2Y-3

機械語命令列の類似度分析を用いた不正コードの分類

碓井　利宣†　　重松　邦彦‡　　水谷　正慶‡　　武田　圭史†　　村井　純†
慶應義塾大学環境情報学部†　　慶應義塾大学大学院政策・メディア研究科‡

1. はじめに

本研究では，静的解析手法を用いて，マルウェアの利用している DLL や API に着目してマルウェア間の類似度を評価し，それを基に分類する手法を提案する．その提案する手法によって，自動的にマルウェアの類似度を評価し，類似性の高いマルウェアや同じソースコードを元にしていると考えられる亜種などを分類するソフトウェアを作成した．

2. 背景と目的

近年，IPA（情報処理推進機構）や，ウイルス対策ソフトを開発している企業の Web サイトで，数多くのマルウェアの報告がされており，その種類も多岐に渡っていると言える．また，パッカーによって，亜種の作成が容易となっており，マルウェアが大量に出現する要因も多い．
そこで，本研究では，静的解析に基づいて自動的にマルウェアの分類を行うソフトウェアを作成することによって，マルウェアの基礎研究の支援や，ウイルス対策ソフトウェアの作成の一助とすることを目的とする．

3. 関連研究

API に焦点を当てて分類を行った例としては，Ismael らの研究[1]や岩本らの研究[2]がある．前者は API のコール・ツリーなどを基にマルウェア同士の比較を行うものであり，後者は API の推移をマルウェアの特徴としたものである．一方で，岩村らの研究[3]では，コール・ツリーが似ていても，関数内での API などの中身が違う場合はうまく機能しないという点を指摘している．また，API の類似性を基にしている研究は多いが，メタデータという形で実際にラベリングを行う例は少ない．そこで，本研究では，関数レベルでの類似度を考慮するとともに，研究者や開発者が検体を探す際の補助となるように，メタデータを用いたラベル付けを行う．

4. 手法

本章では，ソフトウェアの作成にあたって利用した手法について述べる．分類はメタデータを基にして行うものと，2 つのマルウェア間での類似度を算出することによって行うものの 2 段階によって行われる．以下に，分類の際に利用する類似度を算出する手法と，メタデータの付与の仕方，実際に分類を行うための手法のそれぞれについて記す．

4.1 類似度評価

2 つのマルウェアでインポートされている DLL,API をそれぞれ一覧で取得し，比較を行うことで，両者で共通してインポートされているものの個数を得る．これを 2 マルウェア間の DLL，API の全個数の平均で割ることで得られた値を類似度の 1 つとする．また，マルウェアの内部で作成された関数（以下，内部関数と呼ぶ）の特徴によってマルウェア間で関数レベルでの類似度を取得し，それをマルウェアの類似度を評価する 1 つの基準とする．まず，マルウェア内で作成された内部関数の範囲を探す．そしてその中での API 関数の呼び出しを元に，内部関数ごとに呼び出されている API 関数名の一覧を取得する．それを 2 つのマルウェア間で内部関数ごとに比較し，一致した内部関数の個数を内部関数の全個数の平均で割ったものをマルウェア間の内部関数レベルでの類似度とする．得られた 2 つの値はそれぞれ 0.0～1.0 の小数値であり，その 2 つの和をマルウェア間の正味の類似度とする．

4.2 メタデータの付与

メタデータの付与はマルウェアの利用している API を基に行う．働きの方向性が同じ API が一定数

Malware Classification by Similarity Analysis of Machine Instruction Sequence
Toshinori Usui†, Kunihiko Shigematsu‡ , Masayoshi Mizutani ‡ , Keiji Takeda†, Jun Murai†
†Faculty of Environment and Information Studies, Keio University 252-8520, Kanagawa, Japan
‡ Graduate School of Media and Governance, Keio University,252-8520, Kanagawa, Japan
{alc, sigematu, mizutani, keiji, jun}@sfc.wide.ad.jp

