6Z-4

サブルーチンの静的解析に基づくマルウェア分類手法の提案

碓井 利宣† 重松 邦彦‡ 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‡

{alc, keiji, jun}@sfc.wide.ad.jp, sigematu@sfc.wide.ad.jp

1. 背景と目的

インターネット利用の拡大に伴って，多様な種類のマルウェアによる脅威が顕在化してきている．ウイルス対策ソフトウェアベンダである McAfee による脅威レポート[1]によると，毎月平均で約 200 万件の新種および亜種の新しいマルウェア検体が収集されている．日々増加を続けるマルウェアに対して効果的な対策を効率的にとっていくには，マルウェアを解析することが不可欠である．しかし，マルウェアの増加や巧妙化により，出現した新しいマルウェアを迅速に解析することは難しくなっている．

本研究では，以下の 2 つの手法によってマルウェアに関する情報を取得し，それらを提供することで解析を支援することを目的とする．1 つめは，静的解析によって得られた情報を基に分析を行うことで，マルウェアの「種」（＝ファミリー）を特定することで，亜種などの他のマルウェアとの関係性を明らかにすることである．2 つめは，マルウェアの利用しているサブルーチンを解析し，動作の概要を特定することである．これらによって，解析者は解析に先立って事前に情報を得ることで，より迅速に解析を行えるようになることが期待できる．

2. 関連研究

マルウェアの分類や動作特定に関する既存研究の代表的なものについて挙げ，分析する．岩本らの研究[2]では，マルウェアの機械語命令列から Longest Common Subsequence を抽出し，その類似度により分類を行う手法と，その効率化の提案を行い，優れた成果を得た．しかし，動作に着目した分類手法ではないため，マルウェアの動作は把握出来ない．また，勝手らの研究[3]では，マルウェアの使用 API の種類を基にマルウェア間の親近度を算出し，数量化 IV 類によって多次元空間内に配置した．マルウェア間の位置関係によって機能の特定を行っている．しかし，新しい機能が追加された亜種の場合は親近度が低く出る可能性があり，必ずしも全てのマルウェアにおいて正確に機能を推定出来るとは限らない．本研究では，これらの課題の解決を試みる．

3. 提案手法

3.1. サブルーチンの動作の解析

サブルーチンの動作の解析は，サブルーチン内で行われている API の抽象化とその組合せによって行う．ラベルという API を抽象化した概念を提案する．ラベルとは，API の動作の方向性を記述したものである．例えば，RegOpenKeyEX や RegCreateKeyEx などの API が用いられていれば，そのサブルーチンではレジストリの操作を行うことが分かる．したがって，レジストリ操作というラベルをサブルーチンに付与する．このようにしてサブルーチンに付与されたラベルの組合せから，サブルーチンの動作を特定する．動作の特定についても，ラベルの組合せとマルウェアの動作との対応をあらかじめ定義したものを用いて行う．

3.2. 分類によるファミリーの特定

マルウェアの分類には，機械学習手法である Support Vector Machine（以下，SVM）を用いる．まず，ファミリーが既知のマルウェアの各ルーチンが利用している API の種類とその呼び出し回数を特徴ベクトルとして抽出する．そして，それらをファミリーによって分類先クラスを分け，教師信号として多クラス分類器に入力することで，学習モデルをビルドしておく．分類対象が出現した場合，同様に各ルーチンから前述の通り API の情報を抽出し，入力する．これによって，SVM による認識結果として，分類先クラスを得る．学習段階において，分類先クラスはファミリーを基に決定されているため，得られた分類先クラスから，未知のマルウェアであっても，ファミリーを特定することが可能である．

4. 実装

前章で提案した手法に基づいてマルウェアを解析・分類するシステムを実装した．図 1 にシステムの構成を示し，以下に実装の詳細を述べる．本システムは，事前準備部，情報抽出部，サブルーチン解析部，分類部の 4 つの部分によって構成されている．以下に，各部の詳細を示す．

事前準備部　事前準備部では，メモリダンプによるマルウェアのアンパックや IAT の再構築など，静的解析を行う上で必要な処理を行う．

A Proposal for a Method of Malware Classification based on Static Analysis of Subroutines.

