5ZA-09

マルウェアによる情報流出を検知する機械学習モデルの脆弱性の検討

花井 一輝 † 大木 哲史 † 水野 慎太郎 †
† 静岡大学情報学部

1 はじめに

近年，マルウェアによる被害は社会問題になっている．マルウェアは企業や個人のコンピュータに感染し，機密情報などの流出を行う．マルウェアによる情報流出の手法は複雑化しており，収集した重要情報をあたかも正常に見えるパケットに組み込み，偽装することで，情報流出を図る．本稿では，偽装された内部から外部へと発信されるトラフィックのことを偽装トラフィックと呼び，偽装トラフィック検知手法を回避可能なトラフィック作成手法の存在を示すことを目的とする．トラフィック偽装手法は高度化しており，中でも，DNS のホスト名に流出情報を格納する手法は簡易かつ検知が困難な手法として知られている．これに対し，Ahmed らは，異常検知を行う決定木アルゴリズムに基づいて偽装トラフィックを検知する手法を提案している [1]．本アルゴリズムによる検知を回避可能な攻撃手法が存在するかどうかは明らかにされていないが，検知モデルに対する新たな攻撃手法の存在を示すことにより，その対策法を確立する必要性を明らかにすることは，より安全な検知システムの構築に向けた重要なプロセスである．本稿では，決定木を利用した偽装トラフィック検知モデルに対しても攻撃が可能な，Adversarial Examples [2] の作成法を提案し，これにより作成した偽装トラフィックが検知モデルを回避可能であることを示す．

2 関連研究

2.1 DNS を利用したデータ流出の検知手法

DNS のプロトコルなどに流出情報を紛れ込ませることで，C&C サーバとの通信を行うマルウェアが存在する．例えば，攻撃者が example.com というドメインを所持しており，マルウェアの目的は password という文字列の流出であるとする．この時，マルウェアは password.example.com というドメインにクエリを送信する．これにより，DNS サーバの所有主である攻撃者は password という文字列を入手することができる．このような偽装トラフィックは，通常のトラフィックに対して非常に稀であり，大量のデータを用いる二値分類器の学習には適さない．これに対して機械学習による異常検知モデルを用いた検知手法の提案が行われてきた．Ahmed らはリアルタイムに情報流出を行う DNS パケットを検知する手法を Isolation Forest という決定木ベースの異常検知モデルを用いて提案した [1].

2.2 Adversarial Examples

Adversarial Examples(A.E.) とは，入力に対して，微小なノイズを加えることで，機械学習識別器を誤認識させる手法である．A.E. は正解クラスに対する確率を出力する（ソフトラベル）深層学習に対する作成手法が多く提案されてきた．これに対し，Isolation Forest をはじめとする決定木を内部構造とする機械学習モデルは，最も高い確率のクラスの出力が 1 となるハードラベル問題であり，従来手法の適用が困難である．これに対して，Cheng らは Gradient Boosting Decision Tree(GBDT) や Random Forest など，ハードラベルかつブラックボックスな前提での A.E. 作成手法を提案している [2]．既存の検知モデルは，内部で決定木を用いる Isolation Forest により実装されており，Cheng らの手法 [2] に対し脆弱であると考えられる．

3 提案手法

既存の検知手法 [1] に対し，Cheng らが提案した A.E. [2] に基づく A.E. による攻撃を実施し，決定木による偽装トラフィックの検知モデルが脆弱であることを示す．Cheng らが提案したハードラベルモデルに対する A.E. で用いられているアルゴリズムの概略を Algorithm1 に示す．ハードラベルモデル f の出力から局所的な勾配を求める関数 g を算出し，この勾配に基づき入力に対

