G-5

画像特徴量によるマルウェア亜種検知の精度向上
CLBP 特徴量を用いた複合特徴量の検討
Improving the accuracy of Malware Variant detection based on Image Features
Investigation of the compound features using CLBP features

○原田吉樹[1]，泉　隆[2]，滕　琳[2]
*Yoshiki Harada[1], Takashi Izumi[2], Lin Teng[2]

Abstract: Over the past few years, the method of creating malware variants has become the mainstream of cyber attacks. In this research, we study a method to detect variant of malware by image recognition. In this paper, we investigate the compound feature combining Gist, CLBP, HLAC as image feature used for learning and detection.

１．まえがき

近年のサイバー攻撃の主流として，マルウェア亜種を作り出す手法が取られている．FireEye 社が公開したサイバー脅威レポート[1]では，2019 年に検出されたマルウェアファミリのうち未確認の新型は 41%となっており，シグネチャベースのウイルス対策ソフトでは対応が困難になっている．

先行研究[2]では，マルウェア画像を用いた，この亜種検知の精度向上として CLBP 特徴量を利用した手法を提案し, LBP 特徴量に比べ精度の向上が確認された．また，複合特徴量を利用した先行研究[3]では，LBP 特徴量，Gist 特徴量，HLAC 特徴量，3 つの特徴量を統合した複合特徴量を用いた判別によると，複合特徴量が単一特徴量に比べ高い精度を得ている．

本研究では，より検知精度を向上させるために複合特徴量に用いる特徴量を再検討し，精度の比較を行う．

本稿では，CLBP を用いた複合特徴量について検討し，各特徴量との精度比較を行う．

２．マルウェア亜種検知の流れ

本研究における亜種検知の流れを(1)～(3)に示す．

(1)異常検知モデルの構築

亜種検知用の学習データを用意し，同系列のマルウェアファミリごとにグループ分けをする．対象マルウェアファミリの異常検知モデルを構築する．

(2)検知対象データの入力

アノマリスコア算出を行うために，検知対象データとなるマルウェアのグレースケール画像を用いて，画像特徴量の抽出を行う．その後，異常検知モデルへ入力する．

(3)スコアの算出・亜種判定

検知対象データについて，異常検知モデルを用いてアノマリスコアを算出する．

あるファミリの異常検知モデルにおいて，アノマリスコアが閾値未満の場合は、検知対象データは亜種であると判定する．閾値以上の場合では，ファミリの亜種ではないと判断する．全てのファミリで閾値以上だった場合，正常なファイルと判断する．

３．複合特徴量の作成

複合特徴量は，単一特徴量を統合することによって，得ることができる．先行研究[3]を基に，分散を利用した特徴量の評価手法を用いた．

全特徴量について下記評価値を算出し，評価値の高い特徴量を複数選択する．あるファミリ属するマルウェア亜種(正常データ)m を k 個，あるファミリに属さないファイル(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)の 2 つの分散を求める．このとき，あるファミリに属するマルウェア亜種の特徴量 i における平均値をμ_{mi}で示す。

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

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

分散$\sigma_{mi}{}^2$は，あるファミリに属するマルウェア亜種の特徴量 i における分散を表しており，分散$\sigma_{bi}{}^2$は，あるファミリに属さないファイルの特徴量 i における分散を表している．評価値 v_i とし，特徴量 i の評価値は式(3)に示す．

$$v_i=\frac{\sigma_{bi}{}^2}{\sigma_{mi}{}^2} \quad (3)$$

v_i の大きさを基に特徴量選択を行う。

４．亜種検知の実験

本実験では，各ファミリの異常検知モデルにおいて，該当するマルウェアに対する亜種の検知率と，誤って亜種と認識した誤検知率の比較を行う．

特徴量は，Gist 特徴量，LBP 特徴量，CLBP 特徴量，HLAC 特徴量を単一特徴量もしくは複合特徴量として用いる．学習アルゴリズム及びアノマリスコアの算出に Isolation Forest を採用した．モデルの構築及び亜種検知率の評価には，Malimg Dataset[4]内の 25 ファミリ 9339 検体のマルウェアの画像と，誤検知率の評価に 913 検体の正常ファイルを利用した．

各特徴量の平均亜種検知率の結果を Table1，Gist・LBP・HLAC の複合特徴量と Gist・CLBP・HLAC の複合特徴量での検知比較を Table2 に示す．

1：日大理工・学部・情報　2：日大理工・教員・情報

Table 1. 各特徴量の平均亜種検知率[%]

特徴量	平均検知率	平均誤検知率
Gist	85.79	0.34
LBP	87.84	1.50
CLBP	88.56	1.02
HLAC	88.59	4.71
Gist・LBP・HLAC	86.00	0.29
Gist・CLBP・HLAC	89.75	0.17

