組織内ネットワークにおけるハニーポットを備えた動的な機械学習ベースのNIDSの作成と予備的評価

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特定の組織向けの機械学習型 NIDS を作成する。抽出した特徴量で学習を行い採取した通信データで判別をおこなったところ検知率が 99%という高い精度を得ることができた。また、誤検知率は 1%未満という非常に低い結果を得ることができた。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：NIDS, 機械学習, 組織内ネットワーク, honeypot

Creation and preliminary evaluation of dynamic machine learning based NIDS with honeypot in the internal network

Sato Hideya[1,a)]　Hayashi Haruka[1]　Kobayashi Ryotaro[1]

Abstract: we propose NIDS by machine learning to deal cyber-attack that changes day by day. In related researchs, there ara many researches that improve the detection rate by using a publicly available data set labeled with normal and malignant labels. There are some of these problems. First, false detection rate is high because normal communication differs greatly among different organizations. Second, attackers can verify this dataset in advance. Therefore, normal or malignant communication is caputured by using the mirror port or honeypot of the specific organization. Feature values are created by extracting these communication dates. A machine learning based NIDS for a specific organization is created by using these feature value dates. When a NIDS learns dateset caputured on specific organization, the detection rate is as high as 99%, and the false positive rate is less than 1%. This result shows that it is important to capture normal and malignant communication in the installation organization.

Keywords: NIDS, machine learning, Internal network, honeypot

1. 背景

IPA が発表した 2019 年における情報セキュリティ 10 大脅威で組織を狙った標的型攻撃による被害が昨年度ともに 1 位となっている。標的型攻撃とは特定の組織の秘密情報を標的とし様々な攻撃を仕掛けていくことである。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 NIDS はとても重要になる。近年では機械学習型の NIDS が研究され、一般への導入などが徐々に行われようとしている。それら一般に出回る機器の具体的な学習データセットの公開は行われて

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

いないが、その機器を用いて攻撃者が標的型攻撃の事前検証を行うことが可能である。また、多くの研究では KDD Cup 1999 Date といった公開されている通信データセットを使用して研究されており、そのデータセットを用いて同様に攻撃の検証が可能である。そこで上記問題の解決のため、本研究では特定の組織での正常通信とその組織向けの悪性通信を取得し、機械学習に利用する機械学習型の NIDS システムを提案する。

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、評価指標と結果を述べ、5 章で本論文をまとめる。

2. 関連研究

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[6], NSL-KDD Data set[2], Kyoto 2016 Dataset[7] など。NSL-KDD Data set は、KDD Cup 1999 Data set が持ついくつかの問題点を解決したものである。Kyoto 2016 Data set は、KDD Cup 1999 Data set などとは異なり、長期間におよ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。例えば、文献 [1] では、データセットとして KDD Cup 1999 Date を使用しており、文献 [4] では、Kyoto 2016 Dataset を使用している。

しかし、上記の研究はいくつかの問題点を持つ。これらのデータセットは公開されているものであり、攻撃者も中身を確認し、機械学習のモデルを構築することができる。一般に販売されているウイルス対策ソフトにも言える問題であるが、これによって攻撃者が事前に自身の攻撃通信を検証することができる。そのため予兆なく突然攻撃通信がきて組織ネットワークに重大な問題が発生する可能性がある。次に、NIDS の設置組織は国や地域、業種（通信の種類）が一意ではなく攻撃の種類も異なるため、設置組織以外の環境下で作成された機械学習のモデルは設置組織の環境を反映していない可能性がある。また、環境の違いが機械学習のモデルに影響しなかったとしても、攻撃手法や正常通信の内容は時間の経過とともに変化していくため、NIDS を更新しないと検知率や誤検知率が悪化していく可能性がある。既製品の機械学習型 NIDS では誤検知の登録が可能な製品はあるが、新しい機材やソフトを導入するたびに誤検知登録をするのではメンテナンスの負担が重くなる。また、その誤検知率を低くするための研究はいくつか見られるが運用を想定したシステムに関する研究はあまりされていない。実際の運用を考えた時、NIDS においてどのように判別器をメンテナンスし、最新の情報に対応していくかを考えることも重要となる。

3. 提案

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。一般的に出回る機器については機械学習に関する詳細が公開されていないため、本研究では、既存の NIDS として汎用的な判別器を搭載した NIDS を想定する。汎用的な判別器とは、機械学習を使用し、学習には公開された一般的なデータセットあるいは設置組織ではないある一定の条件下で事前に収集したデータセットを使用するものとする。本提案システムでは、まず、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既知の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、それらを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する。そして、そのデータを用いた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。

実験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 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ヤ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既知の悪性通信を取得する。そして、解析ネットワークでは、正常通信と既知の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出を行う。

