Vol.2021-CSEC-95 No.9
Vol.2021-SPT-45 No.9
Vol.2021-EIP-94 No.9
2021/11/8

既存通信データセットに対する中毒攻撃を想定した敵対的学習データ生成の試行

桑山 拓也[1,a)]　嶋田 創[2,b)]　長谷川 皓一[3]　山口 由紀子[2]

概要：近年では，機械学習は幅広い分野で利用され，NIDS にも機械学習を利用する研究が行われてきている．その一方で，中毒攻撃などの機械学習に特有な情報セキュリティ上の脅威が出現している．中毒攻撃対策の研究においては敵対的学習データ，すなわち悪意のある学習データを大量に確保する必要があるが，あらかじめ敵対的学習データが用意されていれば中毒攻撃対策の研究を円滑に開始することが可能となる．そこで，本研究では，NIDS 評価用通信データセットの一つである Kyoto2016 Dataset をもとに，機械学習ベースの NIDS を模した分類器に対する敵対的学習データの生成と評価の実験を行った．実験における中毒攻撃は，Biggio らによる SVM 中毒攻撃アルゴリズムを用いて生成し，クリーン学習データの一部を加える形で実施した．この時に加えたデータの割合を中毒割合とする．評価の結果，生成した敵対的学習データの影響により分類精度が悪化したことが確認できた．また，中毒割合が 25%を超えると分類精度が急激に悪化することが確認された．

Study of Adversarial Training Data Generation Assuming Poisoning Attack on Existing Traffic Dataset

Abstract: Recently, Machine Learning (ML) are used in many area and there are many researches on applying ML to NIDS. On the other hand, there are arising ML specific security issues such as poisoning attacks. To promote counter poisoning attack researches, we have to keep much adversarial training data or malicious training data. If there is a dataset that equips adversarial training data, we can start counter poisoning attack researches rapidly. In this research, we performed adversarial training data generation from Kyoto2016 dataset (traffic dataset for NIDS evaluation) and evaluate them with ML based classifier that imitates NIDS. We generated adversarial training data with Biggios' SVM poisoning attack algorithm and add them as a part of clean training data with varying addition rate (define as "poisoning rate"). We confirmed that the generated adversarial training data degrades classification accuracy and classification accuracy degrades dramatically if poisoning rate overs 25%.

1. はじめに

近年では，機械学習が注目を集め，幅広い分野で活発に利活用され，サイバー攻撃対策にも機械学習が多用されている．不正アクセス等を検知するネットワーク型侵入検知システム (NIDS: Network Intrusion Detection System) に機械学習を利用する研究では，検出率の向上や未知の攻撃に対する検知能力の獲得が試みられている．

一方で，機械学習の利活用の広がりに伴い，敵対的機械学習と呼ばれる，機械学習に特有なセキュリティ上の脅威が出現している．敵対的機械学習とは，モデルとその機能を標的とする攻撃手法の総称である．敵対的機械学習の一種に中毒攻撃がある．中毒攻撃とは，敵対的学習データや中毒データと呼ばれる細工されたデータを加えたりして学習データを汚染することにより，機械学習モデルの分類性能を低下させるなどの悪影響を及ぼす攻撃である．

機械学習の研究を実施するには大量の学習データが必要である．しかし，多くの研究者にとって大量の通信データを収集する環境の整備は困難であるため，NIDS の性能

[1] 名古屋大学大学院情報学研究科
Graduate School of Informtics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-Shi, 464–8601, Japan
[2] 名古屋大学情報基盤センター
Information Technology Center, Nagoya University
[3] 名古屋大学情報連携推進本部情報セキュリティ室
Information Security Office, Nagoya University
[a)] kuwayama@net.itc.nagoya-u.ac.jp
[b)] shimada@itc.nagoya-u.ac.jp

評価を目的とした通信データセットが複数公開されている．その一つに Kyoto2016 Dataset[1] がある．Kyoto2016 Dataset は京都大学に設置されたハニーポットに到達した通信に正常通信，既知攻撃通信，もしくは未知攻撃通信でラベル付けした通信セッションのデータセットである．通信を正常通信または攻撃通信に分類する NIDS の検知能力を向上させる研究であれば Kyoto2016 Dataset のデータのみで実施できる．しかし，今後の脅威となりうる敵対的機械学習やそれに対抗する防御手法の研究を行うには，攻撃を想定した悪意のあるデータも必要である．特に，中毒攻撃においては大量の悪意のある学習データ，すなわち敵対的学習データが必要であるため，大量の敵対的学習データの確保が研究開始におけるハードルとなる．したがって，敵対的学習データがあらかじめ用意されていれば中毒攻撃やそれに対する防御手法の研究を円滑に開始できると期待される．

本研究では，生成した敵対的学習データを公開することを目標とし，機械学習ベースの NIDS を模した分類器に対する敵対的学習データの生成と評価の 2 つの実験を行った．分類器には Support Vector Machine(SVM) を採用し，敵対的学習データを生成する中毒攻撃モデルには，Biggio らによる SVM 中毒攻撃アルゴリズムを採用した [2]．生成実験では，Kyoto2016 Dataset から抽出したデータを中毒攻撃モデルに入力して変換し，敵対的学習データを生成した．評価の結果，本研究で生成した敵対的学習データによって分類器の評価指標を悪化させることができたことを確認した．また，全学習データに占める敵対的学習データの割合が大きいほど評価指標の悪化も大きくなることが確認され，敵対的学習データを追加する割合が 25%を超えると識別精度が急激に悪化することが確認された．