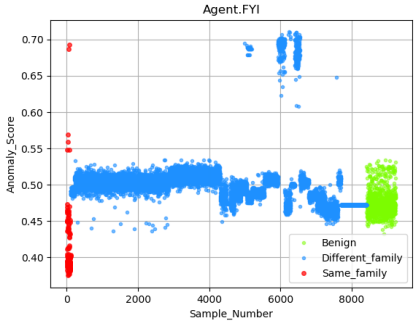
Table 2 .各マルウェアファミリの平均亜種検知率[%]

ファミリ	平均亜種検知率	
	Gist・LBP・HLAC	Gist・CLBP・HLAC
Adialer.C	76.23	76.23
Agent.FYI	64.66	84.48
Allaple.A	98.47	98.34
Allaple.L	99.06	99.13
Alueron.gen!J	88.38	95.96
Autorun.K	69.81	66.98
C2LOP.P	84.25	97.95
C2LOP.gen!g	96.50	99.50
Dialplatform.B	96.61	97.74
Dontovo.A	95.68	98.15
Fakerean	90.29	94.75
Instantaccess	96.29	96.29
Lolyda.AA1	90.14	93.90
Lolyda.AA2	85.33	89.13
Lolyda.AA3	94.31	97.56
Lolyda.AT	83.65	96.86
Malex.gen!J	91.91	97.79
Obfuscator.AD	100.00	100.00
Rbot!gen	93.67	96.20
Skintrim.N	95.00	98.75
Swizzor.gen!E	92.97	95.31
Swizzor.gen!I	81.82	85.61
VB.AT	96.32	97.55
Wintrim.BX	88.66	89.69
Yuner.A	0.00	0.00

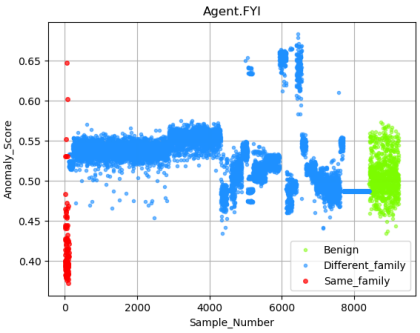
Table1 から，Gist・CLBP・HLAC の複合特徴量が各特徴量に比べ，平均検知率及び平均誤検知率ともに高い精度が得られている．

Table2 から，各マルウェアファミリで比較した場合，基本的に CLBP 特徴量を用いた方が検知率は高い．特に，Agent.FYI や Lolyda.AT といったマルウェアファミリでは，10%以上の平均検知率の向上が確認できた．

Figure 1.を用いて，Agent.FYI におけるアノマリスコアの分布の比較を行う．Figure 1.において，赤い点は同一ファミリ，青い点は別のファミリ，緑の点は正常ファイルを示す．



(a)　Gist・LBP・HLAC 特徴量を用いた場合



(b) Gist・CLBP・HLAC 特徴量を用いた場合

Figure 1. アノマリスコア分布の比較

(a)の複合特徴量では，3 色の点の分布がはっきりと分かれていないが，(b)の複合特徴量では，赤い点の分布とそれ以外の分布が，(a).に比べはっきりと分かれている．そのため，検知率が 85%程度まで向上した．よって，CLBP 特徴量を複合特徴量に用いることでアノマリスコアが向上し，平均検知率が向上することが分かった．

５．まとめ

本稿では，CLBP を用いた複合特徴量の作成し，各特徴量との精度比較を行った結果，CLBP 特徴量を用いた複合特徴量が最も高い検知率となった．

今後は，複合特徴量を用い，かつ学習時のパラメータ変更及び異常データの混入を行い，精度確認を行う．

６．参考文献

[1] FireEye : FireEye，年次レポート「Mandiant M-Trends 2020」日本語版を公開，https://www.fireeye.jp/company/press-releases/2020/fireeye-mandiant-m-trends-2020-report-reveals-cyber

[2] 清宮舜太，泉隆 : 画像特徴量によるマルウェア亜種検知の精度向上-CLBP 特徴量を用いた亜種検知-(2020-02)

[3] 小寺建輝，泉隆，香取照臣:「画像特徴量によるマルウェア亜種検知に関する検討」，平成 30 年度日本大学理工学部学術講演会，G-4，pp.517-518(2018)

[4] Lakshmanan Nataraj, S. Karthikeyan, Gregoire Jacob, B.S. Manjunath : ”Malware Images: Visualization and Automatic Classification”, International Symposium on Visualization for Cyber Security (VizSec)(2011)