4Z-4

データマイニング技術を活用したマルウェア分類法の検討

鈴木　貴之　　宮保　憲治
東京電機大学　情報環境学部

1.はじめに

近年，マルウェアによるサイバー攻撃の被害が増加し，様々な社会問題を引き起こしている．マルウェアを容易に作成出来るツールキットの流通により，マルウェアは日々爆発的に増加し続けており，2013 年第 4 半期で 2000 万件以上の新たなマルウェアが発見されている[1]．この状況の中で，Cuckoo Sandbox[2]などの，マルウェアを自動で動的解析するサンドボックスツールも利用され始めている．今後は,解析された情報を活用することにより，新たに発見されるマルウェアの効果的な推定が可能になると思われる．マルウェアの種類の判別が容易化できれば，詳細な解析作業の効率化にも貢献できる可能性がある．新種のマルウェアの分類手法については,既知のマルウェアとの類似性を比較して分類を行う手法[3]や，教師あり学習を用いる識別手法の一つである SVM(Support Vector Machine)を用いた手法[4]があるが，分類ラベルにアンチウイルスソフト(AV ソフト)のマルウェア名称を利用しているのが現状である．しかしながら，マルウェアの名称規則は AV ソフト間で統一化されておらず，マルウェアに対して AV ソフト毎に違う種類，もしくはファミリー名を命名しているケースが報告されているレベルに留まっている[5]．

本稿では，SVM を用い，マルウェアに対して独自に決定した名称をラベルとして用いる，新しい分類手法を提案する．

2.使用データ及びデータの前処理

以下に，分類に利用するデータの詳細と SVM に用いる際の前処理の内容を述べる．

2.1 利用データ概要

マルウェアの動的解析結果としては，MWS 2013 データセット[6]の一つである，FFRI 社が提供する FFRI Dataset 2013[7]を用いた．本データセットは，FFRI 社が収集した PE 形式かつ実行可能ファイルである 2644 検体を，Cuckoo Sandbox を用いて 1 検体あたり 90 秒間，実行した際の動的解析ログである．動的解析ログには，Cuckoo sandbox と VirusTotal との提携機能により，1 検体で最大 66 種類のアンチウイルスソフトのスキャン結果が記録されている．

2.2 名称の作成と割り振り

本分類手法で用いる独自の名称を決定するために，上述したスキャン結果の中に含まれるマルウェアの名称を分析し，マルウェアの特徴を表す単語を抽出した．マルウェアの名称は，記号や空白文字などで区切られているので，正規表現で単語を抽出し，単語別の頻度を算出した．W32，Win32 といった表記揺れは適宜修正した．その結果，表 1 に示す単語 5 つがマルウェアの特徴を表すことが判明し，比較的,出現頻度が高いため，判別・分類用の名称として採用し，ラベル名称として割り振る処理を実施した．

表 1：マルウェアの特徴を表した単語

Adware
Backdoor
Trojan
Virus
Worm

2.3 特徴量の決定

マルウェアの分類における関連研究では Win32 API の関数名を特徴量とし[4]，近年では，API の関数名及び引数の値が重要であることが報告されている[8]．これらの状況から，本稿では，動的解析ログに含まれる API 関数名及び引数の値に加え，検体が利用したインポート API の関数名を特徴量として抽出し，Bag of Words モデルで特徴ベクトル化を行った．インポート API 関数名を採用した理由は，後述する実験において，インポート API 名を特徴量として採用することにより，識別率が向上したためである．

3.SVM を用いたマルウェア分類実験

2.3 で規定した特徴量を抽出したサンプルに対して，2.2 で決定した名称を分類ラベル名としたサンプルと，Kaspersky の名称を分類ラベル名としたサンプルとの間で，識別精度を比較する実験を行った．比較対象を Kaspersky の名称による分類ラベルとした理由は，FFRI Dataset 2013 内のスキャン結果の中では，最も高い頻度でマルウェアの検知ができたからである．サンプル抽出の際，破損していたログをサンプルから除

Study of malware classification methods by using data mining technology
Takayuki Suzuki, Noriharu Miyaho
School of Information Environment, Tokyo Denki University

外し，更に Kaspersky の分類ラベル作成において，クラスに属する個体数が 2 個体未満のクラスと，そのクラスに該当するサンプルを除外することとした．識別精度の検証に当たっては，パラメータ調整を行ったサンプルを 10 分割交差検証法で評価した．実験諸元を表 2 に示す．

表 2:実験諸元

識別手法	SVM(libsvm)
使用データセット	FFRI Dataset 2013
サンプル数	2007(独自名称ラベル)
	2088(Kaspersky 名称ラベル)

