Computer Security Symposium 2018
22 - 25 October 2018

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

宇根 正志† 井上 紫織†

概要：近年，金融を含む幅広い分野において，人工知能を活用した新しいサービスが開発・提供されはじめている．そうしたサービスを安全に提供するうえで，機械学習の機能を有する IT システム，すなわち，機械学習システムのセキュリティに配慮することが重要である．本稿では，機械学習システムにおける主な脆弱性と対応策にかかる最近の研究動向を説明するとともに，機械学習システムにおけるセキュリティにかかる今後の課題について考察する．

キーワード：機械学習システム，人工知能，脆弱性，セキュリティ

Research Trends on Vulnerability and Countermeasures in Machine Learning Systems

Masashi Une† Shiori Inoue†

Abstract: New services utilizing the artificial intelligence have been recently developed and launched in various fields including the financial sector. In order to securely provide with such services, it is important to take into account the security of IT systems with machine learning functions, i.e., machine learning systems. This paper will describe recent research trends on vulnerability and countermeasures in such systems. It will also discuss and show future research topics relating to the security of machine learning systems.

Keywords: artificial intelligence, machine learning system, security, vulnerability

1. はじめに

近年，金融を含む幅広い分野において AI を活用した新しいシステムやサービスの開発･提供が注目を集めている．AI は，一般に，推論，認識，判断等，人間と同様の知的な処理能力をもつコンピュータ・システムやその技術分野を指すことが多い．AI が人間と同様の知的な処理能力を実現・発揮するためには，画像や音声等を認識し，それに基づいて判断・予測等を行う必要があり，通常そのためのツールとして機械学習（machine learning）が用いられる．現在，深層学習をはじめ，さまざまなタイプの機械学習の手法について実用化に向けた研究開発が活発となっており，技術面の検討のみならず，それらを活用したシステムの開発にかかるガイドラインの策定や，社会・経済に及ぼす影響に関する検討も盛んに行われている．

金融分野において新しい技術を導入し活用する場合，当該技術やそれを実装したシステムのリスクに応じたセキュリティ対策を講じる必要がある．これは，機械学習の機能を実装したシステム（機械学習システム）においても同様である．機械学習では，通常，学習モデルにデータを入力して（学習済みの）判定・予測エンジンを生成するとともに，そのエンジンによってデータの判定・分類や予測を実行する．こうしたシステムで取り扱われるデータ，学習モデルや判定・予測エンジンの機密性や完全性等を分析・評価し，そのシステムのビジネス要件が充足されているかを確認しておく必要がある．そのためには，情報システム一般に存在する脆弱性やそれを悪用した攻撃に加え，機械学習に特有の脆弱性等も把握しておくことが重要である．

本稿では，機械学習システムの主な脆弱性と攻撃手法，それらへの対策手法について最近の研究成果をサーベイする[1]. 2 節では, 機械学習システムの構成モデルを設定し，システム・セキュリティの観点から想定される主な脅威やセキュリティ対策の方向性を整理する．3 節では，最近の研究成果を参照しつつ，機械学習に特有の脆弱性，それらを悪用した主な攻撃手法等を説明する．4 節では，主な対策手法とその有効性にかかる評価手法を説明し，機械学習システムを安全に活用していくうえでの課題等を考察する．

本稿で示されている意見は，筆者ら個人に属し，日本銀行の公式見解を示すものではない．また，ありうべき誤りはすべて筆者個人に属する．

2. 機械学習システムの構成と主な脅威

2.1 機械学習システムの構成

機械学習システムは，一般に，次の 4 つのエンティティによって構成される．①訓練データと学習モデルを用いて判定・予測エンジンを生成する訓練実行者，②訓練実行者から判定・予測エンジンを受け取り，判定・予測を実行する判定予測実行者，③判定・予測エンジン生成やデータの

† 日本銀行金融研究所情報技術研究センター
Center for Information Technology Studies (CITECS), Institute for Monetary and Economic Studies (IMES), Bank of Japan

判定・予測を依頼するシステム利用者，④訓練データを訓練実行者に提供する訓練データ提供者である．判定・予測エンジンの生成や判定・予測における主な処理の流れは次のとおりである（図 1 を参照）。

(A)訓練データ提供者は，訓練データの元になるデータを収集する．その後，システム利用者と協力しつつそのデータを適宜加工するとともに，必要に応じてラベル（訓練データにかかる判定結果等を表すデータ）を付加したうえで訓練実行者に提供する．

(B)訓練実行者は，訓練データを学習モデルに適用し判定・予測エンジンを生成する．

(C)訓練実行者は，生成した判定・予測エンジンを判定予測実行者に提供する．

(D)システム利用者は，判定・予測を行いたいデータを判定予測実行者に提供する．

(E)判定予測実行者は，上記(D)でシステム利用者から受け取ったデータを判定・予測エンジンに適用し，判定・予測を行う．

(F)判定予測実行者は上記(E)の判定・予測結果をシステム利用者に提示する．

(G)システム利用者は，上記(F)の判定・予測結果等を訓練データ提供者に還元する場合がある．例えば，訓練データ提供者は，判定・予測結果が誤っていた場合，それを修正し正しいラベルとして訓練実行者に与え，判定・予測エンジンの改善を図るケースが考えられる．

