)	14.99%	0.29%
OPTIONAL HEADER	*MajorLinkerVersion:MinorLinkerVersion* ∉ H1 (Table 3.2.1)	14.23%	0.41%
	MajorOSVersion:MinorOSVersion ∉ H2 (Table 3.2.1)	6.32%	0.26%
	MajorImageVersion:MinorImageVersion ∉ H3 (Table 3.2.1)	4.78%	0.34%
	SizeOfCode /Sample Size >1	6.36%	0.06%
	SizeOfInitializedData / Sample Size >3	3.58%	0.38%
	SizeOfUninitializedData / Sample Size >1	13.63%	0.23%
	SizeOfImage / Size > 8	5.80%	0.92%
	SizeOfHeaders / Sample Size >0	2.03%	0.04%
	AddressOfEntryPoint / Samples Size >2	12.73%	0.35%
	BaseOfCode / Samples Size >2	4.90%	0.10%
	BaseOfData / Samples Size >4	4.76%	0.05%
	NumberOfRvaAndSizes != 16	2.16%	0%
SECTIONS	*Raw Size* = 0	13.13%	0.62%
	Virtual Size / Raw Size > 10	3.22%	0.71%
	PtrToLineNumber != 0	1.58%	0.02%
	Characteristics (IMAGE_SCN_CNT_UNINITIALIZED_DATA=1)	9.65%	0.46%
	Characteristics (IMAGE_SCN_MEM_SHARED=1)	4.95%	0.23%
	Section Entropy < 1	22.78%	1.13%
	Section Entropy > 7	21.52%	0.96%
	File Entropy > 6.9	56.18%	3.12%
RSRC	*Sub-Language* = 0	36.66%	0.85%
	Resource Size / Sample Size > 0.25	1.05%	0.25%

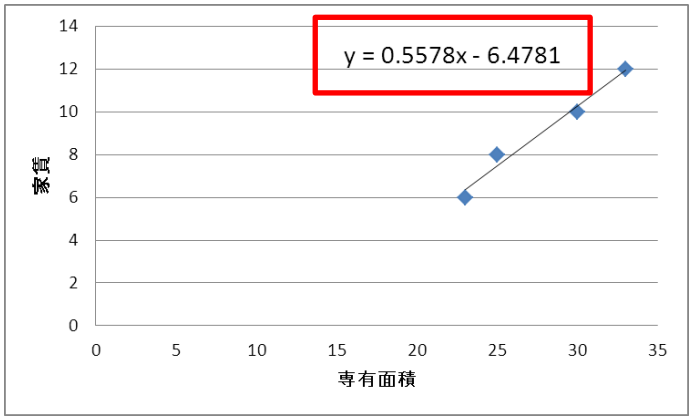
出典：https://www.sans.org/reading-room/whitepapers/malicious/attributes-malicious-files-33979

回帰分析の概要

- 目的変数と説明変数の間の関係式を統計的手法に基づいて推定

- 例えば、家賃（目的変数）、専有面積（説明変数）について下記のデータが存在した場合、回帰分析を行うことで「y.家賃 = a.系数 * x.専有面積 + b.切片」における係数a、切片bを求めることができる

  - データ1:家賃6万円、専有面積23㎡
  - データ2:家賃8万円、専有面積25㎡
  - データ3:家賃10万円、専有面積30㎡
  - データ4:家賃12万円、専有面積33㎡



- 説明変数が複数のケース、非線形のケースに関する分析手法も存在

ロジスティック回帰の概要

- 非線形回帰分析の一手法

- 主に目的変数が質的変数の場合に利用される
  - 例）検査1～検査Nの結果から将来がん発病するか否か予測する
  - 目的変数：がん発病するか否か（0か1か、その確率）
  - 説明変数：検査1～検査Nの結果

- 同様のアプローチを適用し、「Attributes of Malicious Files」で紹介されている PEヘッダーのフィールド値を説明変数として利用し、マルウェアか否かを推定する

ロジスティック回帰に係る検討事項

- 前処理
  - 説明変数の選定
    - 基本的には専門家の知見に基づいて選択
    - 今回は件のレポートに倣う
  - 説明変数の加工
    - 同上（でなければ専門家の知見または統計分析により試行）
- 分析
  - 最適な説明変数の組み合わせ（説明変数の投入方法）
  - 交互作用（説明変数Ａの効果が説明変数Ｂの値に影響を受けて異なる）
    - 今回は簡便のため各説明変数を独立して扱う

- 分析結果
  - 有意確率
  - オッズ比とその信頼区間の評価
  - モデルの適合度
  - モデルの評価

説明変数の加工

- 回帰分析においては非常に重要
  - 例：年齢として（11, 20, 25, 33, 60, 42）と言うデータがあった場合
    - 即値として利用する(11, 20, 25, 33, 60, 42)
    - 10才単位に丸める（10, 20, 20, 30, 60, 40）
    - 40才以上か否かで二値化(0, 0, 0, 0, 1, 1)

