Computer Security Symposium 2020
26 - 29 October 2020

テンソルデータ拡充を用いた組織内ネットワーク攻撃判定方式の回避攻撃に対する防御手法の提案

宍戸 克成[1,a)]　森川 郁也[1]　及川 孝徳[1]　海野 由紀[1]

概要：機械学習はサイバー攻撃の防御にも活用され，サイバー攻撃を自動かつ高精度に発見する防御システムが開発されている．一方で，機械学習に特有な攻撃が発見され，機械学習を用いたシステムのセキュリティに関する研究の重要性が高まっている．特有な攻撃の 1 つである「敵対的入力」は，機械学習の推論時に「元の入力を少し変化させた」入力を作成することにより，本来の推論結果と異なる推論結果を意図的に出力させる攻撃である．我々は，DICOMO2020 で組織内ネットワークの通信から攻撃通信を検知するシステムに対して，攻撃判定システムを意図的に回避できる敵対的入力の作成方法を報告した．Adversarial Training と呼ばれる既存防御手法を適用したが，作成した敵対的入力のうち約 50%しか検知できなかった．本稿では，攻撃判定システムの精度劣化を抑えながら，より多くの敵対的入力を検知する防御手法を提案し，実験結果を報告する．提案手法は，訓練データの重複コマンドに注目し，重複コマンドをそのまま訓練したモデルと重複コマンドを除去して訓練したモデルを用いる Stacking をベースとするアンサンブル学習を行う．結果，提案手法は従来手法と比較して敵対的入力を攻撃通信判定する確率を約 50%から約 90%に改善した．

キーワード：敵対的入力，回避攻撃，Adversarial Training，アンサンブル学習，サイバーセキュリティ

Enhanced Resistance to Evasion Attack of An Attack Detection Method on Internal Networks using Tensor Data Expansion

Shishido Katsunari[1,a)]　Morikawa Ikuya[1]　Oikawa Takanori[1]　Unno Yuki[1]

Abstract: Machine Learning is often applied to Cyber Security. At the same time, attacks specialized in Machine Learning have been observed, and then AI security is necessary to run Machine Learning systems safely. Adversarial examples are inputs to Machine Learning models that an attacker has injected perturbations into original inputs to cause the model to make a mistake. In DICOMO2020, we proposed a method of crafting adversarial examples against an attack detection method on internal networks. In order to detect the adversarial examples, we employed Adversarial Training, which is now most effective defense scheme. However, a probability of classifying adversarial examples as attack is approximately 50%. In this paper we propose a defense scheme that is keeping accuracy for standard data and able to classify more adversarial examples as attack than before. Our scheme is based on an Ensemble Learning called stacking using two different type of data. It contributes to classifying more adversarial examples as attack: the probability is approximately 90%.

Keywords: Adversarial example, Evasion attack, Adversarial Training, Ensemble Learning, Cyber Security

1. 序論

1.1 はじめに

近年，医療・産業・公的サービズなど様々な領域で機械

[1] 株式会社富士通研究所，神奈川県川崎市中原区上小田中 4-1-1
FUJITSU LABORATORIES LTD., 4-1-1, Kamiodanaka, Nakahara-ku Kawasaki, Kanagawa 211–0053, Japan
[a)] k.shishido@fujitsu.com

学習を用いたシステムが開発・実用化されている．サイバー攻撃に対する防御においても，機械学習を活用したマルウェア検知・侵入検知システムが開発され，攻撃活動の早期発見を実現している．

一方で，機械学習に特有な攻撃も多く発見されており，機械学習を用いたシステムに関するセキュリティは非常に重要な研究分野となっている．画像，動画，音声といったメディアデータを訓練する機械学習の潜在的な特徴・脆弱性を利用する攻撃・防御研究が日々進展している．機械学習に特有な攻撃の中でも，敵対的入力と呼ばれる攻撃 [1] は，その他の攻撃と比べ研究が活発である．敵対的入力は，機械学習の推論時に「元の入力を少し変化させた」入力を作成することにより，本来の推論結果と異なる推論結果を意図的に出力させる攻撃である．

本研究では，画像のようなメディアデータでなく，通信トラフィックをターゲットとする機械学習を用いたシステム，特に，組織内ネットワーク攻撃判定方式 [2] に対する敵対的入力の防御技術を提案する．サイバーセキュリティの防御システムは，通信トラフィックやマルウェアシグネチャといった構造データ（テーブルデータ）を訓練する機械学習を利用する．これらデータの敵対的入力が作成できれば，攻撃者は意図的に防御システムの検知から逃れることできる．つまり，敵対的入力の存在は，機械学習を用いたサイバー攻撃の防御システムにセキュリティホールがあることと等価である．そのため，機械学習システムが持つ正確性の劣化を防ぎながら，脆弱性対策を行い，攻撃被害を受けるリスクを低下させることが必要不可欠である．

