マルウェア検知に向けた親プロセスと子プロセスにおける API コール列の類似性の分析

中村 英敏 *†，松田 祥希 †，高橋 健一 †‡，川村 尚生 †‡
(† 鳥取大学大学院 持続性社会創生科学研究科，‡ クロス情報科学研究センター，)

Analysis of similarity of API call sequences in parent and child processes for malware detection

Hidetoshi Nakamura*†, Yoshiki Matsuda†, Kenichi Takahashi†‡, and Takao Kawamura†‡
(†Graduate School of Sustainability Science, ‡Cross-informatics Research Center, Tottori University)

1 はじめに

インターネット利用の増加により，マルウェアを用いたサイバー攻撃が増加している．そのため，正規ソフトウェア（以下，正規ウェア）とマルウェアの判別や，マルウェアファミリーの推定などの研究が行われている．飛山ら [1] は，RNN や CNN などの異なる Deep Neural Network を多段に用いることで，プロセスから得られた API コール列の特徴抽出と分類を行い，マルウェアプロセスを判別している．横瀬ら [2] は，静的解析と動的解析から得られた API コール列の結果を結び付け，静的解析のデータから関連する動的解析のデータを取りだすことで，マルウェアの亜種推定を行っている．しかし，これらの研究では，プロセスの親子関係に着目せず，親プロセスと子プロセスを分けていない．そこで本稿では，親プロセスと子プロセスの類似性に着目し，正規ウェアとマルウェアの分析を行った．

2 データ収集

2.1 検体

正規ウェアは窓の杜 2019 年ランキング，MicrosoftOffice，利用率が高いブラウザ（Google Chrome，Microsoft Edge，Firefox）から収集した．マルウェアは TekDefense，theZoo から Windows 環境向けの検体を収集した．

2.2 データ収集方法

データ収集環境は，ホスト OS として Ubuntu 20.04.4 LTS，ゲスト OS として Windows10 Pro，Ubuntu 20.04.4 LTS を用いた（図 1）．オンライン環境を模倣するためにゲスト OS の Ubuntu 20.04.4 LTS 上で INetSim を起動し，Windows10 Pro 上で起動した検体と通信を行うように設定した．Windows10 Pro 上で検体を実行し，API Monitor を用いて検体が呼び出した API コールを取得した．

API Monitor を起動後，実験検体を起動し，API コールの取得を開始する．60 秒後，検体と API Monitor を停止し，API コールの取得を終了する．

データを収集した結果，子プロセスを起動していた正規ウェア 20 組，マルウェア 23 組が得られた．これらを分析対象として用いる．

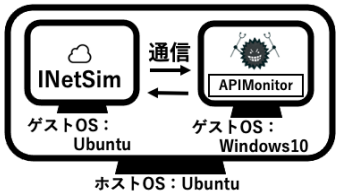

図 1 データ収集環境

3 分析手法

API コール列の類似度を測定するために，1 つの API コールを単語とみなし，自然言語処理手法を用いた．親プロセスと子プロセスの API コール列を BoW，TF-IDF，n-gram を用いて処理し，cos 類似度，一致率を用いて類似度を測定した．その後，類似度を用いて SVM による正規ウェアとマルウェアの判別を行った．学習したモデルを 5 分割交差検証による性能の評価を行った．

3.1 BoW

BoW（Bag-of-Words）とは文書を単語の順番を考慮せずに，出現回数のベクトルで表現したものである．

3.2 TF-IDF

TF-IDF とは，各文書においてその単語がどのくらい出現したのかを意味する TF（Term Frequency）と，その単語の希少性を表す IDF（Inverse Document Frequency）をかけたものであり，文書内の単語の重要度を示す手法の一つである．

3.3 n-gram

n-gram とは任意の文字数 n で文章を分割する手法である．任意の文字数は連続する n 個の単語や文字のまとまりを表し，n が 1 の場合 uni-gram（ユニグラム），n が 2 の場合 bi-gram（バイグラム），n が 3 の場合 tri-gram（トライグラム）と呼ぶ．

3.4 類似度測定

Bag-of-Words と TF-IDF はベクトルであるため，類似度を測定するために cos 類似度を用いる．cos 類似度は-1 から 1 の値をとり，1 に近ければ似ており，-1 に近ければ似ていないことを示す．ベクトル $\boldsymbol{x}$ とベクトル $\boldsymbol{y}$ の cos 類似度は (1) 式で定義される．

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} \tag{1}$$

n-gram は API コールの組であるため，類似度の測定には一致率を用いる．一致率は 1 に近ければ似ており，0 に近ければ似ていないことを示す．一致率は (2) 式で定義される．

$$\text{一致率} = \frac{\text{一致した数}}{\text{総当たり数}} \tag{2}$$

3.5 SVM

SVM（Support Vector Machine）は，機械学習モデルの一種であり，少ない学習データでも安定的な結果を出せることで知られている．SVM はクラス間を分ける超平面を，それぞれのクラス間での最も近いデータ点（Support Vector）との距離ができるだけ大きくなるように決定することでクラスを分類する．