- 一般的にはどういった手法が最適化は専門家の知見に依るところが大きい
  - 長期間に渡る試行錯誤の集積であり比較的公開され難い
  - 適用ドメイン、データに応じて最適な手法は異なる

- 今回は、件のレポートの検知ルールに従い値を二値化（ダミー変数化）
  - Detection Ruleに非マッチ：0
  - Detection Ruleにマッチ：1

最適な説明変数の組み合わせ（説明変数の投入方法）

- 全変数を投入し、スワップワイズ法により最適な組み合わせを模索
  ※R上でstep関数を利用

- 最適なモデルの基準、判定法
  - AIC(赤池情報量基準)
  - モデルの優良性を示す基準
  - 分析対象データに過剰に適応していないかの判断材料
  - 詳細は、割愛するがAIC値が小さい程、優良なモデルと考えることができる
    - http://ja.wikipedia.org/wiki/%E8%B5%A4%E6%B1%A0%E6%83%85%E5%A0%B1%E9%87%8F%E5%9F%BA%E6%BA%96

有意確率 / オッズ比とその信頼区間

- 有意確率
  - その結果が偶然である確率
  - 一般に5%(0.05)未満の場合、有意（偶然ではない）と判断

- オッズ比
  - 説明変数と目的変数の関連性の強さを示す尺度
  - 説明変数の種類によって解釈が異なる
    - 質的変数の場合（例：血液型）
      - 一つの変数を基準とし、他の場合における倍率を数値化
      - 例） A型を基準にすると、 B型は2倍～である
    - 量的変数の場合（例：年代/10代、20代、30代）
      - 他の条件が同一の場合、その変数が1増加した際の倍率を数値化
      - 例）10代から20代になると～は2倍になる
  - 1.0を超える場合は有意、1.0未満であれば有意ではないと判断
  - オッズ比も誤差を含むためＮ％の信頼度を以て取り得る範囲が「N%信頼区間」