我々は，DICOMO2020(マルチメディア、分散、協調とモバイルシンポジウム) にて，通信トラフィックの敵対的入力の作成方法を提案し，Adversarial Training[1] と呼ばれる既存防御手法の適用結果を報告した [3]．Adversarial Training を適用しても，テストデータセットから作成できる敵対的入力に対して，攻撃検知は約半数程度に留まっている．本稿では，通信トラフィックの敵対的入力に強いモデルの訓練方法およびアンサンブル学習の一種である Stacking を応用することにより，組織内ネットワーク攻撃判定方式が持つ本来の正確性の劣化を抑えつつ，より多くの敵対的入力を正しく判定する手法を提案する．

1.2 関連研究

1.2.1 侵入検知システムに対する回避攻撃

侵入検知システムに対する回避攻撃の研究として，ボットネット検知システムに対する回避攻撃 [4], [5] と対策 [6] が報告されている．これらの論文で対象にしているボットネットの検知は，フローベースの侵入検知システムで，インターネットと組織内ネットワーク間の通信トラフィックの分析を行ない，ボットネットマルウェアの感染検知を行う．一般的に，組織内ネットワーク上の感染端末は，インターネット上に存在する Command and Control（C&C）サーバーと通信して攻撃命令を受け取る．機械学習ベースのシステムでは，未感染端末が行う通信（正常通信）と感染端末が C&C サーバーと行う通信（攻撃通信）を学習することで，感染端末を検知している．ボットネット検知システムに対する回避攻撃 [4], [5] は，Targeted exploratory integrity attack[7] と呼ばれる攻撃を利用し，感染端末と C&C サーバー間の通信に小さな変更を加えることで，敵対的入力を作成して検知システムの回避を実現している．Targeted exploratory integrity attack を応用した回避攻撃の対策 [6] は，入力範囲の限定と回避攻撃耐性を強化するアンサンブル学習 [8] を組み合わせ，75%以上の回避攻撃を正しく判定している．

1.2.2 マルウェア検知に対する回避攻撃

機械学習を用いたマルウェア検知システムに対する回避攻撃 [9], [10], [11], [12] は，マルウェアの実行可能性を保ちながら，マルウェア本体のバイナリを書き換えたり，余分なバイナリを追加することで，敵対的入力を作成する．書き換えや追加を効率的に行うために，学習済みモデルの内部情報を利用する Fast Gradient Sign Method (FGSM)[1] や Jacobian-based Saliency Map Approach (JSMA)[13] が使われる．Chen ら [9] は，PE ファイルから得られる Windows API calls を学習してマルウェア検知を行うシステムを対象に，Windows API calls の可能な変更をすべて試して，少ない変更で正常判定される組み合わせを探索している．Kolosnjaji ら [10] は，PE ファイル本体を学習に利用する深層学習の検知システム (MalConv) を対象に，マルウェアのバイナリ本体に余分なバイナリを追加することで，敵対的入力を作成している．加えて，FGSM を利用して追加するバイナリを探索することで，ランダムに選択するよりもバイナリサイズが小さくなることを示した．Huang ら [11] は，独自に構築した深層学習の検知システムを対象に，攻撃者が White-box, Grey-box および Black-box の仮定のもとで，JSMA を利用してマルウェアの敵対的入力の作成が可能であることを示した．Suciu ら [12] は，MalConv を対象に，PE ファイルのヘッダーに記述されているバイナリファイルの本体サイズ（RawSize）とバイナリ実行時にメモリにロードされるサイズ（VirtualSize）の差に着目し，メモリにロードされない，すなわちマルウェアの実行可能性に影響を与えない領域に対して，FGSM を用いてバイナリを書き換えて敵対的入力を作成した．Huang ら [11] は，敵対的入力に対する対策として，Adversarial Training[1], Defensive Distillation[14], Feature Squeezing[15], PCA を用いた次元削減 [16] の 4 つの既存手法を適用し，Adversarial Training が最も敵対的入力を検知できることを実験的に示した．

2. 組織内ネットワーク攻撃判定方式

本節では，コマンド操作の特徴をテンソル化して訓練す

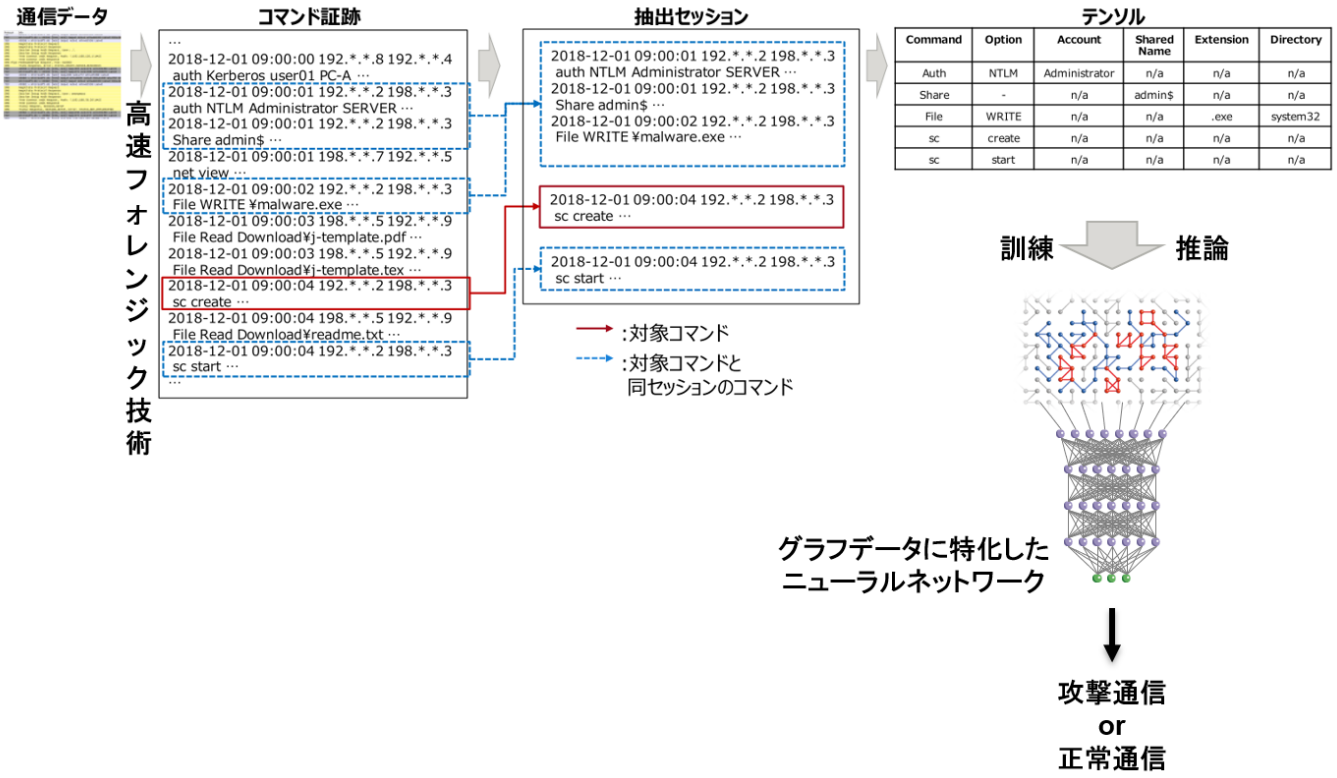

図 1　組織内ネットワーク攻撃判定方式 [2] の概要

Fig. 1　An overview of an attack detection method on Internal Networks[2]

ることで，組織内ネットワークを流れる通信を正常通信と攻撃通信に分類する攻撃判定方式 [2] について説明する．

組織内ネットワーク攻撃判定方式は，ネットワークを流れる膨大な通信データから怪しい通信を発生させたリモート操作コマンドの種類を特定してコマンドの再構成を行う高速フォレンジック技術 [17] とグラフ構造を持つデータの訓練・推論に特化したニューラルネットワーク [18] を組み合わせた機械学習システムである．このシステムの目的は，正常通信（対処不要）と攻撃通信（要対処）を分類し，インシデント対処の工数を削減することである．

組織内ネットワーク攻撃判定方式の概要を図 1 に示す．正常通信と攻撃通信を分類するために，高速フォレンジック技術で通信データをコマンド証跡に変換して再構成したリモート操作コマンドの特徴情報を訓練・推論に利用する．訓練・推論に利用されるデータは，攻撃者が常套的に使用するコマンドと同じセッションの他のコマンドをまとめて 1 セッションとして抽出されたものである．コマンドの特徴は，コマンド名，オプション，使用アカウント，ファイル拡張子名等，多次元かつ離散値の配列（テンソルデータ）で表現される．

高精度な分類を実現するために，大量の正常通信と攻撃通信のサンプルが必要になる．しかし，現実のネットワーク環境において，攻撃通信の発生は珍しく，大量の攻撃通信を集めることが難しい．組織内ネットワーク攻撃判定方式では，モデルの説明機能 (LIME)[19] を応用した攻撃通信のデータ拡充技術を提案しており，少量の攻撃通信を元に，ありうる変動を考慮した攻撃亜種データの生成を可能にした．元の攻撃通信に加えて攻撃亜種データを学習することで，正解率を 89.0%から 94.5%，再現率を 55.6%から 100.0%に改善した．

