講演番号： 205

バイナリデータの画像化を活用したマルウェア分類法の検討

D－19　Study of Malware Classification Method Utilizing Image of Binary Data

鈴木　貴之†　笠間　貴弘†‡　宮保　憲治†

Takayuki SUZUKI†　Takahiro KASAMA†‡　Noriharu MIYAHO†

† 東京電機大学大学院情報環境学研究科

† Graduate School of Information Environment, Tokyo Denki University

‡ 情報通信研究機構

‡ National Institute of Information and Communications Technology

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問題となっている．一部のマルウェアには，難読化やサンドボックス検知など，耐解析機能を備えるマルウェアが存在する．その中で，マルウェアのバイナリデータの画像を基にした分類実験より，難読化されたマルウェアを含んだ場合でも高い識別精度で分類できた事が報告されている[1]．しかしながら，実行可能ファイル形式のマルウェア以外に対する効果に関しては検証されていない．

本稿では，Android マルウェアおよび PDF 文書型マルウェアに対して，バイナリデータを画像に変換し，変換した画像から特徴量を抽出してマルウェアの分類実験を行い，識別精度を検証した結果を述べる．

2　バイナリデータの画像化

画像化の手法は，バイナリデータを 8bit 区切りで 10 進数に変換し，その値をグレースケールと対応付けて画像に変換した後，スクリプトにより正方形のサイズに変換した．これは，後述する実験での特徴量抽出において，正方形の画像であることが望ましいためである．バイナリデータを 256×256 のグレースケール画像に変換する手法の概要を図 1 に示す．

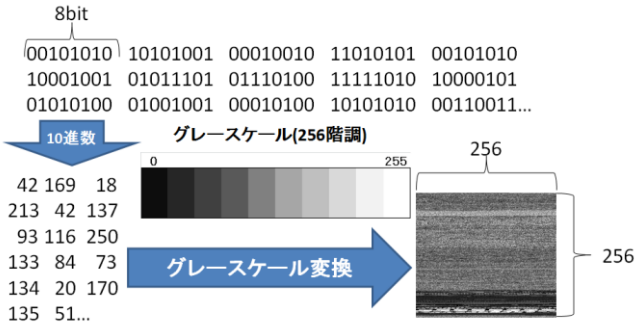

図 1 バイナリ画像化の手法

3　機械学習による分類実験と考察

画像化したマルウェアデータと正常データを用いた分類実験を行った．実験諸元を表 1 に示す．

表 1 実験諸元

機械学習アルゴリズム	K近傍法
特徴量抽出手法	GIST特徴
識別精度算出手法	10-fold cross validation
バイナリデータ画像サイズ	32, 64, 128, 256, 384, 512, 768, 1024
マルウェアデータ(Android)	10ファミリ 4022個
マルウェアデータ(PDF)	10ファミリ 9997個
正常データ(Android)	500個
正常データ(PDF)	500個

評価について，表 1 の画像サイズに変換したバイナリデータ画像を実験に使用し，識別精度の変化を検証した．

Android データを使用した識別精度を図 2 に，PDF データを使用した識別精度を図 3 に示す．図 2 より，Android データを使用した場合，識別精度は概ね 80%以上を確保していることを確認した．図 3 より，PDF データを使用した場合，識別精度が概ね 85%以上を確保していることを確認した．よって，本手法が実行可能ファイル形式以外のファイル形式でも適用可能である事を確認した．PDF データと比較して Android データの識別精度が低い理由としては，Android マルウェアは正常な Android アプリケーションを改変して作成するため，マルウェアデータと正常データとの差異が少ないためだと考えられる．また，画像サイズによる識別精度の変化については，Android データの場合，32×32 と 64×64 のサイズ間で約 3%の改善を確認できたが，それ以降のサイズでは改善の傾向は確認できなかった．また，PDF データの場合，画像サイズによる識別精度の変化はほぼ見られなかった．加えて，特徴量の抽出時間が画像サイズに比例して増加した事から，分類に使用する画像サイズは最小の 32×32 が最適であることを確認した．

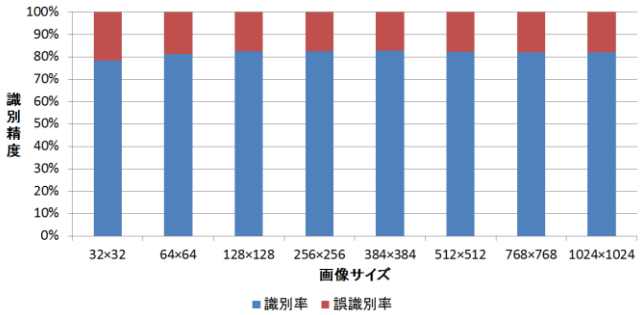

図 2 Android データの識別精度

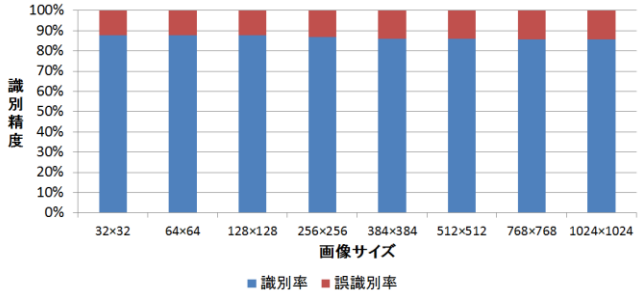

図 3 PDF データの識別精度

4　今後の課題

今後は画像化の手法の改善や，特徴量の抽出手法，機械学習アルゴリズムを検討し，識別精度を改善する予定である．

参考文献

[1] L. Nataraj, et.al.”Malware Images: Visualization and Automatic Classification”, VizSec’11, July 20, 2011

[2] Oliva, A, et.al. “Modeling the shape of a scene: a holistic representation of the spatial envelope”, IJCV, Vol. 42(3), 145-175, 2001