特定マルウェアのみの検知逃れを実現する敵対的学習とその対抗手法

名古屋大学 情報基盤センター 嶋田 創

1. 背景

近年では、近年では機械学習や深層学習を応用したマルウェア検出や悪性通信検出の研究が盛んであり、商用のセキュリティ機器においても、機械学習系の技術の活用をうたっている機材が多数出てきている状況にある。我々も過去研究において、自組織の過去の通信を学習して生成した識別器によるアノマリ型検知など、機械学習系を用いたしかしながら、近年では機械学習系に対して誤判定や誤検知をさせるための学習データを送りつける、敵対的学習という技術がある。現状のサイバーセキュリティ関係の敵対的学習の研究においては、全てのマルウェア等に関しての検知率を下げる話がほとんどである。たとえば、Anderson ら[1]は特徴空間での勾配方向を考慮した誤検知誘発について検討されており、また、Chen ら[2]は Android マルウェア識別用の機械学習系の識別器に細工をしたマルウェアを学習用に送り込むことで検知精度を低下させることが行われて。しかしながら、いずれ攻撃者が、攻撃に用いる特定のマルウェアのみ検知率を下げ、他のマルウェアへの検知率は落とさない形での敵対的学習を実現する、いわば、対機械学習および深層学習検知のマルウェア送付手法を確立することが考えられる。本研究課題において、この脅威について検証を行った。

2. 検証方法

2.1 マルウェア/クリーンウェアの特徴量化

本研究課題における機械学習/深層学習検知では、マルウェアおよびクリーンウェアそれぞれから先行研究[1]と同様に 128 個の API コールの有無のベクトルを特徴量とした。機械学習系マルウェア検知は、この特徴量に対して学習し、マルウェアとクリーンウェアを識別する識別器を作る。本研究課題では特定のマルウェア特徴量の検知精度を下げる、偽学習データとしてのマルウェア特徴量の生成を試みる。

2.2 偽学習データ用マルウェア特徴量生成

図 1 に偽学習データ用マルウェア特徴量生成の方法を示す。本検証では、実運用向け識別器の内部情報を使えない現実のサイバーj 攻撃を想定し、実際にマルウェア/クリーンウェア特徴量を学習した実運用向け識別器の内部情報を用いるのではなく、識別器への入出力から作成した代替識別器に対して偽学習データ注入とその評価を行う形とした。以下、その動きを説明する。

(1) 偽学習データ生成用としてベースとなるベースマルウェア特徴量を設定する。
(2) 強化学習のエージェント側からの変化対象特徴量選択を受け、偽学習データの特徴量を変化させる。特徴量の変化は有効化パターン(値変化 0→1 のみ)と切り替えパターン(1→0 の値変化もあり)の 2 パターンを試みた。
(3) 代替識別器に対して偽学習データをバッチサイズ分複製し学習させる。
(4) 検知を回避させたいターゲットマルウェア特徴量を偽データ学習後の代替識別器に識別させ、悪性度を出力させる。
(5) オリジナルの代替識別器が出力した悪性度と手順(4)で学習した識別器が出力した悪性度を比較し、どのくらい悪性度が低下したかを報酬として出力する。悪性度が上がったか変わらなかった場合は、-1 にさらに上がった分の数値を引く。
(6) 強化学習のエージェント側は報酬を受け取り、最適な行動を選ぶための Q 関数のパラメータを更新して、より正確に最適な行動(より悪性度を低下させる特徴量変更)が選べるように学習する。
(7) エージェント側において、Q 関数の出力と前状態の偽学習データをもとに、次に変化させるべき特徴量

を選択する。

3. 検証方法および検証結果

生成する偽学習データのもととなるベースマルウェア特徴量は、実マルウェア識別器の学習の際に使用したマルウェア特徴量中からランダムに選択し、ターゲットマルウェア特徴量は学習時に使用しなかったマルウェア特徴量からランダムに選択した。今回の検証では偽学習データ生成の全体の試行数を 300 とし、強化学習のループを回すステップ数は、有効化パターンの場合は 25 回、切り替えパターンの場合は 100 回とした。また、識別器には線形回帰、ランダムフォレスト、サポートベクタマシン、MalGAN DNN の 4 種類を利用し、それぞれで偽学習データの効果があるか検証した。