Algorithm 1 決定木に対する A.E. の生成

1: **Input:** 初期特徴量 $\boldsymbol{x_0}$, 変位 $\boldsymbol{\theta_0}$
2: **for** $t = 0, 1, 2, ..., T$ **do**
3: 正規分布からランダムなベクトル $\boldsymbol{u_t}$ を選択
4: $\hat{\boldsymbol{g}} = \frac{g(\boldsymbol{\theta_t}+\beta\boldsymbol{u})-g(\boldsymbol{\theta_t})}{\beta} \cdot \boldsymbol{u_t}$ を計算する
5: $\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \eta_t \hat{\boldsymbol{g}}$
6: **end for**
7: **return** $\boldsymbol{x_0} + g(\boldsymbol{\theta_T})\boldsymbol{\theta_T}$

する変位 $\boldsymbol{\theta}$ を更新する．T 回の繰り返しの後，最終的に求めた変位 $\boldsymbol{\theta}$ と初期特徴量 $\boldsymbol{x_0}$ との和により，検知モデルを回避する特徴量を算出する．β は円滑化パラメータであり，$\beta = 0.005$ に設定してある．なお，決定木の勾配を算出する手法については Cheng らの論文の Algorithm1 に従う．Cheng らの提案は，入力を画像と想定しているが，提案手法ではこれを攻撃対象の決定木へ入力可能な特徴量とする必要がある．本実験では，攻撃対象の手法 [1] で用いられる，ドメイン長，ラベル数，エントロピー，大文字の数，数字の数，サブドメイン長，最長ラベル長，平均ラベル長の 8 つの特徴量をドメイン名から抽出し，入力特徴量とする．また，A.E. を作成する際の初期特徴量 $\boldsymbol{x_0}$ としては，任意の悪性ドメインから抽出した特徴量を用いる．悪性ドメインに変位 $\boldsymbol{\theta}$ を加えることで良性ドメインと識別されるような A.E. を作成することを目標とする．

4 実験

提案手法の検討として，Isolation Forest を用いた既存の悪性ドメイン検知手法 [1] に対して A.E. による攻撃が成功するかを確認する．検知モデルの訓練および評価に利用するデータセットとして，まず良性のドメインを羅列したデータセットである majestic million[*1] を基に，DITL_B_Root-20180410 [3] から 8 万件程度の良性データを抽出した．さらに，悪性データとして Jawad らが作成したデータセット[*2]を用いた．良性データの 70% を訓練データとして使用し，良性データの残り 30% および，全ての悪性データを評価用データとして使用した．訓練データおよび評価データは前処理により検知モデルへの入力となる特徴量に変換される．訓練データを入力として学習を行った Isolation Forest に対して，テストデータによる検証を行う．さらに，作成した検知モデルに対し，いくつかの悪性な入力を基にして A.E. による攻撃を行い，検知モデルを回避できることを確認する．

4.1 結果

既存の検知モデルを再現した．良性ドメインの検知率は 98.2%，悪性ドメインの検知率は 99.5% であった．既存論文 [1] ではそれぞれ 97.99% と 98.49% を記録しており，ほぼ同等の性能であると言える．また，再現した検知モデルに対し A.E. を行った結果，悪性ドメインの特徴量を基に，良性と識別されるサンプルの生成に成功した．

5 議論とまとめ

本実験で，既存の検知モデルの再現を行い，再現したモデルに対して，A.E. の攻撃が成功することを確認した．現状，A.E. のサンプル生成では特徴量の生成しか行えないため，作成した A.E. のサンプルから情報流出を行うドメインを作成し，より現実に則した攻撃シナリオを考え，その対策手法を検討することが今後の課題である．

参考文献

[1] J. Ahmed, H. H. Gharakheili, Q. Raza, C. Russell, and V. Sivaraman. Real-time detection of dns exfiltration and tunneling from enterprise networks. In *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pp. 649–653, April 2019.

[2] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *CoRR*, Vol. abs/1807.04457, , 2018.

[3] Day in the life of the internet (ditl) april, 2015 dataset. PREDICT ID IMPACT ID:USC-LANDER/DITL_B_Root-20180410/rev8803. Provided by USC/B-Root Operations with USC/LANDER project. http://www.isi.edu/ant/lander.

[*1] http://downloads.majesticseo.com/majestic_million
[*2] https://nozzle-data.sdn.unsw.edu.au/dns-exfiltration-dataset