2. 研究背景および関連研究

2.1 敵対的機械学習

敵対的機械学習とは，機械学習モデルとその機能を標的とする手法の総称であり，幅広い手法を含む概念である．

一般的に，機械学習が利活用されるプロセスは学習段階と推論段階の 2 つに分けられる．学習段階は，開発者が機械学習モデルを学習データで学習させ，分類能力を与える準備段階ともいえる段階である．推論段階は，学習段階を経た機械学習モデルをテストデータで性能評価したり，アプリケーションとして実運用に入り，利用者から入力されるデータを分類するなどの推論を行う段階である．教師あり学習の機械学習モデルが行う推論には，分類，回帰の 2 種類があるが，ここでは分類を例にとって説明する．簡単のため，分類問題を解く機械学習モデルを分類器と呼ぶ．

敵対的機械学習による攻撃が加えられるのは，この学習段階と推論段階のいずれかである．敵対的機械学習は攻撃のタイミングや目的に基づき，回避攻撃，中毒攻撃，移転攻撃の 3 種類に分類される．中毒攻撃は学習段階に，回避攻撃と移転攻撃は推論段階に実行される攻撃である．

回避攻撃は，摂動と呼ばれる人間には認識できない大きさのノイズを加え，機械学習モデルに人間による分類とは異なる分類 (誤分類) を引き起こす攻撃である．移転攻撃は，機械学習へのデータ入出力の記録から，モデルの構成や学習データなどの機密情報を抽出する攻撃である [3]．本研究で扱う中毒攻撃については，2.2 節で詳しく解説する．

2.2 中毒攻撃

中毒攻撃とは，学習データへの不正データの追加により，機械学習モデルの決定境界をシフトする攻撃である [3]．この学習データに追加される不正データは中毒データと呼ばれることも多いが，本研究では敵対的学習データと呼ぶ．敵対的学習データを含んだデータの学習によって，学習モデルの決定境界がシフトされることにより，分類器は推論データに対する誤分類を発生しやすくなる．

中毒攻撃には，可用性攻撃とバックドア攻撃の 2 種類が存在する．可用性攻撃は，大量の不正データの注入により機械学習モデルの分類性能を悪化させ，任意のデータに対して誤分類が発生しやすして使用不能にする攻撃である．機械学習を用いた NIDS に対する可用性攻撃であれば，正常通信を攻撃通信であると予測する誤検知や，攻撃通信を正常通信であると予測する見逃しが増加する．一方，バックドア攻撃は少量の洗練されたデータ注入により，バックドアと呼ばれる特定のデータに対してのみ機械学習モデルが誤分類を発生しやすくする攻撃である [3]．

本研究では Biggio らの SVM 中毒攻撃 [2] アルゴリズムを利用するため，3 節でその詳細を説明する．

2.3 通信データセット

1 節で述べた通り，多くの研究者にとって大量の通信データを収集する環境の整備は困難であるため，NIDS の性能評価を目的とした通信データセットが複数公開されている．その一つに Kyoto2016 Dataset[1] がある．その他のデータセットには，DARPA Intrusion Detection Data Sets (DARPA IDDS)[4]，KDD Cup 1999 Data[5]，Kyoto 2006+ Dataset[6] などがある．

Kyoto2016 Dataset は京都大学に設置されたハニーポットに到達した通信を収集/解析し，正常通信，既知攻撃通信，もしくは未知攻撃通信でラベル付けすることで作成されたデータセットである．ハニーポットに到達した通信の生データは，多くの場合は個々のパケットを保存する pcap ファイルとして収集されるが，Kyoto2016 Dataset では通信をパケット単位ではなく，より大きなセッション単位で扱う．そのため，pcap ファイルは日別に分割され，ネットワーク監視用ソフトウェアの Bro IDS による解析でセッションが抽出され，後述の 24 種類の特徴量へとサマライ

ズされる．ラベル付けは IDS，アンチウイルスソフトウェア，シェルコードやエクスプロイトコードを検知するソフトウェア Ashula の検知結果に基づいて行われる．

Kyoto2016 Dataset は，通信セッションをサマライズした 24 種類の特徴量を有し，そのうちの 14 種類は KDD Cup 1999 Data でも使用された特徴量の一部であり，基本特徴量と呼ばれている．残りの 10 種類は Kyoto2016 Dataset の前身となる Kyoto2006+ Dataset で初めて追加された特徴量に一部変更を加えたものであり，追加特徴量と呼ばれている．特徴量のデータ型には，整数，小数，カテゴリ，文字列がある．4.2 節で述べるが，本研究では基本特徴量のうちの 12 種類の数値特徴量を使用した．

2.4 機械学習ベース NIDS

機械学習ベース NIDS とは，通信を正常通信または攻撃通信に分類するために機械学習モデルを利用する NIDS である．一般的な機械学習ベース NIDS には，事前に収集した正常通信や攻撃通信に対してラベルを付けたデータで学習した，教師あり学習アルゴリズムの分類器が使用される．また，正常通信のみで教師なし学習アルゴリズムの分類器を学習し，正常通信のパターンとは異なる通信を攻撃通信と判断するアノマリ型 NIDS として利用する手法もある．

機械学習ベース NIDS が検知の対象とする攻撃通信には様々な種類がある．分類の一例として，DARPA IDDS では，攻撃通信を Denial of Service(DoS)，User to Root，Remote to Local，Probes の大きく 4 つのカテゴリに分類している．それぞれのカテゴリには複数のエクスプロイトのクラスが含まれ，例えば DoS に含まれるエクスプロイトには SYN Flood や Smulf などがある [7].

近年の機械学習フレームワークの継続的な進歩により，多くの機械学習ベース NIDS の正解率は 90%を超える．例えば，Mukkamala らによる SVM ベースやニューラルネットワーク・ベースの IDS の KDD Cup 1999 Data に対する分類性能の比較結果では，正解率は SVM ベース NIDS では 99.50%，ニューラルネットワーク・ベース NIDS では 99.25%が得られている [8]．安藤らは，大学の LAN 宛トラフィックを複製/取得して監視する実ネットワーク上で稼働する機械学習ベース NIDS を構築した [9]．Transductive SVM を使用して実トラフィックを利用した NIDS と DARPA IDDS[4] を使用した NIDS が構築し，実トラフィックを使用した NIDS の方がより低い偽陽性率と偽陰性率を達成することを確認した．機械学習ベース NIDS を利用した製品の実例として，ロジスティック回帰モデルを使用して不正なネットワークトラフィックを検知するトレンドマイクロ社の法人向け侵入防御システムがあり [10]，特に，Domain Generation Algorithm を用いて頻繁にドメインを変更して C&C サーバへの接続を試みるマルウェアに特有のトラフィックの検知に有用であると述べらている．

Algorithm 1 Biggio らの SVM 中毒攻撃 [2]

入力: 学習データ $\mathcal{D}_{\rm tr}$，検証データ $\mathcal{D}_{\rm val}$，1 件の攻撃開始点データの特徴量ベクトルとラベル $\{x_c^{(0)}, y_c\}$，ステップサイズ t
出力: 1 件の敵対的学習データの特徴量ベクトル x_c

1: $\mathcal{D}_{\rm tr}$ で SVM を学習させる．
2: 現在の反復回数 $k \leftarrow 0$.
3: **repeat**
4: 　$\mathcal{D}_{\rm tr} \cup (x_c^{(k)}, y_c)$ で SVM を学習させ直す．
5: 　損失関数の勾配 $\frac{\partial L}{\partial u}$ を $\mathcal{D}_{\rm val}$ を用いて計算する．
6: 　u を $\frac{\partial L}{\partial u}$ に平行な単位ベクトルとする．
7: 　$k \leftarrow k+1$, $x_c^{(k)} \leftarrow x_c^{(k-1)} + tu$ として敵対的学習データを更新する．
8: **until** $L\left(x_c^{(k)}\right) - L\left(x_c^{(k-1)}\right) < \epsilon$
9: **return:** $x_c = x_c^{(k)}$

Apruzzese らにより，テスト用ネットワークを構築して収集した正常通信やマルウェアによる攻撃通信を用いて，機械学習ベース NIDS に対する中毒攻撃の有効性が評価された．攻撃通信データの特徴量の値をランダムに増加させることで作成された敵対的学習データを学習データに注入する中毒攻撃により，ランダムフォレスト/多層パーセプトロン/K-最近傍法による機械学習ベース NIDS の真陽性率がそれぞれ約 60 ポイント悪化した [11].

以上のように機械学習ベース NIDS の研究は活発に行われてきたが，機械学習ベース NIDS に対する敵対的機械学習の研究は現状では少ないことが指摘されている [11].

3. SVM 中毒攻撃

可用性攻撃の一つに，Biggio らの SVM 中毒攻撃が存在する [2]．本研究では，中毒攻撃の手法として SVM 中毒攻撃を採用したため，その詳細を説明する．

SVM は決定境界を学習する際に，マージン最大化とペナルティ最小化という 2 つのルールを適用する．マージンとは，決定境界に対して，それぞれのクラスの最も近いデータとの距離である．そのようなデータをサポートベクトルと呼ぶ．2 クラス分類であれば，サポートベクトルは 2 個存在する．マージンを最大化するように決定境界を定めることで，決定境界付近のデータを誤分類しにくくなり，汎化性能が向上する．しかし，学習データのセットによっては全てのデータを正しく分類 (線形分離) できる決定境界が存在しない場合がある．その場合は，誤分類をある程度許容して決定境界を定めることになる．ペナルティとは，「サポートベクトルよりも決定境界の逆側に位置してしまったデータ」と「サポートベクトル」との距離であり，いわば許容する誤分類の深刻さを表すものである．ペナルティを最小化するように決定境界を定めることで，学習データに対する誤分類を極力減らすことが可能である．

SVM の学習は，学習データ特徴量ベクトルを $\boldsymbol{x}_i$，クラス

Vol.2021-CSEC-95 No.9
Vol.2021-SPT-45 No.9
Vol.2021-EIP-94 No.9
2021/11/8

ラベルを y_i として，次の最小化問題として定式化される．

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} + C\sum_{i=1}^{N}\xi_i \qquad (1)$$

$$\text{subject to } y_i(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i + b) \leq 1 - \xi_i \qquad (2)$$

$\boldsymbol{w}, b$ は，SVM が学習するパラメータであり，SVM がクラスを分類する境界面である決定境界を表す数式の係数であるとする．2 クラス分類を行う SVM の出力である予測クラスは $sign\left(\boldsymbol{w}^{\mathsf{T}}\phi(x) + b\right)$ と表され，-1 または 1 のいずれかの値を取る．この式は，式 2 の条件を満たしながら式 1 内の値が最小となる $\boldsymbol{w}, b, \xi$ を求めることを意味する．式 1 の第 1 項はマージン最大化を表し，第 2 項はペナルティ最小化を表す項である．

SVM 中毒攻撃のアルゴリズムを Algorithm 1 に示す．SVM 中毒攻撃のアルゴリズムは入力の攻撃開始点データ $\{x_c^{(0)}, y_c\}$ を変換して敵対的学習データ x_c を生成し，出力する．このアルゴリズムは勾配上昇法の一種であり，手順を端的に表すと，SVM が検証データを分類した際の誤差を表す損失関数 L を最大化する方向 u へと反復的に敵対的学習データ $x_c^{(k)}$ を更新していく．SVM 中毒攻撃の勾配上昇法は，ニューラルネットワークの学習に用いられる勾配降下法に類似しているが，損失関数を最小化ではなく最大化する点と，更新されていく変数が重みではなくデータ $x_c^{(k)}$ である点で異なる．

Algorithm 1 の入出力と手順について解説する．入力である攻撃開始点データ $\{x_c^{(0)}, y_c\}$ は学習データ $\mathcal{D}_{\mathrm{tr}}$ の中から選ばれるが，そのラベル y_c は本来のラベルを反転させたものを使用する．入力特徴量 $x_c^{(0)}$ には数値型ベクトルのみが使用可能であり，出力 x_c も入力と同じデータ型である．出力 x_c のラベルには，入力のラベル y_c と同じ値を使用すればよい．1 回の実行につき 1 件の攻撃開始点データから 1 件の敵対的学習データを生成するため，複数の敵対的学習データが必要な場合，その件数分だけこのアルゴリズムを実行する必要がある．ステップ 4 で SVM の再学習に用いられるデータは，学習データ $\mathcal{D}_{\mathrm{tr}}$ に作成途中の敵対的学習データ $x_c^{(k)}$ を追加したものである．データの増分学習が可能なインクリメンタル SVM を使用すれば，ステップ 1 で既に $\mathcal{D}_{\mathrm{tr}}$ の学習が済んでいるため，学習ステップ 4 での再学習を高速に行うことができる．ステップ 5 の損失関数の勾配 $\frac{\partial L}{\partial u}$ の計算式はここでは省略するが，文献 [2] で解析的に求められている．ステップ 7 は SVM 中毒攻撃アルゴリズムの要となる部分であり，敵対的学習データ $x_c^{(k)}$ を更新する．ステップ 7 の式中にあるステップサイズ t はニューラルネットワークの学習率に相当するものであり，敵対的学習データの更新量を指定するものである．ステップ 8 では，損失関数 L の変化が小さくなったら収束したと判定し，敵対的学習データを完成とする．

4. 敵対的学習データの生成方法

4.1 生成方法と評価方法の概要

本研究で行った実験は，本節で解説する生成実験と 5 節で解説する評価実験の 2 つに大別される．図 1 は生成実験と評価実験の手順を示した図である．点線より上側の部分が攻撃者が行う処理を想定した生成実験で，下側の部分が分類器の開発者が行う処理を想定した評価実験である．

使用する分類器は SVM，データセットは Kyoto2016 Dataset であり，中毒攻撃モデルには，3 節で解説した Biggio らによる SVM 中毒攻撃アルゴリズムを採用する．SVM 分類器や SVM 中毒攻撃モデルの実装には，Python と機械学習ライブラリである scikit-learn[12] と Adversarial Robustness Toolbox(ART)[13] を使用する．

Kyoto2016 Dataset から抽出したデータを中毒攻撃モデルに入力して変換し，SVM 分類器に対する敵対的学習データを生成する手順は以下の通りである．手順番号は図 1 内の番号と対応している．以下の手順を 10 回繰り返し，敵対的学習データを 10 セット生成した．

(1) 通信データセットに対して前処理を行い，クリーン学習データとテストデータを作成する．実運用時を想定し，クリーン学習データの作成には古い月のデータを利用し，テストデータの作成には新しい月のデータを利用する．前処理として行う処理は，サンプル抽出，特徴量選択，特徴量の標準化，クラスラベルの変換の 4 種類である．

(2) ART に実装されている中毒攻撃モデルを作成する．中毒攻撃モデルが内部で使用する SVM の攻撃対象分類器を与え，生成される特徴量の値の範囲制限も行う．

(3) クリーン学習データの一部を攻撃開始点データとして中毒攻撃モデルに入力して変換し，敵対的学習データを生成する．

(4) 生成した敵対的学習データに後処理を行い，後々，中毒攻撃対策の研究のためのデータセットとして配布することを目的としてファイルに保存する．

生成された敵対的学習データは，特に，攻撃通信を正常通信に誤分類する偽陰性が増加することを意図した実験設定としている．以下，生成実験の各手順の詳細を説明する．

4.2 Kyoto2016 Dataset に対する前処理

2015 年 11 月 (図 1 の古い月に相当する) のデータから無作為抽出した 600 件のサンプルを分割し，100 件を学習データとし，残りの 500 件を検証データとする．検証データは SVM 中毒攻撃中の収束判定のために使用される．2015 年 12 月 (図 1 の新しい月に相当する) のデータから，攻撃通信サンプルと正常通信サンプルの比は 1:1 としつつ，無作為抽出した 4000 件のサンプルをテストデータとする．これらのデータ件数は Biggio らの SVM への中毒攻撃の評価

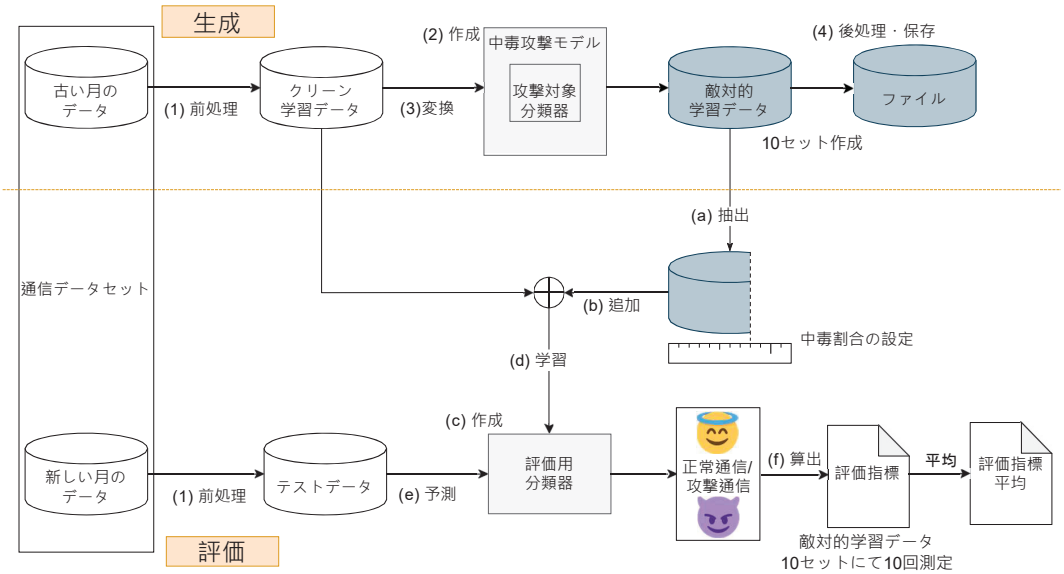

図 1 敵対的学習データの生成と評価の手順
Fig. 1 Procedure of Adversarial Training Data Generation and Evaluation

表 1 使用した特徴量
Table 1 Used Features

属性名	概要
Duration	セッションの長さ (秒)
Src_Bytes	送信バイト数
Dst_Bytes	受信バイト数
Count	過去 2 秒間のセッションのうち現在のセッションと宛先 IP アドレスが同じ数
Same_srv_rate	Count 特徴で該当したセッションのうち現セッションとサービスの種類が同じ割合
Serror_rate	Count 特徴で該当したセッションのうち SYN エラーが起こった割合
Srv_serror_rate	過去 2 秒間のセッションで現セッションとサービス種類が同じセッションのうち，SYN エラーが起こった割合
Dst_host_count	宛先ポートが同じ過去の 100 セッションのうち，現セッションと送信元 IP アドレスと宛先 IP アドレスが同じ数
Dst_host_srv_count	宛先ポートが同じ過去の 100 セッションのうち，現セッションと宛先 IP アドレスとサービス種類が同じ数
Dst_host_same_src_port_rate	Dst_host_count 特徴で該当したセッションのうち現セッションと送信元ポートが同じ割合
Dst_host_serror_rate	Dst_host_count 特徴で該当したセッションのうち SYN エラーが起こった割合
Dst_host_srv_serror_rate	Dst_host_srv_count 特徴で該当したセッションのうち SYN エラーが起こった割合

実験 [2] に倣ったものである．

Kyoto2016 Dataset で提供される 24 特徴量のうち，本研究では数値型である 12 種類の特徴量を使用して正常通信または攻撃通信の 2 クラス分類を行い，同 12 特徴量の敵対的学習データを生成する．表 1 に使用する 12 特徴量とその概要を示す．使用する 12 特徴量はそれぞれ取りうる値の範囲が大きく異なるため，そのまま分類器に入力すると値の大きな特徴量に結果が支配されることになる．そのため，scikit-learn の StandardScaler クラスを用いて各特徴量の平均を 0，分散を 1 に変換する標準化を適用する．標準化の過程で検証データやテストデータの情報が学習データに漏洩しないようにするため，学習データのみから計算された平均と分散に基づいて学習データ，検証データ，テストデータを標準化する．

Kyoto2016 Dataset のクラスラベルには正常通信，既知攻撃通信，未知攻撃通信を表す 1，-1，-2 の 3 種類のクラスラベルが存在する．しかしながら，中毒攻撃向けの敵対的学習データを生成する際には未知攻撃と既知攻撃を区別する必要はないと考え，正常通信，攻撃通信を表す 0，1 の 2 種類のクラスラベルへ変換した．

4.3 中毒攻撃モデルの作成

ART では生成される敵対的学習データの特徴量が取る値の範囲制限が可能である．制限範囲は各特徴量の下限値ベクトルと上限値ベクトルの組で指定する．具体的には，割合を表す特徴量は $[0, 1]$，数量を表す特徴量は $[0, \inf)$ に範囲制限する必要がある．しかし，攻撃対象とする SVM 分類器は標準化された特徴量を入力とするため，制限範囲についても本来の下限値ベクトルと上限値ベクトルを標準化したベクトルの組を指定する．

中毒攻撃の対象となる分類器として線形カーネルを用いた SVM 分類器である LinearSVC を利用した．scikit-learn の LinearSVC クラスのインスタンス作成時のハイパーパラメータとして，Biggio らの SVM への中毒攻撃の評価実

表 2 Kyoto2016 Dataset のサンプルと生成した敵対的学習データのサンプル
Table 2 Samples of Kyoto2016 Dataset and Generated Adversarial Training Data