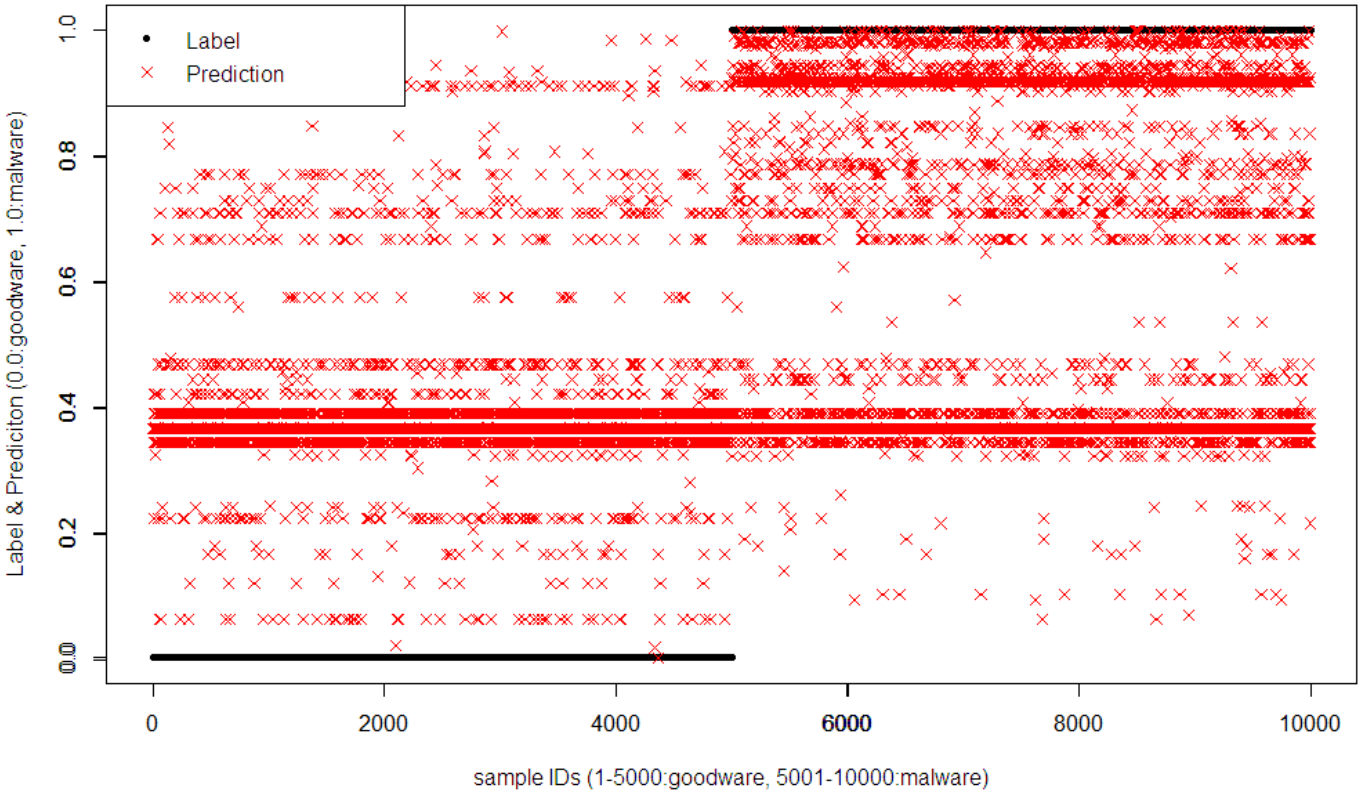
実験結果

- 有意確率が高い変数上位3件のみ抜粋
  - 二値のため非マッチ(0)に対してマッチ(1)した場合のオッズ比に注目
  - TimeDateStamp，SECTION_entropyに関する検知ルールは有意
    （マッチしたか否かでマルウェアらしさが大きく異なる）
  - ImageVersionは、オッズ比が1.0を下回っており有意ではない
    （マッチ有無に関わらずマルウェアらしさへの相関性が低い）

説明変数 / 割当値		オッズ比（95%信頼区間)	有意確率
TimeDateStamp	0	(Reference)	-
	1	19.5 (16.1 - 23.9)	<2E-16
SECTION_entropy	0	(Reference)	-
	1	4.18(3.48 - 5.05)	<2E-16
ImageVersion	0	(Reference)	-
	1	0.174(0.123 – 0.241)	<2E-16

実験結果

- 正解ラベルとモデルによる予測値の対比
  - 横軸：サンプルID(1-5,000:正常系、5,001-10,000:マルウェア)
  - 縦軸：正常系(0.0) – マルウェア(1.0)らしさ

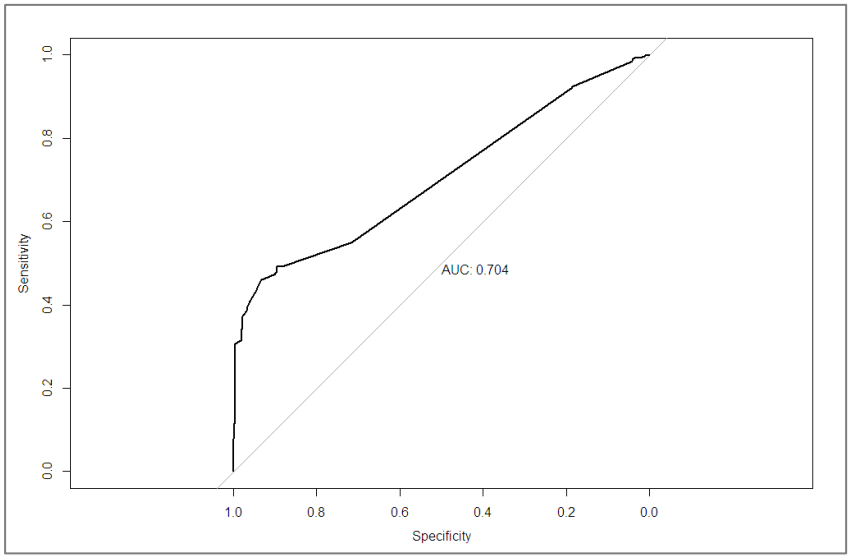


評価及び考察(1/2)

- 前述の3変数以外についても有意確率が5%未満のものに着目し、そのオッズ比を確認することでその変数・検知ルールが有意か判断可能
  - 有意確率が5%未満にも関わらずオッズ比が1.0未満の変数は、加工方法を変更する、その変数自体を除外する等のチューニングを検討

- モデルの適合度
  - 分析対象データに対してモデルがどの程度適合しているかという尺度

  - AUC(Area Under the Curve)
    - 0.0～1.0で数値化
    - 完全な分類では1.0
    - ランダムな分類では0.5
    - 今回の結果は、0.704



ROC曲線とAUC

評価及び考察(2/2)

- 生成されたモデルの評価方法
  - 分析対象データを利用した評価（内的妥当性）
  - 分析対象データとは異なるデータを利用した評価（外的妥当性）

- 今回は、K-fold cross validation(※)により内的妥当性のみ評価(K=13)
  - 全データを13ブロックに分割し、内12セットでモデル構築、残り1セットを評価
  - これを全13ケース繰り返し、その平均から誤差を推定
  - 結果：19.8%（推定誤差）

- 上記の適合度、推定誤差等の数値が改善されることを確認しながら
  モデルのチューニングを行うことが重要

まとめ

- 静的ヒューリスティック検知の研究開発での利用を目的としてマルウェア及び正常系ファイルの静的情報に着目
- 様々な静的情報を変数としてロジスティック回帰を行うことで、どの変数がどの程度効果的か、どの組み合わせが有効か等について定量的に把握することができる
- これを指標に継続的な検知ロジックの改善等が可能と考えられる

Contact Information

- E-Mail: research-feedback@ffri.jp
- twitter: @FFRI_Research