Vol.2021-DPS-186 No.58
Vol.2021-CSEC-92 No.58
2021/3/16

組織内で学習データを採取し定期的に判別器を更新する機械学習ベースのNIDS

佐藤 秀哉[1,a)]　林 はるか[1]　小林 良太郎[1]

概要：本論文では日々変化するサイバー攻撃へ対応するための機械学習型の NIDS を提案する．関連研究では正常悪性のラベルを付与した公開されているデータセットを使用して検知率等を改善する研究が多い．これらの問題点がいくつかあげられる． まず，正常通信が各組織によって大きく異なるために発生する正常通信の誤検知率の高さ．次に攻撃者が事前に検証されるリスクなどが考えられる．そのため本研究では NIDS を設置する特定の組織のミラーポート等を使用し正常通信を回収する．さらに，ハニーポットを設置しその組織に対する悪性通信を回収する．それら通信データから特徴量を抽出して学習にかけることにより，最新の通信データを反映させた特定の組織向けの機械学習型 NIDS を作成する．正常通信と悪性通信の自動収集システムにおいて取得した通信データにおいて，抽出した特徴量で学習を行い取得した通信データで判別を行ったところ，非常に低い誤検知率となった．上記の結果は設置組織内で正常通信と悪性通信を採取することの重要性を示している．

キーワード：機械学習，動的生成，NIDS，組織内ネットワーク

Machine Learning-based NIDS That Collects Learning Data Within an Organization and Updates Discriminators on a Regular Basis

Abstract: According to the IPA’s ”10 Major Security Threats”, the threat of targeted attacks on organizations is Targeted attacks have been recognized as a major threat to Japanese organizations, as it was ranked first in the last year. With this background, machine learning NIDS has been studied in recent years, and we have also proposed a NIDS system that acquires communication data of installed organizations and uses them for machine learning. In this system, the conditions for generating the discriminator (features, algorithm and training data) are it is predetermined for a particular pair. However, the discriminator generation conditions required to achieve higher accuracy and shorter learning time change day by day within the installation organization. Therefore, in this study, we prepared several discriminator generation conditions, selected one of them dynamically appropriate, and we study and preliminary evaluation of a method to generate a discriminator dynamically.

Keywords: Machine learning, Dynamic generation, NIDS, Internal network

1. はじめに

IPA が発表した 2021 年における情報セキュリティ 10 大脅威 [1] で組織を狙った標的型攻撃による被害が 2 位となっている．その手段は多岐にわたり日々新しい攻撃手法が出現している．それらから情報資産を守る方法としてネットワーク侵入検知システム (Network-based Intrusion Derection System: NIDS) があげられる．これまで, 様々な NIDS の研究がなされており, OSS において有名な不正侵入検知・防止システムとして Suricata があげられる．これは不正パターンを登録するシグネチャ型である．様々な攻撃をすることで Suricata の評価研究を Kittikhun Thongkanchorn 氏ら [2] は行った．既製品としてあげられる不正侵入検知・防止システムとしては，Symantec Endpoint Security[3] や TippingPoint Threat Protection System[4] があげられる．Symantec Endpoint Security では，侵入前検知としてファイアウォールによる脅威があると思われる通信を遮断し

[1] 工学院大学 Kogakuin Uniersity
[a)] j117135@ns.kogakuin.ac.jp

ⓒ 2021 Information Processing Society of Japan

たり，よく使用するソフトウェアのうちパッチが適用されていないものへのエクスプロイットやゼロデイ攻撃の無効化などを提供している．また，感染時には，Global Intelligence Network 内の良好ファイルと不良ファイルの例を使用した機械学習を使用し，良好ファイルや不良ファイルなどの判別を行う．TippingPoint Threat Protection System でも機械学習を利用してネットワークトラフィックが不正であるかの判別を行っている．これら既製品などのように従来の不正侵入検知システムに加え機械学習を利用した判別方法が導入されてきている．

近年の研究では，機械学習を主とした侵入検知システムの研究も盛んに行われており，非常に高い精度での検知が可能となっている．これらの研究においては使用した特徴量データセットにおいて高い検知率を出すことが目的となっている．だが，データセットで使われている特徴量と実際の NIDS 設置組織での特徴量は使用しているソフトウェアやサーバーの設置環境により大きく異なると考えられる．また，データセットは取得期間が決まっているため，日々変わる組織への攻撃通信 (以降悪性通信) と組織内の正常な通信 (以降正常通信) へと対応できない可能性がある．従来の研究ではこれら点について考慮されておらず，また実稼働させている例は我々の知る限りでは存在しなかった．

そこで我々は，NIDS の設置組織でリアルタイムに正常通信と悪性通信を取得し，日々継続的に更新を行う機械学習型 NIDS システムの提案をする．これは，常に最新の通信データを使用し，NIDS の設置組織内で特徴量データセットを取得することによりデータセットの取得環境の依存によるスコアの低下を避けることができ，日々変わる正常通信の変化へも対応することができる．また，悪性通信も組織で取得しているため常に最新の攻撃傾向や手法のデータセットを取得でき，極めて高い確率による悪性通信の検知が可能となる．