Label	Duration	Source_ Bytes	Destination_ Bytes	Count	Same_ srv_ rate	Serror_ rate	Srv_ serror_ rate	Dst_ host_ count	Dst_ host_ srv_ count	Dst_ host_ same_ src_ port_ rate	Dst_ host_ serror_ rate	Dst_ host_ srv_ serror_ rate
0	0.000415	44	99	1	1.00	0.00	0.00	84	86	0.00	0.00	0.00
0	2.738384	505	1564	2	0.50	0.00	0.00	90	92	0.00	0.00	0.00
0	0.000000	0	0	2	0.50	0.00	0.03	6	5	0.00	0.00	0.00
0	0.000734	69	111	9	1.00	0.00	0.00	100	100	0.00	0.00	0.00
0	2.724161	505	1564	1	1.00	0.00	0.00	98	99	0.00	0.00	0.00
1	0.000445	43	99	22	1.00	0.00	0.04	97	97	0.00	0.00	0.00
1	0.000000	0	0	0	0.00	0.00	1.00	0	0	0.00	0.00	0.00
1	3.024933	0	0	1	1.00	1.00	0.57	0	33	0.00	0.00	1.00
1	0.000000	0	0	1	1.00	1.00	1.00	3	72	1.00	1.00	1.00
1	1.052390	0	0	1	1.00	1.00	0.57	0	9	0.00	0.00	1.00
3	2.950447	6	20	0	0.06	0.00	0.00	59	61	0.00	0.91	0.90
3	3.190053	5	13	1	0.97	1.00	0.54	1	34	0.00	0.00	1.00
3	0.041797	45	114	2	0.99	0.01	0.52	90	91	0.01	0.01	0.01
3	0.011166	55	103	4	1.00	0.00	0.00	16	30	0.00	0.00	0.00
3	0.006243	11	28	0	0.06	0.01	0.71	6	5	0.00	0.01	0.00

験 [2] のパラメータ設定に準じ，損失関数にヒンジ関数を指定する．その他のハイパーパラメータは明示的に指定せず，既定値を使用した．

ART の PoisoningAttackSVM クラスは，Biggio らの SVM 中毒攻撃アルゴリズムの実装である．攻撃対象分類器，生成処理に関するパラメータ，学習データ，検証データを与えて PoisoningAttackSVM のインスタンスを生成し，これを中毒攻撃モデルとする．生成処理に関するパラメータには eps，steps，max_iter が存在する．eps は Algorithm 1 のステップ 8 における収束判定条件のしきい値 ϵ である．steps は Algorithm 1 のステップ 7 におけるステップサイズ t である．max_iter は Algorithm 1 の反復回数 k の上限であり，k が max_iter に達した時点で敵対的学習データを更新するループ処理が打ち切られる．

4.4 クリーン学習データから敵対的学習データへの変換

攻撃開始点データとして，クリーン学習データ内全ての攻撃通信データを使用する．Biggio らによる SVM 中毒攻撃の評価実験では，攻撃開始点データのラベルに本来のラベルとは反転させたものを使用しているため，本実験では攻撃開始点データのラベルを正常通信とする．

1 件の攻撃開始点が変換されて 1 件の敵対的学習データが生成されるため，100 件のクリーン学習データに含まれる 50 件の攻撃通信データからは最大 50 件の敵対的学習データの生成が可能である．そのため，この実験では 50 件の敵対的学習データを生成する．4.3 節で作成した中毒攻撃モデルに攻撃開始点データを入力して中毒攻撃を実行し，出力として敵対的学習データを得た．

4.5 敵対的学習データの後処理

生成した敵対的学習データは標準化が施されているため，特徴量値を本来のスケールに戻すために標準化の逆変換を行う．逆変換の際に計算誤差によって指定した範囲制限をわずかに超過する場合があるため，超過した特徴量値を範囲内に丸めて修正する．また，敵対的学習データの特徴量は全て浮動小数点数型で生成されるため，本来は整数型である特徴量は四捨五入して整数型に変換する．浮動小数点数型特徴量は小数点以下 6 桁に丸める．

表 2 に Kyoto2016 Dataset の正常通信のサンプル (Label = 0)，Kyoto2016 Dataset の攻撃通信のサンプル (Label = 1)，生成された敵対的学習データのサンプル (Label = 3) を 5 件ずつ示す．生成された敵対的学習データは大幅に外れた値を取ることは無く，中毒攻撃として紛れ込まされても不自然でないものとなっている．

5. 敵対的学習データの評価実験

5.1 評価指標

本研究で取り組む分類問題は，通信サンプル x をクラス正常通信，攻撃通信のいずれかに属していると予測する 2 クラス分類問題である．また，x が攻撃通信クラスに属している場合を陽性 (Positive)，x が正常通信クラスに属している場合を陰性 (Negative) であるとする．予測クラスと

実際のクラスの対応関係は，真陽性 (TP: True Positive)，真陰性 (TN: True Negative)，偽陽性 (FP: False Positive)，偽陰性 (FN: False Negative) の 4 種類に分けられる．SVM 分類器の評価指標として正解率，適合率，真陽性率，偽陽性率，F1 Score を使用し，その定義は以下の通りである．

$$\begin{aligned} \text{正解率 (Acc)} &= (TP+TN)/(P+N) \\ \text{適合率 (Pre)} &= (TP+TN)/(TP+FP) \\ \text{真陽性率 (TPR)} &= TP/P \\ \text{偽陽性率 (FPR)} &= FP/P \\ \text{F1 Score)} &= 2/(\text{適合率}^{-1}+\text{真陽性率}^{-1}) \end{aligned}$$

5.2 実験方法

評価実験では，敵対的学習データを含んだデータで分類器を学習させ，その分類器でテストデータを予測し，正解率などの評価指標を算出する．また，敵対的学習データを追加する割合 (中毒割合) を変化させ，それによる評価指標への影響を確認する．

