Computer Security Symposium 2019
21 - 24 October 2019

敵対的機械学習のための脅威分析手法の提案

森川 郁也[1,a)]　清水 俊也[1]　樋口 裕二[1]　前田 若菜[1]　矢嶋 純[1]

概要：近年機械学習に特有の攻撃が発見され，敵対的環境下での機械学習のセキュリティが重要な研究領域となっている．セキュリティを高めるには脅威分析が重要だが，機械学習に対する攻撃を理解するのは難しく，分析の障害となる．本稿ではこれらの攻撃に特化した脅威分析手法を提案する．提案手法は脅威の尤度を算定する攻撃ツリーに 2 つの特徴がある: (1) 分析対象の攻撃を限定することで容易に尤度を算定できるようにしたこと，および (2) 代理モデルの作成など機械学習特有の攻撃条件を表現できることである．これにより一般の設計者・開発者でも簡易なリスク算定を短時間で行える．脅威分析という性質上継続的な改善は必須だが，現時点での評価についても述べる．

キーワード：敵対的機械学習, 脅威分析, 攻撃ツリー

A Proposal of Threat Analysis for Adversarial Machine Learning

Ikuya Morikawa[1,a)]　Toshiya Shimizu[1]　Yuji Higuchi[1]　Wakana Maeda[1]　Jun Yajima[1]

Abstract: Recently, security for machine learning (ML) in adversarial environments is regarded as an important research area, as many attacks unique to ML have been found. Although threat analysis is indispensable to achieve security, it is hard for designers and developers to understand such attacks, and therefore execution of the analysis is challenging in practice. In this paper, we propose a threat analysis method designed for adversarial machine learning. The method employs an attack tree to estimate likelihood of the attacks, and it has two new features: (1) it simplifies the estimation by focusing on the ML-specific attacks, and (2) it can represent intermediate conditions such as attacks via substitute models. Thus it enables designers/developers without security expertise to estimate security risks within a few hours. We also show a preliminary evaluation of the method.

Keywords: Adversarial machine learning, threat analysis, attack tree

1. はじめに

近年，深層学習を中心とした機械学習技術の進展に伴い，様々なシステムで機械学習を利用する動きが活発である．しかし敵対的サンプルなど機械学習に特有の攻撃が次々と発見され，攻撃者が存在する敵対的環境下での機械学習のセキュリティが問題となっている．システムのセキュリティを高めるには，まず脅威分析によって脅威の存在やリスクの度合いを把握することが重要だが，セキュリティの専門的知識を要する，多大な労力がかかる，など課題も多く，攻撃技術・防御技術ともに未だ確立していない敵対的機械学習については特に困難を伴う．

このような状況を改善するため，本稿では機械学習特有の攻撃に特化した脅威分析手法を提案する．提案手法は脅威の尤度を算定する攻撃ツリーに次のような特徴がある．第 1 に，機械学習特有の攻撃種別に限定することで，分析対象のシステムごとに攻撃ツリーを構築する必要をなくし，尤度算定を容易に行えるようにした．第 2 に，システム上で扱われるデータの操作・入手可能性と訓練済みモデルへのアクセス手段との間の依存関係を明らかにし，代理モデルを介した攻撃など機械学習に特有の条件を考慮した尤度算定を可能にした．

[1] 株式会社富士通研究所 セキュリティ研究所
Security Laboratory, Fujitsu Laboratories Ltd.
[a)] morikawa.ikuya@fujitsu.com

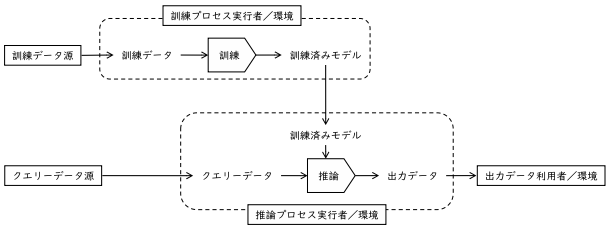

図 1 機械学習利用システムのパイプライン
Fig. 1 Pipeline of systems using machine learning

また提案手法の長所・短所を把握するため，著者ら自身の手で試行・評価を行った．その結果，1 案件あたり 1〜2 時間という短時間で脅威分析を行うことができ，分析結果も分析者本人にとって概ね納得のいくものであった．一方，同じシステムを対象としても分析者によって算定したリスク度にばらつきが見られた．ばらつきの主な原因は対象システムの詳細や想定被害に対する解釈の違いと見られ，客観的・定量的なリスク算定とは現状では言い難いが，一般の設計者・開発者による簡易分析ツールとして，また詳細分析に向けたセキュリティ専門家とのコミュニケーションツールとして，提案手法は機械学習システムにおけるセキュリティ対策に有用と考えられる．

本稿の貢献は次の 2 点である．

- 機械学習利用システムに特有の攻撃種別に対し効率良く実施可能な脅威分析手法を提案する．
- 同分野特有の依存関係やモデルアクセス性を表現した合成攻撃ツリーを示す．