Toshinori Usui†, Kunihiko Shigematsu‡, Keiji Takeda†, Jun Murai†

†Fuculty of Environment and Information Studies, Keio University

‡Graduate School of Media and Governance, Keio University

情報抽出部　情報抽出部では，静的解析手法による情報の抽出を行う．インポートテーブルから API の情報を抽出し，コードセクションを逆アセンブルしていくことで，各 API の呼び出し元位置と回数を取得する．

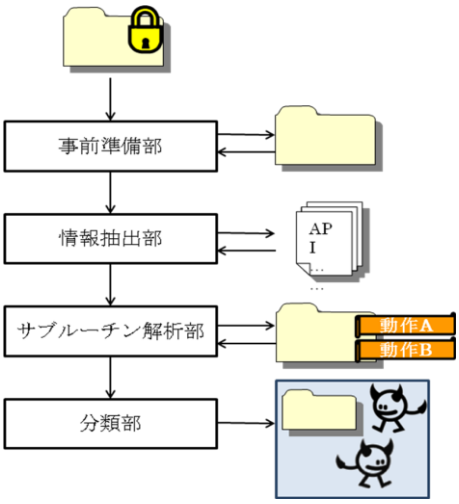

図.1 システム構成図

サブルーチン解析部，分類部　サブルーチン解析部および分類部では，前章で提案した手法によって，それぞれサブルーチンの動作の特定および分類によるファミリーの特定を行う．

5. 実験と評価

5.1. 実験と結果

本システムを用いてマルウェアの動作の概要を特定する実験と，分類を行うことでファミリーを特定する実験を行った．動作特定実験では，研究室の保有するマルウェア収集サーバより特徴的な動作を持つ 10 検体を抽出し，その動作のうちいくつを本システムが正しく特定することが出来るかを，IPA ウイルス情報データベース iPedia[6]で公開されているマルウェアの動作の特徴と比較することによって確認した．分類実験では，同サーバより学習データとして 2000 検体を，分類対象として 1000 検体をそれぞれ重複のないよう注意して無作為抽出し，分類した．分類に先立って，学習データとなる既知のマルウェアの属するファミリーを定義しておく必要がある．ファミリーの定義には，IPA（情報処理推進機構）が公開している届出ウイルス一覧[4]に掲載されている表記を用いた．ある既知のマルウェアがどのファミリーに属すかは，VirusTotal[5]によって検知されるマルウェア名の中に，どのファミリー名が出現するかで判定した．そして，実際に分類することによって，システムの示したファミリーと，前述のファミリーの定義方法によって得たファミリーとが一致するかを基に検証を行った．

動作特定実験の結果，各検体について，動作を特定することが出来た．具体的には，4 検体について 1 つあるいは 2 つの動作が特定出来なかったことを除き，動作を正確に特定することが出来た．一例として，Sasser に属する検体がファイルコピー，レジストリキー作成，ミューテックス作成，バックドア作成，Web リソースアクセス，ファイル作成を行うことに対して，ファイル操作，レジストリキー操作，ミューテックス作成，バックドア作成，ネットワーク利用という動作が得られた．また，分類実験の結果，683 検体を正しい分類先に分類し，ファミリーを特定することが出来た．これらの結果について，次節で議論する．

5.2. 評価

動作特定実験の結果より，マルウェアの動作とサブルーチンで用いられている API の傾向には相関性があると言える．提案手法によって，動作の概要を推定可能である．一方で，実験で特定出来なかった動作もあった．今回の実験では，これはサブルーチンのラベルと動作との対応を定義出来ていなかったものが主であるため，この対応を整備することで改善が期待出来る．一方で，現時点では特定出来るのは概要に過ぎず，例えば，ネットワーク接続を行っていることを特定出来ても，接続先やポート番号などは特定を試みていない．提供出来る情報は多い方が良いため，これらは今後の課題である．

また，分類実験の結果より，分類によるファミリーの特定が可能であることが分かった．今回の認識精度は 68.2%であるが，実験検体を抽出する母集団に偏りがあることなど，精度が低下する要因と考えられるものあり，学習データ量を増やすことなどによっても精度の向上が考えられるため，今後の発展が望める．

6. まとめ

本研究では，マルウェアのサブルーチンから静的解析によって得た API の情報を基にマルウェアの動作特定と分類を行う手法を提案し，それに基づいたシステムの実装を行った．実験結果より，提案手法によってマルウェアの動作の多くを特定することが可能であることを検証出来た．また，ファミリーについても同様に特定することが可能であることが確認された．本研究の成果によって，マルウェアが出現した際に，解析者による解析に先立って事前に情報を提供することが可能であり，解析や対策の立案をより有効かつ効率的に実施出来るようになることが期待される．

参考文献

[1]McAfee 脅威レポート 2011 年第 2 四半期
http://www.mcafee.com/japan/media/mcafeeb2b/international/japan/pdf/threatreport/threatreport11q2.pdf
[2]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，村岡洋一情報処理学会論文誌 51,9 2010,Sep
[3]マルウェアの分類方法とその応用に関する考察 勝手壮馬，安本幸希，伊沢亮一，森井昌克，中尾康二
[4]情報処理推進機構 届出ウイルス一覧
http://www.ipa.go.jp/security/virus/virus_main.html
[5]VirusTotal
https://www.virustotal.com/
[6]情報処理推進機構ウイルス情報データベース iPedia
https://iseclab.ipa.go.jp/zha-virusdb/web/Detail.php
[7]コンピュータウイルスのコード静的解析による特徴抽出と分類について 岩本一樹，和崎克己
[8]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