3.1 正常通信の取得

社内ネットワークでは設置組織における正常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 Router1 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。

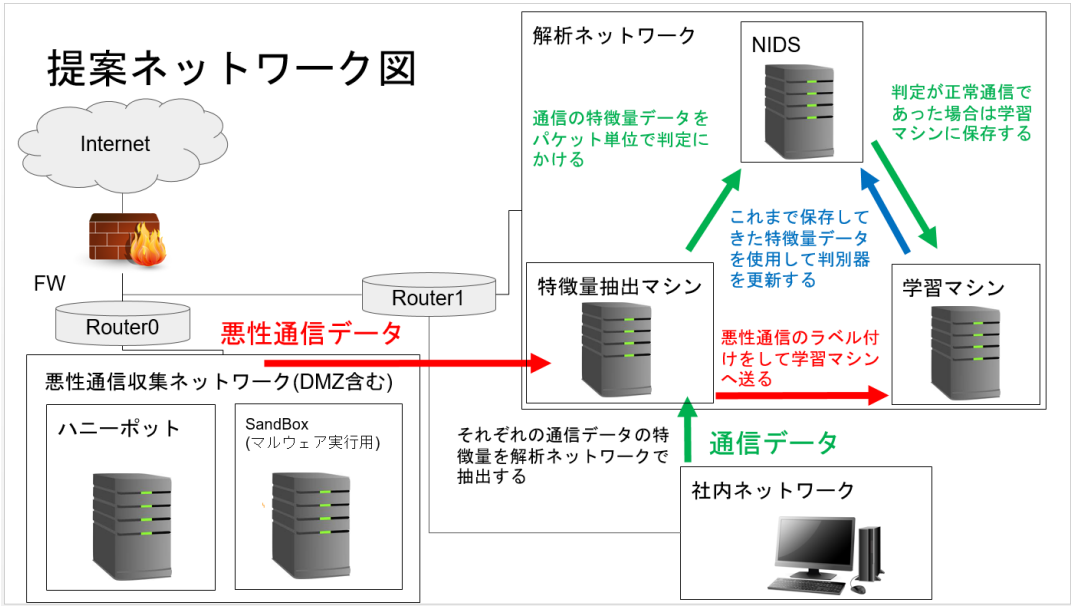

図 1　提案システムのネットワーク図
Fig. 1　Proposed system network

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。

3.2　既知の悪性通信の取得

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。ここでいう悪性通信とは、設置組織の活動または目標達成を外部・内部から妨害または個人情報の流出などその組織にとって不利益に働くことにつながる攻撃通信のことを示す。このネットワークでは、常に外部からの悪性通信の危険性があるため、社内ネットワークなど正常通信が通るネットワークとは別に 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。この時設備費等の問題でその設置が困難な場合には、ルーターを一つにして 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、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 内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のこ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・内部通信はいずれも Router 0 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。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・内部ネットワークから隔離した SandBox 上に設置し実行する。SandBox の内部・外部通信に関しても、ハニーポットと同様、特徴量抽出マシンに送信される。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。

3.3　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、セッション単位ではなく 1 パケットごとに特徴量を抽出する。特徴量とは、パケットの特徴が数値化されたものを示し、例えば、送受信バイト数などがそれに相当する。

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 1 パケットごとに NIDS へと送信し、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を悪性とラベル付けして 1 パケットごとに学習マシンへと送信する。なお、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、通信データの特徴量は学習マシンに送信しない。

3.4　悪性通信の検知

NIDS は、特徴量抽出マシンから送られてきた通信データの特徴量を入力として、1 パケット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。なお、悪性通信収集ネットワークから得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。

3.5 機械学習による判別器の生成

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、判別器を生成する。この判別器は 1 パケット単位で悪性か正常のどちらかの判定を行う。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。

3.6 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分類やメンテナンスの負担軽減のため定期的な判別機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、行うものする。時系列毎の更新の様子を時系列で図 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。取得したパケットは 24 時間おきに判別器再生成のため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時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おこなう。

4. 評価

4.1 評価環境

4.1.1 正常通信の取得

提案システムでは図 1 の社内ネットワークにあたる組織内の正常通信を通信しているネットワークがあるが、本評価環境では著者の使用している PC の通信データを正常通信データとして収集した。使用している PC のスペックは CPU intel core i5-7200U、メモリは 8GB、OS は windows10 home である。通信データを採取するのに使用しているソフトウェアは wireshark ver 3.2.5 (v3.2.5-0-ged20ddea8138) で pcap 形式ファイルとして取得している。wireshark とはネットワーク上で通信されているデータをダンプして解析を行うためのソフトウェアの一種である。提案システム上では正常通信は常時解析ネットワーク上にその通信データを送信し特徴量の抽出と判別に掛けるが、本評価環境では一日分の正常通信のデータがすべて得られた後特徴量抽出・学習マシンへと scp コマンドを使用して手動で送信する。

4.1.2 既知の悪性通信の取得

提案システムでは悪性通信収集ネットワークを FW の下に専用のルーターまたは VLAN などによる専用の区切られたネットワークを組織内に設置する。しかし、本評価環境ではハニーポットを設置する環境として正常通信と採取しているネットワークとは全く別のネットワークを準備した。そこへハニーポット専用の PC として Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz、メモリ 8GB、OS は ubuntu18.04 を設置した。使用したハニーポットは honeytrap である。honeytrap の設定で 11211, 23, 289, 8545, 5555, 27016, 5037, 3890, 9200, 8022, 8080 のポートを解放する。その後 ufw コマンドでそれらポートと設定変更用の ssh 使用ポート以外の通信は遮断する。そして通信データの取得として tcpdump version 4.9.3 を使用して pcap 形式ファイルで取得している。tcpdump とはネットワーク調査ツールの一種である。それら取得した既知の悪性通信データを cron を使って一日おきに特徴量抽出・学習マシンへと送信する。cron は主に UNIX 系の OS で使われている指定した時間に指定したプログラムを実行してくれるプログラムである。また、今回の評価環境では SandBox の設置は行わなかった。

4.1.3 特徴量の抽出および学習

本評価環境では、特徴量抽出・学習マシンへと送られてきたデータの特徴量を cron で 0 時に bro-IDS (version 2.5.3) のプラグインを使用して抽出する。bro-IDS とは wireshark や tcpdump 同様ネットワーク調査ツールの一種である。このプラグインを使用すると pcap ファイルを入力とし csv ファイルとして特徴量を抽出し出力する。正常通信と既知の悪性通信のそれぞれの一日分のデータすべての特徴量 csv データファイルに変換し学習マシンへと保存する。評価環境の解析ネットワークにはこれまで収集したデータをすべて保存するようの NAS が設置されている。そこから学習に使用するデータを取り出している。本来は提案手法のように自動的に学習も開始するために cron へ学習プログラムを登録すべきである。しかし、今回の環境では NAS からの特徴量データの取り出しと学習プログラムの実行は手元の環境でオフラインで行っている。使用している学習アルゴリズムは教師あり学習のランダムフォレストである。ランダムフォレストとは多数の決定木を使用することで予測精度を高めるアルゴリズムである。その他サポートベクター、決定木、SVM、 ロジスティックス回帰等のアルゴリズムについても測定を行ったが、本研究ではもっとも予測精度が高かったランダムフォレストの結果のみ示す。使用する特徴量は bro-IDS プラグインが生成する 47 種類のうち orig_pt、resp_pt、resp_pt(res_pt) などを含む 26 種類である。現在わかっている KDD Cup 1999 Date と bro-IDS プラグインが生成する特徴量で共通しているものを表 1 に示す。

4.2 評価指標

混合行列を表 2 に示す。また、表 2 から誤検知率である FPR (False Positive Rate) を算出した。各指標の定義を以下に示す。

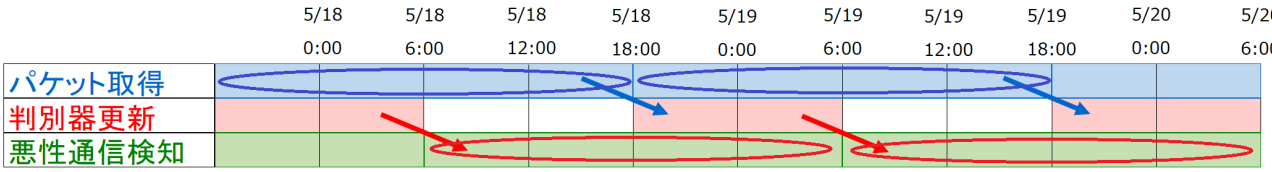

図 2 判別器の更新時間
Fig. 2 Discriminator update time

表 1 共通の特徴量

属性名	概要
duration	セッションの長さ
src_bytes	送信バイト数
dst_bytes	受信バイト数

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

表 2 混合行列

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

5. 結果

表 3 に、提案 NIDS で社内ネットワークから得られた通信を判別した結果を示す。表において、学習データ取得日、判別データ取得日とは、学習、判別にそれぞれ使用した特徴量データの元となった正常・悪性の通信データを取得した日付である。今回、評価環境で取得したデータに関しては、休日に相当する土曜日と日曜日に正常通信が行われないと想定しているため、一部連続した日付となっていない。

表 3 提案 NIDS で社内ネットワークから得られた通信を判別した結果

学習データ取得日	判別データ取得日	FPR
2020-5-15	2020-5-18	0.0692%
2020-5-18	2020-5-19	0.0374%
2020-5-19	2020-5-20	0.0710%
2020-5-20	2020-5-21	0.0441%
2020-5-21	2020-5-22	0.0073%
2020-5-22	2020-5-25	0.0737%
2020-5-25	2020-5-26	0.0092%

6. 考察

表 3 より、提案 NIDS は社内ネットワークに対して誤検知率を 0.1%未満に抑えることがわかる。また、1 日ごとの誤検知率の変動は非常に少ないこともわかる。このことから、提案 NIDS は、組織内で収集されるデータを用いることによって、自組織に適合した機械学習モデルを構築できていることがわかる。

本来、提案 NIDS は社内ネットワークから得られた通信に対してのみ判別を行う。そのため、表 3 では社内ネットワークの通信に対する判別結果しか示されていないが、今回の評価において、社内ネットワークとして収集した通信には悪性通信が含まれていないため、機械学習モデルが悪性通信に対応できているかどうかは不明である。そこで本節では、提案 NIDS が悪性通信に対してどのような判別を行うかどうかを確認するために、提案 NIDS を用いて悪性通信ネットワークから得られた通信に対して判別を行ったとしたら、どのような結果が得られるかを測定した。なお、この測定は社内ネットワークに対する攻撃が成功し、ハニーポットで得られた悪性通信と同様の通信が社内ネットワークに発生してしまった場合、それを検出できるかどうかを検証するという意味を持つ。その測定結果を表 4 に示す。なお、表 4 において検知率である TPR (True Positive Rate) は表 2 の混合表から算出した。 各指標の定義を以下に示す。

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

表 4 より検知率は 99.99%以上となっており悪性通信をほぼ検知できることがわかる。この結果より、組織を狙った攻撃を事前にハニーポットでとらえることができれば、それをほぼ検知することができる。

表 4 提案 NIDS で悪性通信収集ネットワークから得られた通信を判別した結果

学習データ取得日	判別データ取得日	TPR
2020-5-15	2020-5-18	99.9995%
2020-5-18	2020-5-19	99.9985%
2020-5-19	2020-5-20	99.9998%
2020-5-20	2020-5-21	100.0000%
2020-5-21	2020-5-22	100.0000%
2020-5-22	2020-5-25	99.9978%
2020-5-25	2020-5-26	99.9999%

提案 NIDS は学習データを取得し他翌日を判別データの取得日としている。しかし、学習データと判別データの取得日が離れていた場合、誤検知率や検知率に影響を及ぼす

可能性がある。そこで、学習データと判別データの取得日に 10 日の差があった場合の評価を行い、その結果を表 5 に示す。これにより、例えば、ある標的型攻撃をハニーポットにより検出し、その攻撃が数日後に正しく検出されるかを検証することができる。表 5 より、誤検知率は一部 0.1% を超える日付が存在するが、それ以外は 0.1%未満に抑えることができている。また、日付によっては検知率が低下する日も存在するが、ほとんどの日付において 99.9%以上という高い数値を達成している。この結果より、例えある種類の攻撃が一定期間来なかったとしても、それをほぼ検知することが可能であり、誤検知率への影響は小さいと言える。

表 5 1 日前ではなく 10 日前の学習データで判別した結果

学習データ取得日	判別データ取得日	FPR	TPR
2020-5-15	2020-5-26	0.0551%	99.9989%
2020-5-18	2020-5-28	0.0118%	99.9994%
2020-5-19	2020-5-29	0.0917%	99.9989%
2020-5-20	2020-6-1	0.0664%	100.0000%
2020-5-21	2020-6-2	0.1246%	99.9943%
2020-5-22	2020-6-3	0.0196%	99.9783%
2020-5-25	2020-6-4	0.0130%	99.9956%

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誤検知率に大きくかかわってくる問題を提示した。それら問題を解決すべく、正常通信としたデータとハニーポットを設置しその通信データを悪性通信とした 2 種類のデータを取得した。それらデータを使い NIDS の自動更新システムの提案を行った。提案 NIDS を評価した結果、誤検知率は 1%未満となった。

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、検出を回避する [5] 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] 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.
[2] NSL-KDD dataset, https://www.unb.ca/cic/datasets/nsl.html (参照 2020-8-20).
[3] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani: A Detailed Analysis of the KDD CUP 99 Data Set, Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, pp.1-6, 2009.
[4] 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.
[5] Adi Ashkenazy and Shahar Zini: Cylance, I Kill You!, https://skylightcyber.com/2019/07/18/cylance-i-kill-you/ (参照 2020-8-19).
[6] 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).
[7] 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.