情報処理学会論文誌 Vol.62 No.12 1915–1925 (Dec. 2021)

機械学習を用いたNIDSにおける未知の攻撃検知手法の提案

本丸 真人[1] 寺田 真敏[1,a)]

受付日 2021年3月8日, 採録日 2021年9月9日

概要：ネットワークでの攻撃検知は既知の攻撃だけではなく未知の攻撃を検知することも必要である．本論文では，特徴量抽出処理にオートエンコーダ，学習および予測処理に深層強化学習を適用することを特徴とするNIDSにおける未知の攻撃を検知し分別する手法について提案する．評価にはデータセットとしてNSL-KDDを使用し，オートエンコーダとしてDAE，深層強化学習としてDDQNを使用した．提案手法を用いて分別した後，全体のマイクロ平均の正解率，攻撃カテゴリごとの適合率，再現率などを用いて評価を行い，既存手法と比較した．提案手法は既存手法と比較して，マイクロ平均の正解率が高く全体として予測性能が高いという結果が得られた．

キーワード：NIDS，機械学習，オートエンコーダ，強化学習

Proposing a Method for Detecting Unknown Attacks in NIDS Using Machine Learning

Masato Hommaru[1] Masato Terada[1,a)]

Received: March 8, 2021, Accepted: September 9, 2021

Abstract: To detect attacks in networks, it is necessary to detect not only known attacks but also unknown attacks. In this paper, we propose a method for detecting unknown attacks in NIDS using Autoencoder and deep reinforcement learning. For the evaluations, we used NSL-KDD as the dataset, DAE as Autoencoder, and DDQN as deep reinforcement learning. After categorizing the data using the proposed method, we evaluated the results using the micro-average Accuracy, Precision, and Recall for each category, and compared them with the previous methods. The proposed method has higher micro-mean accuracy and better overall prediction performance than the previous methods.

Keywords: NIDS, machine learning, autoencoder, reinforcement learning

1. はじめに

警察庁の資料 [1] によると警察庁の検知システムでの不正な通信の件数は，平成27年で729件（/日・IPアドレス），令和元年で4,192件（/日・IPアドレス）であり，インターネット経由での攻撃は増加傾向にある．セキュリティ対策を考えるうえでネットワーク通信を対象とした攻撃検知は重要な要素で，その攻撃検知の1つとしてネットワーク型侵入検知システム（以下，NIDS）があり，また攻撃検知を改善するためにNIDSを対象とした機械学習の適用についても研究が進められている [2], [3]．機械学習の適用に関する研究の多くは既知の攻撃を対象としたものであるが，インターネット経由での攻撃の中には既知の攻撃だけでなく，未知の攻撃も含まれる．さらに，未知の攻撃は，従来の既知の攻撃を分別するための手法で判別することは難しく，別途，未知の攻撃を検知するためのアプローチが必要となる．これまでの研究では，NIDSにおける既知の攻撃を検知対象としているか，未知の攻撃を検知対象とした場合にも，正常通信か攻撃通信かの判定にとどまり（以下，未知の攻撃検知），未知の攻撃を検知し，どのような攻撃なのかを分別するところまで言及していない．

本論文では，オートエンコーダが過学習を防ぐために利用されることに着目し，この機能を未知の攻撃を検知し分

[1] 東京電機大学
Tokyo Denki University, Adachi, Tokyo 120–8551, Japan
[a)] masato.terada@mail.dendai.ac.jp

© 2021 Information Processing Society of Japan

別する（以下，未知の攻撃分別検知）ための特徴量抽出処理に適用する．さらに学習および予測処理では深層強化学習を適用したNIDSでの未知の攻撃を検知し分別する手法を提案するとともに，データセットNSL-KDDを使用した分別評価を通して提案方式の有効性を示す．

本論文の構成は，2章でNIDSを対象とした機械学習の適用についての既存研究をまとめ，3章で本研究の提案手法の詳細を，4章で提案方式の評価環境と結果について述べ，5章で考察，6章でまとめと今後の課題を示す．

2. 関連研究

本章では，NIDSにおける未知の攻撃検知と深層強化学習の適用に関する研究について述べる．

2.1 未知の攻撃検知

Pérezらはネットワーク型侵入検知において十分なラベルづけされたデータが利用できないことから従来の機械学習での検知が困難である未知の攻撃の増加を1つの課題としてあげ，課題の解決のためにゼロショット学習を適用した[4]．ゼロショット学習は機械学習の1つで学習していないクラスを分類することを目的としている．Pérezらは既存のデータセットをどのように既知の攻撃と未知の攻撃に分けるのかの定義を行い，決定木を用いることで学習を行った．Revathiらは，NSL-KDDデータセットを使って，複数の機械学習の分別精度を評価し，このデータセットが侵入検知方法を比較するのに有効であることを示した[5]．Pervezらは，サポートベクターマシン（SVM）を用いた検知アルゴリズムを提案し，NSL-KDDデータセットで既知の攻撃を検知する精度を評価し，その有効性を示している．一方，未知の攻撃については適用が難しいことを示した[6]．

Zhaoらは，未知の攻撃の分類検知に転移学習を適用し，NSL-KDDデータセットを使って分類検知精度を評価し，その有効性を示している[7]．Yanらは，既知の攻撃の特徴量抽出にスパースオートエンコーダを適用し，サポートベクターマシン（SVM）を用いた攻撃検知を提案している[8]．Alomらは，未知の攻撃検知に特徴量抽出と次元削減のためにオートエンコーダと制限付きボルツマンマシン（RBM）を用いて教師なし深層学習を適用し，KDD-99データセットを使った検知精度を評価し，その有効性を示している[9]．分野は異なるが，Xiaoらは，ブロックリストで検知できないフィッシングサイトのURL判定にURL内部構造を加味した畳み込みニューラルネットワークを適用した[10]．また，Sarkerらは，機械学習を想定した侵入検知ツリー型のセキュリティモデルというアプローチを提案している[11]．

2.2 深層強化学習のNIDSへの適用

NIDSでの攻撃検知の研究にはネットワーク通信のデータセットが用いられる．このデータセットは教師あり学習には適しているが，強化学習には適していない．Martinらは古典的で動的な環境での相互作用に基づく深層強化学習のパラダイムを概念的に変更することによって，データセットを深層強化学習で学習可能とした．深層強化学習のモデルとして4つのモデルで評価を行い，Double Deep Q-Network（DDQN）が最も予測性能が高いことを示した[12]．Hsuらは，深層強化学習モデルのDQN（Deep Q Network）をベースとした攻撃検知システムを提案し，NSL-KDDとUNSW-NB15データセットを使って評価し，その有効性を示している[13]．Benaddiらはセキュア無線LANやIoT向けに[14]，Sethiらはクラウド向けに[15]，DQNをベースとした攻撃検知システムを提案している．

2.3 解決したい課題

関連研究では，NIDSにおける既知の攻撃を検知対象としているか，未知の攻撃を検知対象とした場合にも，正常通信か攻撃通信かを判定する未知の攻撃検知にとどまり，未知の攻撃を検知し，どのような攻撃なのかを分別するところまで言及していない．提案手法は，NIDSにおける未知の攻撃を検知し分別する未知の攻撃分別検知を目的とし，ネットワーク通信の特徴量抽出処理にオートエンコーダ，学習および予測処理に深層強化学習を適用することを特徴とする．

3. 提案手法

3.1 提案手法の概要

深層強化学習を用いたNIDSの未知の攻撃を検知対象とする手法において，特徴量の抽出処理と特徴量の学習および予測処理に異なる機械学習モデルを組み合わせたアプローチは検討されていない．提案手法では，特徴量抽出処理にオートエンコーダ，変換した特徴量の学習および予測処理に深層強化学習を用いる．なお，提案手法の評価では4章で後述する学習していないクラスを所属する攻撃カテゴリへと分別することをNIDSでの未知の攻撃分別検知と定義する．

提案手法は，ステップ1：前処理，ステップ2：特徴量抽出，ステップ3：学習，ステップ4：予測の4段階のステップから構成する．ステップ2以降の流れを**図1**に示す．前処理では，機械学習を行うために入力データを変換

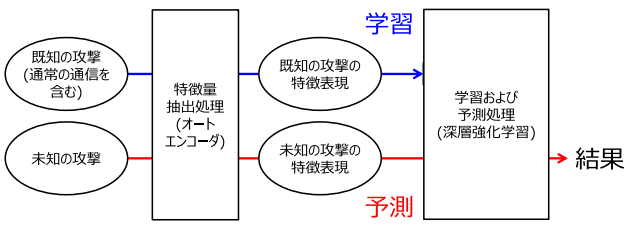


図1 提案手法の概要
Fig. 1 Overview of proposed system.

© 2021 Information Processing Society of Japan

する．特徴量抽出ではオートエンコーダを用いて既知ならびに未知の攻撃データからそれぞれの特徴表現を抽出する．学習では，既知の攻撃の特徴表現を用いて学習する．予測では，学習した既知の攻撃の特徴表現と，入力された攻撃の特徴表現とを用いて未知の攻撃を予測する．

3.2 前処理

(1) カテゴリ特徴量の処理

カテゴリ特徴量とは特徴量が数値ではなく，選択値や文字列で表されるものである．機械学習では入力するデータが数値となっている必要があり，そのままではカテゴリ特徴量を扱うことができないことから，カテゴリ特徴量を数値へ変換する．カテゴリ特徴量の数値への変換には特徴量をベクトルに変換する Feature Hashing や One-Hot エンコーディングなどいくつか手法が存在する．このため，変換にあたっては，サンプリング，標準化と正規化，カテゴリ変数の扱い方，オートエンコーダと深層強化学習のエポック数などに関する予備実験を通して，One-Hot エンコーディングを用いることとした．One-Hot エンコーディングでは，ある特徴量に含まれるデータを新たな特徴量として，新しく生成した特徴量が元の特徴量に含まれる場合は 1 を設定し，含まれない場合は 0 を設定する．

(2) 標準化

機械学習では，入力するデータの特徴量ごとの数値の範囲の差が大きいと予測性能が低下する場合があることから，特徴量ごとの数値の範囲の差を小さくするために標準化を行う．

(3) サンプリング

機械学習では，クラスの数に偏りがあると予測性能が低下する場合があることから，クラスの偏りを軽減するためにサンプリングする必要がある．提案手法では，多数のデータに数を合わせるよう少数データを補完するオーバサンプリングを行った後，少数のデータに数を合わせるよう多数データから抽出するアンダサンプリングを行う．

3.3 特徴量抽出

特徴量抽出ではニューラルネットワークの 1 つであるオートエンコーダを用いて圧縮された特徴表現を取得する．オートエンコーダは他のニューラルネットワークと同様に入力層，隠れ層，出力層で構成され（**図 2**），入力層と出力層は同じ次元数であるが隠れ層は入力層や出力層よりも次元が小さいという特徴を持つ．隠れ層が入力層や出力層よりも小さい次元であり，入力データを 1 度圧縮し．重要な特徴だけを残した後，元の次元に復元することから次元削減を行うことができる．

提案手法では，既知の攻撃データを用いて学習を行い，学習済みのオートエンコーダに既知ならびに未知の攻撃それぞれのデータを入力することで隠れ層から圧縮された特徴表現を取得する．

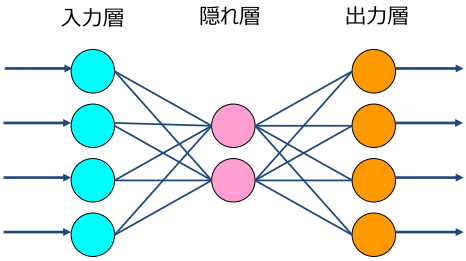


図 2 オートエンコーダの構造
Fig. 2 Structure of Autoencoder.

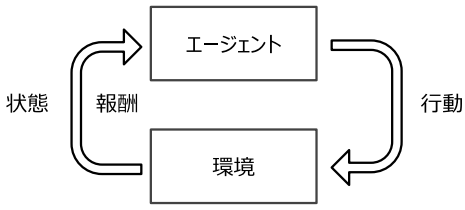


図 3 強化学習の構造
Fig. 3 Structure of reinforcement learning.

3.4 学習および予測

学習および予測で用いる深層強化学習とは関数に深層学習を用いた強化学習であり，強化学習は改善したい指標値が存在するが改善する手法が不明な場合に有効な機械学習の 1 つである．**図 3** に強化学習の簡略化した構造を示す．環境は改善したい指標値を含む空間のことで提案手法におけるネットワーク型侵入検知システムにあたる．エージェントは強化学習において学習を行う主体である．エージェントが環境の状態からその良し悪しを報酬として判断し環境に対して行動を行うということを繰り返すことで学習を進めていく．強化学習では，ある行動を起こしたときの状態と報酬をテーブルとして保持するが深層強化学習ではこのテーブルをニューラルネットワークで代替する．提案手法では，既知の攻撃データを用いて深層強化学習モデルの学習を行い，深層強化学習のニューラルネットワーク部分を用いることにより予測を行う．

4. 未知の攻撃分別検知の評価

提案手法の有効性を示すために，データセット NSL-KDD [16], [17] を用いて，次の 2 つの視点から評価する．

- 未知の攻撃分別検知の予測性能（以降，予測性能）
- 特徴量抽出処理が予測性能に与える影響

4.1 評価条件

(1) データセット

評価に用いる NSL-KDD は 126,620 のデータを持つ 2009 年に作成されたデータセットであり，機械学習を用いた NIDS の研究で広く用いられている．NSL-KDD では，4

© 2021 Information Processing Society of Japan

表 1 NSL-KDD で提供される特徴量
Table 1 NSL-KDD features list.

