修士学位論文要約（令和４年３月）

マルウェア検知システムへのバックドア攻撃に対するオートエンコーダを用いた防御手法に関する研究

松本　悠希
指導教員：菅沼 拓夫

Research on a defense method using autoencoders against backdoor attacks on malware detection systems

Yuki MATSUMOTO
Supervisor: Takuo SUGANUMA

In malware analysis, the increasing number of variants has increased the burden on analysts, and there is growing demand for static analysis using machine learning, which can efficiently detect malware. On the other hand, a typical threat to machine learning is the backdoor poisoning attack. Severi et al. reported that this attack can be applied to malware detection using machine learning. Therefore, in order to detect malware with high accuracy using machine learning, it is necessary to construct an attack-resistant detection model that is not affected by poisoning data. This paper proposes a method to eliminate the effect of poisoning data by applying an autoencoder to training data containing poisoning data, assuming a situation where clean data without poisoning data is not available when constructing a detection model. Through evaluation experiments using real data, we show that the proposed method can significantly reduce the impact of backdoor attacks while minimizing the degradation of malware detection accuracy.

1. はじめに

多くの企業がサイバー攻撃の脅威にさらされており，そのセキュリティ対策が必須となっている．その中でもマルウェアの存在はインターネットに繋がるあらゆる端末に対して悪影響を及ぼす可能性があることから，早い段階で検知，除去することが重要となる．そこで，本論文ではSeveri ら[1)]が提案したマルウェア分類器へのバックドアポイズニング攻撃に対する効果的な対策として，データセットをオートエンコーダに通して毒を除去する方法を提案し，その効果を調べた．

2. マルウェア分類器に対する脅威

静的解析に基づく機械学習を用いた一般的なマルウェア分類に使用されるデータの収集過程について説明する．まず，悪意のあるソフトウェアや通常のソフトウェアのバイナリファイルが脅威情報プラットフォームに収集される．収集後は既存のアンチウイルスエンジンを用いた解析により，自動的にラベル付けされる．こうして収集されたデータを特定の企業や研究者がデータセットとして使用することができる．ここで，収集されるデータは膨大な数となることから，すべてを正確に区別することは困難であり，攻撃者の介入も予想されることから，一定の割合で毒データが混入することが考えられる．このように毒データが混入した状態でデータセットが生成され，機械学習モデルの訓練が行われた場合，生成されたモデルが悪意のあるソフトウェアを害のないソフトウェア（グッドウェア）として誤分類することが起こりえる．

ここで，Severi らは SHAP（SHarpley Additive exPlanations）に基づく攻撃アルゴリズムを提案している．SHAP は，学習済みのモデルにおいて，ある特徴量がそのモデルの予測値に対してどれだけ寄与したかを算出するものである．それにより，ある特徴量のマルウェア分類における貢献度を数値化することが可能となる．実験では，2351 次元から構成される EMBER データ[2)]のうち，機能を損なうことなく変更が可能な 513-2351 次元をSHAP 値に従って操作することで毒データを生成し，訓練データ全体に対して 1%の毒データを注入することにより分類モデルを汚染し，高確率で誤分類を誘発することが可能となる．

3. オートエンコーダによる提案防御手法

攻撃者による毒データの混入そのものを防ぐことは困難であることから，混入後のデータから毒を除去する手法が必要となる．そこで，次に説明する 2 つの手法を適用し，比較検証を行う．第一に次元削減による手法であり，攻撃者が操作可能な 513-2351 次元を削除し，先頭 512 次元のみを分類に使用する手法であり，理論上バックドアによる攻撃成功率を 0 にす

ることが可能であるが，通常のデータの分類精度が劣化する欠点が存在する．第二にオートエンコーダによる手法であり，513-2351 次元をオートエンコーダによって圧縮・復元処理することによりデータの中から主要な特徴のみを抽出し，毒の影響を無効化する．

