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

開発エンジニア向け機械学習セキュリティ脅威分析技術
A Threat Analysis Method on Machine Learning Security for System Development Engineers

矢嶋 純 * 及川 孝徳 * 森川 郁也 * 笠原 史禎 * 乾 真季 *
Jun Yajima Takanori Oikawa Ikuya Morikawa Fumiyoshi Kasahara Maki Inui
吉岡 信和 †
Nobukazu Yoshioka

キーワード 機械学習セキュリティ, 脅威分析, ガイドライン, Adversarial Attack

あらまし

近年，機械学習システムの判断を意図的に誤らせたり，情報を盗み取ったりする機械学習特有の攻撃が指摘されている．このような攻撃に対応するには，機械学習システムにどのような攻撃が実施可能であるかを抽出する脅威分析が必要である．現状，このような脅威分析は機械学習セキュリティの専門家のみが実施できる．機械学習システムの開発プロセスは通常のソフトウェアの開発と異なり，精度確認などを行うための試作フェーズがあることが多い．このようなプロセスに脅威分析を追加する場合，専門家による分析の結果を受けて仕様変更を実施する等の対応を行うため，手戻りが多く発生する可能性がある．この手戻りを削減するには，開発者が自ら脅威分析を行うことが好ましいと考えられる．そこで本論文では開発者が自身で分析を実施可能な脅威分析技術を提案する．同じ目的の技術は昨年筆者らが提案 [1] しているが，この手法は分析内容がブラックボックス的であり，結果の納得性が得にくいという問題があった．これに対し提案手法ではアタックツリーを用いて可視化することで結果の納得性を得やすくなっている．本技術は機械学習工学研究会（MLSE）より発行予定の非強制のガイドラインに含まれる予定である．

提案手法による分析手順は，以下のとおりである．

1. 専門家が事前準備として事前に図 1 のようなアタックツリーを攻撃毎に抽出し，葉に相当する部分に記載される成立条件を満たしているかどうかを明らかにする質問群も作成する．
2. 分析者は想定する攻撃者を決めて，その攻撃者の能力を考慮しながら作成された質問群に回答する．
3. 分析者はさらに回答結果を元に，ツリーの成立条件を満たしているかどうかを確認する．
4. 分析者はツリーの成立状況を確認することでシステムにどの攻撃が実施できるのかを見極め，対応を検討する．

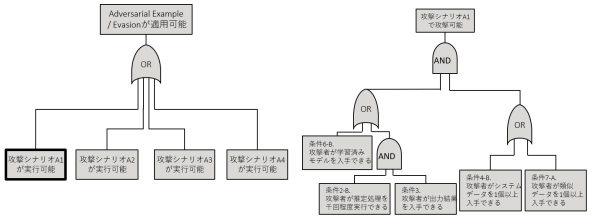

図 1: 専門家が抽出したアタックツリーの一部（例）

我々は提案手法について複数名で試行評価を実施した．本論文では試行評価の結果についても紹介する．

参考文献

[1] 矢嶋, 清水, 森川, 大久保, “機械学習システムに潜む AI セキュリティ脆弱性の分析手法に関する一考察,” 2021 年暗号と情報セキュリティシンポジウム (SCIS2021).

* 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, Kanagawa 211-8588, Japan.
† 早稲田大学, 〒 169-8555, 東京都新宿区大久保 3-4-1, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan