講演番号：87

機械学習の適用によるマルウェアの機能予測

A－7　　Estimating Malware Function by Machine Learning

久保 宏樹† 笠間 貴弘†† 宮保 憲治†
Hiroki Kubo† Takahiro Kasama†† Noriharu Miyaho†
† 東京電機大学 情報環境学専攻 †† 国立研究開発法人情報通信研究機構
† Major in Information Environment, Tokyo Denki University. †† National Institute of Information and Communications Technology

1. はじめに

近年の新型マルウェアの増加によって,そのマルウェアの挙動や目的を明らかにする解析者の負担が増加している.本稿では,機械学習手法の一つである多クラス分類を行う際に使用される,一対他分類によってマルウェアの機能を予測する方法を提案した.さらに複数の特徴量と機械学習アルゴリズムによって作成された分類器の比較実験を行い,機能予測に効果的な特徴量とアルゴリズムの組み合わせを明らかにした.

2. 一対他分類

一対他分類とは 2 クラス分類を行う機械学習アルゴリズムを多クラス分類に適用するための手法である.2 クラス分類器を複数組み合わせクラスの予測を行う.予測する機能が C 個ある場合を考える.i=1,2,...,C としたとき、i 番目の分類器は機能 i を持つマルウェアを 1,i の機能を持っていないマルウェアを0として予測するように学習する.すなわち各々の機能の予測に特化した 2 クラス分類器が C 個作成される.あるマルウェアの機能を予測する際には,C 個の分類器すべてに同一の入力をし,出力が 1 になった分類器に対応する機能を予測結果とする.

3. 特徴抽出

マルウェア作成者は,マルウェアの機能を実装する際に既存のコードを再利用する場合がある.すなわち異なったマルウェア間でも一部バイナリや挙動に類似性が生まれる.そのため機械学習によって機能の予測が可能であると考えた.本稿ではマルウェアから動的解析によって得た各 API の呼び出し回数.静的解析によって得た機械語命令の出現頻度.マルウェアバイナリをグレースケールへ画像化したときの GIST 記述子[1]を特徴量とした.

4. ラベリング

ラベリングのため,マルウェアの機能を判断する.対象のマルウェアがアンチウイルスソフトで検知された際の名前をアンチウイルスベンダーの脅威情報データベースで検索し,対象のマルウェアの持つ機能を判断した.

5. 比較実験

実験に使用するアンチウイルベンダーによって提供されているマルウェアデータセットに対してラベリングを行った結果を表 1 に示す. ラベリングを行ったデータセットに対し特徴抽出を行い,機械学習アルゴリズムであるランダムフォレスト(RF),ロジスティック回帰(LR),SVM でそれぞれ学習させることで作成された分類器を比較した.評価基準には不均衡データによる学習を行った際に用いられる,再現率と適合率の調和平均である F 値を使用した.機能ごとに最も F 値が高くなった分類器を作成した特徴抽出方法とアルゴリズムの組み合わせを表 2 に示す.

表 1 ラベリングの結果

機能	ラベル 0	ラベル 1
adware	7460	204
backdoor	5345	2319
bot	5539	2125
downloader	5681	1983
dropper	7380	284
ransomware	5168	2496
spyware	4990	2674
virus	7616	48
worm	7203	461

表 2　効果的な機械学習アルゴリズムと特徴抽出方法

	特徴抽出法	アルゴリズム	F 値
adware	動的解析	RF	0.73
backdoor	静的解析	RF	0.80
bot	静的解析	RF	0.75
downloader	動的解析	RF	0.76
dropper	動的解析	RF	0.71
ransomware	静的解析	RF	0.91
spyware	静的解析	RF	0.79
virus	静的解析	RF	0.94
worm	静的解析	RF	0.88

6. 評価

表 2 から ransomware,virus,worm 分類器の F 値が比較的高くなった.マルウェアは活動するために WindowsAPI を使用する.例えばransomwareには感染端末のファイルを暗号化する際に使用される「CryptEncrypt」,virus には正常ファイルを悪性の動作をするように改ざんする「WriteFile」,worm にはファイルの複製を行う「CopyFile」などが使用されている. ransomware,virus,worm の機能を持つマルウェアにおいて,WindowsAPI を呼び出す際の引数をスタックに格納する push 命令や,呼び出しを行う call 命令などの出現回数が類似していたと考えられる.そのため該当の機能を予測する分類器の F 値が高くなった.

7. まとめ

本稿では一対他分類によってマルウェアの機能予測を行う方法を提案し,複数の特徴抽出と機械学習アルゴリズムを使用して比較実験を行った際に最も F 値が高くなる組み合わせを明らかにした.

参考文献

[1] Aude Oliva ,Antonio Torralba “Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope” International Journal of Computer Vision Volume