上記(A)～(C)が訓練フェーズ，上記(D)～(G)が判定・予測フェーズにそれぞれ対応する．(A)における訓練データ等の提供に関しては，訓練データが機微な情報の場合，マスキングの実施や暗号化等を行うケースが考えられるが，ここでは分析を単純化するために，そうした処理が完了したデータが訓練実行者に提供されるものとする．

2.2 セキュリティ目標

機械学習システムのセキュリティ目標として，システムで取り扱われるデータや機能の機密性（confidentiality）・完全性（integrity）・可用性（availability）の達成が求められる[2][3]．機械学習システムの場合，保護対象となりうるデータや機能は，①訓練データ，②学習モデル，③判定・予測エンジン，④判定・予測を行う対象となるデータ（判定・予測用データ），⑤判定・予測用データを用いた判定・予測エンジンの出力，⑥システム利用者が訓練データ提供者に還元するデータ（還元データ）である．

例えば，訓練データに着目すると，機密性の観点からは，訓練データ提供者にかかる機微な情報（個人情報等）が含まれているなどの場合，そうしたデータの盗取を防ぐ必要がある．完全性の観点からは，訓練データの改変や不当なモデルの生成（機能の改変）が判定・予測に大きな影響を与える可能性がある場合，それらを防ぐ必要がある．可用性の観点からは，訓練データを訓練実行者に対して大量に送信する攻撃が行われ，訓練実行者の機能が低下する可能性がある場合，そうした攻撃を防ぐ必要がある．各保護対象について 3 つのセキュリティ特性（機密性，完全性，可用性）の要否（およびその達成度合い）を検討したうえで，必要と判断した特性に関してどのようなセキュリティ対策を講じるかを検討することが求められる．

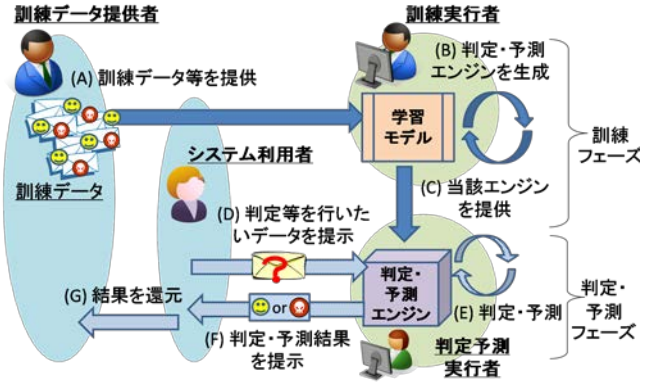

図 1　機械学習システムの構成（イメージ）

2.3 攻撃と対策の方向性

セキュリティ対策を検討するうえで，想定される攻撃を洗い出すことが必要である．各エンティティ，および，エンティティ間の通信路が攻撃箇所となりうる．

(1) 各エンティティへの攻撃

各エンティティへの攻撃として，それぞれが取り扱うデータ，学習モデル，判定・予測エンジンに関する情報の盗取・改変・偽造に加え，学習モデルや判定・予測エンジンを行うサーバを停止させるなどの妨害行為が想定される．

攻撃の手段としては，学習モデル等の脆弱性を悪用することがまず想定される．また，各エンティティの通信相手になりすます，あるいは，外部ネットワークとの接続部分の脆弱性を悪用し，当該エンティティの PC やサーバ等に不正にアクセスすることが想定される．また，機械学習システムに特有の事情として，訓練データ提供者が（意図せずに）不正な訓練データを入手し訓練実行者に送信する（その結果，不正な判定・予測エンジンが生成される）可能性もある．さらに，可用性にかかる攻撃として，大量のサービス要求を各エンティティに送信してサーバをダウンさせることなどが考えられる．

対策の方向性としては，機械学習システムに特有の脆弱性を軽減することが挙げられる．また，一般の情報システムにおいても想定されるものとして，主に機密性と完全性の観点から，①通信相手の認証，②各エンティティの PC やサーバ等（各種データ等を格納）へのアクセス制御，③保護対象の各種データの暗号化，④データベース上の各種データの改変（データベースへの入力前の改変は除く）の

検知等が挙げられる．さらに，可用性の観点から，⑤コンテンツ配信ネットワーク（Contents Delivery Network : CDN）の利用等が考えられる．

通常，上記①～⑤の対策を十分に実施すれば，攻撃者は各エンティティが保有するデータ等にアクセスできず，それらを攻撃に利用できないと想定可能となるほか，可用性を維持することができる．この場合，セキュリティ対策の検討においては，機械学習システムに特有の脆弱性に焦点を当てることとなる．もっとも，各エンティティへの不正侵入，マルウェアによる攻撃，訓練実行者等へのソーシャル・エンジニアリング攻撃等，いわゆるサイバー攻撃が今後一層高度化する可能性は否定できない．そのため，サイバー攻撃の高度化のリスクにも配慮し，各エンティティが保有するデータ等に攻撃者がアクセスするケースも想定して検討する必要もある．