本論文では，この提案システムを実際に稼働させて得られた結果とその調査記録を述べる．2 章では機械学習型 NIDS についての関連研究について述べる．3 章では，提案手法の測定環境を述べる．4 章では，評価システムの構築とその評価指標について述べる．5 章では，作成した評価システムから得られた通信データの分類結果を述べる．6 章では，考察を述べ，7 章で本論文をまとめる．

2. 関連研究

現在，様々なデータセットを使用した機械学習型 NIDS の研究がなされている．ここで使用されているデータセットは様々であり，それぞれ特徴量も異なる．

2.1 既存の機械学習型 NIDS の関連研究

Mrudul Dixit らは，DDos 攻撃へ対策としてナイーブベイズと SVM ベースの NIDS を提案した [5]．この研究では，従来の欠陥があるとするポート番号ベースのシグネチャ型によるパケットフィルタリングの問題点を指摘し，実験用に作成したデータセットを使用して学習・検証を行っていた．Chie-Hong Lee らは，C-ELM を利用した機械学習ベースのネットワーク侵入検知システムを提案した [6]．この研究では，既存のシグネチャ型による悪性通信の検知の問題点を指摘し，NSL-KDD Data set[7] を利用して C-ELM 構成アプローチの最適化を行った．近松康次郎らは，多層パーセプトロンを用いた NIDS の改良手法の検討を行った [8]．この研究では，KDD Cup 1999 Data[9] や NSL-KDD Data set を使用した研究では User-to-Root(U2R) や Remote-toLocal(R2L) の検知率が低いことを指摘し，それらデータセットでの NIDS の改良手法を検討した．平野誠らは，標的型攻撃の対策として多層パーセプトロンを用いた NIDS の学習手法と評価実験を行った [10]．この研究では，近年のマルウェアが大量の攻撃通信を発しなくなったため，従来の手法 (パターンマッチング) では攻撃を検知することができない点を指摘し，Kyoto 2016 Dataset[11] を利用して多層パーセプロンによる評価実験を行った．

2.2 多く研究利用される特徴量データセット

公開されており，かつ，機械学習型の NIDS において訓練データとして用いることのできるデータセットは様々ある：KDD Cup 1999 Data， NSL-KDD Data set，Kyoto 2016 Dataset, CSE-CIC-IDS2018[12] など．NSL-KDD Data set は，KDD Cup 1999 Data が持ついくつかの問題点を解決したものである．Kyoto 2016 Dataset は，KDD Cup 1999 Data などとは異なり，長期に及ぶ悪性通信のデータを収集している．これらのデータセットは，機械学習を利用した侵入検知システムの研究に数多く使用されている．

3. 提案システムの作成

この提案システムの作成にあたり，設置組織内でリアルタイムで正常通信の取得と悪性通信の取得が必要となる．まず，提案システムの作成に必要となる要素の概要を述べる．次に，我々が想定した提案システムの具体的な説明に入る．

3.1 提案システムの概要

そこで本研究では，動的な機械学習ベースの NIDS 運用システムを提案する．提案システムは設置組織内において正常通信と悪性通信を取得し学習を行う．

まず，本提案システムでは，設置組織のエッジルーターから組織内の通信を正常通信として取得し，組織内に設置したハニーポット等から得られる通信を悪性通信として取得する．なお，設置組織内に新しい機材やソフトを導入する場合は，悪性通信として判別される恐れがあるため，正

常通信としてラベル付けをする必要がある．それらすべての通信データから特徴量を抽出し，そのデータを用いて機械学習により判別器を生成する．以上により動的な機械学習ベースの NIDS による悪性通信の判別を可能とする．

3.2 提案システムのネットワーク図

提案システムで使用するネットワークを図 1 に示す．図 1 には社内ネットワーク，悪性通信収集ネットワーク，解析ネットワークの 3 つが存在する．これらのネットワークはルーターとファイアウォールを介し外部ネットワークにつながっている．社内ネットワークにおいて正常通信を取得し，悪性通信収集ネットワークにおいて悪性通信を取得する．そして，解析ネットワークでは，正常通信と悪性通信から特徴量を抽出し，その特徴量を入力とした機械学習により判別器を生成し，この判別器を用いて社内ネットワークにおける悪性通信の検知を行う．