区分	概要
Basic	Protocol_type, Service, Flag など 10 項目
Contents	Num_failed_logins, Logged_in など 12 項目
Time based traffic	Srv_count, Serror_rate など 9 項目
Host based traffic	Dst_host_count など 10 項目

表 2 NSL-KDD のカテゴリ
Table 2 Attack classification in NSL-KDD.

カテゴリ	概要
Normal	攻撃ではない正常な通信
DoS	正当なユーザのアクセスの禁止，リクエストによる過負荷などによりリソースが不足することで正当な要求に応答できなくする通信
Probe	攻撃を行うためにマシンやネットワーク装置をスキャンして脆弱性を発見するための通信
R2L	不正に管理者権限を取得するための通信
U2R	アカウントを持たずにリモートシステムへのアクセスを得ようとする通信

種類 41 個の特徴量（**表 1**）と正常通信を含む攻撃種類を示す 1 個のラベルが付いている．攻撃種類を示すラベルでは正常通信と 39 の攻撃クラスが定義されており，それぞれのクラスが攻撃種類の大枠である 5 つのカテゴリに分類されている（**表 2**）．

(2) 未知の攻撃分別検知

提案手法を用いた評価にあたっては，評価用データセット NSL-KDD の正常通信を除く 4 つの攻撃カテゴリの中からクラス 2 つずつを未知の攻撃，そのほかを既知の攻撃と区分し（**表 3**），未知の攻撃として区分した学習していないクラスを入力データとし，所属する攻撃カテゴリへと分別することを NIDS での未知の攻撃分別検知として評価することとした．

4.2 評価手順

4.2.1 前処理

データセットの前処理は次のとおりである．

(1) 既知ならびに未知の攻撃の分割

4 つの攻撃カテゴリからクラス 2 つずつを未知の攻撃として抽出し，そのほかを既知の攻撃として分割する（表 3）．

(2) カテゴリ特徴量の処理

既知ならびに未知の攻撃の入力データのうち，NSL-KDD のカテゴリ特徴量である Protocol_type，Service，Flag の 3 つに対して One-Hot エンコーディングを行う．これら 3 つ以外の特徴量 38 項目に関しては数値であるため，この処理は行わない．

(3) ラベルの変換

既知の攻撃に関して正解ラベルをクラスから攻撃カテゴリへ変換する．未知の攻撃に関しては特徴表現の分布を比較するため変換しない．

(4) サンプリング

既知ならびに未知の攻撃の入力データに対して，少ないデータを単純にコピーするのではなく近傍にあるデータを用いて増加させる SMOTE（Synthetic Minority Oversampling TEchnique）でオーバサンプリングをした後に，クラスが最近傍ではない多数のクラスのサンプルを削除する ENN（Edited Nearest Neighbor）でアンダサンプリングする．

表 3 NSL-KDD のクラス
Table 3 Class of NSL-KDD.

クラス	数	カテゴリ	既知/未知
neptune	45,871	DoS	既知
smurf	3,311		
back	1,315		
apache2	737		
processtable	685		
mailbomb	293		
pod	242		
worm	2		
udpstorm	2		
teardrop	**904**		**未知**
land	**25**		
normal	77,053	Normal	既知
satan	4,368	Probe	既知
portsweep	3,088		
mscan	996		
saint	319		
ipsweep	**3,740**		**未知**
nmap	**1,566**		
warezmaster	964	R2L	既知
warezclient	890		
snmpguess	331		
snmpgetattack	178		
httptunnel	133		
multihop	25		
named	17		
sendmail	14		
ftp_write	11		
xlock	9		
phf	6		
xsnoop	4		
spy	2		
guess_passwd	**1,284**		**未知**
imap	**12**		
buffer_overflow	50	U2R	既知
ps	15		
xterm	13		
loadmodule	11		
sqlattack	2		
rootkit	**23**		**未知**
perl	**5**		

4.2.2 特徴量抽出

オートエンコーダにはノイズ除去オートエンコーダ（DAE: Denoising Autoencoder）を用いる．DAE はオートエンコーダの入力データにノイズを加えたものであり，ノイズの入ったデータからノイズの入っていないデータを復元するように学習する．DAE のパラメータ一覧（**表 4**）のうち，活性化関数としてはすべての層で relu を使用した．ノイズ率は入力データに対して加えるノイズの割合で，ノイズは平均 0，標準偏差 1 の正規分布で発生させた．

4.2.3 学習

深層強化学習には DDQN（Double Deep Q Network）を

© 2021 Information Processing Society of Japan

表 4 DAE のパラメータ一覧
Table 4 Parameters of DAE.

パラメータ	値
エポック(epoch)	300
活性化関数(activation)	relu
最適スコア関数(optimizer)	adadelta
損失関数(loss)	mse
ノイズ率	0.1

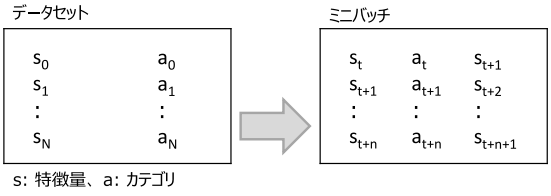


図 4 データセットのミニバッチへの変換
Fig. 4 Transformation from dataset to minibatch.

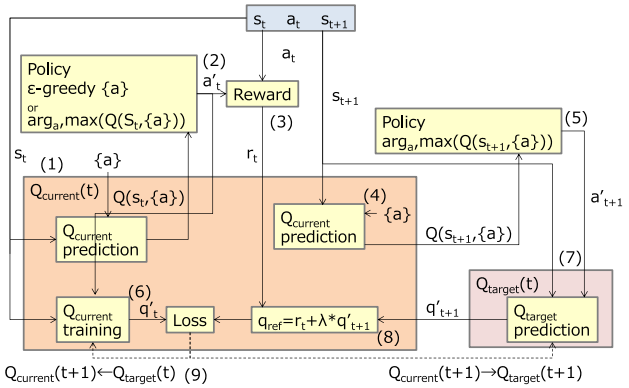


図 5 DDQN の学習の流れ
Fig. 5 Learning diagram of DDQN.