(2) エンティティ間の通信路

エンティティ間の通信路では，両端のエンティティの通信を中継するように通信路にアクセスし（中間侵入攻撃），通信データの盗取，改変・偽造を試行することが想定される．また，大量のサービス要求を各エンティティに送信することによって通信路の帯域を制限するなどの通信の妨害も考えられる．通信データの盗取や改変・偽造への対策方針として TLS（Transport Layer Security）等の暗号プロトコルを活用することが考えられる．サービス妨害への対策方針としては，CDN の利用等が考えられる．これらは，いずれも情報システム一般において広く利用されており，機械学習システム特有のものではない．

3. 学習モデルにかかる脆弱性

3.1 攻撃者の能力にかかる前提

攻撃手法に関する個々の研究では，攻撃者の保有する情報や行動（攻撃者の能力）が異なっている．したがって，想定される攻撃者の能力をあらかじめ分類しておくことは，複数の攻撃手法のインパクトを横並びで比較するうえで有用である[4][5].

攻撃者の能力にかかる分類として，ホワイト・ボックスとブラック・ボックスが広く知られている．ホワイト・ボックスは，攻撃者が，対象とする判定・予測エンジンの構造やパラメータ（損失関数や重み等），エンジンの任意の入出力等，ほぼ完全な情報を得ることができる状況を意味する．一方，ブラック・ボックスは，こうした情報の入手や判定・予測エンジンへのアクセスに一定の制限が課せられている状況を意味する．ホワイト・ボックスとブラック・ボックスの境界は研究論文によって異なっている．

2.3 で説明したように，サイバー攻撃を想定するとホワイト・ボックスの状況を想定した対策が必要となるが，学習モデルや判定・予測エンジンが企業秘密として厳重に管理されている場合等では，ホワイト・ボックスの状況が実現する可能性は相対的に低く，まずは，ブラック・ボックスでの攻撃が焦点となる[4][6]. 攻撃者が利用する情報の種類として，これまでさまざまな分類が示されている．ここでは，最近の代表的な研究として文献[4]の分類を示す．攻撃者が利用する情報の組合せを次の 6 つに分類している．

【分類 1】判定・予測エンジンへのいくつかの入出力ペア（攻撃者が指定可能）
【分類 2】判定・予測エンジンの任意の入力データに対する加工された出力
【分類 3】判定・予測エンジンの任意の入出力ペア
【分類 4】判定・予測エンジンの任意の入出力ペア，訓練データ
【分類 5】判定・予測エンジンの任意の入出力ペア，学習モデルのネットワーク構造
【分類 6】判定・予測エンジンの任意の入出力ペア，学習モデルのネットワーク構造，訓練データ

比較的実現性の高い状況は分類 1～3 であり，まずは，攻撃者がこれらの情報を利用可能と想定したうえで有効な対策を検討することが重要である．

3.2 主な脆弱性と攻撃手法

学習モデル等にかかる脆弱性のうち，セキュリティと密接に関係すると考えられるものを整理すると，①訓練データにかかる情報の漏洩，②判定・予測エンジンにかかる情報の漏洩，③訓練データの変化による判定・予測エンジンの精度低下，④入力の変化による判定・予測の精度低下が挙げられる．また，これらを利用した攻撃については，いずれも，攻撃者が機械学習システムにかかる何らかのデータを入手して実施するものである．

(1) 訓練データにかかる情報の漏洩

学習モデルの構造，判定・予測エンジン，当該エンジンの入出力から，特定のデータが訓練データの一部であったか否かの情報や，訓練データの特性にかかる情報が漏洩しうるほか，訓練データ自体も推定されうる．

イ．特定のデータが訓練データの一部であったか否かの情報の漏洩

画像認識，個人の購買履歴，医療機関受診履歴を利用する一部の機械学習システムにおいて，特定のデータが訓練データの一部であったか否かにかかる情報が漏洩しうることを示す研究事例が知られている．これらは，訓練データにおける特定のデータの有無によって，生成される判定・予測エンジンの入出力関係が異なるという性質を利用している．

文献[7]では，複数の購買履歴等のデータを用いて複数の判定・予測エンジン（特定のデータが訓練データに含まれている場合のエンジンやそうでない場合のエンジン）を生成し，それらのエンジンの入出力を分析することができれ

ば，特定のデータが訓練データに含まれていたか否かを高い確率で推定可能であることを示している．具体的には，（推定対象以外の）訓練データの一部とそれらに対する判定・予測エンジンの出力等を利用し，70～90%の確率で入出力関係が再現可能なエンジンを生成するというものである．攻撃者は，生成した判定・予測エンジンを用いて特定のデータが訓練データに含まれるか否かを判定する．この攻撃は，対象となる判定・予測エンジンや学習モデルが秘匿されていた場合でも有効であり，3.1 の分類 4 に相当する情報を用いた攻撃といえる．また，当該研究は，クラウドが提供する一部の機械学習サービス（学習モデルや判定・予測エンジンの内容は秘匿）に適用可能であることを実証している．

