SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

分布外データに対する脆弱性と検知
Out-of-distribution Vulnerability and Detection

江田 智尊 * 森川 郁也 *
Satoru Koda Ikuya Morikawa

キーワード AI セキュリティ，分布外検知，Out-of-distribution

あらまし

背景: 機械学習アルゴリズムの多くは i.i.d. 仮定の下で学習されるため，分布外 (out-of-distribution, OOD) のサンプルに対して脆弱である．そのようなアルゴリズムは，学習時に想定した分布からのデータには適切に推論を行えるが，OOD サンプルに対して予期しない推論を行うことがある．例えば学習時に想定してないゼロデイ攻撃や新種マルウェアに対して，良性判定を下してしまう．本論文では，マルチクラス分類における OOD 検知を扱う．そのタスクは，既知クラス (学習時にデータに存在するクラス）を正しく判定しつつ，それ以外のクラスの入力を OOD と判定することである．

先行研究: 画像・自然言語処理分野における OOD 検知手法は，特に 2017 年以降多く開発されてきた．これらは，例えば MNIST で学習した判別器に CIFAR10 の画像を入力し，それらを正しく OOD 検知できるかという点を評価する．一方でセキュリティ領域では表形式データがよく用いられる．しかし表形式データにおける OOD 検知を扱う論文は少ない [1, 2]．またこれらも，最新技術の比較が少ない点や，ベースとなる判別器がニューラルネットワークである点が問題である．

論文の貢献:

- 従来無かった表形式データに対する包括的な OOD 検知性能評価を実施した．最新技術を含む既存 10 手法を 7 つの表形式データセットに適用した．その結果，ツリーモデルの一種である Gradient Boosting Decision Tree (GBDT) の OOD 検知性能が高いことを実証した．

- GBDT の検知性能を改善する技術を提案した．本技術は二つの技術要素，1) 疑似 OOD データ生成法, 2) OOD データに耐性をもつ GBDT の学習法，から成る．提案技術が既存技術の OOD 検知精度を 12.6 % 改善した (表 1)．

表 1: OOD 検知性能評価結果の一部 (表中の数字は誤検知 5%時点での OOD 検知成功率)

データ \ 手法	MCDD[1]	CADE[2]	GBDT	提案手法
Avila	0.165	0.100	0.560	0.632
DriveDiag.	0.281	0.448	0.455	0.582
GasSensor	0.785	0.814	0.529	0.537
MNIST	0.373	0.445	0.504	0.492
Segment	0.633	0.412	0.565	0.458
Shuttle	0.987	0.807	0.557	0.984
CICIDS	0.791	0.487	0.898	0.895
平均	0.574	0.502	0.581	0.654

- セキュリティアプリケーションがゼロデイ攻撃に対してどのように脆弱になるかを，説明可能 AI 技術を用いて実証した．具体的には，侵入検知システムのデータを用いて攻撃を判別する判別器を学習する．そして，ある攻撃が訓練データに含まれている/いないときに，判別器が着目する特徴量がどのように変化するかを，説明可能 AI 技術を用いて定量化する．これにより，判別器が未知攻撃に特異な特徴を捉えられないことを実証した．

参考文献

[1] D Lee et al., Multi-Class Data Description for Out-of-Distribution Detection, KDD 2020.

[2] L Yang et al., CADE: Detecting and Explaining Concept Drift Samples for Security Applications, USENIX 2021.

* 富士通株式会社，〒 211-8588 神奈川県川崎市中原区上小田中 4-1-1, Fujitsu Limited, 4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki-shi. (koda.satoru@fujitsu.com)