特定マルウェアのみの検知逃れを実現するデータ中毒攻撃対策向け特徴量探索

名古屋大学 情報基盤センター 准教授 嶋田 創

1. 背景

機械学習技術の発展に伴い、サイバーセキュリティ分野への機械学習の利用も広く見られるようになった。機械学習システムは悪性通信の検知やマルウェア分類などにおいて多数の応用が見られる一方、機械学習システムは細工された入力に対するロバスト性に欠けている。この脆弱性を悪用し、機械学習システムに攻撃を加えることができることも示されている。そのため、機械学習ベースのマルウェア検知手法に対して、攻撃者が攻撃用マルウェアの検知率を下げるために、マルウェア識別器の検知率を下げるように加工したマルウェアを偽学習データとして事前にばらまく攻撃を行うことが考えられる。このような偽学習データが商用の機械学習ベースのマルウェア識別器の訓練データに混入した場合、マルウェア検知精度の悪いマルウェア識別器が世に出てくることになる。そのような攻撃はデータ中毒攻撃と呼ばれ、SVM、ロジスティック回帰、ニューラルネットワークなどの様々な機械学習モデルに対して攻撃可能となると示されている。したがって、このような攻撃を先んじて研究し、対策を立てることが重要である。先行研究[1]では特定マルウェアのみの対策としたために効果的な偽学習データ(中毒攻撃用データ)の生成に失敗した。最終的に本研究では、単純な線形 SVM 識別器を騙す中毒攻撃用データの生成とその混入の検知の研究とした。

2. データ中毒攻撃とその対策の評価

2.1 想定するデータ中毒攻撃シナリオ

図 1 に想定するデータ中毒攻撃シナリオを記す。攻撃者は、セキュリティ対策企業が malware/benignware 識別器の学習に用いる世の中の malware/benignware データセットに対し、その一部を取得し、中毒攻撃モデルを用いて中毒攻撃用データを生成する。この中毒攻撃用データは攻撃者によって世の中の malware/benignware のデータセットに混入される。もし、セキュリティ対策企業が中毒攻撃用データを排除できずに識別器を生成して配布した場合、世の中のマルウェア検知に影響が出る。本研究課題では、識別器を更新する時に、新たに追加した学習用データの中に中毒攻撃用データが混入されたか否かを判別することを目的とする。識別器には単純な線形 SVM 識別器(以下、SVM 識別器)を用い、malware/benignware データセットには MWS Dataset 2020[2]の中から FFRI Dataset 2020 を用いた。データセット中の malware/beningnware から中毒攻撃生成に用いる部分セット、識別器の学習に使われる部分セットを選択した。識別器の学習に使われる部分セットは、当初の学習に用いる部分セットと追加の学習に用いられる部分セットを入れ替えて交差検証を行う。識別に用いる特徴量は 317 種の数値特徴量のみとした。

2.2 中毒攻撃用データの生成

中毒攻撃用データの生成は、Adversarial Robustness Toolbox の poisoningAttacksSVM クラスとして実装されている、Biggio らによる SVM Poisoning アルゴリズムの[3]を利用した。中毒攻撃用データは malware から生成した poisonM データと benignware から生成した poisonB データを生成し、いずれも、中毒攻撃用データの入っていない学習データ(clean データ)から生成した識別器のマルウェア検知精度に対し、clean データに中毒攻撃用データを追加して学習した識別器において検知精度が悪化することを確認した。なお、poisonM データと poisonB データをまとめて学習した場合、検知精度はそれほど悪化しない。

2.3 中毒攻撃用データ混入の検出方法

中毒攻撃用データは識別器の内部状態を混乱させる方向に生成されていると考え、本研究では、追加学習データを追加して学習した後の識別器の内部状態が、追加学習データ無しで学習した識別器の内部状態と一定以上異なる場合、追加学習データに中毒攻撃用データが混入されているものと判断することを試みた。

識別器の内部状態として、本研究で生成された SVM 識別器の係数ベクトルである 317 次元のベクトルを用い、学習前後の係数ベクトルのユークリッド距離で内部状態の変化量とした。