用いる．DDQN は深層強化学習モデルの DQN（Deep Q Network）をベースとしており，DQN は強化学習の Q 学習をベースとしている．DDQN は最適価値関数 Q(s, a) を学習し，そこから適当な方策に従い行動する価値ベースの強化学習である．NSL-KDD の形式では強化学習を行う際，データとして扱うことが難しいため**図 4** のようにデータを変換する．s は状態のことであり特徴量を，a は行動のことでありカテゴリを対応させる．状態，行動，次の状態を 1 つの組とすることでミニバッチを作成する．この際，ミニバッチはデータセットからランダムに抽出する．

次に，作成したミニバッチを用いて DDQN での学習を行う（**図 5**）．手順は次のとおりである．

(1) 最適価値関数 $Q_{current}$ が状態 s_t とすべての行動 {a} を受け取る．
(2) Policy で ε-greedy もしくは項番 (1) の関数で価値が最大となった予測の行動 a'_t を出力する．
(3) Reward で a_t と項番 (2) で出力した a'_t を比較し，同様であれば報酬 r_t を 1，異なれば報酬 r_t を 0 とする．
(4) 最適価値関数 $Q_{current}$ が状態 s_{t+1} とすべての行動 {a} を受け取る．
(5) Policy で項番 (4) の関数で価値が最大となった予測の行動 a'_{t+1} を出力する．
(6) $Q_{current}$ が状態 s_t と項番 (2) で出力した a'_t を受け取り，予測の価値 q'_t を出力する．
(7) Q_{target} が状態 s_{t+1} と項番 (5) で出力した a'_{t+1} を受け取り，予測の価値 q'_{t+1} を出力する．
(8) 項番 (3) の r_t と (7) の q'_{t+1} から実際の価値 q_{ref} を出力する．
(9) 項番 (6) の q'_t と項番 (8) の q_{ref} を用いて $Q_{current}$ と Q_{target} を更新する．

DDQN のパラメータ一覧（**表 5**）のうち，活性化関数としては relu を使用した．

表 5 DDQN のパラメータ一覧
Table 5 Parameters of DDQN.

パラメータ	値
試行回数	10
割引係数	0.01
学習率	0.0001
活性化関数(activation)	relu
最適スコア関数(optimizer)	Adam
損失関数(loss)	huburloss

4.2.4 予測

予測では表 3 で分別した未知の攻撃データを入力とし，DAE を用いて特徴を抽出した後，DDQN を用いて攻撃を分別し判定する．

4.3 評価結果

表 3 に示す未知の攻撃として区分した学習していないクラスを所属する攻撃カテゴリへと分別することを NIDS での未知の攻撃分別検知の予測性能として評価した．この予測性能の評価にあたっては，評価指標として正解率，適合率，再現率，F 値，マイクロ平均（正解率）を用いた．提案手法で扱うデータは多クラス分類であるため，各カテゴリの評価は，そのカテゴリであれば真（Positive），そうでなければ偽（Negative）とした際の真陽性（TP: True Positive），真陰性（TN: True Negative），偽陽性（FP: False Positive），偽陰性（FN: False Negative）を用いる．また，全体の評価はマイクロ平均で行う．

- 正解率
  すべての予測のうち，正解した予測の割合

$$\text{正解率}\ (Accuracy) = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

- 適合率
  真と予測したもののうち，実際に真であるものの割合

$$\text{適合率}\ (Precision) = \frac{TP}{TP + FP} \tag{2}$$

© 2021 Information Processing Society of Japan

- 再現率
  実際に真であるもののうち，正しく真と予測できたものの割合

$$\text{再現率（}Recall\text{）} = \frac{TP}{TP + FN} \tag{3}$$

- F 値
  適合率と再現率はトレードオフの関係にあるため，2 つの調和平均をとった指標（範囲は 0〜1，1 に近いほど予測性能が高い）

$$\text{F 値} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

- マイクロ平均（正解率）
  すべての予測のうち，正解した予測の割合
  なお, FP と FN が同一になることから, マイクロ平均では正解率, 適合率, 再現率, F 値がすべて同じ値となる．

$$\text{マイクロ平均（正解率）} = \frac{TP}{TP + FP} \tag{5}$$

4.3.1 提案手法の予測性能

提案手法の評価結果として，予測結果を**図 6**，カテゴリごとの予測分別数を**図 7** に示す．R2L を除く 3 つの攻撃カテゴリでは再現率の値が高く検知率が高かった．その一方，適合率の値が低く誤検知率も高かった．なお，R2L の F 値が空白となっているのは計算不可能なためである．

4.3.2 既存手法の予測性能

既存手法との比較のために，特徴量抽出処理と学習および予測処理に DDQN，ニューラルネットワーク，サポートベクターマシン（SVM），ランダムフォレストを用いた．提案手法との比較結果を**表 6**，既存手法を用いた予測結果を**図 8**，カテゴリごとの予測分別数を**図 9** に示す．

提案手法との比較結果である表 6 から，既存手法の中で最も予測性能が高い SVM が 59.7%であるのに対し提案手法は 65.7%であり，マイクロ平均の評価では提案手法が最も高い予測性能となった．予測結果である図 6 と図 8 から，Probe，U2R の適合率を比較すると，SVM とランダムフォレストに比べて提案手法は予測性能が低いが，DoS，U2R の再現率を比較すると，SVM とランダムフォレストに比べて提案手法は予測性能が高い．提案手法は既存手法に比べて再現率の点から検知率は高いが，適合率の点から誤検知率も高い．その一方で，カテゴリごとの予測分別数である図 7 と図 9 から，既存手法は提案手法に比べて未知の攻撃を正常通信と予測してしまう可能性が高いという結果が得られた．

4.4 特徴量抽出処理が予測性能に与える影響

提案手法は，特徴量抽出処理にオートエンコーダ，学習

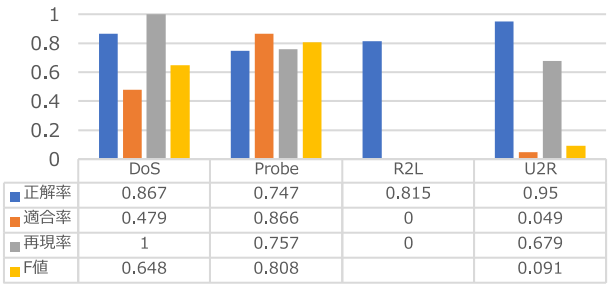


図 6 提案手法による予測
Fig. 6 Prediction using the proposed method.

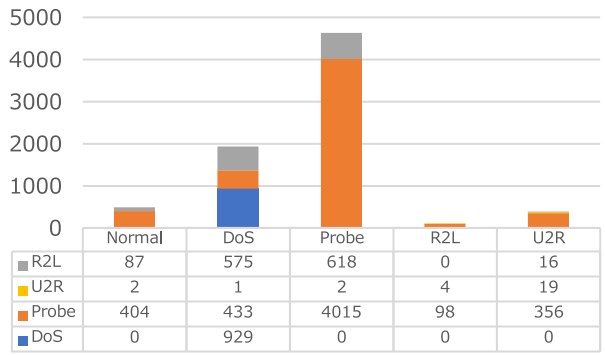


図 7 提案手法による予測分別数
Fig. 7 Prediction classification using the proposed method.

