講演番号： 201

時系列情報を利用したマルウェア検知方式の検討

D－19　　Study of malware detection method by using time series information

山本　貴幸†　宮保　憲治†
Takayuki YAMAMOTO†　Noriharu MIYAHO†
† 東京電機大学情報環境学部情報環境学科
† School of Information Environment, Tokyo Denki University

1. はじめに

近年、マルウェアを作成するツールなどが普及し、マルウェア発生件数は増加の一途をたどっている。それに伴い、シグネチャベースの検出では、新種のマルウェアへの対応が難しくなってきている。本検討では、シグネチャを用いず、機械学習を用いて通信を監視することでマルウェアに感染しているか否かの早期発見が可能である。本稿では、時系列情報を考慮したマルウェア検知方式を検討した事を述べる。

2. 実験内容

本検討では、マルウェア感染後の通信ログに D3M データセット、正常通信の通信ログに 2015 年 11 月 24 日から 27 日に渡って取得した通信ログを使用した。特徴量の抽出方法は早期発見を考慮するため、タイムスロット幅 1 秒に対して、オーバーラップ無しとオーバーラップ 0.5 秒の 2 つを調査した。使用する特徴量として、過去に多く用いられる特徴量[1][2]の中から 15 種類を使用した。本検討で使用する特徴量 15 種類を表 1 に示す。

表1　特徴量 15 種類

TCP パケット数	SYN フラグ数	FIN フラグ数
RST フラグ数	PSH フラグ数	ACK フラグ数
URG フラグ数	SYN/ACK 数	FIN/ACK 数
PSH/ACK 数	RST/ACK 数	通信ポート数
平均通信間隔	合計パケットサイズ	平均パケットサイズ

識別の際に時系列情報を考慮すると識別率の向上に有効であることが確認されている[3]。参考文献[3]では学習器の出力を時系列情報として扱っているが、識別結果が出るまで待つ必要があり、感染の早期発見が難しくなる可能性も考えられる。本稿では、学習データに時系列情報を持たせた場合の識別率の向上効果を評価した。具体的には、未加工データには現在のタイムスロットから抽出した特徴量にラベルを付与したものを使用し、加工済データには現在のタイムスロットから抽出した特徴量に 1 つ前のタイムスロットから抽出した特徴量を結合し、ラベルを付与した。

学習器に、適応的にサンプルの重みを更新して識別器の精度を増強する AdaBoost[1]と、NeuralNetwork を用いて、10 分割交差検証による検討を行った。

3. まとめ

NeuralNetwork と AdaBoost を用いて識別率の検証を行った結果、ともに識別率の向上を確認することができた（図 1、2 参照）。この結果、学習データに時系列情報を付与する操作は、学習器の性能向上に効果的であると考えられる。

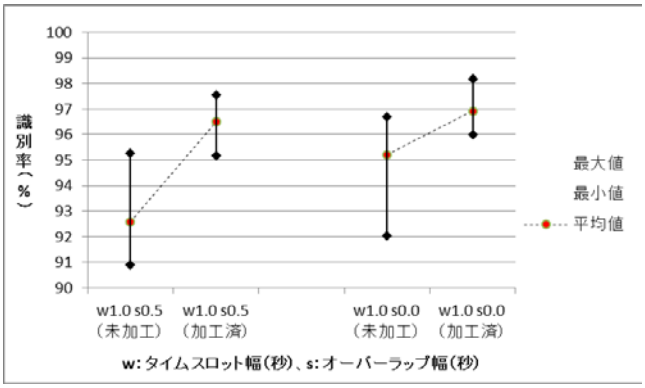

図 1　NeuralNetwork による識別率の変化

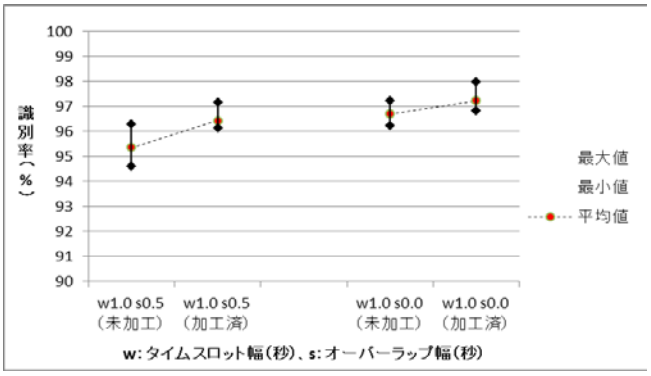

図 2　AdaBoost による識別率の変化

参考文献

[1] 川元研治，市田達也，市野将嗣，畑田充弘，小松尚，"マルウェア感染検知のための経年変化を考慮した特徴量評価に関する一考察"，コンピュータセキュリティシンポジウム 2011 論文集，277-282，2011-10-12
[2] 大島佑典，澁谷優貴，新津善弘，"マルウェア検知における特徴量の組み合わせによる検知率向上法"，電子情報通信学会，2014 年
[3] 市野将嗣，市田達也，畑田充弘，小松尚久，"トラヒックの時系列データを考慮したマルウェア感染検知手法に関する一検討"，情報処理学会，コンピュータセキュリティシンポジウム 2011 論文集，283-288，2011-10-12