ロ．訓練データの特性にかかる情報の漏洩

一部の音声認識のシステムにおいて，訓練データの大半が特定の方言を含む音声データであったか否かが推定されうるという事例や，通信データから通信サービスの種類を判定するシステムにおいて，特定の大手インターネット・サービス・プロバイダーのサーバからの通信データが訓練データの大半を占めていたか否かが推定されうるという事例の研究が知られている[8]．この研究では，学習モデルや判定・予測エンジンにかかる情報を用いて，訓練データの特性を判定する判定・予測エンジンを生成する手法を示すとともに，隠れマルコフ・モデルに基づく一部の音声認識のモデルと，サポート・ベクトル・マシーンに基づく通信サービスを識別するモデルへの適用例（90%程度の確率で判定に成功）を報告している．この攻撃は，3.1 の分類 5 の情報を用いた攻撃に相当する．

ハ．訓練データ自体の推定

判定・予測結果の確からしさを示す確信度（confidence value）が判定・予測エンジンの出力に含まれている場合，訓練データ自体が推定されうる．

一部の顔画像認識のシステム（確信度を出力するもの）において，訓練データとして個人の識別情報や顔画像等が使用されていた場合，判定・予測エンジンの入出力等から顔画像を推定する研究が知られている[9][10]．訓練フェーズでは個人の識別情報（例えば氏名）と顔画像を訓練データとして使用し，判定・予測フェーズでは，顔画像を判定・予測エンジンに入力することで，対応する個人の認識情報と確信度を出力として得る．こうしたシステムを対象に，判定・予測エンジンの複数の入出力から特定の個人の顔画像（あるいはその逆）を推定する．

これらの研究では，ソフトマックス関数を用いたニューラル・ネットワークやパーセプトロンに基づく一部の学習モデルに提案手法を適用している．ソフトマックス関数の場合，その入出力や内部の構造にかかる情報を利用可能なケースでは，提案手法によって推定した特定の個人の顔画像が 80～90%の確率で当該個人の（登録された）顔画像と一致すると判定された旨を報告している．この攻撃は 3.1 の分類 6 の情報を用いた攻撃といえる．

(2) 判定・予測エンジンにかかる情報の漏洩

判定・予測エンジンの入出力から，当該エンジンのパラメータ等にかかる情報が推定される事態が生じうる．例えば，文献[11]では，判定・予測をクラウド上で提供するサービス（判定・予測エンジンへの入力は当該サービスの利用者がネットワーク経由で送信）のうち，判定・予測結果の確信度をエンジンが出力するタイプについて，当該エンジンのパラメータを推定するとともに，ほぼ同一の入出力関係を実現する代替エンジンを生成する手法が提案されている．

ロジスティック回帰や決定木の手法に基づくモデルを利用する実際のサービスに当該手法を適用した場合，数百から数千の入出力ペアを用いて判定・予測エンジンのパラメータを推定することができれば，代替エンジンの生成に90%以上の確率で成功する旨も報告されている．この攻撃は，攻撃者が当該エンジンのパラメータや訓練データに関する知識を有していないことから，3.1 の分類 3 の情報を用いた攻撃に相当する。

(3) 訓練データの変化による判定・予測エンジンの精度低下

訓練データ（の分布）がノイズ等によってわずかに変化した際，それらによって生成される判定・予測エンジンが有意に変化し，誤った判定・予測が出力される場合がある[2][12][13][14]．その結果，判定・予測エンジンの精度が低下することになる．こうした脆弱性を悪用する攻撃として，不正な訓練データ等を学習モデルに入力し，攻撃者にとって都合のよい判定・予測エンジンを生成させるという攻撃がよく知られている[2][6][15][16]．こうした攻撃は，サポート・ベクトル・マシーン，ロジスティック回帰，ニューラル・ネットワークに基づく一部の学習モデル等に対して適用可能であることが示されている．各研究成果では，攻撃者が学習モデル等について事前に知識を有している状況のもとで，高い成功率を達成しうる不正な訓練データを探索する手法の検討に主眼が置かれており，3.1 の分類 5 あるいは分類 6 の情報を用いた攻撃といえる．

例えば，文献[17]は，（攻撃と疑われる）不正な通信か否かを判定したうえで，判定対象の通信データを訓練データとして使用して判定・予測エンジンを順次更新するタイプの不正通信検知のモデル（重心モデル）を攻撃対象としている．この研究では，与えられた訓練データに対して不正なデータを追加することにより，不正な通信か否かを判定する境界を徐々に移動させることができることを示したうえで，当該移動が最大となるようなデータを探索する問題を定式化しその解法を提案している．また，訓練データ全体のうち，どの程度のデータを改変すれば，（当該訓練データによって生成された）判定・予測エンジンにおいてどの

程度の確率で誤判定が発生するかについても関係性を示している．提案手法は，攻撃者が学習モデルと訓練データを知っているという状況を前提としたものであり，3.1 の分類 6 の情報を用いた攻撃に相当する．