以降の構成は次のとおりである．まず 2 節で機械学習とそれに特有の攻撃について整理し，3 節ではセキュリティ対策に不可欠な脅威分析について述べる．4 節で機械学習システムに適した脅威分析手法を提案し，5 節でその試行と評価について述べた後，6 節で考察を加える．7 節で関連研究について述べ，8 節でまとめる．

2. 機械学習とセキュリティ

2.1 機械学習利用システムとデータ

機械学習とは，あるデータ群に対する適切な処理方法を，計算機がデータから自動的に学び取る技術・手法の総称である．機械学習を利用したシステムにおける典型的なデータ処理の流れ（パイプライン）を**図 1** に示す．パイプラインは **訓練**と **推論**の 2 つのプロセスに分けられる．

訓練　用意された訓練データを入力とし，あらかじめ決められた訓練方法（モデルやハイパーパラメーター）によって訓練済みモデルを出力するプロセス．

推論　訓練済みモデルに，クエリーデータを入力し，出力データを得るプロセス．推論処理の主要件はタスクと呼ばれ，個々のクエリーデータの所属を判定する分類 (classification) とクエリーデータから何らかの値を推定する回帰 (regression) に大きく分けられる．

このように機械学習パイプラインで処理されるデータは次の 4 種類に分類できる[*1]．

- 訓練データ
- 訓練済みモデル
- クエリーデータ
- 出力データ

2.2 機械学習利用システムに特有の攻撃

機械学習を回避し欺くようなセキュリティ脅威に関する研究は 10 年以上の歴史があるが [1]，いわゆる人工知能（AI）の研究と応用が深層学習（深層ニューラルネットワーク）の発達に伴って盛り上がるにつれ，2014 年頃からあらためて注目されるようになった．

本稿では **モデル抽出**，**訓練データ推定**，**敵対的サンプル**，**ポイズニング**の 4 つを機械学習システムに特有の攻撃種別と見なす．

モデル抽出　(model extraction) 他人が作った訓練済みモデルを何らかの方法で複製する攻撃である [2]．本稿の目的は攻撃手段によらず脅威の性質を明らかにすることなので，攻撃対象と事実上同じ価値を提供するモデルを意のままに実行できれば手段を問わず脅威と見なす．このためモデルを外部から推定するブラックボックス攻撃（2.3 節を参照）だけでなく，単純にモデルを複製するホワイトボックス攻撃もこの種別に含める．起こりうる直接的な被害は，訓練済みモデルの無断利用・販売などによるモデル所有価値の毀損だが，後述するようにモデルの複製は他の攻撃を容易にすることがある．

訓練データ推定　訓練済みモデルからその元となった訓練データの一部を推定する攻撃である．たとえば，顔認識をタスクとするモデルから訓練に使われた顔画像データを推定する攻撃が知られている [3]．起こりうる被害は主にプライバシー侵害や情報漏洩だが，想定被害を考えるにはまず訓練データの何が推定できると有害かを明らかにする必要がある．なぜなら，そもそも訓練済みモデルは訓練データのもつ情報や特徴を反映するように作られるので，何らかの推定はできて当然だからである．

敵対的サンプル　(adversarial sample or example) 訓練済みモデルが推論・予測を誤るように，意図的に「微小な」変化を加えられたクエリーデータ，およびそれを用いた攻撃を指す [4], [5]．ここで「微小」とは変化前のクエリーデータが本来持つ性質や機能を維持する程度という意味で，具体的には機械学習システムのユースケースに依存する．画像などのメディアデータにお

[*1] 訓練方法（訓練時のモデルの選択やハイパーパラメーター）もデータと見なせるが，外部から操作できることは稀なので本稿では扱わない．

いては「人間が見過ごす，または誤らない程度」の変化に留めるのが一般的である．起こりうる被害として，画像認識や音声認識の誤判定による誤動作や混乱，マルウェア検知や攻撃検知の回避，などが挙げられる．

ポイズニング (poisoning) 訓練データを操作することで訓練後のモデルの精度を低下させたりモデルの出力を意図的に誤らせたりする攻撃である [6], [7]．4 つの攻撃種別の中で唯一訓練プロセスに関与する必要があるため，訓練完了後には（再訓練がない限り）実行できない．起こりうる被害はタスクに強く依存し，訓練に関与するため潜在的な被害の可能性は大きいが，訓練データの操作制約や訓練時の検証など障害もあり，未知の部分が多い．

なお機械学習に対する攻撃は上記の 4 種類だけに限らないことには注意を要する．たとえば，出力データからクエリーデータを導き出す入力データ推定攻撃 [8] や，複数の種別を組み合わせた攻撃（たとえば訓練データ推定がしやすいようにモデルを操作するポイズニング [9]）も考えられている．また新たな種類の攻撃が今後見つかるかもしれない．