3.　これまでの成果

本節では，我々が DICOMO2020 で報告した「組織内ネットワーク攻撃判定方式に対する敵対的入力の作成方法」と「Adversarial Training の適用結果と課題」について述べる．

3.1　組織内ネットワーク攻撃判定方式に対する敵対的入力

組織内ネットワーク攻撃判定方式 [2] に対する敵対的入力の目的は，攻撃通信が攻撃判定器をすり抜け攻撃活動をすることである．つまり，攻撃通信の敵対的入力は攻撃判

Algorithm 1 攻撃通信の敵対的入力作成

Input: 攻撃通信 x_{attack}，正常通信 x_{benign}
Output: 攻撃通信の敵対的入力 x_{adv}
1: **if** 正常通信 x_{benign} は認証系コマンドを含まない **then**
2:　$x_{\text{adv}} \leftarrow x_{\text{attack}} || x_{\text{benign}}$
3:　**return** x_{adv}
4: **else**
5:　Abort
6: **end if**

定器を意図的に正常通信判定させ，かつ組織内ネットワークで攻撃活動ができるように作成しなければいけない．

また，作成した敵対的入力は，現実のネットワーク上で実現できる通信（以下，実現性）である必要がある．上記3つの条件を満たすために，我々は「特定の条件を満たした正常通信を攻撃通信に追加する方法」を提案した．敵対的入力の作成方法を Algorithm 1 に示す．Algorithm 1 内の記号 $x_{\text{attack}}||x_{\text{benign}}$ は，攻撃通信 x_{attack} の後方に正常通信 x_{benign} を追加することを意味する．表 1 の [2]（1 行目）に示すように，Algorithm 1 で作成した敵対的入力は，76.0%の確率で攻撃判定器を意図的に正常通信判定させることに成功している．

3.2 Adversarial Training の適用結果と問題点

Adversarial Training (AT)[1] は，既存の防御手法の中で最も有望な手法である．AT のアイデアは，訓練データから生成できる敵対的入力も訓練することである．訓練データから作成できる十分な数の敵対的入力を訓練することで，敵対的入力に対して強いモデルが獲得できると考えられる．我々は，訓練データのすべての正常通信と攻撃通信に対して Algorithm 1 を適用し，出力された敵対的入力の中から，正常判定されたデータを訓練することで，攻撃判定方式に AT を適用して攻撃耐性を確認した．表 1 の [2] × AT（2 行目），[2] × AT ×アップサンプリング（ 3 行目）に示すように，AT 適用後に発生するデータ不均衡をアップサンプリングで解消したことで，未対策のモデル [2] から正常データに対する精度の向上と敵対的入力の攻撃成功率の減少が認められた．

3.2.1 表 1 の用語

表 1 の用語について説明する．**適合率**は「攻撃通信判定したデータのうち，真の攻撃通信の割合」，**再現率**は「真の攻撃通信のうち，攻撃通信判定されたデータの割合」である．適合率は低いほど過剰検知が発生，再現率は低いほど攻撃通信の見逃しが発生していることを意味する．攻撃通信の敵対的入力のうち，**実験的な敵対的入力**は，Algorithm 1 を用いて自動的に作成したものである．本来，作成した敵対的入力が実現性を満たす通信となっているか確認する必要がある．しかし，膨大なデータ解析が必要で，実現性の確認をしていないため，参考値として扱う．ただし，攻撃通信に追加された正常通信は，認証系コマンドを含まないものに限定されており，現実性を満たす通信となっていることが期待できる．**現実的な敵対的入力**は，有識者が敵対的入力の作成方法に則り，手作業で作成した攻撃通信の敵対的入力である．有識者が手作業で作成しているため，実現性を満たす通信となっている．本研究では，現実的な敵対的入力の結果を主に扱い議論する．**アップサンプリング**は，少ないクラスのデータを増やし，各クラスのデータ数がつり合うようにするサンプリングである．AT では，攻撃通信の敵対的入力を作成・訓練するため，攻撃通信のデータ量が正常通信のデータ量よりも多くなる．具体的に，アップサンプリング前の訓練データの攻撃通信のデータ数の平均値は 3678.2 個に対して，正常通信のデータ数の平均値は 171 個であった．訓練データに含まれる各クラスのデータ量が偏った状態（以下，データ不均衡状態）で訓練を行うと，バイアスのあるモデルが構築されてしまう．データ不均衡状態を解決するために，敵対的入力の作成方法と同じく，「特定の条件を満たす正常通信を正常通信に追加する方法」でアップサンプリングを行った．

問題点

表 1 の [2] × AT ×アップサンプリング（ 3 行目）に示すように，Adversarial Training を適用しても，実験的な敵対的入力の攻撃成功率が 0.610，現実的な敵対的入力の攻撃成功率が 0.444 と高い．

4. 提案手法

本節では，より多くの敵対的入力を攻撃通信と判定するモデルを構築するためのデータ加工法と敵対的入力対策によって性能が落ちてしまう通常データ（敵対的入力を含まない正常通信と攻撃通信からなるデータセット）に対する適合率を改善するアンサンブル学習法を提案する．その準備として，4.1 節でアンサンブル学習の一種である Stacking について説明する．

4.1 Stacking

Stacking は，複数の分類器を用いて精度を改善するアンサンブル学習法の一種である．図 2 に示すように，一段目の訓練データ・クエリデータは，複数の分類器に入力され，それぞれ予測値が出力される．二段目で出力された予測値

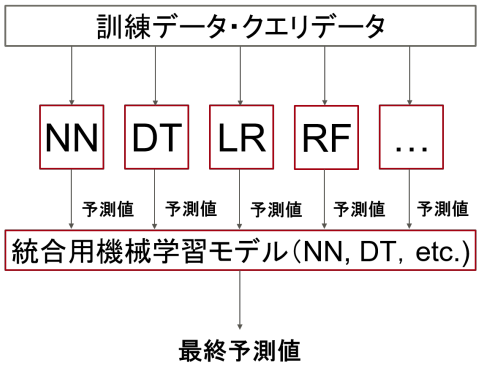

図 2 Stacking の構成例
Fig. 2 An example construction of stacking

は．統合用の分類器に入力され，最終的な予測結果が出力されるパラダイムである．高精度なモデルを獲得するために，ニューラルネットワーク (NN)，Random Forest(RF)，

表 1　Adversarial Training の適用結果 [3]
Table 1　A result of applying Adversarial Training[3]

モデル	通常データ (敵対的入力を含まない正常通信と攻撃通信)				敵対的入力	
					実験的な敵対的入力	現実的な敵対的入力
	正解率	適合率	再現率	F1 score	攻撃成功率	攻撃成功率
[2]	0.945	0.386	1.000	0.557	0.760	1.000
[2] ×　AT	0.895	0.250	1.000	0.400	0.503	0.889
[2] ×　AT × アップサンプリング	0.947	0.397	1.000	0.568	0.610	0.444

ロジスティク回帰 (LR)，決定木 (DT) など得意分野の異なるモデルが利用されることが多い．

4.2　提案手法

4.2.1　訓練データ・クエリデータのデータ加工

より多くの攻撃通信の敵対的入力を攻撃通信判定するモデルを構築するために，テンソルデータに存在する重複コマンドを除去するデータ加工法を提案する．訓練・推論時に入力するテンソルデータは，攻撃者が常套的に使用するコマンドと同じセッションの他のコマンドをまとめて 1 セッションとして抽出したコマンドである．1 セッション内で同一のコマンドが実行されれば，テンソルデータにも同一のデータが現れる．本研究では，1 セッション内で実行された同一のコマンドを重複コマンドと定義する．図 3 に示すように，訓練・推論時に入力するテンソルデータの重複コマンドを除去し，1 セッションに各コマンドがたった 1 つ存在する状態に変換してから訓練・推論を行う．

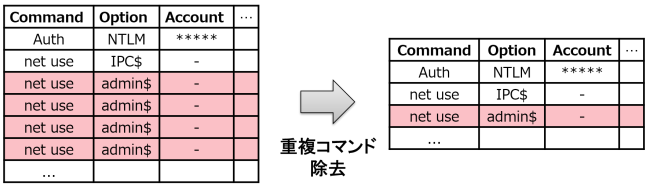

図 3　テンソルデータの重複コマンド除去の 1 例
Fig. 3　An example of duplicate elimination on tensor data

4.2.1.1　データ加工の新たな課題

テンソルデータの重複コマンドを除去した後に，Adversarial Training とアップサンプリングを行って訓練を行うことで，表 3 の重複コマンド除去モデル（3 行目）に示すように，敵対的入力の攻撃成功率がほぼゼロになる．しかし，通常データに対する適合率，および正解率が低下する．すべての入力を攻撃通信判定すれば，敵対的入力の攻撃成功率はゼロになるが，同時に適合率も下がる．重複コマンドを除去して訓練・推論を行う手法は，多くの入力に対して攻撃通信判定を出している状態に近いため，適合率の低下を抑えながら攻撃成功率をゼロに近づけることが訓練データ・クエリデータのデータ加工の新たな課題となる．