文献[18]は，攻撃用の訓練データと本来の訓練データの差分を一定以下にするという制約のもとで，攻撃者が目標とする判定・予測エンジンとの差分を最小化するエンジンを生成するように，攻撃用の訓練データを探索する問題とその効率的な解法を提案し，サポート・ベクトル・マシーン，ロジスティック回帰，線形回帰に基づく一部の学習モデルへの適用事例を示している．文献[17]と同様に，攻撃者が学習アルゴリズムと訓練データを知っていることが前提であり，3.1 の分類 6 の情報を用いた攻撃といえる．

(4) 入力の変化による判定・予測エンジンの精度低下

判定・予測エンジンへの入力がノイズ等の影響によってわずかに変化した際，誤った判定・予測結果が出力される場合がある．こうした脆弱性を悪用して判定・予測エンジンに誤判定等を引き起こす攻撃が数多く提案されている．このような入力は adversarial example と呼ばれる．最近では，深層ニューラル・ネットワークに基づく機械学習のアルゴリズムを対象とした研究成果の発表が目立つ．

文献[19]では，深層ニューラル・ネットワークに基づく画像認識や手書き文字認識の一部のモデルを対象に，誤った判定結果が出力される入力データを探索する手法を提案している．こうした入力データは，例えば，訓練データ（画像）に一定のノイズが付加されたデータとして表現されたりする．提案手法は，目標とする（誤った）判定結果の出力を実現しつつ，その判定・予測結果の誤差を最小化する入力の近似値を探索するという問題を定式化するとともに，その近似解を効率的に求めるものである．深層ニューラル・ネットワーク等に基づく 4 種類の学習モデルに提案手法を適用したところ，訓練データに微小なノイズを付加した（誤判定を引き起こす）入力データを探索することができた．攻撃者は，学習モデルや判定・予測エンジンにかかる情報を入手することが前提とされており，3.1 の分類 6 の攻撃に相当する．

文献[15]では，深層ニューラル・ネットワークを用いた手書き文字認識の一部の学習モデルを対象に，攻撃用の入力データを探索する手法を提案している．提案手法では，攻撃者が訓練データ，学習モデル，判定・予測エンジンを事前に知っている状況を想定しており，3.1 の分類 6 の情報を用いた攻撃に相当する．そのうえで，入力データに付加されるノイズや改変が判定・予測エンジンの出力に及ぼす影響を示す関係式を構成し，意図した誤判定を引き起こす攻撃用の入力データを探索している．提案手法の有効性を実験で確認したところ，正規の入力データ（文字画像）を構成するピクセルのうち，平均で約 4%のピクセルに一定の改変を加えると，攻撃者が意図したクラスに約 97%の確率で誤判定させることができたとしている．

このほか，ある判定・予測エンジンにおいて誤判定等を引き起こしやすい入力は，同一の学習モデルによって生成された別のエンジンにおいても誤判定等を引き起こしうることも知られている．例えば，画像データの判定を行う一部の機械学習システムにおいて，判定・予測エンジンを複数の手法によりそれぞれ生成したうえで，訓練データとして使用した画像データに一定の処理を施して入力すると，複数のエンジンにおいて有意な確率で誤判定が発生した事例がある．

文献[19]では，ある特定の判定・予測エンジンにおいて誤判定等を引き起こす入力が，訓練データ，レイヤー数，重み減衰のパラメータ等を変更して生成した他のエンジンにおいても比較的高い確率で誤判定等を発生させる場合があることを示している．例えば，ソフトマックス関数を用いたニューラル・ネットワークにおいて，重み減衰のパラメータを変化させつつ複数の判定・予測エンジンを生成したうえで，あるエンジンにおいて誤判定等を引き起こす（攻撃用の）入力データを探索してそれを他のエンジンに適用したところ，10%～80%の確率で誤判定が発生した旨を報告している．

4. 攻撃への対策手法

本節では，対策の有効性を評価するための主な尺度を説明するほか，3. で整理した攻撃への主な対策手法とその有効性評価にかかる代表的な研究成果を紹介する．

4.1 評価尺度

既存の研究論文では，判定・予測エンジンの出力の正確性にかかる評価尺度として，①『不正』と判定された入力データのうち正しく判定したものの割合を示す適合率（precision），②不正な入力データ全体のうち正しく『不正』と判定したものの割合を示す再現率（recall），③入力データ全体のうち正しく判定したものの割合を示す正解率（accuracy）を用いるケースが多い．

また，ノイズや改変を加えた入力データによって誤判定等を引き起こす攻撃の場合には，ノイズや改変の度合いを評価の尺度とするケースもある．例えば，ノイズ等を加えた入力データと元の入力データとの距離（例えば，両データ間のユークリッド距離の平均値）を尺度とすることがある．当該距離が小さいほど不正な入力データとしての検知が困難になると考えられることから，攻撃としての有効性がより高いと考えることができる．逆に，対策を講じる側からみると，対策によって当該距離がより拡大するほど，攻撃を成功させるためにより多くのノイズや改変を入力データに加える必要が生じるという意味で，対策の効果が相対的に大きいといえる．

もっとも，研究論文によっては，これらの指標がすべて記載されているとは限らず，攻撃による判定・予測エンジ

ンへの影響度合いを横並びで比較することが困難な場合が少なくない．さらに，対策手法を講じた結果，判定・予測の精度が有意に低下してしまうと，対策実施の意味が失われることとなる．