2.3 攻撃者のモデルへのアクセス性

機械学習利用システムへの攻撃については，攻撃対象の訓練済みモデルに対する攻撃者のアクセス性を考慮することが重要である．

まず機械学習に限らないセキュリティ一般と同様に，ホワイトボックス攻撃とブラックボックス攻撃に分けられる．ホワイトボックス攻撃はモデル内部の情報や状態まですべて入手可能な者による攻撃を意味する．一方でブラックボックス攻撃は，モデル内部への直接的なアクセスは不可能だが，通常の入出力を介してクエリーを行い出力を得ることが可能な者による攻撃を指す．ブラックボックス攻撃のほうが行う機会は多いが，攻撃を成功させるのは難しいとされる．

これらに加え機械学習システムにおいては，代理モデルを介したブラックボックス攻撃を考慮する必要がある．対象モデルの内部へアクセスできないブラックボックス攻撃者は，対象モデルと同様に動作する代理モデル（substitute, surrogate などと呼ばれる）を自分の手元に作成することで，代理モデルに対してホワイトボックス攻撃を実行できる．代理モデルに対する攻撃が対象の訓練済みモデルに対しても有効とは限らないが，少なくとも敵対的サンプル攻撃についてはある程度有効だと示されている [10]．このように類似のモデル間で同じ攻撃が可能な性質は転移性 (transferability) と呼ばれる．代理モデルがどの程度の転移性をもちうるかは一概には言えないが，同一のタスクに対して同程度の推論性能を示すモデルは転移性をもちやすいことが指摘されている [11]．

代理モデルを作成するには，攻撃者自ら訓練データを用意して訓練することもできるが，対象のモデルにクエリーして得た出力データを利用したほうが効率が良い．これは前述のモデル抽出攻撃と本質的に同じであり，大量にクエリーできることはそのモデルに特化した最適な訓練データを得るのと等価と考えることもできる．

3. 脅威分析

脅威分析とは，情報処理システムに潜むセキュリティ脅威の存在やその規模・性質を明らかにする作業である．次のような理由で，脅威分析はシステムのセキュリティ確保に不可欠である．

- 一般にシステムは継続的に運用・改修されるため，セキュリティを維持するには想定脅威や対策に関する情報や見解を明確化・文書化しておく必要がある．
- 未知の脅威には対処できないので，脅威の存在を知る必要がある．
- 一方で，攻撃手段や脆弱性の可能性は無数にある．攻撃や脆弱性への対処に金銭や時間，労力などのコストを無尽蔵にかけることはできないので，少なくとも優先度付けのために，想定被害などリスクの大きさを概算する必要がある．

脅威分析は，運用中に攻撃が行われたり脆弱性が見つかったりしたときに実施してもよいが，運用開始前やできれば設計・開発の早い段階で実施するのが望ましい．システムが実際に攻撃にさらされないうちに被害を未然に防ぐことができ，対処のための設計や機能の変更にも低コストで対応できるからである．

3.1 脅威分析の手順

典型的な脅威分析は次のような手順で実施される[*2]．

(1) システムのモデル化：まず分析対象のシステムをモデル化する．一般のシステム分析と同様にプロセス・機能を分離し整理するが，脅威分析においてはデータおよび関与するアクター（主体）にも着目する．データの操作や流出・漏洩が攻撃の手段でもあり被害でもあること，および異なるアクターの間の信頼境界をまたぐときに攻撃を受けやすいことが理由である．保護資産およびその価値の明確化もしばしば行われる．

(2) 脅威の同定：システムモデルに基づいて潜在的な脅威のリストを作成する．すなわち，アクターやプロセス，データが騙される・改変される・漏洩するといった，もしも起きたならばセキュリティ侵害になりそうな事象を（攻撃が可能か・現実的かはさておき）列挙する．

(3) リスク算定：列挙した個々の脅威についてリスクの度

[*2] ここでは脅威モデリング手法 [12], [13] に基づいて手順を記すが，他の手法でも概ね同様の手順を経る．本稿の提案もベースとなる脅威分析手法に依らず適用可能である．

合いを算定する．典型的には，脅威の尤度（likelihood）と影響度（impact）の積をリスク度とする．

$$\text{リスク度} = \text{尤度} \times \text{影響度}$$

尤度は攻撃の容易さ・実現可能性や被害が起きる確率・条件などから，影響度は失われる保護資産の価値や想定される被害の大きさから，それぞれ算定する．

(4) 対処検討：最後に個々の脅威への対処方法を検討する．算定したリスク度が対処すべき脅威の優先度や対処にかけるコストの基準となる．対処方法には，攻撃を可能にしている脆弱性を修正する，回避策・緩和策を取る，マニュアル等で注意喚起する，などの選択肢がある．

脅威の尤度を検討・算定する際には**攻撃ツリー** がしばしば使われる．攻撃ツリーとは，個々の脅威（たとえば攻撃のゴール）を根ノードとし，親ノードを達成する条件や攻撃方法を子ノードとしてまとめた木構造データである．攻撃ツリーの例を **図 2** に示す．攻撃ツリーを構築することで脅威の達成条件を定性的に検討・表現することができ，各ノードの尤度を何らかの尺度・算定方法で決めることで最終的に根ノードの尤度を算定できる．