以上インポートされていた場合，その働きに準じたメタデータをマルウェアに付与する．例えば，ネットワークを利用する API が一定数以上インポートされていた場合，ネットワークというメタデータが付与される．

4.3 分類

初段階の分類として，付与されているメタデータの種類が同じもの同士をクラスタとして分ける．さらにその中で第二段階の分類として，あらかじめ閾値を設定しておいた閾値を基に，類似度が閾値以上のマルウェア同士を同じクラスタとしてまとめていくという分類の手法を用いる．

5. 実装

実装したソフトウェアの機能は，情報抽出，メタデータ付与，類似度算出，分類の 4 つに大きく分けられる．本章では，それぞれの実装手法について述べる．

5.1 情報抽出

DLL,API の取得は，PE ヘッダの Import Table Address の値を基にインポートセクションを参照することで行う．インポートセクションを構成する IMAGE_IMPORT_DESCRIPTOR 構造体を辿り，Name メンバから参照できる DLL 名と，IMAGE_IMPORT_BY_NAME 構造体に格納されている API 名を得る．

内部関数の範囲とその中で呼び出されている API については，call 命令に着目することで取得する．PE ヘッダの BaseOfCode の値を基にコード領域の先頭を参照し，逆アセンブラライブラリである HDE32 を利用して逆アセンブルしていく．そして，near call 命令を探し，その即値を参照することによって内部関数の展開されているアドレスを取得する．さらに，その内部関数の中での far call 命令を探し，そのアドレス指定部分を参照することによって，内部関数で呼び出している API の展開されているアドレスを取得する．API のアドレスと API の対応付けは，インポートセクションの Import Address Table に含まれるメモリ上のアドレスと Import Name Table に含まれる API 名を紐付けすることで行う．

5.2 メタデータ付与

メタデータの付与は，情報抽出の段階で取得した DLL,API の一覧を，メタデータと API の対応を格納したデータベースと照らし合わせることで行う．あらかじめメタデータを付与する API の個数の閾値を決めておき，その個数以上の API がインポートされていた場合に，当該メタデータを付与する．

5.3 類似度算出

情報抽出部分によって抽出された情報を基に，先に述べた手法を用いて，類似度を算出する．算出した類似度は，0.0～2.0 の小数値で保持する．

5.4 分類

4.3 で述べた分類手法を用いて，前項での類似度算出部分で算出した類似度を基に分類を行う．

6. 今後の課題

本手法は，マルウェアを 1 対 1 で比較した結果を用いることを基準としており，検体数が増えれば計算量は幾何級数的に増加する．より効率的な手法を模索する必要であると言える．また，メタデータと類似度を基にしたクラスタリングにも改善の余地があり，NN や SVM，k-means 法など，様々な手法との親和性を考察し，よりよい結果が生まれるよう考えるべきである．

7. まとめ

マルウェアがインポートしている DLL,API に焦点を当て，類似度を評価する手法を提案し，それとともに，API を基にして付与したメタデータやマルウェア同士の類似度によって分類を行うソフトウェアを作成した．これを用いて分類することにより，機能を基にしたマルウェア検体の選出や，同種・亜種などのマルウェア間の関係の把握も容易になった．

参考文献

[1] GRAPHS, ENTROPY AND GRID COMPUTING: AUTOMATIC COMPARISON OF MALWARE Ismael Briones and Aitor Gomez

[2] 静的解析によるマルウェアの API 推移の抽出とクラスタ解析 岩本 一樹，和崎 克己 CSS2009

[3] 機械語命令列の類似性に基づく 自動マルウェア分類システム 岩村 誠，伊藤 光恭，村岡 洋一

[4] マルウェアコードの類似度判定による機能推定 安本幸希，森井昌克，中尾康二

[5] Malware Detection Based on Mining API Calls Ashkan Sami, Babak Yadegari, Hossein Rahimi, Naser Peiravian, Sattar Hashemi, Ali Hamze