表 6 提案手法と既存手法の比較
Table 6 Comparison between the proposed and the existing method.

検知分別手法	マイクロ平均
提案手法(DAE+DDQN)	0.657
DDQN のみ	0.537
ニューラルネットワーク	0.551
サポートベクターマシン(SVM)	0.597
ランダムフォレスト	0.572

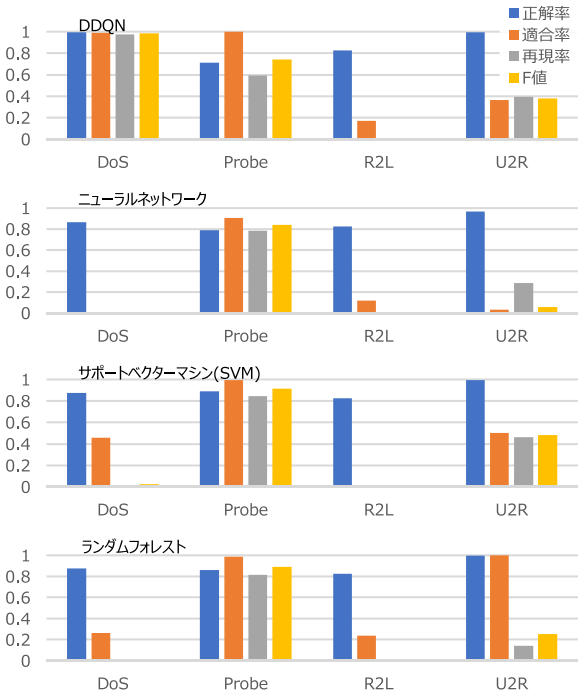


図 8 既存手法による予測
Fig. 8 Prediction using the existing method.

© 2021 Information Processing Society of Japan

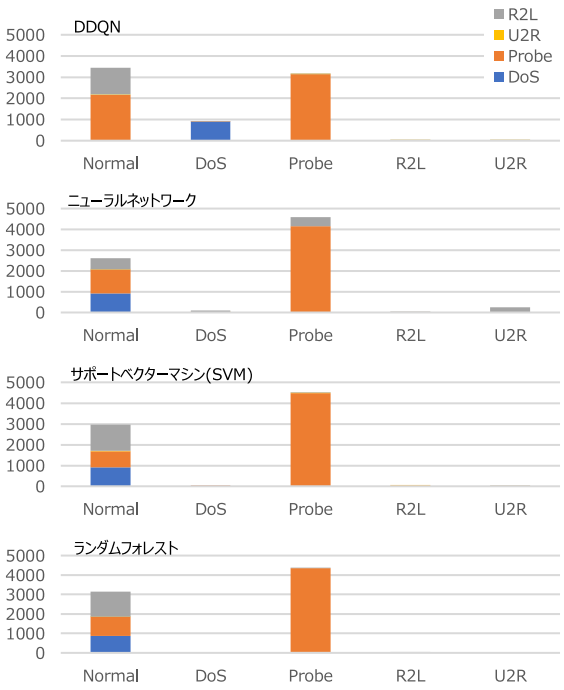


図 9 既存手法の予測分別数
Fig. 9 Prediction classification using the existing method.

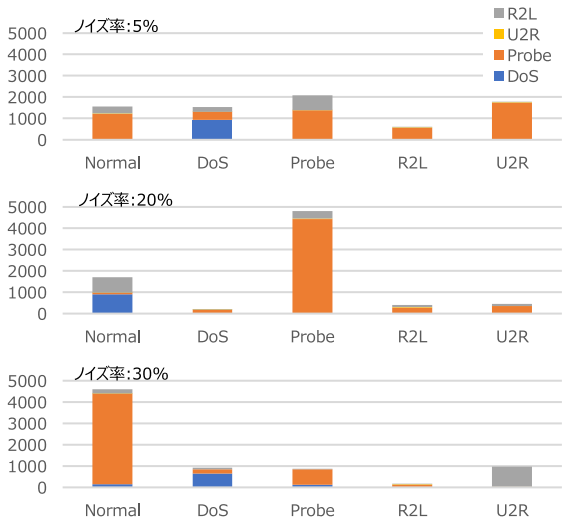


図 10 DAE のノイズ率変更による予測分別数
Fig. 10 Prediction classification using the proposed method due to noise rate change on DAE.

表 7 DAE のノイズ率の影響
Table 7 Effect of noise rate on DAE.

ノイズ率	マイクロ平均
5%	0.309
10% 提案手法採用値	0.657
20%	0.604
30%	0.187

表 8 深層距離学習を用いた予測
Table 8 Prediction using deep metric learning.

パラメータ λ	マイクロ平均
0.01	0.507
0.5	0.48
1.0	0.612

および予測処理に深層強化学習を適用しており，異なる機械学習モデルを組み合わせていることを特徴としている．ここでは，特徴量抽出処理としてのオートエンコーダの妥当性を，ノイズ除去オートエンコーダ（DAE）のノイズ率が予測性能に与える影響と，特徴量抽出処理に深層距離学習という別の機械学習モデルを適用することで検証する．

4.4.1 DAE のノイズ率の変更

DAE はノイズの乗った入力から元の入力を復元することで汎化能力が高くなる特徴を持つが，ノイズ率が高くなることにより元の入力と乖離していくことになるから，オートエンコーダの妥当性を，DAE のノイズ率が予測性能に与える影響の視点から検証した．

DAE のノイズ率を変更した場合の予測結果については，提案手法で採用したノイズ率 10%のマイクロ平均が最も良い（**表 7**）．また，カテゴリごとの予測分別数である**図 10** から，ノイズ率が高いと特定のカテゴリに偏り，低いとカテゴリの偏りがないという予測結果となった．

4.4.2 学習モデルの変更

オートエンコーダの妥当性を，特徴量抽出処理に深層距離学習という別の機械学習モデルを適用することで検証した．深層距離学習ではパラメータ λ を変更することによりパラメータ間の距離を変更できることから，特徴量抽出処理において，同じカテゴリを近い距離に，異なるカテゴリを遠い距離に変換する．λ は 0 から 1 の範囲をとり，1 に近いほど距離を遠ざける．

特徴量抽出処理を DAE の代わりに深層距離学習に置き換えた場合の予測結果については，λ の値を大きくした方が，すなわちカテゴリ間の距離を大きくした方がマイクロ平均の予測性能が若干高くなったが，提案手法には及ばなかった（**表 8**）．また，カテゴリごとの予測分別数である図 7 と**図 11** から，提案手法に比べて未知の攻撃を正常通信と予測してしまう可能性が高いという結果が得られた．