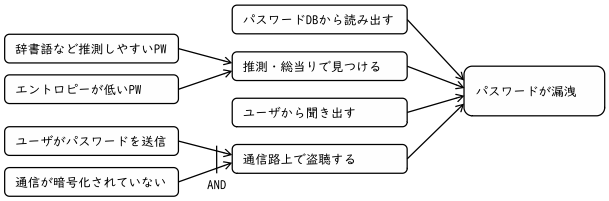

図 2 攻撃ツリーの例
Fig. 2 Example of attack tree

3.2 脅威分析における課題

一般に脅威分析を行う際には主に次のような課題がある．

- セキュリティの専門的知識やスキルが必要．攻撃ツリーを構築するには，そもそも攻撃手段を知っているか思いつく必要があり，攻撃者視点の知識やスキルがないと難しい．
- 対象システムに対する知識が必要．特に脆弱性の可能性やセキュリティ対策の十分性まで分析するには，設計や実装の詳細レビューやペネトレーションテストが必要である．また実施できたとしても定量化は難しい．
- 被害・影響度の想定や解釈が難しい．保護資産の価値や想定被害の大きさは機械的には決められず，社会的・組織的な視点で考える必要がある．特に評判やプライバシーなどの価値を客観的に定めるのは非常に難しい．
- 多大な時間と労力を要する．特に脅威や攻撃手段は潜在的には無数に考えられるため，何らかの基準で限定しないと非常に手間がかかってしまう．

4. 提案手法

現実に開発・運用される機械学習利用システムのセキュリティを高めるには，機械学習特有の攻撃の脅威を把握することが重要である．なぜならそうした攻撃は発見されて日が浅く，見落とされがちだからである．

提案手法の特徴は，分析で扱う脅威（攻撃）を 2.2 節で挙げた機械学習特有の 4 つの攻撃種別に限定することで，3.2 節で挙げた課題のうち「セキュリティの専門的知識やスキルが必要」と「多大な時間と労力を要する」の 2 つを緩和する点にある．我々は 4 つの攻撃種別に限定して攻撃ツリーの構築を行い，1 つの**合成攻撃ツリー**によって尤度の算定を可能にした．また攻撃ツリーの構築にあたっては，代理モデルを介した攻撃の可能性などモデルへのアクセス性を取り入れ，機械学習システムへの攻撃に関する専門的な知識がなくても特有の攻撃条件を分析に組み入れられるようにした．提案手法は機械学習システムに対する脅威分析のすべての課題を解決するものではないが，分析対象システムの設計者や開発者が比較的容易に大まかな分析結果を得ることを可能にする．

4.1 機械学習システム向け合成攻撃ツリー

この節では提案手法の最大の特徴である，機械学習システムに特化した合成攻撃ツリーについて述べる．

2.2 節で挙げた機械学習特有の攻撃種別とデータ種別との関係を **表 1** にまとめる．表の左半分の攻撃手段は攻撃尤度に，右半分のゴールは攻撃影響度に，それぞれ寄与する．

このとき訓練済みモデルへのアクセス性が問題となる．2.3 節で述べたように，モデルへのアクセス性にはホワイトボックス，ブラックボックスに加えて，代理モデルを介したアクセスを考慮する必要がある．我々はこれらのアクセス性を達成するための条件について考察し，攻撃者の能力とアクセス性の間に**表 2** に示す依存関係があると考えた．

これらの攻撃特性およびアクセス性に基づき，4 つの攻撃種別をまとめて扱える合成攻撃ツリーを構築した．**図 4** にその概略を示す．ただし，実際の分析に使う攻撃ツリーの一部詳細が図では省略されている．省略したのは，たとえば攻撃者による訓練の難易度などの諸条件やモデルアクセス性の違いによる尤度への影響などである．

4.2 脅威分析の手順

提案手法による脅威分析は「モデル化」と「リスク算定」の 2 段階に分けて実施される．

4.2.1 対象システムのモデル化

脅威分析の最初の工程では，対象システムをモデル化する．分析者は，2.1 節で述べたパイプライン（図 1）に対象システムで使われるデータと関与するアクターを明らかに

表 1　機械学習システムに特有の攻撃の特性（括弧内は必須ではないが重要な攻撃手段を表す）
Table 1　Characteristics of attacks specific to machine learning systems. (Secondary conditions in parentheses)

攻撃種別	攻撃手段			ゴール		
	訓練データ	訓練済みモデル	クエリーデータ	訓練データ	訓練済みモデル	出力データ
モデルの複製・抽出		アクセス	（操作）		漏洩	
訓練データ推定		アクセス	（操作）	漏洩		
敵対的サンプル		（アクセス）	操作			変更
ポイズニング	操作	（アクセス）			変更	変更

表 2　攻撃者の能力から導出するモデルへのアクセス性
Table 2　Model accessibility from attacker capability

攻撃者の能力		モデルへのアクセス性
訓練済みモデルを入手可能	→	ホワイトボックス攻撃が可能
↓		
モデルへ多量のクエリーが可能	→	ブラックボックス攻撃が可能
↓		
訓練データを入手・模倣可能	→	代理モデルを介した攻撃が可能