文献[5]は，こうした点を指摘したうえで，攻撃手法や対策手法を提案・評価する論文においては，少なくとも，適合率に加えて，実用性とのトレードオフを評価する観点から，「判定・予測エンジンに入力されたデータのうち，正規の入力データを不正と誤って判定したものの割合」を示す偽陽性率（false positive rate）も研究論文に明記すべきであると提案している．また，判定のしきい値を変化させたときの適合率と偽陽性率の関係を表す ROC 曲線（Receiver Operating Characteristics Curve）を明記することによって，対策手法の有効性を示すことができればより望ましいとしている．

4.2 対策手法

対策としては，各攻撃手法を実行するうえで必要とされる情報を攻撃者に入手させないようにする，あるいは，そうした情報が攻撃者に入手されたとしても攻撃が成功しないように学習モデルや判定・予測エンジンを改良することが考えられる．前者は，攻撃者が利用できる情報を制限し，ブラック・ボックスの状況を実現するという対応である．後者は，ホワイト・ボックスの状況を前提として，学習モデル等のセキュリティを向上させるという対応である．

(1) 判定・予測エンジンや訓練データにかかる情報の盗取への対策

判定・予測エンジンのパラメータ等の情報の盗取・推定に対しては，攻撃に必要な（判定・予測エンジンの）出力や確信度等を攻撃者が入手できないようにすることが考えられる．

そうした手法の 1 つとして，確信度の値を丸めたうえで出力する，あるいは，確信度を出力しないようにすることが挙げられる[9]．もっとも，確信度を出力せず，判定結果として最も確からしいクラスのみを出力するように構成したとしても，より多くの入出力を攻撃者が入手することができる場合，判定・予測エンジンのパラメータ等を推定することができる場合があるという分析結果もある[11]．

また，暗号化したデータを入力として学習アルゴリズムや判定・予測エンジンに適用し，その出力（判定・予測結果）も暗号化したまま得られるようにするという手法が提案されている．こうした手法における暗号として，データを暗号化したまま加算・乗算が可能な準同型暗号が利用されている．例えば，文献[20]では、暗号化したデータのまま訓練や判定・予測を実行可能なニューラル・ネットワークのアルゴリズム（CryptoNets と呼称）が提案されている．文献[21]では，準同型暗号によって，暗号化したデータのまま深層学習を実現する手法が提案されている．これらの研究では，画像データ等を用いた実験により，一定の処理性能と判定・予測の精度を実現可能であることが示されている．

訓練データを復元・推定する攻撃に関しては，既存の攻撃を実施するうえで確信度が必要となることから，上記と同様に，確信度が出力されないように判定・予測エンジンを構成することが考えられる．

(2) 訓練データや判定・予測用データの操作への対策

主な対策手法は，学習モデルや判定・予測エンジンに入力される不正なデータを検知・排除するものと，不正なデータによる学習モデルや判定・予測エンジンへの影響を軽減・解消するものに大別することができる．これらの対策は，概ね，訓練データと入力データの両方に共通している．

不正な入力データによる攻撃への対策手法については，文献[5]において網羅的に検討されている．検討の対象として，ニューラル・ネットワークを用いた画像認識のモデルへの適用が想定される 10 種類の代表的な対策手法が挙げられている．これらのうち，不正な入力データを検知・排除する手法が 8 件であり，①ニューラル・ネットワークを利用するもの（3 件），②主成分分析を利用するもの（3 件），③入力データの分布の差異を利用するもの（2 件）に分類される．

各手法の有効性の評価においては，攻撃者の能力として，次の 3 種類が想定されている．すなわち，①攻撃者が当該対策手法にかかる情報を一切有していない（ゼロ知識攻撃＜zero knowledge adversary＞），②攻撃者が，対策の存在を知っているが，そのパラメータや対策手法のモデルの入出力を入手することができない（限定知識攻撃＜restricted knowledge adversary＞），③攻撃者が対策手法のパラメータやその入出力を入手することができる（完全知識攻撃＜perfect knowledge adversary＞）というものである．完全知識攻撃はいわゆるホワイト・ボックスの攻撃に類似したものといえる．

各対策手法の有効性は，各攻撃者の能力に応じて，各対策手法が実装された判定・予測エンジンに最新の攻撃手法（文献[22]で提案されているもの）を適用することによって評価している．評価結果は，当該攻撃がどの程度軽減されるかによって示されている．こうした評価のもとでは，偽陽性率を小さく抑えて実用性を確保すると同時に，適合率の向上と，入力データ間のユークリッド距離の拡張を実現する対策手法が高い評価を得ることになる．

評価の結果，多くの対策手法が既存研究で示されている攻撃手法に対して十分に効果を発揮しているとは言い難い状況であることが判明している（表 1 を参照）．完全知識攻撃の場合，いずれの対策手法も不正な入力データを十分に検知することができなかったほか，ゼロ知識攻撃の場合も，一部の手法（入力データの正規化）を除き，高い適合率と低い偽陽性率の両立が困難であるという結果が示された．

表 1　不正な入力データを利用する攻撃への主な対策と評価結果の概要

対策方針		（画像認証のモデルの場合） 対策手法の概要	各対策手法の有効性評価（概要）
不正な入力データの検知・排除	ニューラル・ネットワークの利用	不正な入力データを生成し，当該データを検知するための判定・予測エンジンを別途生成[23][24].	対策手法にかかる情報を用いることなく訓練データから不正な入力データを生成（ゼロ知識攻撃）．適合率が約 70%，偽陽性率が約 40%[23].
		学習途中の処理データから，不正な入力データを検知する判定・予測エンジンを別途生成[25].	ゼロ知識攻撃による入力データによって，適合率が約 80%，偽陽性率が約 30%.
	主成分分析の利用	入力データや学習途中の処理データから主成分を抽出．各成分の重み等を検知に利用[26][27].	ゼロ知識攻撃によって，適合率が約 60%，偽陽性率が約 40%[27].
		画像データの主成分を抽出し，不正な入力データを検知する判定エンジンを生成[28].	完全知識攻撃による入力データでは，入力データ間のユークリッド距離は拡張されず.
	入力データの分布の差異の利用	最大平均差異（maximum mean discrepancy）を利用[23].	ゼロ知識攻撃による入力データと正規の入力データの間に，分布の有意な差異はみられず（検知困難）.
		隠れ層の出力の尤度を算出．しきい値以下の場合，不正な入力データと判定[29].	一部のデータセットによる評価では，ゼロ知識攻撃によって，適合率が 20%以下.
学習モデル等への影響を軽減	入力データの正規化	ドロップアウトを適用．判定・予測エンジンの出力の分散の和がしきい値以上の場合，不正な入力データと判定[29].	ゼロ知識攻撃によって，適合率が 75%以上．限定知識攻撃と完全知識攻撃の場合，適合率がそれぞれ約 10%，約 2%.
		画像データに平均値フィルターを適用し入力データとする[27].	ゼロ知識攻撃によって，適合率が約 80%．完全知識攻撃では，入力データのユークリッド距離は拡張せず.

（備考）文献[5]の内容を基に作成.

5. 今後の課題と考察

4. で示したとおり，機械学習システムに対する主な攻撃手法への対策としてさまざまなものが提案されているが，現時点では，十分な有効性が確認された対策手法はほとんど存在していない．また，有効性を評価するにあたり，いくつかの定量的な評価尺度が提案･使用されているものの，どの尺度を使用するかは研究論文により異なっているなど，複数の対策手法の評価結果を横並びで比較することも容易ではない．これらの点については，既にいくつかの研究論文で指摘され課題として認識されており，今後の研究の進展が期待される.

こうした状況を踏まえると，機械学習システムを今後活用するうえでユーザーが留意すべき事項として，以下の 3 つが挙げられる.

第一に，機械学習システムの脆弱性や攻撃手法について，最新の研究動向を随時フォローし把握しておくことが必要である．最近の AI や機械学習への注目度の高まりを受けて，これらの分野の研究論文の発表数は増加傾向にある．そうしたなか，脆弱性や攻撃手法を指摘する研究論文も今後増えていく可能性が高い．機械学習システムを利用する側，あるいは，それを利用して自社の顧客にサービスを提供する側としては，最新の脅威や攻撃手法をフォローし，機械学習システムの利用や顧客へのサービス提供にどのような影響が及ぶ可能性があるかを確認していくことが求められる.

第二に，既存研究で提案されている機械学習システムの脆弱性や攻撃手法が，金融分野等、個別の利用分野における機械学習システムにどの程度当てはまるかを明らかにしていくことが求められる．本稿では，最近の主な研究成果を説明したが，それらの攻撃手法は，画像・文字・音声認識に機械学習を適用する分野に焦点を当てたものが多かった．特に，大量のデータを利用する深層ニューラル・ネットワーク等の学習モデルを対象としつつ,「人間にとっては同一の画像のようにみえるが，判定・予測エンジンは異なる画像と判断する」といった人間の知覚感度の限界による脆弱性を利用したものが目立つ．こうした脆弱性が金融分野のアプリケーションにどの程度当てはまるかは定かでない．既存研究における脆弱性が金融分野での機械学習システムにどの程度当てはまるか検討していくことが重要である.

また，既存研究における攻撃手法では，攻撃者が学習モデルや判定・予測エンジンの内容に関する情報を利用可能

であるという状況を前提としているものが大半である．こうした前提条件が金融分野における機械学習システムの利用環境において成立するか否かを個別に評価し，何らかの対策が必要か否かを評価していくことが求められる．そのうえで，対策が必要であることが判明したものに関して，どのように対処すべきかを他のエンティティと協議しつつ決定していくことになる．

第三に，機械学習システムのセキュリティや対策手法の有効性にかかる評価手法を検討・確立していくことが重要である．機械学習システムにおけるセキュリティや対策手法の有効性にかかる評価手法は研究途上の段階にあり，学界でも重要な課題として認識されている．今後，評価手法にかかる研究成果をフォローし，それらをどのように活用するか検討することが求められる．

