機械学習を用いたネットワーク攻撃検知システムにおける説明可能性について

2020SE062 住山和茂
指導教員：沢田篤史

1　はじめに

機械学習技術を用いたタスクにおいて，ブラックボックス化が指摘されている．ブラックボックス化とは，一般に高い精度が期待できるモデルほど，結果を出力するまでの過程を人間が理解することが難しいという問題である．

この問題を解決する技術に説明可能な AI（XAI）があるが，単一の XAI で得られる情報には限界がある．

本研究の目的は，機械学習を用いた攻撃検知システムの説明可能性を向上させることである．本研究では，個々の結果に対してどんな根拠に基づいて出力を決定したかを説明する能力の高さを，説明可能性の定義として採用する．

目的達成に向けて，次の技術的課題を設定する．

1. システム導入前に必要な情報を整理し，効果的な XAI 併用方法を明らかにする．
2. 得られた情報の効果的な活用方法を明らかにする．
3. XAI 併用方法と情報の活用方法が，説明可能性の向上に与える効果を明らかにする．

本研究では，XAI 技術の併用方法に類似検索を組み合わせる方法を提案する．従来よりも多角的な情報をセキュリティ分析者に提示することで，説明可能性の向上を試みる．

提案する方法の妥当性は，簡単なプロトタイプによる実験で，定性的に評価する．

2　機械学習を用いた攻撃検知システムにおける XAI 適用とその問題点

近年，機械学習はセキュリティの分野でも注目を集めている．例えば，データセットから自動で異常を学習することで，高い精度のアノマリ型ネットワーク攻撃検知システムを構築することが可能である．

都留ら [1] の研究では，Kyoto2016 データセットで学習された XGBoost を用いて攻撃検知システムを構築し，2 値分類で 96.04 %（平均）の正解率を記録した．XGBoost は，ブースティングと決定木を組み合わせた機械学習モデルで，ブラックボックス化が指摘されている．

攻撃検知システムが誤った検知結果を出力することで経済的に重大な問題を引き起こす可能性がある．よって，現場への導入には入念な確認作業が必要であるが，AI のブラックボックス化はその妨げとなっている．

Patel[2] の研究では，攻撃検知システムに XAI を実装して説明可能性の向上を試みた．最も説明可能性が高い決定木を除き，1 つの機械学習モデルに 1 つの XAI を実装して分析を行っており，十分な説明可能性を確保できていない．攻撃検知システムの説明可能性を向上させるために，効果的な XAI 併用方法を考察する必要がある．

Keshk ら [3] の研究では，攻撃検知システムの説明可能性を向上させる目的で，複数の XAI を統合したフレームワークが提案された．局所的説明には，SHAP と ICE を併用している．SHAP は，Lundberg ら [4] によって提案された各特徴量の結果への寄与度を用いた根拠説明を行う XAI である．また，ICE は，Goldstein ら [5] によって提案された各入力ごとに特徴量と検知結果の関係をグラフ化して説明を行う XAI である．この研究では，ICE による説明がすべての局所的説明を重ね合わせた状態の大局的説明になっており，各検知結果に対応したグラフが不鮮明になっている．よって，局所的説明として，適切に機能する ICE の実装方法を考察する必要がある．

3　説明可能性向上に向けた技術的課題

本研究の目的は，機械学習を用いた攻撃検知システムの説明可能性を向上させることである．

この目的を達成するために，次の技術的課題を設定する．

1. システム導入前に必要な情報を整理し，効果的な XAI 併用方法を明らかにする．
2. 得られた情報の効果的な活用方法を明らかにする．
3. XAI 併用方法と情報の活用方法が，説明可能性の向上に与える効果を明らかにする．

4　攻撃検知結果の説明可能性を向上させる XAI 技術の併用方法の提案

図 1 は，本研究で提案するシステムの概要である．

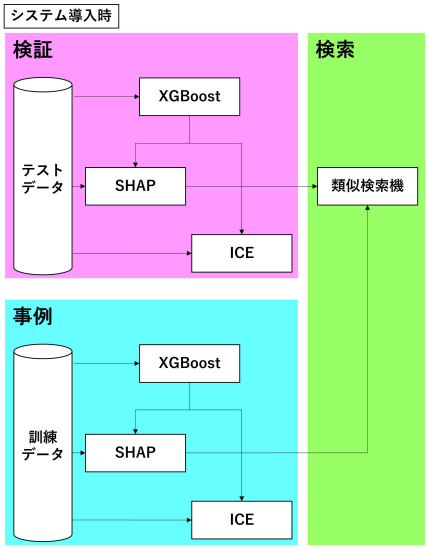

図 1　システムの概要

本研究では，SHAP/ICE の併用に，2 種類の類似検索を組み合わせる方法を提案する．想定する場面は機械学習を用いた攻撃検知システムの導入前で，信頼可否を判断するのに役立つ情報をセキュリティ分析者に提供する．