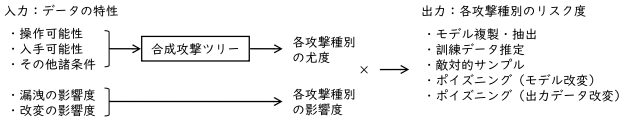

図 3　リスク算定の流れ
Fig. 3　Risk derivation flow

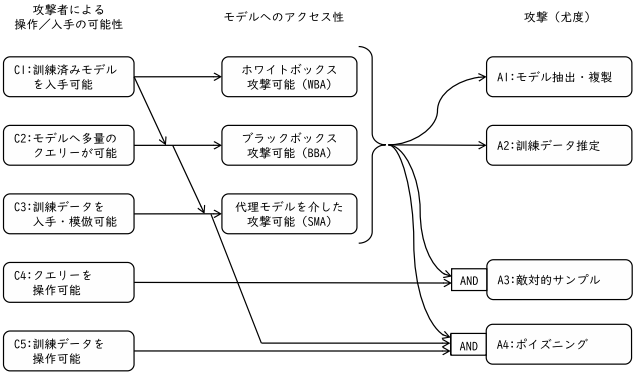

図 4　機械学習システム向け合成攻撃ツリー
Fig. 4　Composed attack tree for machine learning systems

する．

4.2.2　攻撃ツリーによるリスク算定

次に 4 つの攻撃種別に対するリスク算定を行う．リスク算定の入出力と処理の流れを**図 3** に示す．

分析者は，各データ項目に対し，攻撃者による操作／入手の可能性，および漏洩／改変の影響度をそれぞれ採点する．いずれも 1（低い）から 5（高い）の 5 段階で採点するものとした．できるだけ安定して素早く採点できるように採点基準のガイド文書も用意した（ガイド文書の例を付録 A.1 に示す）．

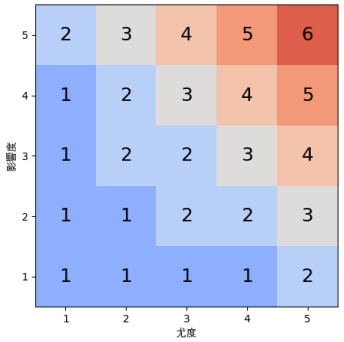

図 5　リスク度の算定
Fig. 5　Calculation of risk score

採点が終われば，4 つの攻撃種別のリスク度は自動的に算定できる．攻撃尤度の算定には 4.1 節の攻撃ツリー（図 4）を使用する．その後，攻撃尤度とその攻撃によって侵害されるデータの影響度からリスク度を算定する．尤度の条件結合およびリスク度の算定は次の式により行うものとした．

$$\text{尤度}_{OR} = \max(\text{尤度}_1, \ldots, \text{尤度}_n) \tag{1}$$

$$\text{尤度}_{AND} = \min(\text{尤度}_1, \ldots, \text{尤度}_n) \tag{2}$$

$$\text{リスク度} = \lfloor \text{尤度} \times \text{影響度}/5 \rfloor + 1 \tag{3}$$

$$\text{ただし, 尤度} \in \{1, \ldots, 5\}$$

$$\text{影響度} \in \{1, \ldots, 5\}$$

各尤度と影響度から求められるリスク度の値は **図 5** のようになる．リスク度は 1 から 6 の 6 段階となるが，尤度・影響度ともに高い場合を際立たせるため敢えてこのような算定法を選んだ．[*3]

ポイズニング攻撃については，同じ尤度に対して 2 つの影響度（訓練済みモデルの改変と出力データの改変）を区別し，それぞれのリスク度を算定するものとした．

各攻撃種別のリスク度は，3 種類のモデルへのアクセス性（ホワイトボックス，ブラックボックス，代理モデル媒介）のそれぞれについて算出されるが，本稿では単純化のため省略し，3 種のうち最大のものを代表として記載した．

攻撃種別とアクセス性の組み合わせによっては脅威と見

[*3] なおリスク度を比例尺度として計算するのは便宜的な扱いであり，あくまで順序尺度として扱うべきである．

表 3　異なる対象に対する初回試行の評価
Table 3　Evaluation of the first trials on different targets.

分析者	A	B	C	D		
評価項目	評価結果				μ	σ
作業時間（分）	60	70	80	110	80	21.6
モデル化の理解度	3	5	5	4	4.25	0.96
モデル化の容易さ	4	2	3	5	3.50	1.29
リスク算定の理解度	4	3	5	3	3.75	0.96
リスク算定の容易さ	4	2	3	4	3.25	0.96

μ は相加平均，σ は標本標準偏差．

表 4　異なる対象に対する分析結果
Table 4　Result of the first trials on different targets.