3.3 正常通信の取得

社内ネットワークでは，正常か悪性かわからない通信が流れているものとする．ここでいう正常通信とは組織内で使用されているソフトウェア・機材がその使用目的のために必要としている通信のことである．また，それらの通信がないとその組織活動または目標達成のために支障をきたす可能性があるものとする．このネットワークにおける外部との通信はすべて Router 1 を通る．社内ネットワークの外部・内部通信はいずれも Router 1 のミラーポートを介して，解析ネットワークの特徴量抽出マシンに送信される．使用しているルーターにミラーポートが存在しなかったり，設備費用に余裕がありネットワークの分岐が複雑である場合などはこのルーターを複数個使用することも可能である．その場合でも社内ネットワークで発生した通信はすべて解析ネットワークの特徴量抽出マシンに送信する必要がある．なお，社内ネットワークから送信される通信は，外部からの攻撃により悪性通信を含む可能性があるため，NIDS によって正常と判別されたもののみ正常通信としてラベル付けされる．

3.4 悪性通信の取得

悪性通信収集ネットワークは，設置組織をターゲットとした悪性通信を収集するためのネットワークである．ここでの悪性通信とは，設置組織における外部または内部からの不正な通信や普段は通信を行っていないようなネットワークへの通信など，その組織が普段行っている通信から逸脱している通信のことを示す．この悪性通信収集ネットワークでは，常に外部からの悪性通信による危険がある．そのため，社内ネットワークなど正常通信が通る Router 1 とは別に Router 0 を設置し，その下に悪性通信収集ネットワークを構築していく．このとき，設備費等の問題でその設置が困難な場合には，ルーターを 1 つにして VLAN などでネットワークを仮想的に区切ることでそれを実現することも検討してよい．VLAN とは，ルーター内部で仮想的な LAN セグメントを作成することによってネットワークの分離を図るものである．このいずれかの方法で分離した悪性通信収集ネットワークは，社内ネットワークとの通信は不可能である．なお，提案システムにおける解析ネットワークはミラーポートを介して正常通信と悪性通信を取得しているため，社内ネットワークや悪性通信収集ネットワークからその存在を認識されることはない．

悪性通信を取得するための環境構築としてまず Router 0 に作成した DMZ などの他ネットワークに影響を及ぼさないように設定をしたネットワークにハニーポットを設置する．ハニーポットとは，あえて脆弱性を残し攻撃者からの不正アクセスを受ける前提で設置されマルウェア検体を入手するためなどに使用されるシステムのことである．ハニーポットの外部通信はいずれも Router 0 のミラーポート等を介して，解析ネットワークの特徴量抽出マシンに送信される．また，ハニーポット内にバイナリファイルが設置された場合，それを外部ネットワークから隔離した SandBox 上に設置し実行する．SandBox の通信も特徴量抽出マシンに送信するが，SandBox は外部との通信を遮断する．そして，ハニーポット同様 Router 0 のミラーポート等を使用し，外部へ通信を漏らさないように設定をし，解析ネットーワークに通信データを送信する．なお，悪性通信収集ネットワークから送信される通信は，すべて悪性通信としてラベル付けされる．

3.5 特徴量の抽出

特徴量抽出マシンはルーターを介して送られてきた通信データから，セッション単位ごとに特徴量を抽出する．特徴量とは，セッションの特徴が数値化されたものを示す．

特徴量抽出マシンは社内ネットワークから得られた通信データの特徴量をセッションごとに NIDS へと送信するが，悪性通信ネットワークから得られた通信データの特徴量は悪性とラベル付けしてセッションごとに学習マシンへと送信する．これは，特徴量抽出マシンでは社内ネットワークから得られた通信データが正常かどうかを判断することができないため，NIDS による判別を行ってから学習マシンへの送信を行う必要があるためである．

3.6 悪性通信の検知

NIDS では，特徴量抽出マシンから送られてきた社内通信における通信データの特徴量を入力として，セッション単位で悪性通信を検知する．判別が正常通信であった場合は，正常とラベル付けして学習マシンに送信する．判別が悪性通信であった場合は，悪性とラベル付けして学習マシンに送信する．なお，悪性通信収集ネットワークから得

