5ZA-06

機械学習による分類に向けたマルウェア特徴量抽出手法の検討

厚地紗江香†　　平川 豊‡
†芝浦工業大学大学院理工学研究科　‡芝浦工業大学工学部情報工学科

１．研究の背景と目的

近年、社会の情報化が推し進められ、様々な情報がインターネットを介して共有される機会が多くなり、サイバー攻撃、特にそれに用いられるマルウェア（悪意のあるソフトウェア）の脅威と被害は年々増加している。このことから、従来のシグネチャ（攻撃を識別するルール）方式によるマルウェア解析はコストが高くなっている。そのため、より効率的な手法が求められる。現在は、機械学習を用いたマルウェア解析手法が活発に研究されている。

機械学習には入力となる特徴量（説明変数）と出力となる予測値（目的変数）が必要となり、特に特徴量の選択はモデルの精度を左右する。先行研究には、API（アプリケーションインターフェース）呼び出し列からマルウェアの特徴量を抽出する手法が多くある。しかし、API の順序関係に着目して特徴抽出を行っている研究は少ない。API の順序関係はマルウェアの処理プロセスを表現し、挙動を定義する上では重要なポイントである。

そこで本研究では、API 呼び出し列の順序関係に着目し、これを利用した特徴量抽出手法を検討する。

２．関連研究

既存研究[1]では、自然言語処理で多く活用されている n-gram、Doc2Vec、TF-IDF などを用いてマルウェアの API 呼び出し列をベクトルに変換後、機械学習モデルであるサポートベクタマシン(SVM)、ランダムフォレスト(RF)、K 近傍法(KNN)、多層パーセプトロン(MLP)のそれぞれに特徴を学習させることでマルウェアの分類モデルを構築し、分類精度を比較している。

既存研究[2]では深層学習を用いて複数のマルウェア分類手法を提案している。深層学習の代表的な手法である Recurrent Neural Network(RNN)と Echo state networks(ESN)のデータに対して時系的に更新される隠れ層のパラメータ（以降状態ベクトルと呼ぶ）を利用している。その１つとして、API 呼び出し列の中間時点の状態ベクトルと最終時点の状態ベクトルを結合したものをマルウェアの特徴量とする二分割手法を提案しており、抽出後はロジスティック回帰と多層パーセプトロンを用いて分類モデルを構築している。

Method for Feature Extraction toward Malware Classification Using Machine Learning
†Saeka Atsuchi,　‡Yutaka Hirakawa
†Electrical Engineering and Computer Science, Shibaura Institute of Technology, Tokyo, Japan
‡Computer Science and Engineering, Shibaura Institute of Technology, Tokyo, Japan

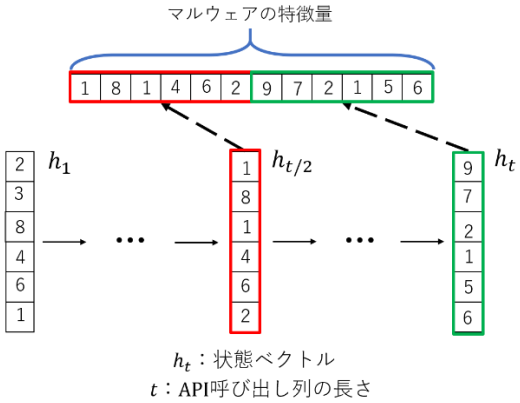

図 1.　二分割手法の概略図

３．検討手法

３－１．手法概要

既存研究[1]では API の出現頻度や出現確率を考慮しているが、その順序を反映した特徴量ではない。既存研究[2]では時系列データの学習に特化した RNN と ESN が利用されたが、この２つは時系列データの長期的な依存関係を出力に反映できない。よって API 呼び出し列が長い場合、正確な特徴量を抽出することが出来ない。

そこで、長期的な依存関係を記憶することのできる Long Short Term Memory(LSTM) の状態ベクトルを活用することで特徴量を抽出する。また既存研究[2]の二分割手法が中間と最終の状態ベクトルのみで妥当なのかを、新たに三分割手法を実装することで比較検討を行う。

３－２．Long Short Term Memory

Long Short Term Memory（以下 LSTM）とは入力ゲートi_t、出力ゲートo_t、記憶セルc_t、忘却ゲートf_tで構成される深層学習の代表的なモデルである。それぞれは以下の式で表される。

$$i_t = \sigma(W_x^{(i)} x_t + W_y^{(i)} y_{t-1} + b^{(i)})$$
$$o_t = \sigma(W_x^{(o)} x_t + W_y^{(o)} y_{t-1} + b^{(o)})$$
$$f_t = \sigma(W_x^{(f)} x_t + W_y^{(f)} y_{t-1} + b^{(f)})$$
$$c_t = f_t \cdot c_{t-1} + g_t \cdot i_t$$

なお記憶セル内のc_tと出力y_tは以下である。

$$y_t = o \cdot \tanh(c_t)$$
$$g_t = tanh(W_x^{(g)} x_t + W_y^{(g)} y_{t-1} + b^{(g)})$$