3. 評価結果

評価は、10,800 件の malware/benignware 混在のクリーンな訓練データを学習して生成した識別器の係数ベクトルと、10,800 件のクリーンな訓練データに加えて 300/150/75 個の追加学習データを追加して学習した識別器の係数ベクトルのユークリッド距離を算出し、明確な識別用の閾値が設定できるか確認する形で行った。追加学習データは複数のデータ塊を準備し、あるデータ塊から生成した閾値が、別のデータ塊で有効か否かで判別した。追加学習データ塊は、300 個追加の場合は 4 種類、150 個追加の場合は 8 種類、75 個追加の場合は 16 種類となる。図 2 に各追加学習データ塊追加後の係数ベクトルの変化のユークリッド距離の最小値/最大値/平均値を示す。図の横軸は追加学習データの種類(clean/poisonM/poisonB)とその数(300/150/75)を示し、図の縦軸はユークリッド距離を示す。

図より、追加学習データが少ないほどユークリッド距離の変化が小さいことが分かるが、一方で、全ての追加学習データにおいて clean データと poisonM/poisonB データを追加した時のユークリッド距離の変化量には明白な差があり、閾値設定が可能なことが分かる。閾値設定の方法としては、セキュリティ対策企業側でも中毒攻撃用データを生成し、生成した中毒攻撃用データと混入させた時のユークリッド距離の変化量の最小値とクリーンなデータを追加した時のユークリッド距離の最大値を事前に評価し、その中間値を閾値とする方法が考えられる。この方法で設定した閾値は、図 2 のデータの場合、300 個の追加時に 0.083、150 個の追加時に 0.109、75 個の追加時に 0.145 となる。

参考文献

[1] 高木聖也ら, "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," 信学技報, Vol. 119, No. 288, ICSS2019-62, pp. 13-18, 2019 年 11 月.

[2] 寺田真敏ら, "マルウェア対策のための研究用データセット MWS Datasets ～コミュニティへの貢献とその課題～", 情処研報, Vol.2020-IFAT-139 No.8, 2020 年 7 月.

[3] Biggio, et al., "Poisoning Attacks against Support Vector Machines," ICML '12, pp. 1807-1814, Jun. 2012.

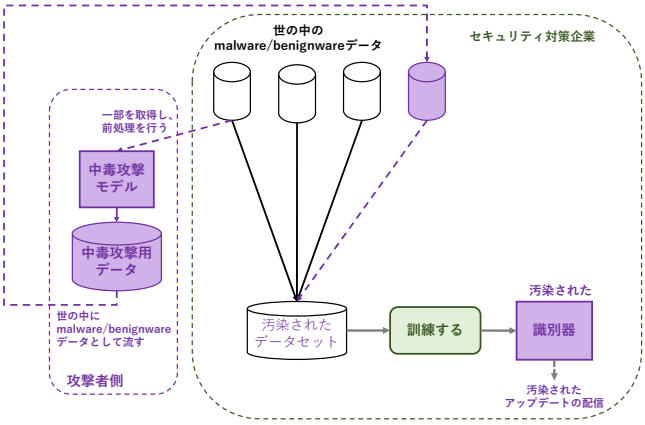

図 1: 想定するデータ中毒攻撃シナリオ

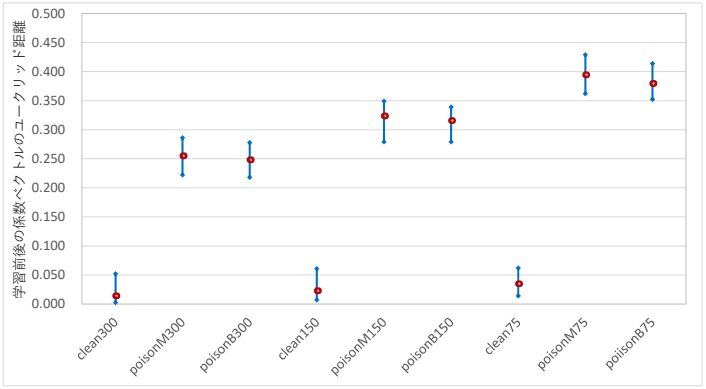

図 2: SVM 識別器の勾配係数変化量