外部の機械学習システムをネットワーク経由で利用するクラウド等のような形態の場合も，こうした評価手法が重要となる．そのシステムを運営する外部事業者が，対策手法の実装やセキュリティ管理を適切に実施していることがセキュリティ確保の前提条件となる．ユーザーとしては，外部事業者における対応の適切性をいかに確認・確保するかについても検討する必要があると考えられる．

4.2 で説明したように，準同型暗号等を用いて，データを暗号化したまま秘密に学習や予測・判定を行う手法の研究が活発化している．こうした先端的な研究開発によって，訓練データ等を秘匿したまま学習を行うことができるようになれば，クラウドを運営する外部事業者のセキュリティ管理への要求レベルを低下させたとしても，安全な訓練や判定・予測が実現する可能性がある．こうした先端的な技術の研究動向にも注目していく必要があろう．

参考文献

[1] 宇根正志. 機械学習システムのセキュリティに関する研究動向と課題. IMES Discussion Paper Series. 2018, no.2018-J-16.

[2] Barreno, M. et al., The Security of Machine Learning. Machine Learning. 2010, 81 (2), pp.121-148.

[3] Papernot, N. et al., Towards the Science of Security and Privacy in Machine Learning. 2016, arXiv: 1611.03814v1.

[4] 先崎佑弥, 大畑幸矢, 松浦幹太. 深層学習に対する効率的な Adversarial Examples 生成によるブラックボックス攻撃とその対策. 2018 年暗号と情報セキュリティ・シンポジウム予稿集. 2018 年.

[5] Carlini, N. and Wagner, D., Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. Proceedings of the 10th ACM Workshop on Security and Artificial Intelligence (AISec). 2017, pp.3-14.

[6] Suciu, O. et al., When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks. 2018, arXiv: 1803.06975v1.

[7] Shokri, R. et al., Membership Inference Attacks Against Machine Learning Models. Proceedings of the 2017 IEEE Symposium on Security and Privacy. 2017, pp.3-18.

[8] Ateniese, G. et al., Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. 2013, arXiv: 1306.4447v1.

[9] Fredrikson, M. et al., Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communication Security. 2015, pp.1322-1333.

[10] Fredrikson, M. et al., Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. Proceedings of the 23rd USENIX Security Symposium, USENIX. 2014, pp.17-32.

[11] Tramèr, F. et al., Stealing Machine Learning Models via Prediction APIs. Proceedings of the 25th USENIX Security Symposium. 2016, pp.601-618.

[12] Biggio, B. et al., Poisoning Attacks against Support Vector Machines. Proceedings of the 29th International Conference on Machine Learning, 2012.

[13] Biggio, B. et al., Support Vector Machines under Adversarial Label Noise. JMLR Workshop and Conference Proceedings, Asian Conference on Machine Learning. 2011, vol.20, pp.97-112.

[14] Biggio, B. et al., Evasion Attacks against Machine Learning at Test Time. Machine Learning and Knowledge Discovery in Databases. LNCS 8190. Springer. 2013, pp.387-402.

[15] Papernot, N. et al., The Limitations of Deep Learning in Adversarial Settings. Proceedings of 2016 IEEE European Symposium on Security and Privacy. 2016, pp.372-387.

[16] Chio, C. and Freeman, D., Machine Learning and Security. O'Reilly Media. 2018.

[17] Kloft, M. and Laskov, P., Online Anomaly Detection under Adversarial Impact. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS). 2010, pp.405-412.

[18] Mei, S. and Zhu, X., Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners. Proceedings of the 29th AAAI Conference on Artificial Intelligence. 2015, pp.2871-2877.

[19] Szegedy, C. et al., Intriguing Properties of Neural Networks. Proceedings of 2014 International Conference on Learning Representation. 2014.

[20] Dowlin, N. et al., CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. JMLR Workshop and Conference Proceedings, The 33rd International Conference on Machine Learning. 2016, vol.48, pp.201-210.

[21] Phong, L. T. et al., Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. IEEE Transactions on Information Forensics and Security. 2018, 13 (5), pp.1333-1345.

[22] Carlini, N. and Wagner, D., Towards Evaluating the Robustness of Neural Networks. Proceedings of 2017 IEEE Symposium on Security and Privacy. 2017, pp.39-57.

[23] Grosse, K. et al., On the (Statistical) Detection of Adversarial Examples. 2017, arXiv: 1702.06280v2.

[24] Gong, Z. et al., Adversarial and Clean Data Are Not Twins. 2017, arXiv: 1704.04960v1.

[25] Metzen, J. H. et al., On Detecting Adversarial Perturbations. Proceedings of 2017 International Conference on Learning Representation. 2017.

[26] Hendrycks, D. and Gimpel, K., Early Methods for Detecting Adversarial Images. Proceedings of 2017 International Conference on Learning Representation. 2017.

[27] Li, X. and Li, F., Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics. 2017, arXiv: 1612.07767v2.

[28] Bhagoji, A. N. et al., Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers. 2017, arXiv: 1704.02654v2.

[29] Feinman, R. et al., Detecting Adversarial Samples from Artifacts. 2017, arXiv: 1703.00410v3.