検証モジュールでは，テストデータからセキュリティ分析者が検知結果を選択して，その検知結果に対応した判断根拠が SHAP によって説明される．その後，特徴量を選択して，ICE で判断傾向を提示する．

事例モジュールでは，検索モジュールで出力されたサンプル番号に対応した検知結果について，検証モジュールと同様の情報を提示する．

検索モジュールでは，検証モジュールで着目した検知結果と事例モジュールの全検知結果の SHAP の情報を基に，類似データのサンプル番号を取得する．

データセットは，UNSW-NB15[*1]を表形式データで採用する．現代の豊富な攻撃に対応でき，KDD99 や NSL-KDD よりも有用性がある．

機械学習モデルは，XGBoost を採用する．Grinsztajn ら [6] により，表形式データにおいて木構造による分類モデルは深層学習モデルよりも優れていることが示された．本研究では，Patel[2] の研究で学習された XGBoost を，正解率が 89.87 %で再現して活用する．

SHAP は，TreeSHAP[7] を採用する．寄与度計算に木構造を利用し，迅速で正確な説明が得られる．

ICE は，Github 上で公開されたクラス[*2]を採用する．各検知結果ごとに独立した図を作成し，Keshk ら [3] の研究における問題の解決を試みる．

本研究では，SHAP，ICE の順番で XAI を併用する．攻撃検知システムの判断根拠に基づいて，特定の特徴量と検知結果の関係性分析を可能にすることを試みる．

類似検索について，SHAP の寄与度は値の範囲が揃えられており，ユークリッド距離のみで実現可能である．

1 つ目の類似検索は，最も近いデータを 1 つだけ出力する．類似データとの判断根拠の違いが，特定の特徴量と検知結果の関係性に与える影響の提示を試みる．

2 つ目の類似検索は，着目する特徴量を除いて類似データを指定個数だけ検索し，着目する特徴量の寄与度が大きい順に出力する．着目する特徴量の寄与度の違いが，検知結果の関係性に与える影響の提示を試みる．

5　実験結果と考察

SHAP，ICE の順番で XAI を併用することで，セキュリティ分析者は自身の知識や勘だけでなく，機械学習を用いた攻撃検知システムの判断根拠に基づいて特定の特徴量と検知結果の関係性を分析できることがわかった．

1 つ目の類似検索では，類似データとの判断根拠の違いが特定の特徴量と検知結果の関係性に与える影響を見ることができた．

2 つ目の類似検索では，他の特徴量の影響を完全に無視できず，期待する情報を得ることができなかった．

データセットのような限られたデータ群から，着目する特徴量以外の寄与度が一致するデータを検索できなかった．よって，着目する特徴量以外の寄与度が一致するデータを生成する方法を明らかにする必要がある．

各特徴量が攻撃または安全な方向に寄与することが，セキュリティ分析者の専門知識では何を意味するのかが不明確である．知識ベースとの融合は，セキュリティ分析者に XAI の情報をわかりやすく提示するうえで重要な課題であり，その方法を明らかにする必要がある．

本研究での妥当性評価は，本稿の著者がプロトタイプを利用して定性的に行ったことから，主観評価による影響は避けられない．今後は，セキュリティ分野の専門家などによるより客観性のある評価方法を検討する必要がある．

6　おわりに

ブラックボックス化は，機械学習を用いた攻撃検知システムの導入前における確認作業の妨げとなっている．

本研究では，SHAP/ICE の併用に，2 種類の類似検索を組み合わせる方法を提案した．

実験から，SHAP，ICE の順番による XAI の併用や 1 つ目の類似検索で，期待する効果を得た．一方で，2 つ目の類似検索で，期待する効果を得ることができなかった．

参考文献

[1] 都留 悠哉ら, ”勾配ブースティング決定木を用いたネットワーク侵入検知システムの提案”, 研究報告電子化知的財産・社会基盤 (EIP), 2021.

[2] Harshil Patel, “IoT Network Intrusion Detection and Classification using Explainable (XAI) Machine Learning Algorithms”, ISE 5194 Human-Centered Machine Learning Spring 2021.

[3] Marwa Keshk et al.,“An explainable deep learning-enabled intrusion detection framework in IoT networks”, Information Sciences Volume 639, 2023.

[4] Scott M. Lundberg et al., ”A Unified Approach to Interpreting Model Prediction”, Advances in Neural Information Processing Systems 30, NIPS 2017.

[5] Alex Goldstein et al., ”Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation”, Journal of Computational and Graphical Statistics Volume 24, 2015.

[6] Leo Grinsztajn et al., ”Why do tree-based models still outperform deep learning on tabular data?”, Advances in Neural Information Processing Systems 35, NeurIPS 2022.

[7] Scott M. Lundberg et al.,“Consistent feature attribution for tree ensembles”, arXiv:1706.06060, 2017.

[*1] https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15

[*2] https://github.com/ghmagazine/ml_interpret_book