G-4

画像特徴量によるマルウェア亜種検知に関する検討

A Study on Detection of Variant of Malware by Image Features

○小寺 建輝[1]， 泉 隆[2]，香取 照臣[2]
*Tateki Kodera[1], Takashi Izumi[2], Teruomi Katori[2]

Distribution of malware automatic generation tool and outflow of source code of malware simplify and speed up variant of malware generation. In this research, we study a method to detect variant of malware by image recognition. In this paper, we investigate the compound feature combining LBP, Gist, HLAC as image feature used for learning and detection.

１．はじめに

近年，マルウェア亜種自動生成ツールの流通や，マルウェアのソースコード流出により，亜種生成が簡易化・高速化され，ウイルス定義ファイル等のパターンファイルの作成・配布が追い付かない現状となっている．例えば，IoT マルウェアである「Mirai」は作成者によってソースコードが公開されており，それを改変した亜種が大量に作成され，多くの IoT デバイスが感染の被害にあった．このような亜種検知に関する問題を解決するため，機械学習により亜種を検知・分類する研究が現在取り組まれている．その中でも，Windows 系マルウェアを画像化し，画像認識によって亜種を該当するファミリに分類する先行研究[1]では，高い識別精度でマルウェアを分類できたことが報告されている．これは，亜種が元のコードの一部のみを改変して作成されるため，元のマルウェアとその亜種，つまり同一ファミリでは視覚的に類似した画像が得られるためである．しかし，先行研究では，各ファミリの亜種と正常ファイルを識別することが検討されていない．

これらを踏まえて本研究では，同一ファミリの類似したマルウェアの画像をグループ化してグループごとに亜種検知モデルを構築し，各モデルで亜種か否かを識別することで亜種の検知及び正常ファイルを識別する手法を検討する．

本稿では，亜種検知モデルの学習時及び亜種検知・識別時に利用する画像特徴量として，LBP(Local Binary Pattern)特徴量[2]，Gist 特徴量[3]，HLAC (Higher-order Local Auto-Correction)特徴量[4]を組み合わせた複合特徴量を検討し，各特徴量を単一で用いた場合との亜種検知率・誤検知率の比較を行う．

２．マルウェア亜種検知アルゴリズム

亜種検知モデルの学習及びそのモデルを利用して各ファミリの亜種と正常ファイルを識別するアルゴリズム(Fig. 1)を以下の４つのフェーズに分けて説明する．

(1) 類似した画像のグループ化
　学習データであるマルウェアにファミリ名のラベルを割り当て，ファミリごとにグループ化する．

(2) 異常検知モデルの構築
　検知対象の画像が(1)で作成したグループに属するか否かをアノマリスコアにより判定する，異常検知モデルを各グループで構築する.また，判定のために各モデルでアノマリスコアに閾値を設定する．

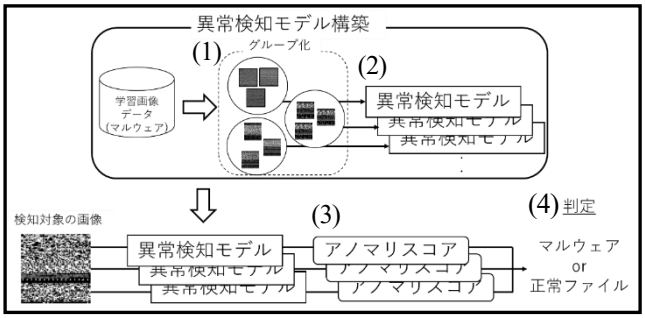

Figure1. Variant of malware detection algorithm

(3) アノマリスコアの算出
　検知対象の画像の特徴量を各異常検知モデルに入力し，アノマリスコアを算出する．

(4) マルウェアの判定
　検知対象の画像のアノマリスコアがあるモデルで閾値未満であった場合，そのモデルに該当するマルウェア(亜種)であると判定する.また，アノマリスコアが閾値以上であった場合，そのモデルに非該当のファイルと判定する．
　全てのモデルにおいて非該当のファイルと判定された画像を正常ファイルと識別する．

