侵入検知システムの検知精度向上のための敵対的サンプルによるデータ拡張法の検討

侵入検知システムの検知精度向上のための敵対的サンプルによるデータ拡張法の検討

川畑　魁星 [a)]・金丸　和樹 [b)]・油田　健太郎 [c)]・山場　久昭 [d)]・岡崎　直宣 [e)]

Data Expansion Using Adversarial Examples to Improve the Accuracy of Intrusion Detection Systems

Kaisei KAWABATA, Kazuki KANEMARU, Kentaro ABURADA,
Hisaaki YAMABA, Naonobu OKAZAKI

Abstract

In recent years, cyber-attacks such as unauthorized access and malware have been increasing along with the increasing use of network systems and the spread of new network technologies. Intrusion detection systems (IDS) have been attracting attention as one of the security technologies to protect systems from these cyber attacks. The accuracy of intrusion detection using deep learning is highly dependent on the data used for training, and a large amount of labeled training data is required. It is difficult to prepare a large amount of training data while taking into account the bias of the data. In this research, we propose a computationally efficient data expansion method using Jacobian-based Saliency Map Attack (JSMA), one of the adversarial example generation methods, and investigate how to improve the detection accuracy of signature-based IDS using deep learning models. To evaluate the proposed method, we built a small-scale model, extracted data with low classification accuracy, and extended the data with adversarial samples that were perturbed to bring them closer to the correct class, and compared the detection accuracy before and after the data extension. As a result of the experiment, the detection accuracy after the data expansion using the adversarial example was found to be better than that before the data expansion in terms of Accuracy, Recall, and F1-score. Although the proposed method improves the detection performance against attacks, it also increases the possibility of false positives, which requires improvements to reduce the degradation of Precision.

Keywords: Intrusion detection system, Data expansion, Adversarial example, Jacobian-based saliency map attack

1. はじめに

近年、ネットワークシステムの利用の増加や新たなネットワーク技術の普及と伴に不正アクセスやマルウェアなどのサイバー攻撃が増加している。国立研究開発法人情報通信研究機構（NICT）は、2019 年の観測レポートで調査目的とみられるスキャン活動や IoT 機器を狙った攻撃活動が大幅な増加傾向にあることを報告している [1)]。

これらのサイバー攻撃からシステムを保護するセキュリティ技術の 1 つとして侵入検知システム（IDS）が注目されている。IDS とは、ネットワーク上を流れるトラフィックを監視することで、不正なアクセスや通信をリアルタイムに検出して管理者に通知するシステムである。

IDS の検知手法の 1 つとしてシグネチャ型検知がある。シグネチャ型検知は攻撃の特徴的なパターンを事前に登録しておき、監視対象の通信とパターンマッチングを行うことで不正な通信や異常を検知する。

近年では、インターネットトラフィックの増加や深層学習技術の発展により、シグネチャ型検知に深層学習技術を応用した幅広い研究が行われているが、大量の学習用データが必要になることが問題となっている。深層学習を用いた侵入検知の精度は学習に使用するデータに大きく依存し、ラベリングされた大量の学習用データが必要である。十分なデータがある場合には深層学習の精度は機械学習アルゴリズムよりも高くなるが、データに偏りがある場合には性能が大きく低下する。また、ラベリングは一般的に手動で行うため作業コストが高く、データの偏りを考慮しつつ大量の学習用データを用意するのは困難である。

大量のデータを用意できない場合、データを現実的にあり得る範囲で変化させてデータ数を増加させるデータ拡張が行われる。データ拡張は限られたデータからより多くのデータを生成することを可能にして学習データの多様性を増加させることができる。

しかしながら、サイバー攻撃は日々多様化、進化している

[a)] 工学専攻機械・情報系コース大学院生
[b)] 情報システム工学科学部生
[c)] 情報システム工学科准教授
[d)] 情報システム工学科助教
[e)] 情報システム工学科教授

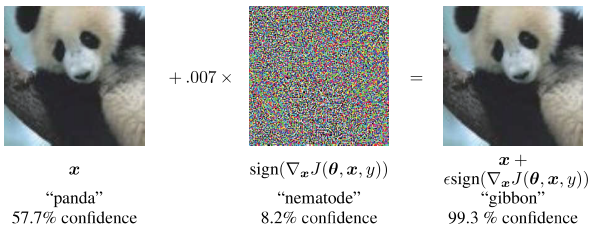

図 1. 画像分野における敵対的サンプル（[3] より引用）

ため、シグネチャ型検知ではシグネチャの登録や機械学習モデルの再学習を頻繁に行う必要があり、データ拡張の計算量を小さく抑える必要がある。

そこで、本研究では敵対的サンプル生成手法の 1 つである JSMA を用いて計算量を抑えたデータ拡張を提案し、深層学習モデルを用いたシグネチャ型 IDS の検知精度向上の検討を行う。データ拡張の有効性を検証するために、全結合層とドロップアウト層のみで構成したシンプルなディープニューラルネットワーク（DNN）を用いて、データ拡張前とデータ拡張後の検知精度の比較を行った。

以下、本概要の構成を述べる。第 2 章では先行研究の「敵対的サンプル」について述べ、第 3 章では提案手法について述べる。第 4 章では提案手法の有効性を評価する実験について述べ、第 5 章ではまとめと今後の課題について述べる。

2. 先行研究

2.1 敵対的サンプル

敵対的サンプル（Adversarial Example）[2] は主に画像分野で使用される機械学習モデルに対する攻撃手法の 1 つであり、データに小さな摂動を加えることで作為的に機械学習モデルの判断を誤らせるサンプルである。図 1 は画像分野における敵対的サンプルで、機械学習モデルに 57.7%の確率で panda（パンダ）と認識されていた画像に、小さな摂動を加えることで 99.3%の確率で gibbon（テナガザル）と誤認識されることを示している。敵対的サンプルは、標的の機械学習モデルの誤分類を誘発して攻撃の検知を回避する Evasion や学習用データに細工を施すことで予測性能を低下させる poisoning 等の攻撃に利用される。また、敵対的サンプルには転移性という性質があり、あるモデルに対して生成した敵対的サンプルは、アーキテクチャの異なる別のモデルに対しても高確率で有効に作用することが判明している [4]。

敵対的サンプルの生成手法は生成時のモデルの事前情報や標的クラスの有無によって分類される。まず、対象となるモデルのアーキテクチャやパラメータ等の内部情報をすべて所有している場合はホワイトボックス、所有していない場合をブラックボックスと呼ぶ。一方、敵対的サンプルの生成時に特定のクラスに分類するように摂動を加える場合は標的型回避、標的のクラスを指定せずに誤ったクラスに分類するように摂動を加える場合は非標的型回避と呼ぶ。

2.2 Jacobian-based Saliency Map Attack

Jacobian-based Saliency Map Attack（JSMA）は Papernot ら [5] によって提案されたホワイトボックスでの標的型回避に分類される敵対的サンプルの生成手法である。

この手法ではモデルの入力に対する出力のヤコビ行列を分析して、入力特徴量の値を変化させた場合の出力の変化を推測することで、摂動を加える特徴量を選択する。具体的な選択は以下のように行う。まず入力 x に対する出力 $Z(x) = (Z_1(x), \cdots, Z_k(x))$ から勾配 $\nabla Z(x)$ を計算する。次に入力 x の各特徴量がクラスを予測するのにどれだけ影響力を持つかを表現した Saliency Map を式（1）によって求め、$S[x,t][i]$ が最大となる特徴量 x_i に $theta$ の摂動を加える。入力 x のクラスが変更されるか、加えた摂動の総量が指定された値に達するまでこのプロセスを繰り返すことで敵対的サンプルを生成する。

$$S(x,t)[i] = \begin{cases} 0 \; if \; \dfrac{\partial Z_t(x)}{\partial x_i} < 0 \; or \; \displaystyle\sum_{j \neq t} \dfrac{\partial Z_j(x)}{\partial x_i} > 0 \\ \dfrac{\partial Z_t(x)}{\partial x_i} \left| \displaystyle\sum_{j \neq t} \dfrac{\partial Z_j(x)}{\partial x_i} \right| \; otherwise. \end{cases} \tag{1}$$

ここで $\frac{\partial Z_t(x)}{\partial x_i}$ 及び $\sum_{j \neq t} \frac{\partial Z_j(x)}{\partial x_i}$ は、入力特徴量 x_i の値が変化した場合に $Z_t(x)$ がどの程度増加し、$\sum_{j \neq t} Z_j(x)$ がどの程度減少するかを定量化している。

この手法では入力データのうち効率よく出力を変化させられる特徴量を探し出して摂動を加えることで、摂動の総量を減らすことができる。

3. 提案手法

深層学習を用いたシグネチャ型 IDS の検知精度を向上させるための、敵対的サンプルによるデータ拡張を提案する。敵対的サンプルを生成するためのサブモデルと侵入検知を行うメインモデルには、全結合層とドロップアウト層から構成された DNN を使用する。