対象システム	W	X	Y	Z		
脅威項目	リスク度				μ	σ
モデルの複製・抽出	5	4	3	5	4.25	0.96
訓練データ推定	5	2	4	5	4.00	1.41
敵対的サンプル	1	4	6	1	3.00	2.45
ポイズニング（モデル）	2	4	5	2	3.25	1.50
ポイズニング（出力）	1	4	5	2	3.00	1.83

μ は相加平均，σ は標本標準偏差．
W: SNS メッセージからの商品推奨
X: 所有写真からの関心推奨
Y: 外観画像からの製品欠陥検査
Z: クレジットカード使用履歴からの不正検知

なさないものもある．その代表例は代理モデルを介した訓練データ推定で，代理モデル作成の時点でブラックボックス攻撃での訓練データ推定が達成できていると考えられるため，新たに脅威とは見なさないものとした．

5.　試行・評価

本節では提案する脅威分析手法の試行とその評価について述べる．

試行には提案手法を表計算ソフトウェア上で実装した分析シートを利用した．対象システムをモデル化してパイプラインとデータを明らかにした後，分析シート上で各データ種別に対する入手可能性，操作可能性，および影響度を採点すると，各攻撃種別のリスク度が算出される．

試行は，著者ら自身のうちセキュリティや機械学習については一定の専門性をもつが脅威分析については経験の浅い複数名が行った[*4]．このため客観的な評価とは言えないが，予備的な評価と理解されたい．

5.1　異なる対象に対する試行

最初の試行ではそれぞれの分析者が対象システムを任意に選び，分析した．対象としたシステムについては詳述しないが，機械学習利用システムとしてよく見られる種類のものである．

試行作業にかかった時間と作業後に回答したアンケートの主観評価項目 4 つを**表 3** に示す．初めての分析にも関わらず作業時間はおよそ 80 分 ± 20 分であり，1 日から数日程度を要する一般的な脅威分析よりずっと短い時間で分析が可能であった．主観評価は 1 (最低) から 5 (最高) の 5 段階である．平均はいずれも 3 を上回っており，分析作業は概ね理解しやすく容易といえるが，モデル化よりリスク算定のほうが難しい傾向がみられる．

分析結果を**表 4** に示す．分析対象のシステムは表の下部に示したとおりだが，その詳細は割愛する．「自分の分析結果に納得できるか」というアンケートに対して全分析者から「納得できる」との回答を得ており，主観的には納得性のある分析ができたといえる．

[*4] 以降では分析者 A〜D および E〜H としているが，先入観を排するため 5.1 節と 5.2 節で順序を入れ替えた．

内容を見ると，W と Z に顕著なように，訓練データの機微性が高い事例ではモデル抽出や訓練データ推定のリスクが高く算定されている．また敵対的サンプルやポイズニングについては，攻撃の可能性もさることながら，改変された訓練済みモデルや出力データを誰がどのように使うかによって改変影響度が大きく異なり，欠陥の大規模な見逃しがそのまま不良品の出荷につながる Y ではリスクが大きく，推奨商品のユーザへの提示に留まる W ではリスクが小さくなっている．

ただし，分析内容を詳細に検討すると個々の分析者による考え方の違いも影響を与えていることがわかる．たとえば W と X はどちらも関心推定やそれに基づく推奨だが，X の分析者は誤った推定や推奨がサービス自体の価値や存続に大きな影響を与えうると評価したため，敵対的サンプルやポイズニングのリスクが W より高く算定されている．また Y の分析者は，データ送信経路での改竄の可能性を最大限高く見積もったため，データ操作によって生じる攻撃のリスクが高く算定されている．

5.2　同一対象に対する試行：商品推奨システム

次に同一の対象システムに対しての分析結果が分析者によってどう変わるかを確認するための試行を行った．

分析対象は前節の試行における W であり，ソーシャルネットワークサービス (SNS) へ投稿したメッセージ群に基づいて投稿者の好みそうな商品を推薦する仮想的なシステムである．自然言語処理によって投稿者の特性を特徴量ベクトル化し，協調フィルタリングによって類似した投稿者の好みから商品を予測する．訓練や予測は SNS とは独立したクラウドサービスとして運用され，ユーザはインターネットを介して推薦を受ける．過去の商品選択は必ずしも公知ではないが，投稿メッセージ群は公知である（少なくとも攻撃者は知ることができる）ものとした．

分析結果を**表 5** に示す．この結果からは，分析者によって算定したリスク度に大きなばらつきが見られる．分析

シートを精査したところ，ばらつきの原因は対象システムの詳細や影響度に対する解釈の違いと考えられる．たとえば分析者 E のみ突出して訓練データ推定のリスク度が高い．その原因は，訓練データのうちの過去の商品選択データは機微である可能性やクエリーデータの機微メッセージが再訓練に使われる可能性を考慮し，訓練データの漏洩影響度を高く採点したことであった．この考えが妥当かどうかの判断はより詳細な検討を要するが，いずれにせよ現時点の提案手法が客観的・安定的にリスクを算定できるとは言い難いことを本試行は示している．

表 5　同一の対象に対する試行結果
Table 5　Result of trials on the same target.