３．複合特徴量

本稿では，LBP 特徴量(256 次元)，Gist 特徴量(320 次元),HLAC 特徴量(35 次元)を組み合わせた複合特徴量について検討する.しかし，全特徴量をそのまま組み合わせた計 611 次元の複合特徴量の中には識別精度の向上に寄与しない，もしくは識別精度を低下させる特徴量が含まれる可能性がある．そのため，全特徴量を結合後，各特徴量の評価値を算出し，評価値が高い特徴量のみを選択する必要がある．そこで，マルウェア亜種検知(異常検知)に有効な特徴量を評価するために，分散にもとづいた特徴量の評価手法を検討する．

まず,あるファミリのマルウェア亜種(正常データ)m を k 個，その他のファイル(異常データ)b を l 個としたとき，それらの i 番目の特徴量に関して式(1), (2)により２つの分散を求める．

$$\sigma_{m_i}^2 = \frac{1}{k}\sum_{j=1}^{k}(m_{ij} - \mu_{m_i})^2 \quad (1)$$

$$\sigma_{b_i}^2 = \frac{1}{l}\sum_{j=1}^{l}(b_{ij} - \mu_{m_i})^2 \quad (2)$$

１：日大理工・院（前）・情報　２：日大理工・教員・情報

分散$\sigma_{m_i}^2$は，あるファミリのマルウェア亜種における特徴量 i の分散を表しており，分散$\sigma_{b_i}^2$は，その他のファイルにおける特徴量 i の分散を表している．同一ファミリ内のマルウェア亜種は互いに類似している必要があるため，分散$\sigma_{m_i}^2$は小さくなることが理想である．一方，その他のファイルはあるファミリのマルウェア亜種のデータ群からみて異常である必要があるため，分散$\sigma_{b_i}^2$は大きくなることが理想である．これらを踏まえて，特徴量 i の評価値 v_i を，式(3)のように$\sigma_{m_i}^2$と$\sigma_{b_i}^2$の分散比であらわす．

$$v_i = \frac{\sigma_{b_i}^2}{\sigma_{m_i}^2} \tag{3}$$

4．実験

異常検知モデルの構築時及び亜種検知・識別時に利用する画像特徴量として，LBP 特徴量，Gist 特徴量，HLAC 特徴量，及びそれらを全て組み合わせた 611 次元の複合特徴量，さらに 3 章で示した手法により特徴量選択を行った複合特徴量を採用し，各種特徴量を用いた場合の亜種検知率・誤検知率の比較を各ファミリのモデルごとに行った．

ここで，モデルの構築及び亜種検知率の評価に Malimg Dataset[5]内の 25 ファミリ 9339 検体のマルウェアの画像を，誤検知率の評価に 913 検体の正常ファイルをそれぞれ利用して 10 分割交差検証を行う．Table 1 に Malimg Dataset の内訳を示す．

また，モデル構築の学習アルゴリズム及びアノマリスコアの算出に Isolation Forest[6]を採用した．なお、各モデルにおけるアノマリスコアの閾値は，線形判別分析法により決定した．

実験結果を Figure 2，Table 2 に示す．

Table 1. Malimg Dataset breakdown

Family	Type	Number of samples
Adialer.C	Dialer	122
Agent.FYI	Backdoor	116
Allaple.A	Worm	2949
Allaple.L	Worm	1591
Alueron.gen!J	Trojan	198
Autorun.K	Worm:AutoIT	106
C2LOP.gen!g	Trojan	200
C2LOP.P	Trojan	146
Dialplatform.B	Dialer	177
Dontovo.A	Trojan Downloader	162
Fakerean	Rogue	381
Instantaccess	Dialer	431
Lolyda.AA1	PWS	213
Lolyda.AA2	PWS	184
Lolyda.AA3	PWS	123
Lolyda.AT	PWS	159
Malex.gen!J	Trojan	136
Obfuscator.AD	Trojan Downloader	142
Rbot!gen	Backdoor	158
Skintrim.N	Trojan	80
Swizzor.gen!E	Trojan Downloader	128
Swizzor.gen!I	Trojan Downloader	132
VB.AT	Worm	408
Wintrim.BX	Trojan Downloader	97
Yuner.A	Worm	800

