3T-9

アセンブリ言語レベルでの未知ウィルス検出方法

Detecting unknown viruses in assembly language domain

田中秀和　中井斉　堀田貴之　下江啓次郎　直江健介　武藤佳恭
慶應義塾大学環境情報学部　慶應義塾大学総合政策学部　慶應義塾大学政策・メディア研究科

概 要

現在存在するポリモーフィックウィルスやメタモーフィックウィルスの検出は、既存のアンチウィルスソフトを用いても困難だとされている。これはウィルスコードが感染するたびに自分自身のコードを書き換えるためである。また未知のマルコードに対しても、単純なパターンマッチングによる検出では回避されてしまう。本論文ではこれら未知ウィルスを検出する方法を示す。具体的にはウィルスバイナリコードを逆アセンブルして、アセンブラコードレベルでの検出と検出されたパターンからの学習、学習結果から分類判定による検出を行う。

1　はじめに

近年ウィルスやワームといったマルコードによる被害は増加の一途を辿っている。その背景には多種化、巧妙化したウィルスの蔓延が起因していると考えられる。従来からのウィルス検出の手法として、定義ファイルにウィルスの特徴パターンを記述し、それを用いてパターンマッチングをする方法が多く用いられてきた。しかし、メタモーフィックウィルスや未知のウィルスのパターンは定義ファイルに格納されていないため、この手法ではウィルスを特定することが出来ない。実際にはこのようなウィルスを検出するために、ヒューリスティック探索などを利用している。しかし、誤警報や誤検出を発生するという問題、さらにこの手法を用いた場合においても総合的な検出率は70%〜80%に留まるという問題を抱えている [1]。本論文の目的は、完全な防御方法の無いメタモーフィックウィルスのような自分自身を書き換えてしまうコードや亜種を生じさせるウィルスに対する検出精度を向上させることにある。このような問題を解決するために、静的にウィルスのバイナリコードを逆アセンブルし、アセンブリ言語レベルによるウィルスのパターンの解析を行う。また、ウィルスパターンについて学習アルゴリズムを組み込んだ検出手法を用いることで、未知ウィルス、及び亜種ウィルスの総合的な検出精度の向上を目指す。

2　既存のウィルス検出手法の問題点

メタモーフィックウィルスの特徴として、以下の三点が挙げられる。

1．　ウィルス本体をいくつかのブロックに分割し、そのブロックの順番をジャンプ命令などで入れ替える

2．　実行しても意味が無い、ジャンクコードを挿入する

3．　ある機能を実現するコード中の命令を複数の命令で実現するようにする。あるいは複数の命令を1つの命令で実現できるようにする。

これらのことからメタモーフィックウィルスに対しては、定義ファイルに基づくパターンマッチング検出での検出率が低くなることが知られている。そのため、既存のアンチウィルスソフトではファイルのコピー回数、レジストリ操作などの処理といったウィルスに類似する特定の動作の組み合わせなどからウィルスであるかを判断するようなヒューリスティック探索を用いて検出している [2]。しかしヒューリスティックスキャンではウィルスの感染は発見できても、何のウィルスかは特定できない。さらには完全な検出を約束しないため、その疑わしいコードを隔離することは出来ても、ウィルスコード部のみを削除するといった処理は非常に難しいとされる。結果的にウィルスを特定するためには、パターンマッチング手法との複合的な検出手法を用いる必要があるのが現状である [3]。

3　提案手法

3.1　アセンブラレベルでの解析

ヒューリスティック手法を用いた検出手法の問題を解決するために、ウィルスの挙動であると特徴付けるルール群をより細かくすることを提案する。すなわち、コードが実行する動作というルールだけでなく、ウィルスのバイナリコードを静的に逆アセンブルし、アセンブリ言語レベルで対象ファイルを解析するアプローチをとる。この操作によって、従来の静的ヒューリス

ティックスキャンのルールに、さらに細かいパターンを定義出来ることとなる。具体的には、メタモーフィックコードによく見られるようなファイル内の規則的なジャンプ命令やジャンクコードの羅列などが発見された場合、それはメタモーフィックウィルスであると判断することとする。また自分自身を書き換えるルーチンは必ず存在するため、このルーチンを検出することでウィルスの判断材料とする。さらにここで動的なビヘイビア検出を行うことによって、より高い確率でそのファイルがメタモーフィックウィルスであると判断できる。アセンブリ言語レベルでの解析は後述するウィルスコード学習の際必要となるウィルスコードの特徴語群の抽出に必要不可欠なプロセスとなっているため非常に重要なプロセスとなる。