4.2.2　適合率を改善するアンサンブル学習法

敵対的入力の攻撃成功率の低下と適合率の改善を同時に達成するために，重複コマンドを除去しないデータを訓練したモデルと重複コマンドを除去したデータを訓練したモデルを統合する Stacking を応用したアンサンブル学習を提案する．図 4 に，提案手法の概要を示す．重複コマンドを除去しない訓練データを用いて訓練したモデルは，表 1 の [2] ×　AT ×アップサンプリングのモデルである．このモデルは，通常データに対する適合率が高いが，敵対的入力の攻撃成功率も高い．一方，重複コマンドを除去した訓練データを訓練したモデルは，[2] ×　AT ×アップサンプリングの訓練データの重複コマンドを除去したデータを訓練したモデルである．このモデルは，敵対的入力の攻撃成功率がほぼゼロと低いが，通常データに対する適合率も低い．本研究が提案するアンサンブル学習法は，訓練データの形式が異なる 2 つのモデルを用いる Stacking を応用したアンサンブル学習である．また，4.1 節で述べた公知の Stacking と異なり，各モデルの予測値に加えて，2 つの異なる訓練データの違いを表す特徴を入力している．具体的に，入力している特徴量は，以下の 2 つである．

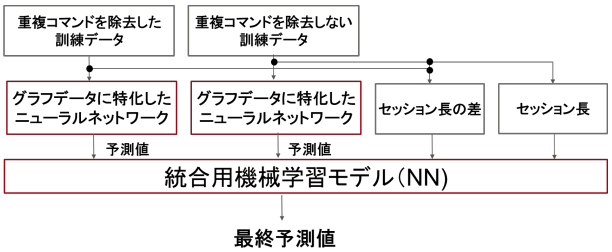

図 4　適合率を改善するアンサンブル学習法の概要
Fig. 4　An overview of our method for improving precision

(1) 重複コマンドを除去しないデータのセッション長
(2) 重複コマンドを除去しないデータのセッション長と重複コマンドを除去したデータのセッションの差

2 つの特徴量は，重複コマンドを除去しないデータと重複コマンドを除去したデータの差異を表すと考えられる．Algorithm 1 で示したように，敵対的入力は，攻撃通信の後方に正常通信を追加して作成されるため，セッション長が長く，かつ重複コマンドが発生しやすいと考えられる．こ

れらの特徴を Stacking する際に与えることで，入力データの特徴を統合用機械学習モデル（NN）が学習し，入力に応じて特徴の異なる 2 つのモデルを選択できるようになる．入力に応じて最適なモデルを選択できるようになれば，通常データに対する適合率の低下を抑えながら，敵対的入力の攻撃成功率を低下することが期待できる．

5. 実験と評価

5.1 実験データ

本研究で使用した攻撃通信と正常通信はデータ [2] で使われたデータである．本稿では概要のみ説明する．

攻撃通信のうち，訓練データは「ATA Suspicious Activity Playbook」[20] 等を参考に作成し，実際に模擬環境において得られた通信観測データを使用した．

攻撃通信のうち，評価データには動的活動観測 BOS(Behavior Observable System) の研究用データセット (以下，BOS Dataset) とサイバー攻撃誘引基盤 STARDUST[21] の通信観測データの一部を使用した．なお、BOS Dataset はマルウェア対策のための研究用データセット (MWS Datasets)[22] に含まれている．具体的には，STARDUST の通信観測データのうち，日本や台湾の組織を標的とした DragonOK[23] と PowerShell Empire[24] による攻撃の通信観測データを使用した．

正常通信は 2018 年 3 月から 8 月までの 6 ヶ月間に富士通株式会社が管理するネットワークから収集したデータである．そのうち，前半 3 ヶ月分を訓練データ，後半 3 ヶ月分を評価データとして使用した．

5.2 訓練

組織内ネットワーク攻撃判定方式 [2] は，訓練データセットから 10 個のサブデータセットを作成し，各データセットを学習して 10 個の分類器を構築して正常通信と攻撃通信を分類する．我々は，各サブデータセットに対して，Adversarial Training とアップサンプリングを適用したデータセットとそのデータセットから重複コマンドを除去して得られたデータセットの計 20 個のデータセットを作成した．表 2 に，訓練データ数と評価データ数を示す．

表 2 提案手法の訓練データと評価データ数
Table 2 the number of training dataset and test dataset

	訓練データ	評価データ
攻撃通信と実験的な敵対的入力	3678.2(平均)	5346
現実的な敵対的入力	—	9
正常通信	3678.2(平均)	748

5.3 評価

重複コマンドを除去しないデータで訓練したモデル（重複コマンド除去しないモデル）[3]，重複コマンドを除去したデータを訓練したモデル（重複コマンド除去モデル（提案手法 1））, 重複除去しないモデルと重複除去したモデルの Stacking（Stacking モデル）と提案手法である Stacking 応用モデル（提案手法 2）の通常データに対する精度と敵対的入力の攻撃成功率を表 3 に示す．通常データは敵対的入力を含まない正常通信と攻撃通信からなるデータセットである．重複コマンド除去しないモデルは適合率が高い（正常通信を攻撃通信と判定する過剰検知が少ない）が，敵対的入力の攻撃成功率が高い．重複コマンド除去モデルは，敵対的入力の攻撃成功率がほぼゼロと低いが，適合率も低いことがわかる．これら 2 つのモデルに Stacking を適用すると，通常データの適合率および正解率は改善するが，敵対的入力の攻撃成功率の低下しない．つまり，Stacking は，敵対的入力の攻撃成功率を下げる効果が弱いことを示している．一方，提案手法である Stacking 応用モデルは，敵対的入力の攻撃成功率が 0.111 まで改善され，適合率の低下も重複コマンド除去モデルと比べると抑えられている．4.2.2 節で述べた 2 つの特徴量も入力して訓練することで，Adversarial Training のみを適用した既存手法 [3] よりも敵対的入力の特徴を捉えていると考えられる．

5.4 システムとしての対策

これまで述べた方法は，訓練に利用するデータや訓練方法を工夫する敵対的入力の対策だった．現実で利用される機械学習システムは，学習済みモデルと専門家が決めたヒューリスティックなルールを組み合わせてシステム化される．そのため，システムの一部である学習済みモデルがもつ潜在的な脆弱性を小さくすることには，大きな意味がある．本研究の提案手法は，9 個ある現実的な敵対的入力のうち，1 個を正常通信に誤判定している．誤判定を起こした敵対的入力を分析すると，ルールベースで補えば攻撃通信に判定できることがわかった．つまり，学習済みモデルが敵対的入力を正しく分類できなくても，分析から得られたルールを適用することで，システム的に弾くことができる．このように，学習済みモデルが受ける攻撃リスクを小さくすることで，機械学習システムとして敵対的入力に対して効果的な対策ができるようになる．

6. まとめ

本稿では，画像などのメディアデータと異なる特性を持つ通信データの敵対的入力を扱い，アンサンブル学習の一種である Stacking を応用した敵対的入力の防御手法を提案した．既存防御手法である Adversarial Training は，敵対的入力の攻撃成功率が約 50%と高い．本研究では，通信データ特有の重複コマンドに着目し，重複コマンドを除去しないデータと除去したデータを訓練したモデルと 2 つのデータの違いを表す特徴量を用いて Stacking を応用したア

表 3　提案手法の適用結果
Table 3　A result of applying our methods

モデル	通常データ				敵対的入力	
					実験的な敵対的入力	現実的な敵対的入力
	正解率	適合率	再現率	F1 score	攻撃成功率	攻撃成功率
敵対的入力の対策なしモデル [2]	0.945	0.386	1.000	0.557	0.760	1.000
重複コマンド除去しないモデル [3]	0.947	0.397	1.000	0.568	0.610	0.444
重複コマンド除去モデル (提案手法 1)	0.875	0.218	1.000	0.358	0.003	0.000
Stacking モデル	0.956	0.443	1.000	0.614	0.520	0.444
Stacking 応用モデル (提案手法 2)	0.914	0.293	1.000	0.450	0.124	0.111

ンサンブル学習を行うことで，分類器の精度劣化を抑えながら敵対的入力の攻撃成功率を下げることができた．すべての敵対的入力を正しく分類することはできなかったが，提案手法とシステム的な補助ルールを組み合わせて守ることで，提案手法が検知できなかった敵対的入力も正しく分類することが可能になる．

謝辞　本研究を進めるにあたり，貴重な攻撃データの提供およびサイバー攻撃誘引基盤“STARDUST”を開発した独立行政法人情報通信研究機構（NICT）に感謝いたします．

参考文献

[1] Goodfellow, I. J., Shlens, J. and Szegedy, C.: Explaining and Harnessing Adversarial Examples, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

[2] 及川孝徳，西野琢也，矢野翔太郎，海野由紀，古川知快，鳥居　悟，伊豆哲也，金谷延幸，津田　侑，井上大介：テンソルデータ拡充を用いた組織内ネットワーク攻撃判定方式，暗号と情報セキュリティシンポジウム (SCIS) (2019).

[3] 宍戸克成，森川郁也，及川孝徳，海野由紀：テンソルデータ拡充を用いた組織内ネットワーク攻撃判定方式の回避攻撃に対するロバスト性の向上，マルチメディア、分散、協調とモバイルシンポジウム (DICOMO) (2020).

[4] Apruzzese, G., Colajanni, M. and Marchetti, M.: Evaluating the effectiveness of Adversarial Attacks against Botnet Detectors, *18th IEEE International Symposium on Network Computing and Applications, NCA 2019, Cambridge, MA, USA, September 26-28, 2019* (Gkoulalas-Divanis, A., Marchetti, M. and Avresky, D. R., eds.), IEEE, pp. 1–8 (2019).

[5] Wu, D., Fang, B., Wang, J., Liu, Q. and Cui, X.: Evading Machine Learning Botnet Detection Models via Deep Reinforcement Learning, *2019 IEEE International Conference on Communications, ICC 2019, Shanghai, China, May 20-24, 2019*, IEEE, pp. 1–6 (2019).