特徴量抽出処理の DAE の前段に深層距離学習（$\lambda = 1.0$）を挿入した場合，すなわち深層距離学習（$\lambda = 1.0$）と DAE を組み合わせた場合の予測結果（**図 12**）は，マイクロ平均 0.635 で，カテゴリごとの予測分別数（**図 13**）の点からも提案手法には及ばなかった．

5. 考察

本章では，評価結果をふまえ，提案手法を用いた予測と，特徴量抽出処理としてのオートエンコーダの妥当性について考察する．

5.1 予測性能

既存手法がほとんどの未知の攻撃を Normal もしくは

© 2021 Information Processing Society of Japan

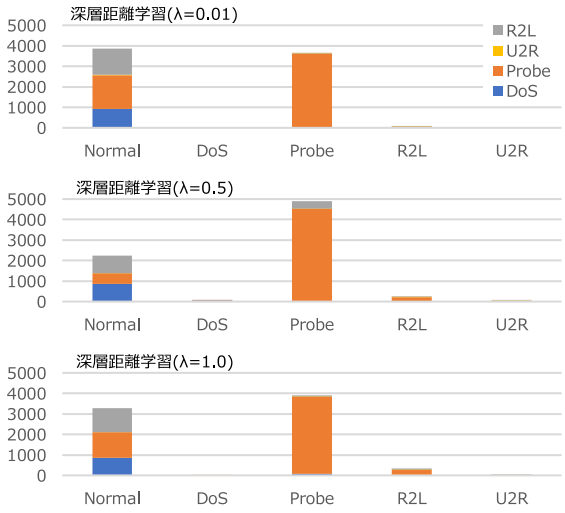
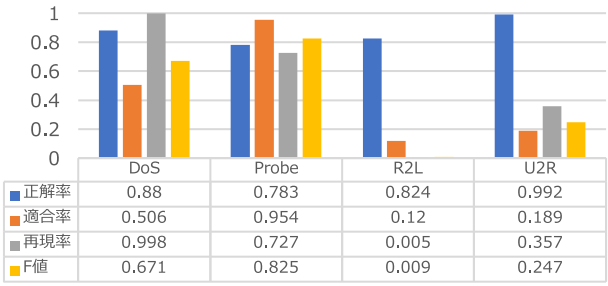


図 11 深層距離学習による予測分別数

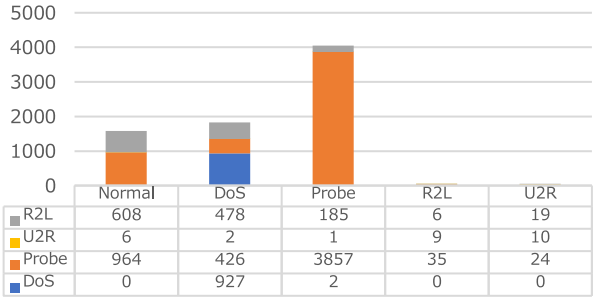
Fig. 11 Prediction classification using the deep metric learning.



	DoS	Probe	R2L	U2R
正解率	0.88	0.783	0.824	0.992
適合率	0.506	0.954	0.12	0.189
再現率	0.998	0.727	0.005	0.357
F値	0.671	0.825	0.009	0.247

図 12 深層距離学習（$\lambda = 1.0$）と DAE による予測

Fig. 12 Prediction using the deep metric learning ($\lambda = 1.0$) and DAE.



	Normal	DoS	Probe	R2L	U2R
R2L	608	478	185	6	19
U2R	6	2	1	9	10
Probe	964	426	3857	35	24
DoS	0	927	2	0	0

図 13 深層距離学習（$\lambda = 1.0$）と DAE による予測分別数

Fig. 13 Prediction classification using the deep metric learning ($\lambda = 1.0$) and DAE.

Probe と予測しているのに対して，提案手法は既存手法と比べてマイクロ平均での予測性能が高く，予測分別数の点でも検知できる攻撃カテゴリも多いことから，R2L を除いて，提案手法は未知の攻撃分別検知に対する予測性能は高いと考える．ここでは，R2L に区分される未知の攻撃が，他の攻撃カテゴリの予測性能に比べて低くとどまってしまったことを特徴表現の可視化から考察する．

機械学習で決定境界が作られる際にカテゴリごとに特徴量が集まっている方が予測性能は高くなる．提案手法における特徴量の集まり方の度合いを比較するために可視化した特徴表現を**図 14**，**図 15** に示す．なお，二次元へのデータの変換には主成分分析を使用した．

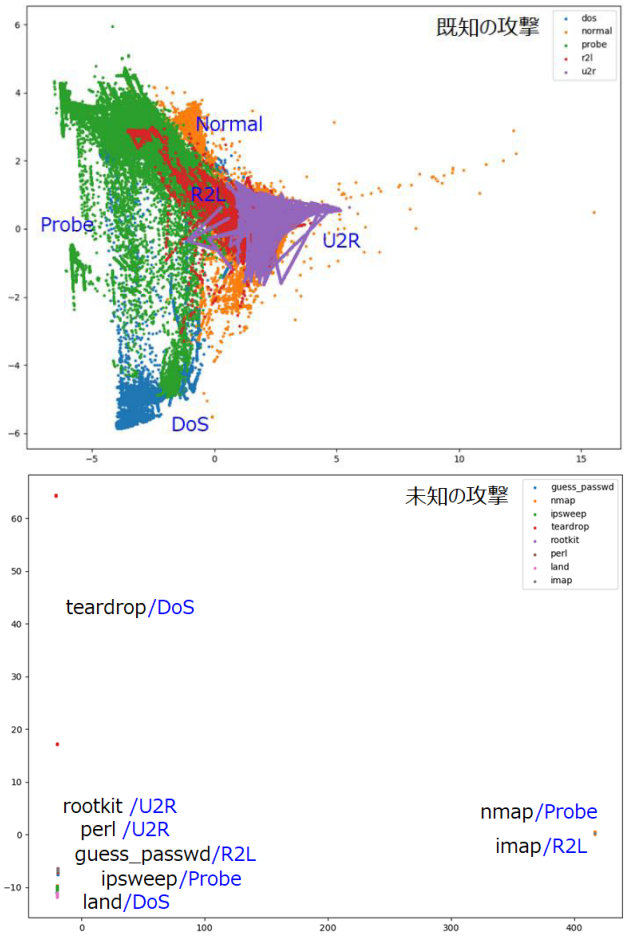


図 14 DAE 処理前の特徴表現

Fig. 14 Feature expression before DAE process.