4. 実験結果と考察

独自に決定した分類ラベル名と，Kaspersky の分類ラベル名での識別結果を表 3 に示す．

表 3：分類実験結果

採用ラベル	独自名称ラベル	Kaspersky 名称ラベル
TP Rate	89.9%	72.8%

TP(True Positive) Rate は分類ラベルに正しく分類された割合を表す．表 3 より独自に決定した分類ラベルでは 90%近くの識別率が得られ，AV ソフトの名称による分類法よりも優れた識別率を達成できたが，1 割ほどの誤識別が確認された．分類ラベル別の識別率を図 1 に，図 1 の識別結果を混合行列で表した結果を図 2 に示す．

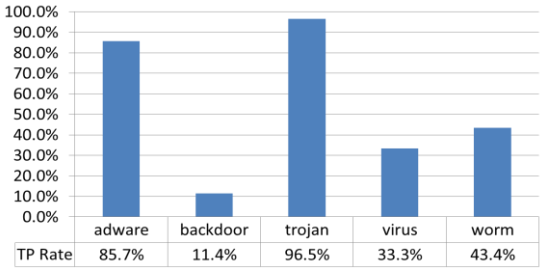

図 1:クラス別の識別率

adware	backdoor	trojan	virus	worm	
66	0	11	0	0	adware
0	4	31	0	0	backdoor
9	3	1734	5	46	trojan
1	0	22	12	1	virus
0	0	81	0	62	worm

図 2:クラス別の分類結果の混合行列

図 1 より，Adware は 85%以上，Trojan は 95%以上の識別率を持つが，Virus，Worm は 50%以下の識別率，backdoor においては約 10%の識別率であることが判明した．図 2 に示す誤識別の要因は Trojan へ誤分類された結果と推認できる．この理由はマルウェアの振る舞い及び機能が，backdoor，Virus，Worm と Trojan で重複した部分が存在するため誤識別に至ったと考えられる．

この検証を行うため，分類クラスから Trojan を除いて再分類した．この結果を図 3,4 に示す．

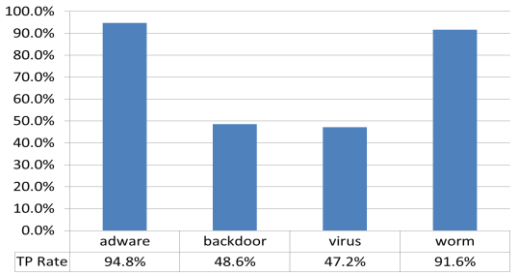

図 3:trojan を除く分類によるクラス別の識別率

adware	backdoor	virus	worm	
73	1	1	2	adware
5	17	3	10	Backdoor
6	3	17	10	virus
3	5	4	131	worm

図 4：Trojan を除く分類結果の混合行列

図 4 から，図 2 では識別率の低い backdoor，Virus，Worm の識別率が向上することが判明し，上記の仮定が正しいことが検証できた．以上のことから，マルウェアの振る舞い及び機能に着目したラベルを付与する事によって，識別精度が更に向上できる事が判明した．

5. まとめ

独自に命名した名称ラベルでの分類精度を比較・評価し，有効性を検証した．今後は API 関連以外のログに記録された内容に着目して，検体の細かい振る舞いや機能に着目した名称による分類実験を行う予定である．

参考文献

[1] McAfee 脅威レポート:2013 年第 3 四半期 http://b2b-download.mcafee.com/products/japan/pdf/threatreport/threatreport13q3.pdf
[2] Cuckoo Sandbox http://www.cuckoosandbox.org/
[3] 堀合啓一，今泉隆文，田中英彦，マルウェア亜種の動的挙動を利用した自動分類手法の提案と実装，情報処理学会論文誌，Vol50 No4 1321-1333(2009-04)
[4] Konrad Rieck, Thorsten Holz, Carsten Willems, Patrick Dussel, and Pavel Laskov, ”Learning and Classification of Malware Behavior”, DIMVA-2008(2008)
[5] 星澤裕二，太刀川剛，山村元昭，マルウェアの亜種等の自動化，IEICE Technical Reort，ISEC2007-54(2007-07)
[6] 神園雅紀他，マルウェア対策のための研究用データセット～MWS 2013 Dataset(2013-10)，http://www.iwsec.org/mws/2013/about.html
[7] FFRI Dataset 2013, http://www.iwsec.org/mws/2013/files/FFRI_Dataset_2013.pdf
[8] 藤野朗稚，森達哉，自動化されたマルウェア動的解析システムで収集大量 API コールログの分析，computer security symposium 2013(2013-10)