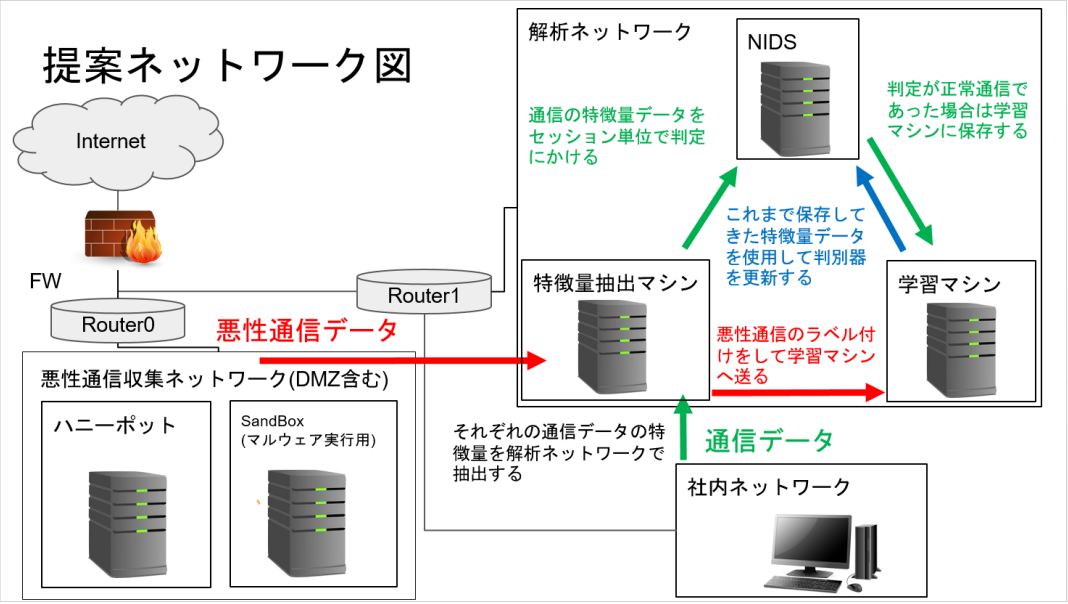


図 1 提案システムのネットワーク図
Fig. 1 Network diagram of the proposed system

られた通信データは悪性であることがわかっているため，NIDS による判別の対象とはならない．

3.7 機械学習による判別器の更新

学習マシンは，NIDS から送られてくる正常とラベル付けされた特徴量と特徴量抽出マシンから悪性とラベル付けされた特徴量を入力として機械学習を行い，判別器を更新する．この判別器はセッション単位で悪性か正常のどちらかの判別を行う．学習マシンは，生成した判別器を NIDS に送信し，NIDS 上の判別器と置き換える．

3.8 提案システムの動作タイミング

本システムでは，時間の経過とともに変化する攻撃手法や正常通信のために定期的な判別器の再生成を行う必要がある．更新は深夜などの作業を行う人が居ない時間帯に行うものとする．時系列毎の更新の様子を図 2 に示す．パケット取得とマルウェア検知は常に行うものとする．取得した通信データは 24 時間おきにまとめて特徴量抽出・学習マシンへと送信する．判別器更新にかかる時間を 6 時間と過程し，その更新が終了するまでは前日の判別器を使用する．その更新が終了したときにマルウェア検知に使用している判別器を新しい判別器へと変更を行う．

4. 評価

提案システムは，設置組織に適した機械学習型 NIDS の生成を行うことを目的としたシステムである．そこで，提案システム構成を元にほぼ同等の評価システムを構築し，その分類精度を提案システムの指標として示すことができると考えられる．NIDS における通信を正常か悪性かを判断する判別器を機械学習手法であるランダムフォレストを使用して作成した．ランダムフォレストを使用する理由としては，測定時最も判定率が高かったためである．

4.1 提案システムと評価システムの差異

評価システムは提案システムとほぼ同じにできているが，いくつか異なる点が存在する．まず，提案システムでは社内ネットワークで得られた通信は 1 度 NIDS の判定にかけ正常通信と判定されたもののみを正常通信の学習データとして解析ネットワーク内の学習マシンに保存するが，評価システムでは社内ネットワークで得られた通信はそのまま正常通信と決めつけ特徴量を取り出しラベル付けを行っている．次に，正常通信に関してだが本来は人が行った通信を取得してそれを社内ネットワークの通信データとして使うが今回は社内ネットワークに正常通信生成用スクリプトを設置し，その通信を取得する．次に，提案システムではハニーポットは DMZ 上に設置しているが，DMZ では攻撃とは関係ない通信が多く来ることにより判定率が下がることがわかったので，ポートフォワーディングによるパケットルーティングを行った．

4.2 ソフトウェア構成

今回，評価システムでの通信取得にはすべて tcpdump を使用している．悪性通信の取得には，マルチハニーポットプラットフォームの 1 つである T-Pot を使用した．また，そこで取得できるマルウェアは Docker 上で動作させ通信の取得を行った．それら取得した通信データは bro-IDS と呼ばれるオープンソースのソフトウェアネットワーク分析フレームワークのプラグインを使用して特徴量を抽出し

Vol.2021-DPS-186 No.58
Vol.2021-CSEC-92 No.58
2021/3/16

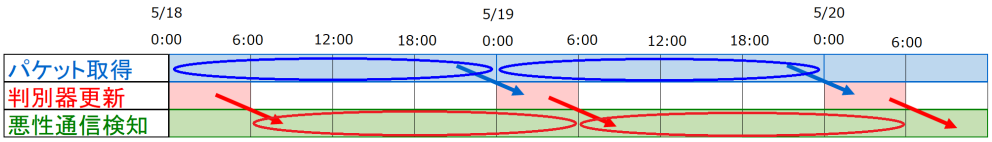


図 2 判別器の更新時間
Fig. 2 Discriminator update time

た．今回使用した特徴量は抽出した 47 種類のラベルのうち，整数データの 26 種類である．それを機械学習ライブラリの 1 つである scikit-learn を使用して学習モデルの作成を行った．