DAE 処理前の可視化（図 14）に示すとおり攻撃カテゴリが異なっていても重なってしまっているものが多く，特に，R2L は既知ならびに未知の攻撃において他の攻撃カテゴリとの重なりの傾向が顕著である．DAE 処理後の可視化（図 15）でも R2L は他の攻撃カテゴリと重なっているが，その他については処理前に比べて攻撃カテゴリごとの特徴量が集まっている．このことから，未知の攻撃の分別においても特徴量が攻撃カテゴリごとに集まっていることが重要であることが示唆される．

特徴表現の可視化から見た攻撃カテゴリごとの予測性能に関する考察は，次のとおりである．

- DoS
  teardrop は既知の攻撃の DoS とおおよそ同じところに位置するが，land は異なるところに位置している．teardrop のみ同じところに位置しているにもかかわらず DoS の再現率が高いのは teardrop と land が両者とも同方向に位置しており，この空間が DoS の空間として学習されたためと考える．

© 2021 Information Processing Society of Japan

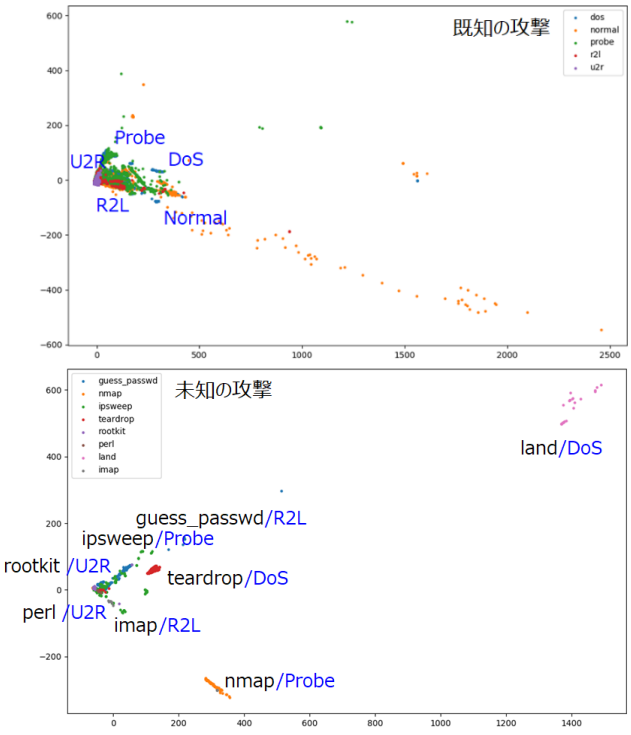


図 15 DAE 処理後の特徴表現
Fig. 15 Feature expression after DAE process.

- Probe
  ipsweep は既知の攻撃の Probe とおおよそ同じところに位置するが，nmap は異なり，さらに Normal の空間に位置している．このことから ipsweep が DoS や U2R として，nmap が Normal として分別されたと考える．
- R2L
  既知の攻撃において他の攻撃カテゴリと重なっているものが多く（図 14），guess_passwd，imap も他の攻撃と重なっていることが（図 15），他の攻撃カテゴリの予測性能に比べて低くとどまっている要因であると推察している．
- U2R
  rootkit，perl ともに既知の攻撃の U2R とおおよそ同じ位置に存在している．未知の攻撃の予測性能が低いのは，既知の攻撃において他の攻撃カテゴリと重なっていることが要因であると推察している．

5.2 特徴量抽出処理としてのオートエンコーダの妥当性

ノイズ率変更による予測結果である表 7，カテゴリごとの予測分別数である図 10 に示す結果から，ノイズ率を変更することにより予測性能，特に，カテゴリごとの予測分別数を調整できることを明らかとした．特徴量抽出処理を深層距離学習に置き換えた場合については，既存手法と同様に未知の攻撃を正常通信と予測してしまう可能性が高く，DAE の前段に深層距離学習（$\lambda = 1.0$）を挿入した場合の結果から，DAE を用いることでカテゴリの境界は分離されていると判断できる結果を得た．

6. まとめ

本論文では，特徴量抽出処理と学習および予測処理に異なる機械学習モデルを組み合わせたアプローチとして，オートエンコーダと深層強化学習を用いた NIDS における未知の攻撃を検知し分別する手法について提案した．提案手法の評価には，データセットとして NSL-KDD を使用し，オートエンコーダとして DAE，深層強化学習として DDQN を使用した．全体のマイクロ平均の正解率，攻撃カテゴリごとの適合率，再現率などを用いて既存手法と比較した結果，提案手法は既存手法と比較して，マイクロ平均の正解率が高く全体として未知の攻撃分別検知の予測性能が高いという結果が得られた．また，特徴表現の可視化から攻撃カテゴリごとの空間が明確に分離されていると予測性能が高くなることを示した．

本論文では，未知の攻撃を既存手法と比較して高い予測性能で分別できることを示したが，分別できない攻撃カテゴリが存在し，誤分別も少なくなかった．今後は，攻撃カテゴリごとの空間の分離に着目し，予測性能の向上や分別精度の改善のためにパラメータの調整や異なる機械学習アルゴリズムの組合せを考えている．

参考文献

[1] 警察庁：令和元年におけるサイバー空間をめぐる脅威の情勢等について，入手先 ⟨https://www.npa.go.jp/publications/statistics/cybersecurity/data/R01_cyber_jousei.pdf⟩（参照 2021-01-10）．
[2] 高原尚志，櫻井幸一：KDD CUP 99 Data Set を用いた異なる学習データによる機械学習アルゴリズムの評価，コンピュータセキュリティシンポジウム 2015 論文集 (2015)．
[3] Xin, Y. et al.: Machine Learning and Deep Learning Methods for Cybersecurity, *IEEE Access*, Vol.6, pp.35365–35381, DOI: 10.1109/ACCESS.2018.2836950 (2018).
[4] Pérez, J.L.R. and Ribeiro, B.: Attribute Learning for Network Intrusion Detection, *Advances in Big Data* (*INNS 2016*), pp.39–49, DOI: 10.1007/978-3-319-47898-2_5 (2017).
[5] Revathi, S. and Malathi, A.: A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection, *International Journal of Engineering Research and Technology*, Vol.2 (2013).
[6] Pervez, M.S. and Farid, D.M.: Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs, *The 8th International Conference on Software, Knowledge, Information Management and Applications* (*SKIMA 2014*), pp.1–6, DOI: 10.1109/SKIMA.2014.7083539 (2014).
[7] Zhao, J., Shetty, S., Pan, J., et al.: Transfer learning for detecting unknown network attacks, *EURASIP J. Info. Security*, Vol.1, DOI: 10.1186/s13635-019-0084-4 (2019).
[8] Yan, B. and Han, G.: Effective Feature Extraction via Stacked Sparse Autoencoder to Improve Intrusion Detection System, *IEEE Access*, Vol.6, pp.41238–41248 DOI:

© 2021 Information Processing Society of Japan

10.1109/ACCESS.2018.2858277 (2018).
[9] Alom, M.Z. and Taha, T.M.: Network intrusion detection for cyber security using unsupervised deep learning approaches, *2017 IEEE National Aerospace and Electronics Conference* (*NAECON*), pp.63–69, DOI: 10.1109/NAECON.2017.8268746 (2017).
[10] Xiao, X., Zhang, D., Hu, G., Jiang, Y. and Xia, S.: CNN–MHSA: A Convolutional Neural Network and multi-head self-attention combined approach for detecting phishing websites, *Neural Networks*, Vol.125, pp.303–312, ISSN 0893-6080, DOI: 10.1016/j.neunet.2020.02.013 (2020),
[11] Sarker, I.H., Abushark, Y.B., Alsolami, F. and Khan, A.I.: IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model, *Symmetry*, Vol.12, 754 (2020).
[12] Lopez-Martin, M., Carro, B. and Sanchez-Esguevillas, A.: Application of deep reinforcement learning to intrusion detection for supervised problems, *Expert Systems with Applications*, Vol.141, 112963, ISSN 0957-4174, DOI: 10.1016/j.eswa.2019.112963 (2020).
[13] Hsu, Y.-F. and Matsuoka, M.: A Deep Reinforcement Learning Approach for Anomaly Network Intrusion Detection System, *2020 IEEE 9th International Conference on Cloud Networking* (*CloudNet*), pp.1–6, DOI: 10.1109/CloudNet51028.2020.9335796 (2020).
[14] Benaddi, H., Ibrahimi, K., Benslimane A. and Qadir, J.: A Deep Reinforcement Learning Based Intrusion Detection System (DRL-IDS) for Securing Wireless Sensor Networks and Internet of Things, Deng, D.J., Pang, A.C. and Lin, C.C. (Eds.), *Wireless Internet*, *WiCON 2019*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol.317, Springer, DOI: 10.1007/978-3-030-52988-8_7 (2020).
[15] Sethi, K., Kumar, R., Prajapati, N. and Bera, P.: Deep Reinforcement Learning based Intrusion Detection System for Cloud Infrastructure, *2020 International Conference on COMmunication Systems & NETworkS* (*COMSNETS*), pp.1–6, DOI: 10.1109/COMSNETS48256.2020.9027452 (2020).
[16] Dhanabal, L. and Shantharajah, S.: A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms (2015).
[17] University of New Brunswick: NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB, available from ⟨https://www.unb.ca/cic/datasets/nsl.html⟩ (accessed 2021-01-10).

付　　録

付録では，図 8〜図 11 のデータテーブルを**表 A·1**〜**表 A·4** に示す．表 A·1 の F 値の空白セルは計算不可能なためである．

表 A·1　既存手法による予測
Table A·1　Prediction using the existing method.

		DoS	Probe	R2L	U2R
DDQN	正解率	0.996	0.713	0.826	0.995
	適合率	0.991	0.999	0.172	0.367
	再現率	0.973	0.591	0.004	0.393
	F 値	0.982	0.743	0.008	0.379
ニューラルネットワーク	正解率	0.866	0.79	0.827	0.966
	適合率	0	0.905	0.118	0.033
	再現率	0	0.783	0.002	0.286
	F 値		0.84	0.003	0.059
SVM	正解率	0.877	0.888	0.823	0.996
	適合率	0.458	0.994	0	0.5
	再現率	0.012	0.845	0	0.464
	F 値	0.023	0.914		0.481
ランダムフォレスト	正解率	0.876	0.862	0.827	0.997
	適合率	0.263	0.989	0.238	1.0
	再現率	0.005	0.812	0.004	0.143
	F 値	0.011	0.892	0.008	0.25

表 A·2　既存手法の予測分別数
Table A·2　Prediction classification using the existing method.

		Normal	DoS	Probe	R2L	U2R
DDQN	R2L	1271	4	3	5	13
	U2R	6	0	1	10	11
	Probe	2145	4	3137	14	6
	DoS	25	904	0	0	0
ニューラルネットワーク	R2L	552	73	433	2	236
	U2R	4	1	0	15	8
	Probe	1138	13	4155	0	0
	DoS	928	0	1	0	0
SVM	R2L	1262	0	21	0	13
	U2R	6	0	3	6	13
	Probe	774	13	4486	33	0
	DoS	916	11	2	0	0
ランダムフォレスト	R2L	1286	0	5	5	0
	U2R	14	0	0	10	4
	Probe	976	14	4310	6	0
	DoS	881	5	43	0	0

表 A·3　DAE のノイズ率変更による予測分別数
Table A·3　Prediction classification using the proposed method due to noise rate change on DAE.

		Normal	DoS	Probe	R2L	U2R
ノイズ率 5%	R2L	338	224	718	6	10
	U2R	3	1	1	4	19
	Probe	1222	387	1379	571	1747
	DoS	0	929	0	0	0
ノイズ率 20%	R2L	736	11	353	97	99
	U2R	3	1	6	13	5
	Probe	69	157	4441	291	348
	DoS	903	26	0	0	0
ノイズ率 30%	R2L	217	77	35	18	949
	U2R	3	0	0	14	11
	Probe	4236	193	731	127	19
	DoS	155	655	117	2	0

© 2021 Information Processing Society of Japan

表 A·4 深層距離学習による予測分別数
Table A·4 Prediction classification using the deep metric learning.

		Normal	DoS	Probe	R2L	U2R
深層距離学習 (λ=0.01)	R2L	1277	0	14	4	1
	U2R	12	0	1	10	5
	Probe	1635	0	3616	54	1
	DoS	928	0	1	0	0
深層距離学習 (λ=0.5)	R2L	851	20	369	32	24
	U2R	5	0	3	11	9
	Probe	508	9	4526	223	40
	DoS	873	56	0	0	0
深層距離学習 (λ=1.0)	R2L	1171	0	68	47	10
	U2R	12	0	1	8	7
	Probe	1242	8	3776	273	7
	DoS	862	0	66	1	0

本丸 真人

2021 年東京電機大学大学院未来科学研究科情報メディア学専攻修士課程修了．同年（株）ハウテレビジョン入社．

寺田 真敏 （正会員）

1986 年千葉大学大学院工学研究科写真工学専攻修士課程修了．同年（株）日立製作所入社，研究開発グループ，Hitachi Incident Response Team でサイバーセキュリティの研究に従事．博士（工学）．現在，東京電機大学大学院未来科学研究科教授．2004 年 JPCERT コーディネーションセンター専門委員，（独）情報処理推進機構セキュリティセンター研究員，2007 年日本シーサート協議会運営委員，2008 年中央大学大学院客員講師，2015 年より ICT-ISAC 運営委員を兼務．

© 2021 Information Processing Society of Japan