3.1 問題点と解決のアイディア

深層学習は学習用データの偏りに弱く、訓練データでの出現頻度が低いデータは分類が難しいという問題がある。この問題を解決するために、しばしばデータ拡張が行われるが、IDS 分野では計算量を小さく抑える必要がある。計算量を抑えるために、データ拡張の対象となる範囲を、出現頻度の低いデータのみに絞り込む。そこでまず、小規模なモデル（サブモデル）を構築して学習・分類を行い、分類精度の低いデータの抽出を行う。この時、分類精度が低いデータが出現頻度の低いデータであると予想できる。抽出したデータから JSMA を用いて新たな類似のトラフィックデータを生成することでデータ拡張を行う。

また、敵対的サンプルを生成するためのサブモデルの各レイヤーのノード数やエポック数を、侵入検知に使用するメインモデルより少なく設定し、敵対的サンプルの特徴の 1 つである転移性を利用して計算量を削減する。

3.2 提案手法の流れ

具体的な手順を以下に示す。

1. 訓練データの前処理 (特徴選択、標準化、One-hot Encoding) を行う。
2. 前処理済みの訓練データを用いてサブモデルを構築する。

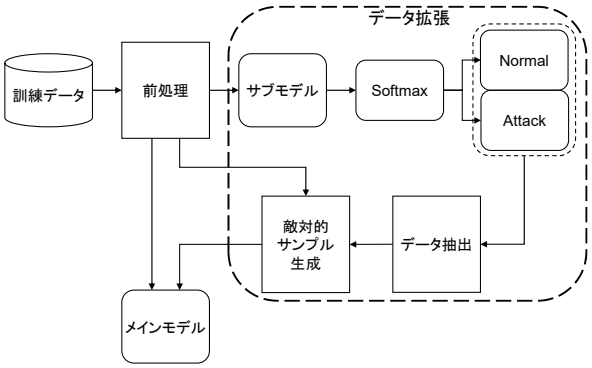

図 2. 提案手法のアーキテクチャ

3. 訓練データをサブモデルに入力して二値分類を行う。サブモデルからは各クラスに属する確率が出力される。
4. サブモデルの出力のうち正しいクラスに属する確率が閾値以下のデータを抽出する。
5. 抽出したデータから敵対的サンプルを生成する。
6. 訓練データと生成した敵対的サンプルを結合したデータからメインモデルを構築する。

4. 評価実験

4.1 実験の目的

データ拡張前と提案手法によるデータ拡張後の検知精度の比較を行うことで JSMA を用いたデータ拡張の有効性の検証を行う。

4.2 データセットと前処理

評価には IDS の研究に広く使用されている公開データセットの 1 つである UNSW-NB15[6] を使用した。UNSW-NB15 は 49 個の特徴量と 2 種類のラベルを持つ約 254 万件のネットワークのトラフィックデータで構成されたデータセットである。また、各データが 43 個の特徴量と 2 種類のラベルを持ち、175,341 件のレコードを持つ訓練データと 82,332 件のレコードを持つテストデータが公開されている [7]。

UNSW-NB15 は通常のトラフィックと 9 種類の攻撃トラフィックで構成されるが、今回は訓練データとテストデータを使用して Normal と Attack の 2 つのクラスで二値分類を行った。

まず前処理フェーズでは、計算量の削減及び検知精度向上のために特徴選択を行い、今回は 10 個に削減した。特徴選択には Recursive Feature Elimination（RFE）を用いた。RFE はモデルを繰り返し作成し、特徴量の数が指定された数になるまで各イテレーションで最も分類に寄与しない特徴量の削除を繰り返す手法であり、先行研究 [8] により優れた結果が得られている。次に各特徴量のスケールを統一するために特徴量の標準化を行い、値の平均が 0、分散が 1 になるように変換した。最後に深層学習で特徴量の値がテキストとなるカテゴリカル特徴を扱うため、One-hot Encoding により数値特徴に変換した。One-hot Encoding では各カテゴリカル特徴をユニークな要素数の次元のベクトルで表現し、対応する座標のみ 1 で他が 0 になる One-hot ベクトルに変換する。

表 1. モデルの構成

レイヤー	サブモデル	メインモデル
Dense	64	128
Dropout	0.2	0.2
Dense	128	256
Dropout	0.3	0.3
Dense	64	128
Dense	32	64
softmax	2	2

4.3 敵対的サンプルの生成

まず敵対的サンプルを生成するために訓練データを用いてサブモデルを構築した。サブモデルの構成を表 1 に示す。また、サブモデルのパラメータは最適化アルゴリズムにデフォルトパラメータの Adam、バッチサイズを 256、エポック数を 30 に設定した。

