4X-01

実行モジュールの特徴量と通信パターンを併用したマルウェア分類手法

蛭田 将平† 山口 由紀子†† 嶋田 創†† 高倉 弘喜††
†名古屋大学工学部電気電子情報工学科 ††名古屋大学情報基盤センター

1 はじめに

近年急増している標的型攻撃には，未知のマルウェアが使用される場合が多い．またマルウェアの種類も多様化しており，C&C サーバを利用し悪意あるプログラムをダウンロードさせるマルウェアや，PC の機能の一部をロックし解除させるために身代金を要求するマルウェアが特に急増している．日々増加するマルウェアに対しマルウェア解析者の人数は不足し，さらに未知のマルウェアを解析するコストは大きい．そのため，マルウェア解析のコストを軽減させる高速なマルウェア分類手法が求められている．

本稿ではマルウェアの通信パターンからパケット単位で特徴量を抽出し，既知マルウェアとの類似性を元にマルウェアの分類を行う手法を提案する．複数の種類のマルウェアと通信パターンが酷似しているマルウェアに対しては，マルウェアの実行モジュールから特徴量を求め，既知マルウェアと比較することでマルウェアの分類精度を向上させる．

2 マルウェア分類の提案手法

提案手法の流れを図 1 に示す．なお本研究で使用するパケットキャプチャファイルは，2013 年 10 月から 2014 年 10 月までに，NTT のマルウェア動的解析システム BotnetWatcher[1] において別途収集した 1049 のマルウェア検体を動作させて取得したものである．以下に提案手法の具体的な手順を示す．

2.1 特徴抽出

同じツールで生成された同一の目的を持つマルウェアは部分的に似たような通信を行うことが推測される．したがって，マルウェアの通信パターンからマルウェアを分類する有益な情報が得られると期待できる．なお，マルウェアによって生成されたパケットの出現パターンを通信パターンと呼ぶ．

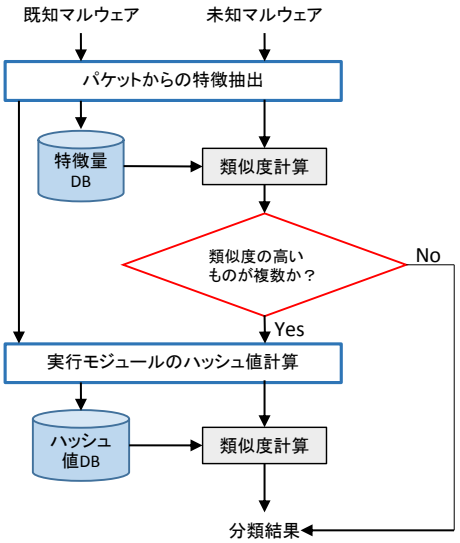

図 1: マルウェア分類の提案手法のプロセス

本手法では，まずマルウェアを実際に実行させ得られたパケットキャプチャファイルからパケット単位で特徴を抽出し，クラスタリングに使用する特徴量ベクトルを作成する．

特徴量はサービスポート番号，パケット長，通信プロトコルで構成される．通信プロトコルは ICMP，ARP，TCP，UDP に加えて，マルウェアの活動と関連性が高い DNS と HTTP を個別のプロトコルとして扱うこととした．

2.2 クラスタリング

特徴抽出によって得られた特徴量ベクトルをクラスタリングする．本手法では，クラスタリング手法として非階層クラスタリングである *k*-means++[2] を使用する．*k*-means++は与えられたデータから初期値として *k* 個のクラスタ中心を距離ができるだけ遠くなるように選択し，各クラスタに割り当てられるデータとの距離の分散を最小化するように動作する．

2.3 クラスタ列生成

各マルウェアの通信パターンについてクラスタリングによるラベリングを行いクラスタ列を生成する．生成されたクラスタ列の例を表 1 に示す．ここでは $k = 20$ として各クラスタに A から T のラベルを割り当てている．

Malware classification method by analyzing traffic patterns with executable module features
Shohei HIRUTA† Yukiko YAMAGUCHI†† Hajime SHIMADA†† Hiroki TAKAKURA††
†Department of Electrical and Electronic Engineering and Information Engineering School of Engineering, Nagoya University
††Information Technology Center, Nagoya University

表 3: ssdeep による実行モジュールのハッシュ値の例 (太字は同じ種類のマルウェア内で共通な部分)