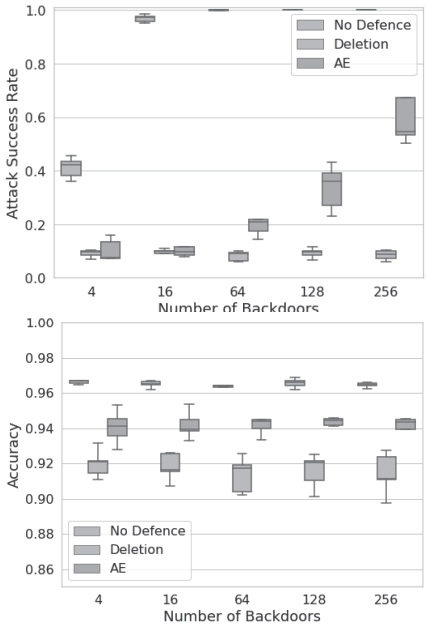

図１　バックドア数を変えた場合の攻撃成功率と分類精度の推移（Independent Selection）

分類モデルに対して Independent Selection 攻撃を行った場合の攻撃成功率と精度に関する結果を図 1 に示す．対策なしの場合は攻撃成功率はバックドア数が 16 を超えた時点で 100%に達するが，次元削減を行った場合は 10%前後まで低減できることを確認できた．一方オートエンコーダによる対策の場合，対策をしない場合よりも効果があることが確認できたが，バックドア数の増加に伴い攻撃成功率が上昇する結果となった．また，分類精度についてはオートエンコーダによる対策が次元削減と比較して約 3%の改善がみられた．

4. 内部結合オートエンコーダによる提案防御手法

オートエンコーダによる防御手法では一定の効果がある一方で，毒データによる影響の緩和に限界があることが確認されたため，より攻撃成功率を抑えつつも精度を維持可能な手法として，内部結合オートエンコーダを使用した疑似データへの置換による毒の除去を提案する．この手法では，513-2351 次元をオートエンコーダで処理する過程で毒が混入することのない 1-512 次元を挿入することでクリーンな次元の割合を増やし，毒をノイズとして除去することを目標としている．

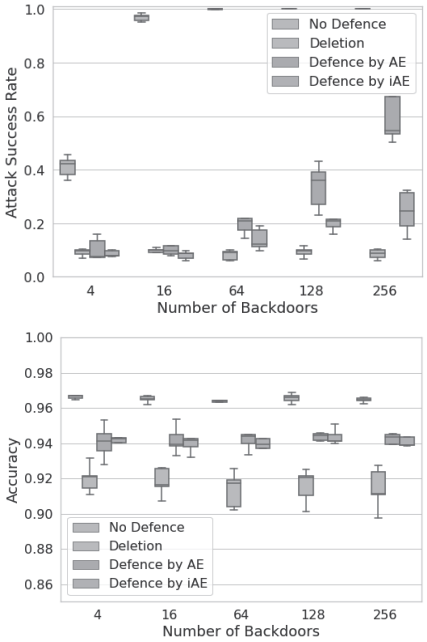

図２　バックドア数を変えた場合の攻撃成功率と分類精度の推移（Independent Selection）

分類モデルに対して Independent Selection 攻撃を行った場合の攻撃成功率と精度に関する結果を図 2 に示す．その結果，内部結合オートエンコーダを用いた対策手法では，通常のオートエンコーダと同等の分類精度を維持しつつ攻撃成功率を最大で約 40%低減させる効果があることが確認された．

5. まとめ

本論文では Severi らが提案したマルウェア分類器に対する攻撃の対策として，オートエンコーダ及び内部結合オートエンコーダによる防御手法を提案し，次元削減との比較により効果を検証した．その結果，Severi らが提案した攻撃の成功率を，オートエンコーダ及び内部結合オートエンコーダが次元削減に匹敵する精度で大きく低減させることに成功した．

文献

1) SEVERI, Giorgio, et al. Exploring backdoor poisoning attacks against malware classifiers. arXiv e-prints, 2020, arXiv: 2003.01031.
2) ANDERSON, Hyrum S.; ROTH, Phil. Ember: an open dataset for training static pe malware machine learning models. arXiv preprint arXiv:1804.04637, 2018.