最終的に、検証結果では、ターゲットマルウェア特徴量の悪性度スコアを当初よりも低下させるに至らなかった。しかしながら、図 2 に示すように、強化学習の Q 関数が悪性度を下げる方向に学習されることは検証できた。図 2 の横軸は有効化パターンにおけるステップ数であり、縦軸は悪性度スコアの削減量(負の値は悪性度が上がっていることを示す)、2 つのグラフは 300 回の試行の最初と最後における変化である。グラフより、last score 側では、悪性度スコアの削減量が増える方向に学習が進んでいることが示されている。

本検証結果をもとに、利用する特徴量の増加、変化させる特徴量の改善などを推進するとともに、注入された偽学習データの検出手法についても研究を推進する予定である。

参考文献

[1] H. S. Anderson, et. al., “Evading Machine Learning Malware Detection,” Black Hat USA 2017, Jul. 2017.

[2] S. Chen, et. al. “Automated Poisoning Attacks and Defenses in Malware Detection Systems: An Adversarial Machine Learning Approach,” arXiv:1706.04146, Jun. 2017.

発表文献

- 高木聖也, 長谷川皓一, 山口由紀子, 嶋田創, ″機械学習を用いたマルウェア検知システムに対する強化学習による敵対的サンプル生成の課題,″ 電子情報通信学会研究報告, Vol. 119, No. 288, ICSS2019-62, pp. 13-18, 2019 年 11 月.

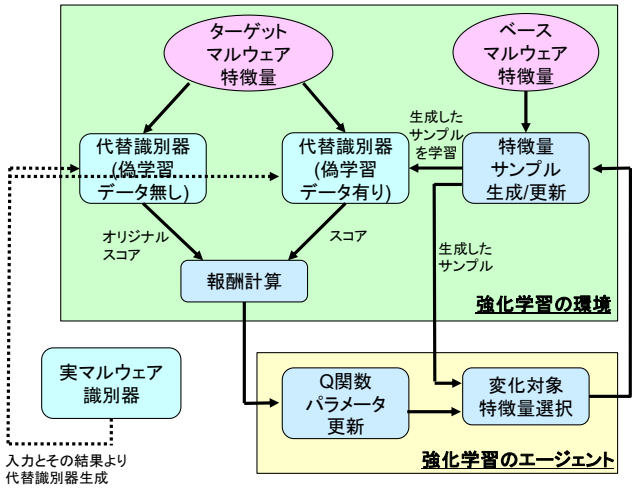

図 1: 強化学習による偽学習データ生成

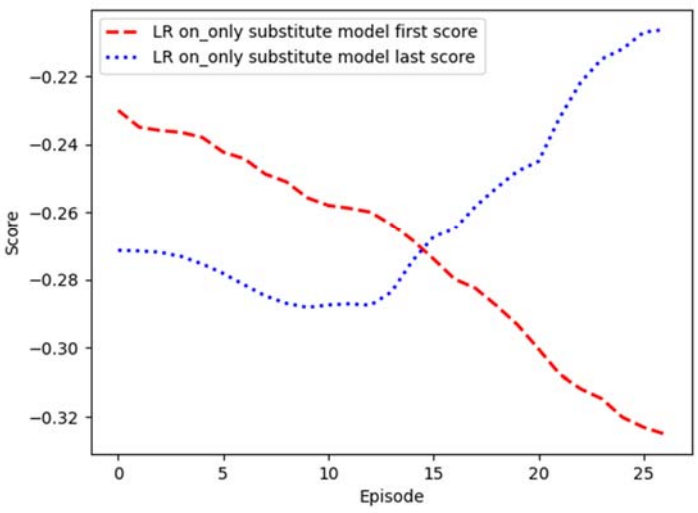

図 2: 悪性度スコア低下量(線形回帰、有効化パターン)