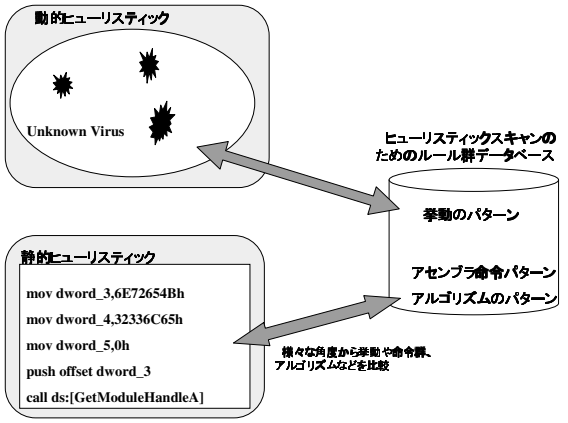

図 1: アセンブラレベルによって検出をするイメージ

3.2 パターン学習による検出

疑わしいコードをアセンブリ言語に直した後、ニューラルネットワークを用いた学習アルゴリズム [4][5] によるパターン検出およびその学習結果に基づく分類判定による検出を行う。オペランド、オプコードといったそれぞれのニーモニックを単位としたベクトルを作る必要があるため、アセンブリ言語による解析の後にはじめて学習を行うこととなる。学習アルゴリズムには教師付き学習と教師なし学習に大きく分けることが出来る。本論文では、教師なし学習を用いる。この理由として、学習対象がどのように変化するのかが推測できないため、教師付き学習では教師値を設定することが出来ないため、自動的に分類できる教師なし学習を用いる。学習による検出が必要な理由として未知なウィルス、または亜種のウィルスが出現した場合に高い確率でそのファイルがウィルスであると特定付ける必要があり、多くの場合既知のウィルスから、その挙動パターンやアルゴリズムパターンなど様々なファクターからパターンを学習するといった必要性が要求される。これによって、未知ウィルスや亜種ウィルスが出現したとしても、学習による効果からそのファイルがウィルスかどうかを高い確率で特定付けることが可能である。

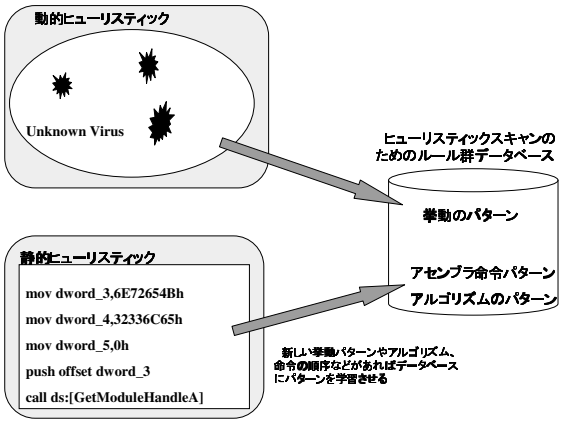

図 2: 学習アルゴリズムのイメージ

4 まとめ

アセンブリ言語に逆アセンブルしたコードからパターンを抽出して学習することによって、現存のヒューリスティック検出率の 70%〜80%とされる値から、さらに高い検出率まで向上させることが期待できる。また、新しいパターンを追加してヒューリスティックスキャンを用いる際のルールやパターンを増やすことによって、誤警報や誤検出を減少させることが期待できるであろう。

参考文献

[1] Symantec ホワイトペーパーシリーズ Vol XXXIV ヒューリスティック手法解説:シマンテックの Bloodhound 技術 http://www.symantec.com/region/jp/sarcj/reference/heuristc.pdf

[2] コンピュータウィルス脅威のメカニズム 2003 勝村幸広 日経 BP 社

[3] ウィルス対策マニュアル SOFTBANK 2003 David Harley,Robert Slade, Usr E.Gattiker

[4] ニューラルネットワークの設計と応用 1999 Bahman Kermanshai

[5] データマイニング 1998 山本英子 共立出版

[6] Ruo Ando, Hideaki Miura, Yoshiyasu Takefuji,"File system driver filtering against metamorphic viral coding",WSEAS TRANSACTIONS ON INFORMATION SCIENCE AND APPLICATIONS,Issue 4, Volume 1, October 2004,ISSN: 1790-0832