以下は評価実験の手順であり，番号は図 1 内の番号に対応している．以下の手順を中毒割合 r_p を 0%以上 33.3%以下の 2%刻みで変化させて繰り返した．評価は，使用するクリーン学習データ，敵対的学習データ (4 節で 10 セット作成したもの)，テストデータを変更して 10 回繰り返し，10 回計測した評価指標の平均値を計算し，これを中毒割合 r_p における結果とする．

(a) 4.4 節で生成した敵対的学習データから中毒割合 r_p に応じた件数を抽出する．
(b) 抽出した敵対的学習データをクリーン学習データに加える．
(c) 4.3 節で攻撃対象とした SVM 分類器と同じハイパーパラメータを持つが，未学習である SVM 分類器を新たに生成する．この SVM 分類器を評価用分類器とする．
(d) 敵対的学習データを含んだデータで評価用分類器を学習させる．
(e) 評価用分類器でテストデータ中の個々の通信に対して正常通信または攻撃通信であるかを予測するテストを行う．
(f) 予測結果から正解率などの評価指標を算出する．

4.3 節で前述したように，生成した敵対的学習データは 50 件である．この 50 件の敵対的学習データが 100 件のクリーン学習データに加えられることで学習データは 150 件となる．それゆえ，中毒割合は最大で約 33.3%となる．

5.3 結果と考察

表 3 に各中毒割合における評価指標の 10 回平均値を示す．図 2 に中毒割合を変化させた際の評価指標の 10 回平均値の推移を示す．図 2 は表 3 に掲載されているデータをグラフ化したものである．

図 2 から確認できることとして，中毒割合の増加に伴って各評価指標が悪化すること，中毒割合が 25%を超えたあたりから評価指標の悪化量が増大し始め 30%を超えるとその悪化が急激なものとなること，F1 Score 以外の評価指標のうち真陽性率の悪化が最大であり，中毒割合が最大と最小の場合で比較して 65 ポイントも悪化していることが分かる．また，表 3 より，偽陽性率が 0%に近づいていることが確認できる．10 回の計測のうち，偽陽性率が 0%となった，すなわち，陽性と予測されたサンプルが 0 件であった計測回も存在した．

表 3 各中毒割合における評価指標
Table 3 Evaluation Metrics on Individual Poisoning Rate

中毒割合	Acc	Pre	TPR	FPR	F1 Score
0%	0.7992	0.8559	0.7206	0.1222	0.7820
2%	0.7976	0.8539	0.7193	0.1242	0.7804
4%	0.7879	0.8429	0.7114	0.1356	0.7705
6%	0.7893	0.8471	0.7087	0.1301	0.7710
8%	0.7881	0.8508	0.7015	0.1253	0.7681
10%	0.7920	0.8583	0.7015	0.1175	0.7713
12%	0.7906	0.8592	0.6971	0.1160	0.7690
14%	0.7880	0.8626	0.6870	0.1110	0.7643
16%	0.7840	0.8632	0.6772	0.1091	0.7582
18%	0.7807	0.8634	0.6682	0.1068	0.7530
20%	0.7796	0.8669	0.6618	0.1027	0.7503
22%	0.7752	0.8670	0.6521	0.1016	0.7436
24%	0.7743	0.8631	0.6544	0.1058	0.7437
26%	0.7658	0.8679	0.6299	0.983	0.7291
28%	0.7457	0.8702	0.5815	0.901	0.6845
30%	0.7230	0.8742	0.5278	0.817	0.6376
32%	0.5907	0.8782	0.2196	0.382	0.3071
33.3%	0.5300	0.6991	0.704	0.103	0.1247

上記より，敵対的学習データの影響によって分類器の決定境界が移動して攻撃通信クラスと予測される領域が小さくなったことが予想される．真陽性率の低下は，分類器が攻撃通信を見逃す機会が多くなることを意味するため，悪意ある攻撃者の観点から見て都合の良い結果である．そのため，真陽性率は中毒攻撃の効果を判断する基準として適すると考えられる．この評価実験では，真陽性率が最大で 65 ポイント悪化したため，中毒攻撃の効果が十分に発揮されのではないかと考えられる．

6. おわりに

本研究では機械学習ベースの NIDS を模した分類器に対する敵対的学習データの生成と評価の 2 つの実験を行った．生成実験では，Kyoto2016 Dataset から抽出したデータを SVM 中毒攻撃モデルに入力することで変換し，敵対的学習データを生成した．評価実験では，中毒割合を変化させつつ敵対的学習データを含んだデータで分類器を学習させ，その分類器でテストデータを正常通信または攻撃通信であるか予測し，敵対的学習データよる評価指標への影響を確認した．