Table2.　Average of variant detection rate and false detection rate

特徴量	複合 (特徴量選択)	複合 (611 次元)	LBP	Gist	HLAC
平均亜種検知率(%)	96.5	95.4	91.3	95.0	91.6
平均誤検知率(%)	0.38	0.56	1.13	0.66	1.14

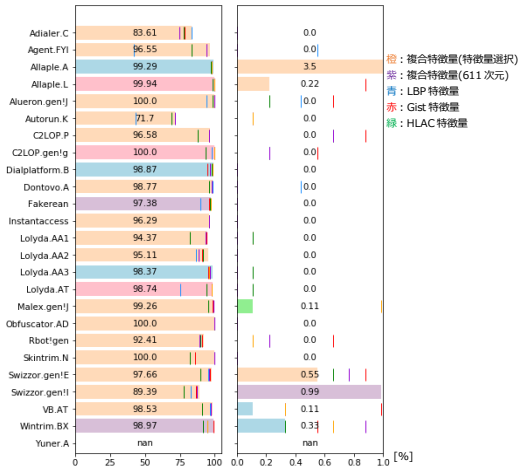

Figure2. Experimental result

(Left: variant detection rate, Right: false detection rate)

Figure 2 で示したグラフは，各ファミリのモデルにおいて最も高い亜種検知率(Left)・最も低い誤検知率(Right)を示した画像特徴量をあらわしており，画像特徴量ごとに色付きの棒グラフで表現している．例えば，Allaple.A というファミリの亜種検知率は青色のグラフになっていることから，LBP 特徴量を用いたときに亜種検知率が最も高くなったことを示している．これを踏まえ全体の結果では，複合特徴量(特徴量選択)を用いた場合において，25 ファミリ中 16 ファミリで最も高い亜種検知率，17 ファミリで最も低い誤検知率を示し，他の特徴量を用いたときよりも安定した識別精度が確認された．また Table 2 からも，複合特徴量(特徴量選択)を用いた時の平均亜種検知率・平均誤検知率が最も良い結果となった．分散にもとづいた特徴量選択を行うことで，異常検知に有効な特徴量を適切に選択できたことが考えられる．また，特徴量選択を行う前の複合特徴量(611 次元)においても，平均亜種検知率・平均誤検知率ともに，各単一特徴量よりも優れた結果となった．このことから複合特徴量は，単一特徴量のみではとらえられない特徴量を補うことができ，識別精度の向上に有効と考えられる．

5．まとめ

本稿では，亜種検知モデルの構築時及び亜種検知・識別時に利用する画像特徴量において，各単一特徴量を組み合わせた複合特徴量を検討し，さらに分散にもとづいた徴量選択手法を示した．各特徴量による亜種検知率・誤検知率を比較した結果，複合特徴量(特徴量選択)が最も高い識別精度を示すことを確認した．

今後は，本稿で利用した特徴量選択手法と異なる特徴量選択手法を利用した場合との精度の比較を行う．

6．参考文献

[1] L. Nataraj, et al. : ”Malware Images:Visualization and Automatic Classification”, VizSec’11(2011-07)

[2] DC.He and L.Wang : “Texture Unit, Texture Spectrum, And Texture Analysis”, IEEE Transactions on Geoscience and Remote Sensing, Vol.28, pp.509-512(1990)

[3] A. Olivia and A. Torralba : ” Modeling the shape of a scene: a holistic representation of the spatial envelope”, International Journal of Computer Vision, Vol.42, No.3, pp.145–175(2001)

[4] N.Otsu and T.Kurita : ”A new scheme for practical flexible and intelligent vision systems”, Proc. IAPR Workshop on Computer vision, pp.431-435(1998-10)

[5] Abien Fred Agarap : “Malimg Dataset”, https://www.kaggle.com/afagarap/malimg-dataset (2018-09)

[6] Fei Tony Liu, et al. : “Isolation-Based Anomaly Detection”, ACM Transactions on Knowledge Discovery from Data(TKDD), Vol.6, No.1, pp.1-39(2013-03)