5. 作成した評価システムから得られた通信データの分類結果

本評価システムでは，システムネットワーク内でリアルタイムで得られた通信データを使用し，動的に更新を行うことによって最新の攻撃への対応やデータ取得環境の依存によるスコアの低下を避けることができると想定している．そこで，評価システムを稼働し得られた結果から正常通信の誤検知や悪性通信の検知の割合を取得する．それによって従来システムとの比較用データや提案システムの評価指標を示すことができると考える．

5.1 評価システムにおける評価指標

評価システムから得られた分類結果より，表 1 のような混同行列を定義する．そして，その混同行列から今度は偽陽性率である FPR(False Positive Rate) と真陽性率である TPR(True Positive Rate) を定義した．偽陽性率は正常通信を悪性通信と誤って判断してしまう確率であり，真陽性率は悪性通信を悪性通信と正しく判別できた確率である．

$$FPR = \frac{FP}{TN + FP}$$

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

表 1 混同行列

	データの結果	
判別結果	TP(True Positive)	FP(False Positive)
	FN(False Negative)	TN(true Negative)

5.2 判別器更新

本評価システムでは更新は最も高頻度で 1 日単位での更新を行っている．また，そのほかにも学習データを増加させた場合や判定データと学習データの期間が開いてしまった場合の影響などを調べるために，1 週間前と 2 週間前からそれぞれメンテナンスや不具合発生日を含む 7 日分の学習データを使用していく．メンテナンスや不具合が起きた日における通信データは存在しない．または，データ量が通常の日と比べて低下している．今回の評価システムでは，1/13[*1]に通信量増大のための正常通信生成スクリプトの変更と 1/24〜1/27 に学習マシンのパーツのメンテナンスを実施した．それにより，1/24〜1/27 のデータは今回存在しない．

5.3 今回得られた正常と悪性の特徴量データの調査

特徴量データとして寄与率の高いレスポンスポート番号を示していく [13]．まず，正常通信生成スクリプト変更前の期間である 1/1〜1/12 までのデータをまとめた図 3 とスクリプト変更後の 1/13〜1/31 までのデータをまとめた図 4 を示す．スクリプト変更前に多く見られる 57482 番ポートは正常通信の再現のために ssh 通信として使用しているポートである．変更後は 443 番ポートへの通信を増やしたため，割合が高くなった．また，443 番ポートへの通信のみを増やしたためその他のポート番号への通信量の総和は低下していない．次に，1/1〜1/31 の期間にハニーポットへ攻撃が行われたポート番号のうち上位 5 つの割合を図 5 示す．最も攻撃通信が多かった通信は 22 番ポートだった．よく使用される ID やパスワードの試行の記録も観測されたため ssh 接続を機械的に試みているものだと考えられる．それに続けて 445，3389 番ポートが多い結果となった．今回の図では，上位 5 つのポート番号のみ示しているがそのほかの多くのポート番号が攻撃を受けていた．

5.4 異なる学習データセット区分を使用して得られた分類結果

異なる学習データセット区分を使用し，ランダムフォレストにより作成した判別器から得られた分類結果を示す．表 2 では，前日から 1 日分のデータを使用している．表 3 と表 4 では，それぞれ 1 週間分のデータを使用している．これら表の違いは表 3 は 1 週間前から，表 4 では 2 週間前からのデータを始点としてデータを使用している．1/24〜1/27 のデータはメンテナンスの実施のため判定データがなく，各表に載せていない．

前日の 1 日分のデータを学習に使用して判定した結果で

[*1] 使用しているデータはすべて 2021 年度のものです．

ⓒ 2021 Information Processing Society of Japan

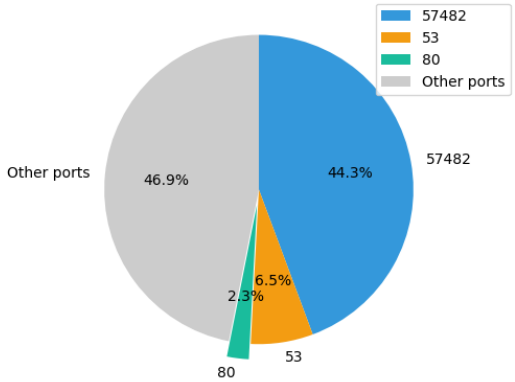


図 3 スクリプト変更前の正常通信の特徴量 1/1〜1/12
Fig. 3 Features of normal communication before script change 1/1〜1/12