分析者 / 脅威項目	E	F	G	H	μ	σ
	リスク度					
モデルの複製・抽出	5	3	5	5	4.50	1.00
訓練データ推定	5	2	2	2	2.75	1.50
敵対的サンプル	1	5	4	1	2.75	2.06
ポイズニング（モデル）	2	5	4	3	3.50	1.29
ポイズニング（出力）	1	5	4	3	3.25	1.71

μ は相加平均，σ は標本標準偏差．

6.　考察と提言

提案手法は機械学習への攻撃特有の知識を必要とせず，試行結果によれば短時間での脅威分析を可能にするが，分析結果にはばらつきが見られ客観的・安定的ではない．その理由は，脅威分析の課題（3.2 節）のうち「セキュリティの専門的知識やスキルが必要」と「多大な時間と労力を要する」の 2 つは緩和できたが，「被害・影響度の想定や解釈が難しい」や「対象システムに対する深い知識が必要」の 2 つは未解決であるためと考えられる．しかし，後の 2 つは脅威分析の本質的な課題であり，解決は難しい．

提案手法は，具体的な機械学習アルゴリズムやデータ内容などの詳細には踏み込まず，外形的な情報だけで分析を可能にする．しかし実際に攻撃がどの程度可能か調べたり，機械学習固有の対策（たとえば敵対的サンプルに対して頑健なモデルにするなど）を実施したりするには，機械学習セキュリティの専門家が詳細に踏み込んで分析・検討する必要がある．

こうした事情を踏まえると，機械学習利用システムに対する脅威分析は次のようにすべきと考えられる[*5]：脅威分析は，一般の設計者・開発者が自ら行う簡易分析と専門家が行う詳細分析の 2 段階に分ける．簡易分析はリスク上限の把握と非専門的な対処（機微なデータを扱わない，モデルへのアクセスを制限する，など）のために行い，その後詳細分析や対策によってリスクを下げていく．提案手法は主に簡易分析に利用するとともに，詳細分析に向けて設計者・開発者とセキュリティ専門家の間でのコミュニケーションツールとして活用する．

最後に，提案手法の設計や試行を通じてわかったこととして，以下の 3 点はリスクに広汎な影響を与えるのでリスク抑制に効果的といえる．

- （機微でなく，かつ操作されにくい）クリーンな訓練データを使うべき．訓練データやモデルの漏洩影響度を下げるため．
- 出力データの利用にはフェールセーフを設けるべき．モデルや出力データの改変影響度を下げるため．
- 訓練済みモデルへのアクセスを制限するべき．ホワイトボックス攻撃や代理モデルを介した攻撃の尤度を下げるため．

7.　関連研究

機械学習のセキュリティ脅威を概観する研究は近年盛んに発表されている．Biggio と Roli [1] は攻撃者のゴール，知識，および能力に基づいて敵対的サンプルやポイズニングなどの攻撃を分類し，評価した．このような分類方法は後のサーベイ・体系化研究 [11], [14] でも使われている．ただし，これらの分析は攻撃手法を分類・体系化することが目的で，定性的または部分的であり，機械学習利用システム全般に対し脅威分析の手法を提示するものではない．

機械学習利用システムの品質に関し工学的なアプローチも試みられているが，発展途上である．AI プロダクト品質保証コンソーシアム（QA4AI コンソーシアム）は 2019 年 5 月に AI プロダクト品質保証ガイドライン [15] を公開した．しかしセキュリティに関する言及は少なく，注意喚起に留まる．セキュアな深層学習工学の提案 [16] も存在するが，具体的な手法は述べられていない．

機械学習利用システムに対する脅威分析について高橋 [17] は，脅威モデリング手法 [12], [13] を機械学習システムに適用した結果，ある程度の分析が可能であったと報告し，同手法における脅威の 6 分類 (STRIDE) に加えて不確定性を考慮することを提案した．また宇根と井上 [18], [19] は金融機関で見られる機械学習の典型的なユースケースをいくつか挙げ，定性的な脅威分析を行った．この分析では，機械学習システムのパイプラインを大まかにモデル化した上で，データ提供者や訓練実施者などのアクターを明示し，アクターの利害関係によるいくつかのパターンに分類して考慮すべき脅威を論じている．

このように，我々の知る限り機械学習システムに対する脅威分析を扱った研究は少なく，具体的な脅威分析手法の提示，および機械学習特有の攻撃ツリーによる分析は本提案手法の特筆すべき点といえる．

[*5] ただし機械学習に限らない一般的な攻撃の脅威分析や対策は別途必要である．

8. むすび

本稿では機械学習特有の攻撃に特化した脅威分析手法を提案した．その特徴は，攻撃手段の間の依存関係やモデルへのアクセス性を明らかにし，それを表現した合成攻撃ツリーを示した点である．また暫定的な試行評価を行い，少なくとも主観的に納得できる分析結果を短時間で導き出せることがわかった．一方で分析結果にはばらつきが見られるため，手法の改善や利用方法の工夫が今後の課題である．

謝辞 本稿の関連研究に関し，国立情報学研究所の吉岡信和准教授に貴重な助言をいただいた．

参考文献

[1] Biggio, B. and Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning, *Pattern Recognition*, Vol. 84, pp. 317 – 331 (online), DOI: https://doi.org/10.1016/j.patcog.2018.07.023 (2018).