[6] Apruzzese, G., Andreolini, M., Marchetti, M., Colacino, V. G. and Russo, G.: AppCon: Mitigating Evasion Attacks to ML Cyber Detectors, *Symmetry*, Vol. 12, No. 4, p. 653 (2020).

[7] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G. and Roli, F.: Evasion Attacks against Machine Learning at Test Time, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III* (Blockeel, H., Kersting, K., Nijssen, S. and Zelezný, F., eds.), Lecture Notes in Computer Science, Vol. 8190, Springer, pp. 387–402 (2013).

[8] Biggio, B., Corona, I., He, Z.-M., Chan, P. P., Giacinto, G., Yeung, D. S. and Roli, F.: One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time, *International Workshop on Multiple Classifier Systems*, Springer, pp. 168–180 (2015).

[9] Chen, L., Ye, Y. and Bourlai, T.: Adversarial Machine Learning in Malware Detection: Arms Race between Evasion Attack and Defense, *European Intelligence and Security Informatics Conference, EISIC 2017, Athens, Greece, September 11-13, 2017* (Brynielsson, J., ed.), IEEE Computer Society, pp. 99–106 (2017).

[10] Kolosnjaji, B., Demontis, A., Biggio, B., Maiorca, D., Giacinto, G., Eckert, C. and Roli, F.: Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables, *26th European Signal Processing Conference, EUSIPCO 2018, Roma, Italy, September 3-7, 2018*, IEEE, pp. 533–537 (2018).

[11] Huang, Y., Verma, U., Fralick, C., Infante-Lopez, G., Kumar, B. and Woodward, C.: Malware Evasion Attack and Defense, *49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, DSN Workshops 2019, Portland, OR, USA, June 24-27, 2019*, IEEE, pp. 34–38 (2019).

[12] Suciu, O., Coull, S. E. and Johns, J.: Exploring Adversarial Examples in Malware Detection, *2019 IEEE Security and Privacy Workshops, SP Workshops 2019, San Francisco, CA, USA, May 19-23, 2019*, IEEE, pp. 8–14 (2019).

[13] Papernot, N., McDaniel, P. D., Jha, S., Fredrikson, M., Celik, Z. B. and Swami, A.: The Limitations of Deep Learning in Adversarial Settings, *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, IEEE, pp. 372–387 (2016).

[14] Papernot, N., McDaniel, P. D., Wu, X., Jha, S. and Swami, A.: Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks, *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, IEEE Computer Society, pp. 582–597 (2016).

[15] Xu, W., Evans, D. and Qi, Y.: Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks, *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*, The Internet Society (2018).

[16] Bhagoji, A. N., Cullina, D., Sitawarin, C. and Mittal, P.: Enhancing robustness of machine learning systems via data transformations, *52nd Annual Conference on*

Information Sciences and Systems, CISS 2018, Princeton, NJ, USA, March 21-23, 2018, IEEE, pp. 1–5 (online), DOI: 10.1109/CISS.2018.8362326 (2018).

[17] 海野由紀，森永正信，及川孝徳，古川和快，金谷延幸，津田　侑，遠峰隆史，井上大介，鳥居　悟，伊豆哲也：標的型攻撃の被害範囲を迅速に分析するネットワークフォレンジック手法の改良，コンピュータセキュリティシンポジウム (CSS) (2018).

[18] Maruhashi, K., Todoriki, M., Ohwa, T., Goto, K., Hasegawa, Y., Inakoshi, H. and Anai, H.: Learning Multi-Way Relations via Tensor Decomposition With Neural Networks, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, AAAI Press, pp. 3770–3777 (2018).

[19] Ribeiro, M. T., Singh, S. and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016* (Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D. and Rastogi, R., eds.), ACM, pp. 1135–1144 (2016).

[20] Harris, A. and Levitz, G.: *ATA Suspicious Activity Playbook*.

[21] 津田　侑，遠峰隆史，金谷延幸，牧田大佑，丑丸逸人，丑丸逸人，神宮真人，高野祐輝，安田真悟，三浦良介，太田悟史，宮地利幸，神薗雅紀，衛藤将史，井上大介，中尾康二：サイバー攻撃誘引基盤 STARDUST，コンピュータセキュリティシンポジウム (CSS) (2017).

[22] 荒木粧子，笠間貴弘，千葉 大紀充弘，寺田真敏：マルウェア対策のための研究用データセット～ MWS Datasets 2019 ～，情報処理学会, Vol.2019-CSEC-86, No.8, 2019 年 7 月 (2019).

[23] Haq, T., Moran, N., Vashisht, S. and Scott, M.: *Operation Quantum Entanglement* (2014).

[24] 石川芳浩：PowerShell Empire を利用した標的型攻撃 (2017).