マルウェア名	ハッシュ値
HERU:Trojan.Win32.Generic 1	CMQ+SAN7uprgvM5OSUswZXg69gbm4hfpF**mRvR6JZlbw8hqIusZzZ**Ie
HERU:Trojan.Win32.Generic 2	gY324bcgPiJLQrfARGSRUJsbY6ZgvSMBD3t8**mRvR6JZlbw8hqIusZzZ**x2
Backdoor.Win32.DarkKomet 1	**W9HFJ9rJxRX1uVVjoaWSoynxdO1FVBaOiRZTERfIhNkNCCLo9EkNC/**
Backdoor.Win32.DarkKomet 2	6**9HFJ9rJxRX1uVVjoaWSoynxdO1FVBaOiRZTERfIhNkNCCLo9EkNC/**

表 1 の例より，同じ種類のマルウェアについて類似したクラスタ列が得られることから，クラスタ列を適切に扱うことで，マルウェアを分類できることを示している．

表 1: 通信パターンのクラスタ列の例

マルウェア名	クラスタ列
Trojan.Win32.Yakes 1	QQRRRSRRSRRLLLLLLRRL...
Trojan.Win32.Yakes 2	QQRRRRSRRSRRLLLLLLRR...
Adware.Win32.Agent 1	RRRRSRRRRRSRRQQRRERS...
Adware.Win32.Agent 2	RRRRSRRRRRSRRRRESBBB...

2.4 クラスタ列の類似度計算

各マルウェアのクラスタ列を比較し類似度を計算して未知マルウェアを分類する． 類似度計算にはゲシュタルトパターンマッチングを使用する [*]．

ゲシュタルトパターンマッチングは比較する 2 つのクラスタ列に対し，最長共通部分列を再帰的に求め，すべての共通部分列の長さの和を 2 つのクラスタ列の長さの合計で除算した結果を類似度とする．完全に一致するなら類似度は 1 となり，いずれの部分列でも一致しないなら類似度は 0 となる．

2.5 実行モジュールのハッシングによる類似度計算

一部のマルウェアでは通信パターンが類似しているものが複数存在し，正しく分類することができない．表 2 に，通信パターンによるマルウェア分類の予備評価で見られた例を示す．

表 2: Trojan.Win32.Yakes との類似度計算の例

マルウェア名	類似度
Trojan.Win32.Yakes	0.835
Trojan-Ransom.Win32.Foreign	0.802
Trojan-FakeAV.Win32.FakeSysDef	0.859

この問題を解決するため，あるマルウェアと高い類似度を示す既知マルウェアが複数存在した場合，Fuzzy Hash を用いてそれぞれのマルウェアのハッシュ値を取得し，これを元に類似度計算を行う．

Fuzzy Hash には CTPH(Context Triggered Piecewise Hashes) を計算する ssdeep を使用する [†]．データの先頭から一定のデータ長に対して部分ハッシュ値を生成して連結する手法であるため，同一の部分データを含むデータからは類似するハッシュ値が得られる．ssdeep の本特性により，同じ種類の実行モジュールを持つマルウェア間では高い類似度を示すと期待できる．類似度の計算には前述のゲシュタルトパターンマッチングを使用する．

表 3 にマルウェアの実行モジュールを ssdeep でハッシングした例を示す．表 3 の例より，同じ種類のマルウェアの ssdeep によるハッシュ値は類似する文字列を含んでいるため，マルウェアを分類する上で有用な手がかりになることが期待できる．

3 おわりに

本稿ではマルウェア解析コストを軽減させるため，通信パターンと実行モジュールを利用したマルウェア分類手法を提案した．

現時点では，提案手法の性能評価に十分なマルウェアの検体を収集できておらず，Fuzzy Hash によるマルウェア分類の有効性を確認できていない．提案手法の今後の課題として，同様の通信パターンを持つ複数の種類のマルウェアに対して，実際に分類することが挙げられる．

謝辞

共同研究を通じて研究用データセットを提供してくださった NTT セキュアプラットフォーム研究所に深く感謝する．

参考文献

[1] 青木一史ら，”半透性仮想インターネットによるマルウェアの動的解析 ”，CSS 2009，2009.

[2] David Arthur，et al.， ”k-means++: The advantages of careful seeding,” SODA07，pp. 1027-1035，2007.

[*] Ratcliff/Obershelp pattern recognition, http://www.nist.gov/dads/HTML/ratcliffObershelp.html

[†] Fuzzy Hashing and ssdeep，http://ssdeep.sourceforge.net/