以上の各式の重みW_x、W_yを LSTM が学習している。

３－３．　特徴量抽出の流れ

検討手法は以下の手順で行う。

① マルウェアの API 呼び出し列を収集する（本研究では既存のデータセットを使用）
② 説明変数を n 番目の API、目的変数を n+1 番

目の API として LSTM に API 呼び出し列を学習させる

③ 学習済みの LSTM モデルに調査対象となるマルウェアの API 呼び出し列を入力する

④ 設定した時点数ごとの状態ベクトルを保存し、それぞれを順に結合する

４．評価・実験

４－１．使用するデータセット

本実験では Ki Y[3]によって公開、共有されているデータセットを使用する。このデータセットには 23145 個のマルウェアサンプルから得られたマルウェア毎の API 呼び出し列が含まれている。この詳細を以下の表に示す。

表１．データセット詳細（[3]より抜粋）

Category	Subcategory	Ratio (%)
Backdoor		3.37
Worm	Worm	3.32
	Email-Worm	0.55
	Net-Worm	0.79
	P2P-Worm	0.3
Packed		5.57
PUP	Adware	13.63
	Downloader	2.94
	WebToolbar	1.22
Trojan	Trojan (Generic)	29.3
	Trojan-Banker	0.14
	Trojan-Clicker	0.12
	Trojan-Downloader	2.29
	Trojan-Dropper	1.91
	Trojan-FakeAV	18.8
	Trojan-GameThief	0.63
	Trojan-PSW	3.79
	Trojan-Ransom	2.58
	Trojan-Spy	3.12
Misc.		5.52

４－２．実験方法

本研究の特徴量抽出手法の有効性を確認するため、抽出した特徴量を用いた機械学習による分類実験を行う。既存研究[2]と同様のデータセットであるため、同研究の 4 値分類の実験を行い、比較を行う。この実験ではトロイの木馬(Trojan)、アドウェア(PUP)、ワーム(worm)、その他に分類を行うため、この 4 カテゴリーの API 呼び出し列とそれに対応するラベル付けを行うことでデータセットを再構成する。

このデータセットの３０％である 6924 個のマルウェアの API 呼び出し列を LSTM に学習させ、特徴抽出モデルを構築する。さらに分類モデルのための訓練データを 80%、テストデータを 20%の割合でランダムに用意する。最後に訓練データを前述した特徴抽出手法によりベクトル化、これを用いて機械学習により分類モデルを構築する。この機械学習の手法はサポートベクタマシン(SVM)、ランダムフォレスト(RF)、K 近傍法(KNN)を用いた。

４－３．評価

評価指標は既存研究[1]と同様、正解率を用いる。また LSTM の入力層は終端記号を含む 1166 次元に、隠れ層は 200 次元、出力層は 1166 次元に設定し、最適化アルゴリズムに Adam、損失関数に交差エントロピー、エポック数は 50 という条件で学習を行った。

この特徴抽出モデルを用いて 4 値分類の分類モデルを構築する。サポートベクタマシンのカーネルには rbf カーネル、ランダムフォレストの max_depth は 4、K 近傍法の最近傍個数(k)は 3 という条件で３－１節で述べた二分割手法と三分割手法との比較実験を行った。結果は以下である。

表２．二分割手法での分類の正解率

手法	訓練データ	テストデータ
SVM	80.99	75.96
KNN	83.35	76.20
RF	64.41	63.46

表３．三分割手法での分類の正解率

手法	訓練データ	テストデータ
SVM	82.53	76.63
KNN	85.60	76.22
RF	65.24	64.77

既存研究[2]で提案された二分割手法よりも、三分割手法の方が数パーセント正解率が高いことが明らかになった。

５．考察

本節では実験結果について考察する。まず、既存研究ほどの正解率が得られなかった原因は、データセットの３０％しか用いず、API 呼び出し列の特徴の傾向をとらえるには十分ではなかったのではないかと考える。また三分割手法の方の正解率が高い理由は、時点数を多くしたためであると考える。これにより、API 呼び出し列の特徴を細かに反映した特徴量を再現できたと考える。

６．まとめ・今後の課題

本研究では、機械学習によるマルウェアの分類に向けた特徴抽出手法の検討を行った。そして実験を行うことでその有効性を確認し、既存研究ほどの精度を得られることはできなかったが、考察により最適な時点数の更なる検討が必要であることがわかった。今後の課題として、LSTM の精度向上のために訓練データを増やし、より時系列データに対して、より長期的な依存関係の獲得を可能にする Attention 機構の導入を検討、また時点数の増減による正解率の変動を調査し、最適な時点数を見つけることが挙げられる。その他、より適切な特徴量が抽出可能なモデルを随時検討していく。

参考文献

[1] Tran,T.K and Sato,H “NLP-based approaches for malware classification from api sequences. 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES) in 2017, pages 101-105. IEEE, 2017.

[2] Pascanu, R., Stokes, J.W., Sanossian, H., Marinescu, M., Thomas, A.: “Malware classification with recurrent networks.”, Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),pp.1916-1920. ,2015

[3] Y. Ki, E . Kim, H . Kang Kim, “A Novel Approach to Detect Malware Based on API Call Sequence Analysis”, International Journal of Distributed Sensor Networks - Special issue on Advanced Big Data Management and Analytics for Ubiquitous Sensors, Volume 2015, January 2015