[2] Tramèr, F., Zhang, F., Juels, A., Reiter, M. K. and Ristenpart, T.: Stealing machine learning models via prediction apis, *25th USENIX Security Symposium*, pp. 601–618 (2016).

[3] Fredrikson, M., Jha, S. and Ristenpart, T.: Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures, *ACM CCS*, pp. 1322–1333 (online), DOI: 10.1145/2810103.2813677 (2015).

[4] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R.: Intriguing properties of neural networks, *arXiv preprint arXiv:1312.6199* (2013).

[5] Goodfellow, I., Shlens, J. and Szegedy, C.: Explaining and Harnessing Adversarial Examples, *ICLR*, (online), available from ⟨http://arxiv.org/abs/1412.6572⟩ (2015).

[6] Biggio, B., Nelson, B. and Laskov, P.: Poisoning Attacks Against Support Vector Machines, *ICML 2012*, pp. 1467–1474 (online), available from ⟨http://dl.acm.org/citation.cfm?id=3042573.3042761⟩ (2012).

[7] Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C. and Roli, F.: Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization, *CoRR*, (online), available from ⟨http://arxiv.org/abs/1708.08689⟩ (2017).

[8] Kusano, K., Takeuchi, I. and Sakuma, J.: Privacy-preserving and Optimal Interval Release for Disease Susceptibility, *ACM Asia CCS*, ACM, pp. 532–545 (2017).

[9] Hidano, S., Murakami, T., Katsumata, S., Kiyomoto, S. and Hanaoka, G.: Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes, *Annual Conference on Privacy, Security and Trust (PST)*, IEEE, pp. 115–11509 (2017).

[10] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B. and Swami, A.: Practical Black-Box Attacks Against Machine Learning, *ACM Asia CCS*, pp. 506–519 (online), DOI: 10.1145/3052973.3053009 (2017).

[11] Papernot, N., McDaniel, P., Sinha, A. and Wellman, M. P.: SoK: Security and Privacy in Machine Learning, *2018 IEEE European Symposium on Security and Privacy (EuroS P)*, pp. 399–414 (online), DOI: 10.1109/EuroSP.2018.00035 (2018).

[12] Howrad, M. and LeBlanc, D.: *Writing Secure Code*, chapter 4, Microsoft Press, 2nd edition (2002).

[13] Swiderski, F. and Snyder, W.: *Threat Modeling*, Microsoft Press (2004).

[14] Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S. and Leung, V. C. M.: A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View, *IEEE Access*, Vol. 6, pp. 12103–12117 (online), DOI: 10.1109/ACCESS.2018.2805680 (2018).

[15] AI プロダクト品質保証 (QA4AI) コンソーシアム（編）：AI プロダクト品質保証ガイドライン 2019.05 版 (2019).

[16] Ma, L., Juefei-Xu, F., Xue, M., Hu, Q., Chen, S., Li, B., Liu, Y., Zhao, J., Yin, J. and See, S.: Secure Deep Learning Engineering: A Software Quality Assurance Perspective, *arXiv preprint arXiv:1810.04538* (2018).

[17] 丸山宏，高橋正和ほか：機械学習システムのセキュリティ，情報処理学会第 81 回全国大会 パネル討論 (2019).

[18] 宇根正志，井上紫織：機械学習システムの脆弱性と対応策にかかる研究動向について，コンピュータセキュリティシンポジウム 2018 論文集， Vol. 2018, No. 2, pp. 193–200.

[19] 井上紫織，宇根正志：金融分野で活用される機械学習システムのセキュリティ分析，IMES Discussion Paper 2019-J-1, 日本銀行金融研究所 (2019).

付　　録

A.1 データ項目に対する採点ガイドの例

本稿で述べた提案手法の試行において，関連データ項目の漏洩影響度を採点するにあたり分析者に示すガイド文章を下記に示す． 操作影響度，入手可能性，操作可能性についても同様のガイド文章を用意した．

漏洩の影響度

もしもこのデータが流出・漏洩したとしたら，その影響度はどの程度ですか？
すなわち，このデータが漏洩や推定により攻撃者の手にわたったら，その結果起きうる被害はどの程度の大きさですか？下記のスコアを目安に記入してください．（複数の個人・組織に被害が及びうる場合は合計して考えてください．）

5 人間の健康や生命への被害，大きな物理的被害，または個人や組織への甚大な経済的被害（目安：個人 1 千万円，組織 1 億円以上）

4 個人のプライバシーへの被害，軽微な物理的被害，または個人や組織への経済的被害（目安：個人 10 万円，組織 100 万円以上）

3 個人の軽微なプライバシーへの被害，または個人や組織の機微な情報漏洩・推定等の被害（目安：個人 1000 円，組織 1 万円以上）

2 個人・組織内の一般的な情報の漏洩・推定

1 公開データや容易に入手・類推が可能なデータで，機微度は低い

（よくわからない場合や自信がない場合は想定範囲内で最高のスコアを付けてください．）