次に、サブモデルに訓練データを入力して二値分類を行い、サブモデルの出力のうち正しいクラスに属する確率が閾値以下のデータを抽出した。本実験では、抽出するサンプル数を考慮して確率の閾値を 0.75 に設定した。これにより、ラベルが Normal のデータを 14,143 件、ラベルが Attack のデータを 17,259 件抽出した。最後に、抽出したデータから敵対的サンプルを生成し、訓練データに加えた。抽出したデータのラベルごとに以下のパターンで敵対的サンプルを生成した。

1. ラベルが Normal のデータを Normal のクラスに近づけて生成
2. ラベルが Attack のデータを Attack のクラスに近づけて生成
3. 両方のラベルのデータをそれぞれの正しいクラスに近づけて生成

JSMA を用いて敵対的サンプルを生成する際の 3 つのパラメータ *theta*、*gamma*、*target* について説明する。*theta* は各イテレーションで加える摂動の大きさを決める値であり、*gamma* は摂動の総量を決める値、*target* は摂動を加える際の標的となるクラスで、クラスを *target* に近づけるように摂動を生成する。誤分類したデータから敵対的サンプルを生成する場合は *theta* を 0.001、*target* を正しいクラスに設定した。JSMA では正しいクラスに分類されたデータに摂動を加えることができないため、正しく分類出来ているが正しいクラスの属する確率が閾値以下のデータから敵対的サンプルを生成する場合には *theta* を-0.001、*target* をラベルと異なるクラスに設定することで疑似的に敵対的サンプルを生成した。また、*gamma* の値を 0.01、0.05、0.1 の 3 パターンで実験を行っている。

4.4 評価方法と評価指標

敵対的サンプルを用いたデータ拡張の有効性の検証は、次のようにして行う。JSMA のパラメータ *gamma* の値を変動させた 3 パターンでデータ拡張を行い、データ拡張を行わない場合とどの程度検知精度に違いがあるかを比較する。侵入検知に使用するメインモデルの構成を表 1 に示す。またメインモデルのパラメータは最適化アルゴリズムにデフォルトパ

表 2. *gamma*=0.01 の検知精度

ラベル	Accuracy	Precision	Recall	F1-score
ベースモデル	0.80	0.92	0.74	0.84
Normal	0.83	0.90	0.78	0.86
Attack	0.84	0.83	0.84	0.86
Normal+Attack	0.85	0.80	0.89	0.86

表 3. *gamma*=0.05 の検知精度

ラベル	Accuracy	Precision	Recall	F1-score
ベースモデル	0.80	0.92	0.74	0.84
Normal	0.82	0.91	0.76	0.86
Attack	0.85	0.83	0.84	0.86
Normal+Attack	0.85	0.80	0.87	0.86

表 4. *gamma*=0.1 の検知精度

ラベル	Accuracy	Precision	Recall	F1-score
ベースモデル	0.80	0.92	0.74	0.84
Normal	0.80	0.92	0.74	0.84
Attack	0.80	0.92	0.74	0.85
Normal+Attack	0.81	0.91	0.76	0.85

ラメータの Adam、バッチサイズを 256、エポック数を 50 に設定した。

実験の評価指標として、Accuracy、Precision、Recall、F1-score の 4 つの指標を用いる。TP（True Positive）は正しく Attack に分類できたデータ、TN（True Negative）は正しく Normal に分類できたデータ、FN（False Negative）は誤って Normal に分類された Attack のデータ、FP（False Positive）は誤って Attack に分類された Normal のデータを示す。Accuracy、Precision、Recall、F1-score は以下の式で定義される。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

4.5 実験結果と考察

3 パターンの *gamma* の値について、データ拡張を行っていないベースモデルと提案手法によるデータ拡張を行ったモデルの検知精度を表 2〜4 に示す。

まず $gamma = 0.01$ 及び 0.05 の場合、ベースモデルと比較して全てのパターンで Accuracy、Recall、F1-score が向上しているが、Recall とトレードオフの関係にある Precision は低下している。ベースモデルでの F1-score が 0.84 だったのに対し、最も向上した組み合わせでは 0.86 と 0.02 向上した。両方のラベルのデータから敵対的サンプルを生成した場合は、検知精度は最も向上しているが、Precision が Recall 以下に低下していることから、最も優れているわけではないと考える。また、Precision の低下の要因として、特徴量の分布の偏りが考えられる。図 3 は、特徴量の 1 つである ‘dbytes’ の度数分布であり、x 軸が標準化後の特徴量 ‘dbytes’ の値、y 軸がデータ数を対数スケールで表している。今回の実験では前処理の段階で標準化を行ったが、いくつかの特徴量に偏りがあったために特徴量の値によって摂動の影響が変動したと考える。

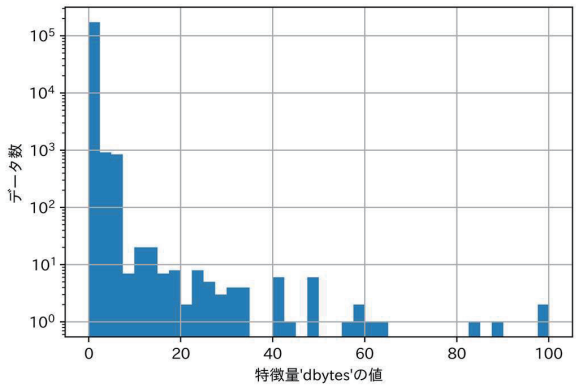

図 3. 特徴量 ‘dbytes’ の度数分布

次に $gamma = 0.1$ の場合、全てのパターンで殆ど検知精度に変化はない。この理由として摂動が大きい場合には敵対的サンプルは元のデータから大きく離れ、出現頻度の低いデータのデータ拡張として機能していないと考えられる。

今回の実験において最も検知精度が向上した組み合わせでは、Accuracy が 0.05、F1-score が 0.02 向上しているため、敵対的サンプルを用いたデータ拡張は有効だと考える。しかし、データ拡張を行った場合に Precision が低下し、誤検知の可能性が高まるため、改良を行う必要があると考える。

5. まとめ

本研究では深層学習を用いたシグネチャ型 IDS の検知精度向上のための、敵対的サンプルによるデータ拡張の提案を行った。

提案手法の評価として、小規模なモデルを構築して分類精度の低いデータを抽出し、正しいクラスに近づけるように摂動を加えた敵対的サンプルによるデータ拡張を行い、データ拡張前後で検知精度の比較を行った。実験の結果、敵対的サンプルを用いたデータ拡張後の検知精度は、データ拡張前の検知精度よりも Accuracy、Recall、F1-score の観点では優れていることが分かった。この提案手法によるデータ拡張では攻撃に対する検知性能は向上するが、誤検知の可能性も高まることから、Precision の低下を軽減するための改善が求められる結果となった。

今後の課題としては、まず 1 つに正しいクラスに分類されているデータの敵対的サンプルの生成方法の改善がある。JSMA の特徴から、ラベルと異なるクラスから遠ざけるように摂動を加えることで疑似的に敵対的サンプルを生成しているため、正しいクラスに近づけるように摂動を加えることで無駄な摂動を削減できると考える。また、特徴量のスケーリングの改善と JSMA のパラメータの最適化を行う必要がある。今回の実験では各特徴量の値を標準化したが、分布に偏りがあるため特徴量毎に摂動の影響が異なっていたと考えられる。そのため、スケーリングやパラメータの最適化で各特徴量の分布の偏りの影響を小さくすることで、検知精度はさらに改善されると考える。

参考文献

1) Nicter 観測レポート 2019, https://www.nict.go.jp/cyber/report/NICTER_report_2019.pdf, 2019.

2) C. Szegedy, W. Zaremba, J. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus : Intriguing properties of neural networks, Intriguing properties of neural networks, ICLR, abs/1312.6199, 2014.

3) I. Goodfellow, J. Shlens, C. Szegedy: Explaining and harnessing adversarial examples, In international Conference on Learning Representations, 2015.

4) N. Parpernot, P. McDaniel, I. Goodfellow: Transferability in Machine Learning: From Phenomena to Black-box Attacks Using Adversarial Samples, arXiv:1605.07277, 2016.

5) N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, B. Celik, A. Swami: The Limitations of Deep Leaning in Adversarial Setting, IEEE European Symposium on Security and Privacy(EuroS& P), IEEE, pp. 372-387, 2016.

6) ACCS. UNSW-NB15 Dataset. Available online: https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/, (accessed on 2020/11/10)

7) N.Moustafa, J. Slay: A comprehensive data set for network intrusion detection systems(UNSW-NB15 Network Data Set), In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), pp. 1-6, 2015.

8) R. Abou Khamis, A.Matrawy: Evaluation of Adversarial Training on Different Types of Neural Networks in Deep Leaning-based IDSs, arXiv preprint arXiv:2007.04472, 2020.