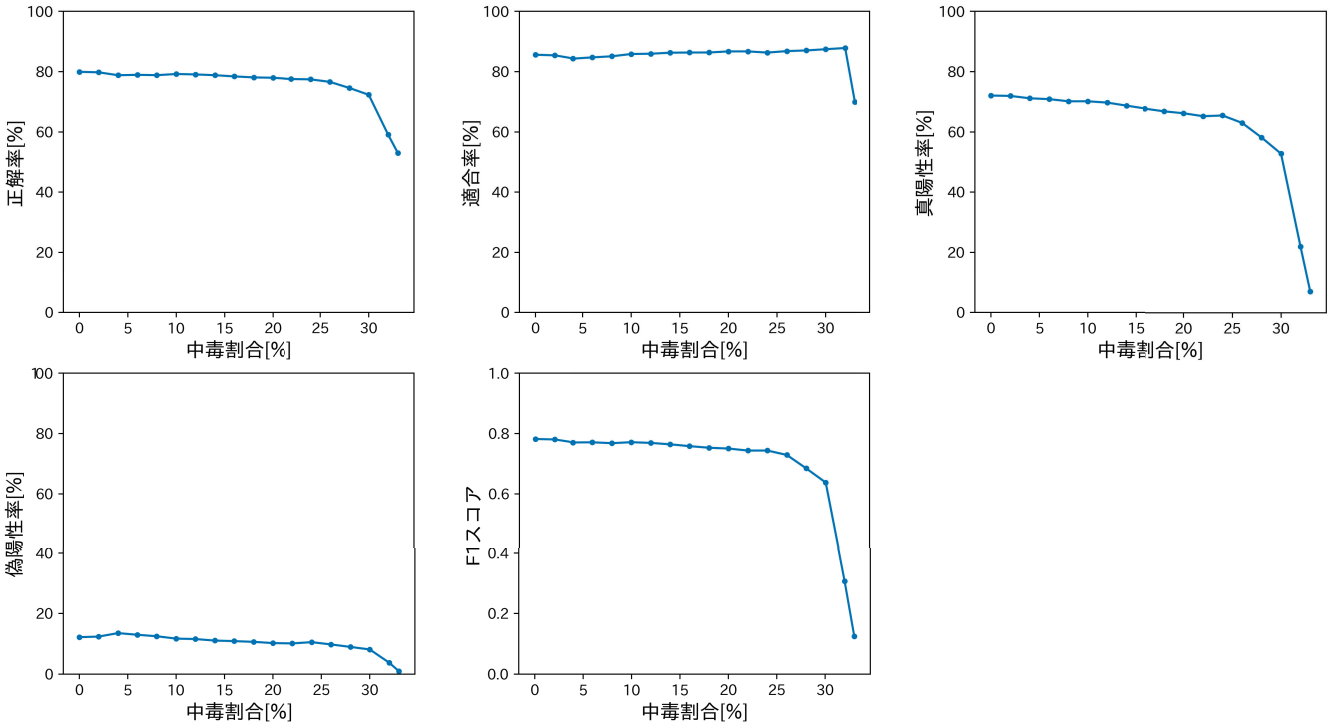

図 2 中毒割合を変化させた際の評価指標の推移
Fig. 2 Transistion of Evaluation Metrics on Varied Poisoning Rate

本研究では Kyoto2016 Dataset を入力とする SVM 分類器に対して SVM 中毒攻撃が有効であり，敵対的学習データの影響により分類器の評価指標を悪化させられることを示した．特に，真陽性率は中毒割合が最大と最小の場合で比較して最大 65 ポイントも悪化し，中毒割合が大きいほど評価指標の悪化も大きくなり，中毒割合が 25%を超えると評価指標が急激に悪化したことが確認した．

本研究では攻撃者がクリーン学習データや分類器のパラメータを取得できているという仮定のもとで生成実験を行った．しかし，クリーン学習データや分類器のパラメータは公開されるものではないため，実際には攻撃者がそれらを取得することは困難である．また，ART による SVM 中毒攻撃の実装はインクリメンタル SVM を使用しておらず，実行時間を短縮する余地が残っている．よって，クリーン学習データに含まれない攻撃開始点データから生成した敵対的学習データの生成，決定木など SVM 以外の機械学習アルゴリズムの分類器に対する敵対的学習データの生成，実行時間の短縮の 3 点を今後の課題とする．

謝辞 本研究は JSPS 科研費 JP19H04108 の助成を受けたものである。

参考文献

[1] 多田竜之介ら：NIDS 評価用データセット：Kyoto 2016 Dataset の作成，情報処理学会論文誌， Vol. 58, No. 9, pp. 1450–1463 (2017).
[2] Biggio B., et al.: Poisoning Attacks against Support Vector Machines, *ICML '12*, pp. 1807–1814 (2012).
[3] 石川冬樹ら：「AI セキュリティ」その脅威と対策を考える，`https://www.jnsa.org/seminar/nsf/2020/data/A2_IoTSecurityWG.pdf` (2020). (参照 2021-02-02).
[4] Laboratory, M. L.: 1998 DARPA Intrusion Detection Evaluation Dataset, `https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset`. (参照 2021-02-02).
[5] Archive, T. U. K.: KDD Cup 1999 Data, `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`. (参照 2021-02-02).
[6] Song, J., et al.: Statistical Analysis of Honeypot Data and Building of Kyoto 2006+ Dataset for NIDS Evaluation, *BADGERS '11*, pp. 29–36 (2011).
[7] Laboratory, M. L.: MIT Lincoln Laboratory: DARPA Intrusion Detection Evaluation, `https://archive.ll.mit.edu/ideval/docs/attacks.html`. (参照 2021-02-10).
[8] Mukkamala, S., et al.: Intrusion Detection Using Neural Networks and Support Vector Machines, *IJCNN'02*, Vol. 2, pp. 1702–1707, (2002).
[9] 安藤滋ら：実際のネットワークトラフィックから作成したデータを用いた知的侵入検知システムの構築，日本知能情報ファジィ学会ファジィ システム シンポジウム 講演論文集， Vol. 23, pp. 827–827, (2007).
[10] トレンドマイクロ：侵入防御システム，`https://www.trendmicro.com/ja_jp/business/products/network/intrusion-prevention.html`. (参照 2021-02-10).
[11] Apruzzese G., et al.: Addressing Adversarial Attacks Against Security Systems Based on Machine Learning, *CyCon 2019*, Vol. 900, pp. 1–18, (2019).
[12] Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830 (2011).
[13] Nicolae, M.-I., et al.: Adversarial Robustness Toolbox v1.0.0 (2019).