Random Forest を用いたボットネットの検出

呂, 良
九州大学大学院システム情報科学府

馮, 尭鍇
九州大学大学院システム情報科学研究院 : 助教 | 九州先端科学技術研究所

川本, 淳平
九州大学大学院システム情報科学研究院 : 助教 | 九州先端科学技術研究所

櫻井, 幸一
九州大学大学院システム情報科学研究院 : 助教 | 九州先端科学技術研究所

出版情報：電子情報通信学会総合大会講演論文集. 2016, pp.A-7-11-, 2016-03-15. 電子情報通信学会
バージョン：
権利関係：

Random Forestを用いたボットネットの検出

Botnet Detection Using Random Forest

呂　良*　　フォン　ヤオカイ♣†　　川本　淳平♣†　　櫻井　幸一♣†
Lu Liang　　Yaokai Feng　　Junpei Kawamoto　　Kouichi Sakurai

*九州大学大学院システム情報科学府
♣†九州大学大学院システム情報科学研究院、（公財）九州先端科学技術研究所（兼）

1　はじめに

近年、ボットネットの使用により DDoS 攻撃の被害が増大しており、ボットネットの検出が世間に注目されている。攻撃者が感染したコンピューター、いわゆる、ボットを制御するため、一般的に Control and Command(C&C)サーバーを利用している。ボットと C&C サーバーの通信は攻撃が始まる前の段階ので、C&C 通信を検出することはボットネットの早期発見と考えられる。本研究では、C&C 通信の検出に着目、多数の特徴量から成り立ち高次元ベクトルを定義、それを対応する機械学習手法 Random Forest を用いて分類モデルを作り、識別率の向上の結果が得られた。

2　既存研究

機械学習手法を用いて C&C 通信を検出する既存研究としては、[1]がある。[1]で、山内らはボットネットのトラフィックの観察結果により、C&C 通信のトラフィック特徴を代表できると思われる 7 次元の特徴ベクトルを定義した。Naïve Bayes(NB)、Logistic 回帰(LR)、Support Vector Machine(SVM)三つの機械学習手法を用い、それぞれ分類モデルの性能を評価した。

3　提案手法

3.1 特徴量選択

既存研究[1]では 7 つの特徴量を利用したことに対して、本研究では、通信のトランスポート層とネットワーク層の常見な特徴量 55 個を選んだ。提案で使われた特徴量は表１のように示した。

表 1 本研究で使われた特徴量

特徴量	説明
Spc/Rpc	送信/受信パケットの数
Sd/Rd	送信/受信パケットの総サイズ
Smin/Rmin	送信/受信パケット最小サイズ
Smax/Rmax	送信/受信パケット最大サイズ
Savg/Ravg	送信/受信パケット平均サイズ
Svar/Rvar	送信/受信パケットサイズの分散
Sr0~14/Rr0~14	送信/受信パケットの中、パケットサイズが 0~100kb、100~200kb…1400kb~の比率
Flag S, P, R, F	一つフローでフラグ S, P, R, F の出現比率
Sitmin/Ritmin	送信/受信パケットの最小間隔時間
Sitmax/Ritmax	送信/受信パケットの最大間隔時間
Sitavg/Ritavg	送信/受信パケットの平均間隔時間
Sitvar/Ritvar	送信/受信パケット間隔時間の分散
time	フローの接続時間

3.2 機械学習手法

既存研究[1]では NB、LR、SVM 三つの機械学習手法を利用したことに対して、本研究では Random Forest 機械学習手法(RF)を導入して性能を測った。Random Forest 手法は沢山の決定木モデルの集団である。モデル作り際、ランダムでデータ集と特徴量集から子集を選び、同時に多数の学習モデルを生成する。従って、この機械学習は特徴ベクトルの次元数が高くても、処理時間が短い、そして過学習しにくいという利点を持っている。

4　実験

4.1 データセット

今回の実験では[1]と同じデータセットを用いた。即ち、不正通信データセットは CCCdataset'09'10 と CCCpractice13 を用いた。正常通信データは 2012 年 8 月から 9 月までの間、我々の研究室にあるサーバーを監視して収集した HTTP 通信データと IRC 通信データを用いた。二つのホスト間のすべての通信を一つのフローとして、データからフロデータを抽出した。

4.2 実験結果

今回の実験では、フロデータを Weka に導入、交差検証の方法でモデルの性能を検証し、[1]の結果と比較した。その結果は表２のように示した。

表２　実験結果と比較

	学習手法	特徴量の数	検知率	誤検知率
[1]での検知結果	SVM	7	92%	10%
	NB	7	92%	19%
	LR	7	94%	26%
本研究での検知結果	RF	55	98.9%	1.1%

5　まとめ

本稿では、C&C 通信の検出よりボットネットの検出方法に注目し、TCP/IP プロトコルに関する新たな特徴ベクトルを使用、Random Forest 機械学習手法で分類モデルを作った。検証実験の結果より、既存研究より高い識別率と低い誤識別率が得られた。

謝辞

この研究の一部は、科学研究費(基盤研究 No. 25330131)の支援を受けています。ここに記して謝意を表す。

参考文献

[1] 山内一将、川本淳平、堀良彰、櫻井幸一、" C&C トラフィック分類のための機械学習手法の評価"、情報処理学会論文誌　56(9), 2015 年 9 月