3.6 評価指標

正解率（Accuracy），適合率（Precision），再現率（Recall），F 値（F1-score）を評価指標として用いる．各指標の計算式を (3)〜(6) 式に示す．ここで，TP は True Positive，FP は False Positive，FN は False Negative，TN は True Negative を表す．TP はマルウェアのデータを，正しくマルウェアだと予測したことを表す．

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 - score = \frac{2Recall * Precision}{Recall + Precision} \tag{6}$$

4 分析結果

4.1 比較

それぞれの特徴量の関係をみるために，散布図行列を作成した．その一部を図 2 に示す．Label の Leg が正規ウェアの値，Mal がマルウェアの値となっている．左上から右上にかかる対角線上のグラフは，有限の標本点から全体の分布を推定する手法であるカーネル密度推定の結果を表す．これをみると，たしかに正規ウェアとマルウェアとではグラフの形が異なることがわかる．また，それぞれの特徴量との散布図を見ると，正規ウェアの値は右上に集まっているが，マルウェアの値は左側に集まっていることがわかる．それぞれの値が 1（右上）に近づくほど類似度が高く，値が 0（左下）に近づくほど類似度が低いといえるため，正規ウェアのほうがマルウェアより親プロセスと子プロセスの API コール列の類似度が高いといえる．

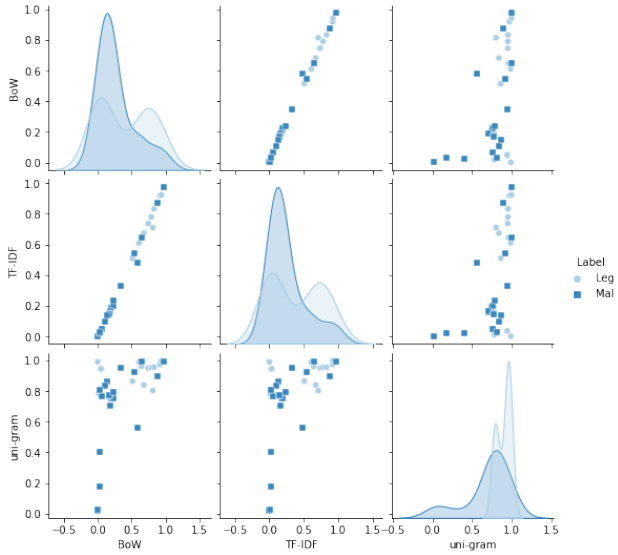

図 2　散布図行列

表 1　SVM による判別結果

	正解率	適合率	再現率	F 値
BoW, TF-IDF	0.59	0.67	0.68	0.60
BoW, uni-gram	0.75	0.83	0.71	0.75
BoW, tri-gram	0.75	0.79	0.68	0.71
TF-IDF, uni-gram	0.73	0.78	0.71	0.73
TF-IDF, tri-gram	0.78	0.83	0.68	0.73
uni-gram, tri-gram	0.72	0.91	0.67	0.75
BoW, TF-IDF, uni-gram	0.78	0.83	0.76	0.78
BoW, TF-IDF, tri-gram	0.70	0.76	0.71	0.72
BoW, uni-gram, tri-gram	0.77	0.88	0.71	0.76
TF-IDF, uni-gram, tri-gram	0.77	0.88	0.71	0.76
BoW, TF-IDF, uni-gram, tri-gram	0.75	0.84	0.67	0.73

4.2 判別

類似度を用いて SVM による判別を行った（表 1）．結果，BoW，TF-IDF，uni-gram の組み合わせを用いたとき，正解率 0.78，適合率 0.83，再現率 0.76，F 値 0.78 を得た．

5 まとめ

本稿では親プロセスと子プロセスの API コール列の類似性に着目した分析を行った．正規ウェアとマルウェアの親プロセスと子プロセスがどれだけ似ているかを BoW，TF-IDF，n-gram を用いて測定した．また，得られた類似度を用いて SVM による学習を行った．その結果，正規ウェアとマルウェアで違いが見られることが分かった．

今後は検体数を増やしたり，Word2Vec や Doc2Vec などの他の自然言語処理手法を用いた分析を行っていきたい．

参考文献

[1] 飛山駿, 山口由紀子, 嶋田創, 秋山満昭, & 八木毅. (2016). Deep Neural Network 多段化によるプロセスの挙動に着目したマルウェア推定手法. コンピュータセキュリティシンポジウム 2016 論文集, 2016(2), 310-317.

[2] 横瀬謙信, & 大久保潤. (2019). word2visualvec を応用したマルウェア亜種推定の枠組みの提案. コンピュータセキュリティシンポジウム 2019 論文集, 2019, 1047-1051.