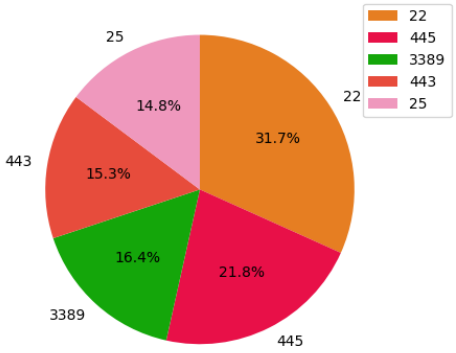


図 5 ハニーポットの攻撃傾向 [レスポンスポート]
Fig. 5 Attack tendency of honeypot[response port]

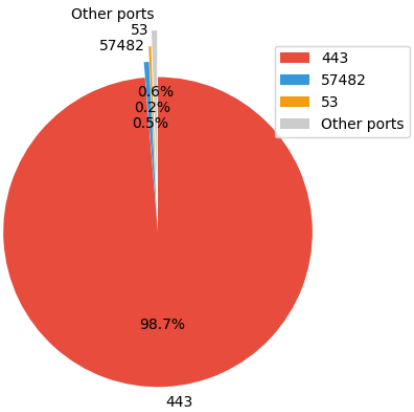


図 4 スクリプト変更後の正常通信の特徴量 1/13〜1/31
Fig. 4 Features of normal communication after script change 1/13〜1/31

ある表 2 の結果に注目すると，1/1〜1/13 における FPR はほとんどにおいて 0.5%を切っている結果となった．1/14 以降においては，FPR は 0.05%以下となる場合が多く，最も精度が低くても 3%程度となった．また，TPR は 1/1〜1/31 の多くの場合で 99.8%以上と高い精度で悪性通信を検知することができた．表 2 の FPR において最も精度が高いのは 1/20 の 0.01820%で，TPR では 1/10 の 99.98684%となった．1 週間前から前日分までのデータを学習に使用して判定した結果である表 3 の結果に注目すると，1/1〜1/13 における FPR はほとんどにおいて 0.4%を切っている結果となった．1/14 以降においては，FPR は 0.04%以下となる場合が多く，最も精度が低くても 0.058%程度となった．また，TPR は 1/1〜1/31 の多くの場合で 99.9%以上と高い精度で悪性通信を検知することができた．表 3 の FPR において最も精度が高いのは 1/22 の 0.02248%で，TPR では 1/11 の 99.99211%となった．2 週間前から 1 週間前までのデータを学習に使用して判定した結果である表 4 の結果に注目すると，1/1〜1/13 における FPR はほとんどにおいて 0.4%を切っている結果となった．しかし，1/14 以降においては FPR は大きく乱れる場合が多く，最も精度が高い場合 1/28 の 0.02233%なるが，最も精度が低い場合 1/18 の 96.51315%となった．しかし，TPR は安定して 1/1〜1/31 の多くの場合で 99.9%以上と高い精度で悪性通信を検知することができた．表 4 の FPR において最も精度が高いのは 1/28 の 0.02233%で，TPR では 1/3 の 99.98563%となった．

3 つの実験結果を比較した結果，FPR の平均値が 1 番良かったのは表 3 で，TPR の平均値では表 4 となった．いずれにおいても共通して正常通信生成スクリプトの変更を行った 1/13 において FPR のスコアが急激に低下している．しかし，表 2 と表 3 に関してはその後平均してスクリプト変更前より FPR で高い精度を出している．表 4 においては 1/13 以外にも精度が急激に低下している日付がいくつか確認できた．

6. 本評価システムの考察

表 2，表 3，表 4 のいずれにおいても正常通信の大きな変更があった 1/13 の正常通信の判定データは精度が大きく下がることが判明した．すなわち，通常とは異なる通信が突発的に発生した場合対応できないことがわかった．しかし，表 2，表 3 においては 1/13 を過ぎた後は安定して高い精度で正常通信の判定ができているため，そのような通信を起点として学習データ区分を決めるか，あらかじめ用意したその通信のデータセットを用意することでその通信にも対応できると考えられる．また，今回の評価環境では途中で通信内容を変えてしまったため大きく判定精度が下

ⓒ 2021 Information Processing Society of Japan

表 2 判別器 (One Day)
Table 2 Classifier(One Dataset)

