2017
年 度

類似トラフィックを用いた機械学習の初期学習にかかるコスト軽減

吉田健太

木下
研究室

[修士論文]

類似トラフィックを用いた機械学習の初期学習にかかるコスト軽減

（指 導 教 員）　木下 俊之

コンピュータサイエンス専攻　木下研究室

学籍番号 G2116032

吉田健太

[2017 年度]

指定様式 1：内表紙

東京工科大学大学院バイオ・情報メディア研究科

修士論文

論文題目

類似トラフィックを用いた機械学習の初期学習にかかるコスト軽減

指導教員

木下 俊之	印

提出日　　2018 年　　1 月　　23 日

提出者

専　　攻	コンピュータサイエンス専攻
学籍番号	G2116032
氏　　名	吉田健太

指定様式 2：概要

修士論文概要

論 文 題 目	類似トラフィックを用いた機械学習の初期学習にかかるコスト軽減
執 筆 者	吉田健太
指 導 教 員	木下 俊之 教授

現行 IntrusionDet tionSystem (IDS) にはアノマリ型とシグネチャ型が存在している。パターンファイルをセッティングするシグネチャ型は既存攻撃には強いが未知攻撃や亜種攻撃に対しては検知できない課題がある。そこで未知攻撃や亜種攻撃に対して検出するためのアノマリ型が IDS が注目を集めている。しかし実環境において稼働に耐える学習データを生成するデータ量を確保するには多大な時間的コストがかかることが問題となっている。本研究では類似するネットワークからキャプチャーしたネットワークデータを用いて学習にかかる時間的コストの削減を提案する。より実環境に近い状態を想定して脆弱性スキャナ及び実ネットワークのトラフィックを用いて評価を行う。なお、検知対象として HTTP における攻撃に注目しており、プロトコルは HTTP 限定する。

注 1：A4 サイズ、和文は 800 字以内、英文は 500words 程度、横書きで作成

指定様式 3：英文論旨

Abstract

Title	Cost reduction for initial learning using analogy traffic of machine learning
Author	Kenta Yoshida
Supervisor	Professor Kinoshita Toshiyuki

Write an abstract of your Paper.

注 1：英語要旨—300 ワード程度。

目 次

図 目 次

表 目 次

第1章

はじめに

1.1 はじめに

近年、インターネットが爆発的に普及し、個人間だけではなく企業間の取引・軍事的にも重要な役割を担っている。だが、普及と同時に多数の脅威も現れた。
その脅威の一部がマルウェアと呼ばれる。マルウェアの定義とは文脈によって様々であるが、不正かつ有害な動作をするもの、自己増殖を行い感染を拡大するものなどが代表的な定義とされている。
マルウェアの歴史は古く、まだインターネットが一部の大学・企業間のみの通信の時代から生成された。その後様々なマルウェアが作られ、OS 開発元やセキュリティソフトを開発してる企業などはいたちごっこを繰り返してきた。
インターネットにおける脅威はマルウェアだけではなく、ハッキング・クラッキングによる脅威も依然猛威を振るっている。
それらの脅威の行動源となっているものとして、初期は自己のハッカーとしての技術を自慢する承認欲求や感染対象の反応をみて快感を得る愉快犯が多かったが、近年では敵対企業の信用を落としたりランサムウェアによる身代金を得たりする利益を得ようとする傾向が出ている。
以上のようにインターネットの普及率や技術の向上から様々な脅威が発生する中、各国ではサーバー空間を第 5 の戦場として対応しているというのが現状である。

1.2 研究の背景

脅威の対応は初期からあまり変わらず、侵入に対するパターンマッチを侵入口に設置してある Firewall(FW)・Intrusion Detection System(IDS) にその脅威が発生するネットワークかパターンをセットし、それらで検知した通信を脅威をみなし検知・拒否を行う。そういったパターンはセキュリティアナリスト達の手動で生成されていたが、脅威の発生からパターン作成・配布のタイムラグによるゼロデイ攻撃が後を絶たなかった。だが 2000 年前半からマルウェアの検体などを機械学習にかけることによってパターンを自動生成する研究が進み実用化されたことによって、パターン生成の速度が上がっており、そういっ

た攻撃のリスクを減らすことが可能となった。
だが、これらの対策は後手に回ってしまい既存の攻撃には強くとも、未知の攻撃もしくは既存の攻撃にオリジナリティを加えた亜種攻撃に対しては脆弱であった。
そして、近年上記のようなパターンマッチを主とするシグネチャ型から、通常の通信を正常な通信とみなしそれ以外の通信を以上とみなすアノマリ型の研究が進んでいる。
だが、アノマリ型はシグネチャ型の特質と相反するものであり、未知攻撃・亜種攻撃には強くとも既存の攻撃には弱いという点が問題として挙げられている。更に、アノマリ型は正規の通信のデータの蓄積が精度と比例するため、大量の学習データが必要となり時間的コストがかかってしまうという点がある。

1.3 研究の目的

本研究ではアノマリ型 IDS に対して時間的コストの削減を目指す。
近年では様々な手法によってこの時間的コストや精度向上を目指す研究が行われているが、本研究ではそういったアルゴルズム的な視点ではなく、技術が進化しても対応できる時代的視点が長く使えるようなシステムの構築を目的とする。

第2章

関連技術

2.1 IDS

IDS とは Intrusion Detection System、つまりは不正侵入検知システムの略称である。IDS の機能として、ネットワーク上に流れるパケットを解析し、不正なアクセスの痕跡を見つけた場合管理者に通報するというものである。
通常 IDS は内部ネットワークと外部ネットワークとの境界に置き、両間の通信を読み取り解析するためサイバーセキュリティの要とも言えるシステムである。
似た機能を搭載したシステムとして FW・IPS が存在する。
FW は通常 IP アドレス・プロトコル・ポート番号のみをチェックし、中身を見ることができないため通常のアクセスと偽装したアクセスとの区別がつくことができない。例えば web 観閲のための Port80 は通す設定にしていると、マルウェア通信で Port40 通信している場合に検知することができない。
IPS は IDS と機能自体はほぼ同じなのだが、大きく違う点は IPS は異常を検知して管理者に通報するのみだが、IDS はアクセス自体を遮断するということである。IPS と IDS はセキュリティポリシーや設置場所によって使い分けることが重要となる。

2.2 scikit-learn

近年機械学習開発が積極的に進んでおり、様々なアルゴルズムが使用できるよう、ライブラリの開発が進んでいる。
scikit-learn はそのような複数のライブラリをまとめて使いやすくしたライブラリである。サポートベクターマシン、ランダムフォレスト、Gradient Boosting、k 近傍法、DBSCAN などを含む様々な分類、回帰、クラスタリングアルゴリズムを備えており、Python の数値計算ライブラリの NumPy と SciPy とやり取りするよう設計されている。

2.3 metasploit

2.4 Nessus

第3章

提案

3.1 システムの全体説明

3.1.1 正規通信の収集・加工方法

3.1.2 学習データの統合

3.2 クライアント側での運用

第4章

実装

章立ては指導教員の方針に従ってください.

第5章

評価

5.1 脆弱性スキャナ Nessus による外部からの攻撃

5.2 metasploit による内部からの通信

5.3 シグネチャ型 IPS との比較

第6章

評価

6.1　結果概要

謝辞

本論文の作成にあたり、終始適切な助言を賜り、また丁寧に指導して下さった木下 俊之先生にこの場を借りて感謝の意を表します。また、励ましやデータサンプルの協力・精神的支えになってくれた友人各位には深く感謝致します。本当に有難うございます。

参考文献

付録 A

ソースコード

A.1 CONTENT