ファイル内文字列情報を活用したマルウェア検知手法の検討

Study of a malware detection method using strings inside an executable file

石渡 純一† 小出 遼†† 笠間 貴弘††† 宮保 憲治†
Junichi ISHIWATA† Ryo KOIDE†† Takahiro KASAMA††† Noriharu MIYAHO†
† 東京電機大学 †† 東京電機大学大学院 ††† 国立研究開発法人情報通信研究機構
† School of System Design and Technology, Tokyo Denki University †† Graduate School of Information Environment, Tokyo Denki University ††† National Institute of Information and Communications Technology

1. はじめに

近年，マルウェア感染の社会問題が顕在化しており，この状況を鑑みると，迅速かつ容易にマルウェア検出を行うことのできる技術を確立する必要がある．

マルウェアの解析方法は主に表層解析，動的解析，静的解析の 3 つに分けられる．表層解析とはファイルの情報からマルウェアを解析する手法で，他の解析方法と比較して短時間で情報を得ることが可能である．

本稿では，表層解析で取得対象とした文字列情報を有効に活用し，マルウェアを迅速に検知する手法を提案する．

2. 提案手法

2.1 提案手法の概要

本提案の概要を図 1 に示す．実行ファイルから文字列情報の抽出を行った後に，抽出文字列情報を空白区切りで単語に分割する．次の段階では分割して得られた単語（the, ZYYd 等）を使用して文字列情報を TF-IDF を用いてベクトルに変換した後に L2 正規化（ベクトルの大きさを 1 にする処理）を適用する．その後，機械学習システムを用いて分類を行い，正解率と F 値を用いて性能評価を行った．

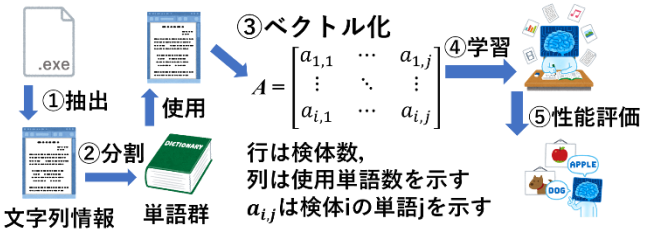

図 1　提案概要

2.2 機械学習システム

機械学習システムに適用するために，文字列情報を TF-IDF を用いてベクトルに変換した後に，L2 正規化を適用した．TF-IDF の算出式を下記に示す．

$$\mathrm{tf}-\mathrm{idf}=\mathrm{tf}(t,d)\times(\mathrm{idf}(t,d)+1) \quad (1)$$

$$\mathrm{idf}(t,d)=\log\frac{1+n_d}{1+\mathrm{df}(d,t)} \quad (2)$$

tf(*t*,*d*)は検体 *d* における単語 *t* の出現回数，n_dは検体の総数，df(*d*, *t*)は単語 *t* を含んでいる検体 *d* の個数を表す．出現回数の多い単語を 1000 個から 10000 個まで 1000 個ずつを増加させたベクトルとすべての単語（1,745,325 個）を使用したベクトルの作成を行った．機械学習アルゴリズムは Random Forest を使用した．

3. 性能評価

3.1 性能評価に活用したデータセット

本検討ではFFRIDataset2019[1]に含まれる文字列情報を抽出して使用した．当該データセットは株式会社 FFRI セキュリティが提供をしている PE 形式のマルウェアおよび正規ファイルの表層解析ログのデータセットである．本検討では FFRI Dataset2019 から正規データ 500 件，マルウェア 500 件をランダムに抽出して使用した．

3.2 性能評価の方法と評価結果

機械学習システムの分類精度の評価に当たっては 5 分割交差検証を実施した．分類結果の評価には正解率と F 値を使用した．性能評価実験の測定結果を図 2 に示す．

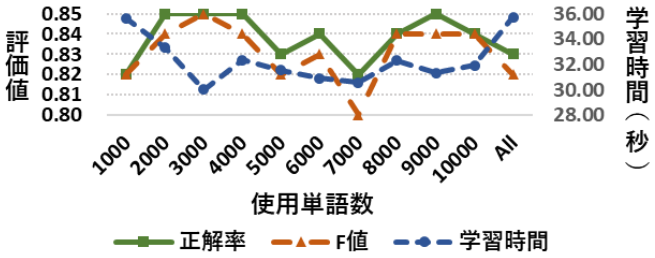

図 2　性能評価実験結果

4. 考察

ベクトル化に使用した単語数は 1000 個から 10000 個とすべての単語であったが，3000 個使用した場合が，正答率，F 値が最も高かった．使用単語数を削減しても分類精度は大きくは低下しないことが分かった．この理由は検体における単語の出現回数に対して検体の総数が少ないため，tf(*t*,*d*)と比べて idf(*t*,*d*)は分類精度の評価に関わる寄与度が小さいことが考えられる．

5. まとめと今後の予定

本稿では文字列情報を活用したマルウェア検知手法の提案をした．今後は使用データ量を増やし，未知マルウェアに対しても分類精度を維持できる手法の検討に取組む予定である．

参考文献

[1] 荒木粧子，他：マルウェア対策のための研究用データセット MWS Datasets 2019~，情報処理学会，Vol.2019-CSEC-86, No.8, 2019.7