判定データ取得日	FPR	TPR
2021/01/01	0.45758%	99.95712%
2021/01/02	0.37648%	99.96780%
2021/01/03	0.48369%	99.98620%
2021/01/04	0.38068%	99.96489%
2021/01/05	0.53004%	99.88964%
2021/01/06	0.48178%	99.97846%
2021/01/07	0.64667%	99.93992%
2021/01/08	0.91209%	99.95948%
2021/01/09	0.62837%	99.98634%
2021/01/10	1.16060%	99.98685%
2021/01/11	0.38143%	99.97399%
2021/01/12	0.38148%	99.95317%
2021/01/13	68.69951%	99.94491%
2021/01/14	0.47039%	99.35123%
2021/01/15	3.08815%	99.97906%
2021/01/16	0.02227%	99.37897%
2021/01/17	0.02344%	99.82614%
2021/01/18	0.02755%	99.88130%
2021/01/19	0.02594%	99.87129%
2021/01/20	0.01821%	99.76280%
2021/01/21	0.02178%	99.85297%
2021/01/22	0.02444%	99.87826%
2021/01/23	0.02731%	99.91886%
2021/01/24	メンテナンス日	メンテナンス日
2021/01/25	メンテナンス日	メンテナンス日
2021/01/26	メンテナンス日	メンテナンス日
2021/01/27	メンテナンス日	メンテナンス日
2021/01/28	前日データなし	前日データなし
2021/01/29	0.04680%	99.98096%
2021/01/30	0.05176%	99.90526%
2021/01/31	0.05795%	98.37694%
平均値	3.05486%	99.82510%

表 3 判別器 (Seven Day)
Table 3 Classifier(Seven Dataset)

判定データ取得日	FPR	TPR
2021/01/01	0.46952%	99.96594%
2021/01/02	0.42007%	99.98178%
2021/01/03	0.41629%	99.98423%
2021/01/04	0.41637%	99.97794%
2021/01/05	0.48258%	99.96499%
2021/01/06	0.67528%	99.96554%
2021/01/07	0.45624%	99.96060%
2021/01/08	1.01982%	99.95630%
2021/01/09	0.47576%	99.97337%
2021/01/10	0.46424%	99.98869%
2021/01/11	0.44166%	99.99212%
2021/01/12	0.42164%	99.98219%
2021/01/13	73.07012%	99.95278%
2021/01/14	0.46516%	99.96363%
2021/01/15	0.05198%	99.98342%
2021/01/16	0.03951%	99.95927%
2021/01/17	0.03359%	99.91157%
2021/01/18	0.03179%	99.89929%
2021/01/19	0.03075%	99.90007%
2021/01/20	0.02743%	99.83710%
2021/01/21	0.02417%	99.87032%
2021/01/22	0.02249%	99.87058%
2021/01/23	0.02316%	99.90314%
2021/01/24	メンテナンス日	メンテナンス日
2021/01/25	メンテナンス日	メンテナンス日
2021/01/26	メンテナンス日	メンテナンス日
2021/01/27	メンテナンス日	メンテナンス日
2021/01/28	0.02835%	99.84684%
2021/01/29	0.03675%	99.91887%
2021/01/30	0.05834%	99.92506%
2021/01/31	0.05862%	99.97825%
平均値	2.96895%	99.94125%

がってしまったが，増えた通信である 443 番ポートを除いた判定結果は，前日分のデータでは 99.84894%，1 週間分のデータでは 1 週間前が 99.88528%，2 週間前が 99.87520% となった．すなわち，新しい通信の判定精度は下がるが既存の通信に対する判定精度は低下していなかったことが分かった．評価システムにおいて判定精度を向上させるためには，より正常通信生成の再現性を高めて行く必要があると考える．攻撃通信に関してはいずれの学習データセット区分においても高い精度で判別可能であった．これは，ハニーポットの攻撃傾向が短い期間内では大きく変動しないことが原因であると考えられる．しかし，0.01%以下の単位においては数値の変動が大きかった．今回の評価システムではサンドボックスで得られた攻撃通信もひとえに悪性通信とラベル付けを行ったが，内部からの攻撃通信としてラベル付けを行うこと内部でマルウェアの動作が起こった時の分類も可能となる．

7. まとめ

本研究では，機械学習ベース NIDS の学習において公開されているデータセットを使用した際，データ取得組織と設置組織の違いが正常通信や悪性通信の判定率の低下につながる問題点とデータセットではリアルタイム性に欠け，日々変化する通信内容に対応できないという問題点について指摘した．そこで我々は，NIDS の設置組織でリアルタイムに正常通信と悪性通信を取得し，日々継続的に更新を行う機械学習型 NIDS システムの提案を行った．提案ネットワークを模した評価ネットワークを構築し，実際に日々の通信の判定を行った結果，複数条件の学習データセット区分において正常通信においては通信内容の急激な変化には対応できないものの学習データセット区分の工夫などで

表 4　判別器 (Fourteen Day)
Table 4　Classifier(Fourteen Dataset)

判定データ取得日	FPR	TPR
2021/01/01	0.45360%	99.97595%
2021/01/02	0.43196%	99.97945%
2021/01/03	0.43215%	99.98564%
2021/01/04	0.41637%	99.98020%
2021/01/05	0.46675%	99.97222%
2021/01/06	0.74241%	99.97282%
2021/01/07	0.43244%	99.97702%
2021/01/08	1.11319%	99.95253%
2021/01/09	1.53501%	99.97092%
2021/01/10	0.43222%	99.98054%
2021/01/11	0.38545%	99.98082%
2021/01/12	0.40557%	99.97041%
2021/01/13	77.00958%	99.94353%
2021/01/14	1.69864%	99.96852%
2021/01/15	88.80062%	99.98167%
2021/01/16	93.81644%	99.98007%
2021/01/17	93.50900%	99.98081%
2021/01/18	96.51315%	99.97751%
2021/01/19	71.82603%	99.98351%
2021/01/20	95.52928%	99.97020%
2021/01/21	0.04166%	99.97538%
2021/01/22	0.04107%	99.97279%
2021/01/23	0.03889%	99.96536%
2021/01/24	メンテナンス日	メンテナンス日
2021/01/25	メンテナンス日	メンテナンス日
2021/01/26	メンテナンス日	メンテナンス日
2021/01/27	メンテナンス日	メンテナンス日
2021/01/28	0.02233%	99.83696%
2021/01/29	0.02324%	99.79653%
2021/01/30	0.03268%	99.78904%
2021/01/31	0.03356%	99.89165%
平均値	23.19197%	99.95229%

対応が可能であると考えられる．悪性通信ではいずれにおいても 99%以上という高い精度で検知が可能であった．また，ハニーポットへのペネトレーションテスト用ツールなどを使った攻撃などによるツール類の攻撃通信の特徴をとらえることで検知率の向上につながるのではないかと考えられる．

標的型攻撃への対策のために文字列のリストを悪意のあるファイルに追加することで，検知を回避する [14] との報告もある．これらはの学習データ汚染や事前学習モデル汚染などの機械学習ベースシステムのセキュリティ的な問題点から発生する問題である．本提案システムにおいて考えられる具体的な攻撃手法としてアドバーサリアルアタック (Adversarial Attack) と呼ばれるものがある．これは学習データにノイズデータを混ぜることにより本来の判別されるであろう期待されていたデータが全く異なる結果として判別されてしまうことである．提案システムで言えば外部との通信データを利用することで学習データにノイズを混ぜ本来悪性通信であるはずのものを正常通信であると判別してしまうということである．現段階ではこれに対する直接的な対抗策ができていない．これを今後の対応で処理していきたい．また，大企業などの大規模通信への対応のため FPGA による負荷分散を行う必要がある．

謝辞　本研究の一部は，JSPS 科研費 20K11818，19K11968，19H04108 の支援により行った．

参考文献

[1] 情報セキュリティ 10 大脅威 2021, https://www.ipa.go.jp/security/vuln/10threats2021.html (参照 2021-1-29).
[2] Kittikhun Thongkanchorn, Sudsanguan Ngamsuriyaroi, Vasaka Visoottiviseth: Evaluation Studies of Three Intrusion Detection Systems under Various Attacks and Rule Sets, 2013 IEEE International Conference of IEEE Region 10 (TENCON 2013), 2013.
[3] Symantec Endpoint Security, https://jp.broadcom.com/products/cyber-security/endpoint/end-user
[4] TippingPoint Threat Protection System, https://www.trendmicro.com/ja_jp/business/products/network/intrusion-prevention/tipping-point-threat-protection-system.html
[5] Mrudul Dixit, Ankita Moholkar, Sagarika Limaye, Devashree Limaye: Naive Bayes and SVM based NIDS, 2018 3rd International Conference on Inventive Computation Technologies (ICICT), pp.527-532, 2018.
[6] Chine-Hong, Yann-Yean Su, Yu-Chun Lin, Shie-Jue Lee: Machine learning based network intrusion detection, 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), pp.79-83, 2017.
[7] NSL-KDD dataset, https://www.unb.ca/cic/datasets/nsl.html (参照 2020-8-20).
[8] 近松康次郎，平川 豊: ニューラルネットワークを用いた侵入検知システム改良手法の検討, 第 81 回全国大会講演論文集, pp.455-456, 2019.
[9] UCI KDD Archive: KDD Cup 1999 Data, UCI KDD Archive (online), http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (参照 2020-8-19).
[10] 平野 誠, 八槇 博史: 機械学習を用いた攻撃検知に関する学習手法の精度評価, 第 81 回全国大会講演論文集, pp.461-462, 2019.
[11] 多田竜之介, 小林良太郎, 嶋田創, 高倉弘喜: 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.
[12] Universly of New Brunswick: CSE-CIC-IDS2018, https://www.unb.ca/cic/datasets/ids-2018.html (参照 2021-1-22).
[13] 林はるか, 佐藤秀哉, 小林良太郎: 機械学習ベースの NIDS における動的な判別器生成に関する検討と予備評価, 第 91 回 情報処理学会コンピュータセキュリティ研究会, 2020 年 11 月
[14] Adi Ashkenazy and Shahar Zini: Cylance, I Kill You!, https://skylightcyber.com/2019/07/18/cylance-i-kill-you/ (参照 2020-8-19).

ⓒ 2021 Information Processing Society of Japan