日銀レビュー

2019-J-2

金融分野における機械学習システムの活用とセキュリティ対策

金融研究所　井上紫織、宇根正志

Bank of Japan Review

2019 年 3 月

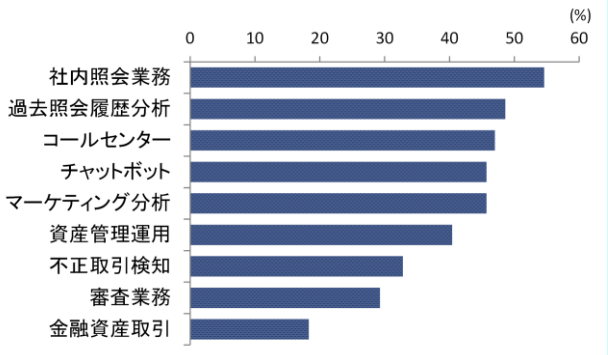
近年、金融を含むさまざまな分野において、人工知能（AI：artificial intelligence）、とりわけ、機械学習を実装したシステム（機械学習システム）の活用にかかる検討が進んでいる。こうした新たな技術・システムを導入する際には、そのメリットだけでなく、セキュリティ面のリスクに対しても十分に目を向ける必要がある。本稿では、機械学習システムのセキュリティをめぐる動向を紹介し、金融分野において活用される機械学習システムのセキュリティ対策上のポイントを考察する。

はじめに

近年、人工知能（AI：artificial intelligence）の実社会における活用にかかる検討が急速に進んでいる。AI は、一般に、推論や認識、判断等、人間と同様の知的な処理能力を持つコンピュータ・システムやその技術分野を指す。AI の機能を実現するツールとして用いられる代表的な技術が機械学習であり、それを実装したシステム（機械学習システム）にかかる検討が盛んに行われている。

AI や機械学習システムは、金融分野においても注目を集めている。金融情報システムセンターのアンケート調査によると、AI を「導入中」、「準備段階」または「検討中」と回答した金融機関の割合は、2015 年は 12.7%であったが、2016 年は 36.8%、2017 年は 49.3%と急増している[1]。AI の活用目的は、照会関連業務（社内照会業務、過去照会履歴分析、コールセンター、チャットボット等）、マーケティング分析や資産管理運用等、多岐にわたり、事務の効率化や精度の向上、新サービスの提供による収益の向上、経営リスクの低減等につながることが期待されている（図表 1 参照）。

【図表 1】AI の活用目的



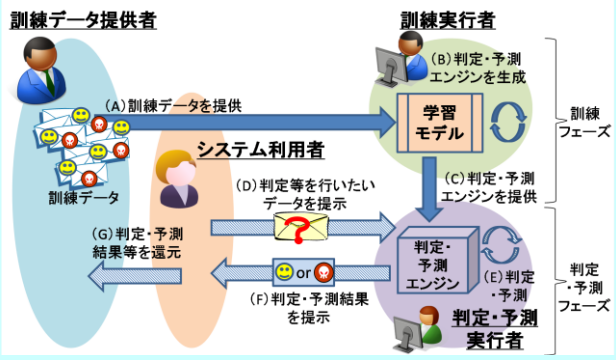
（注）AI を「導入中」、「準備段階」または「検討中」と回答した金融機関のうち、AI の活用目的として各項目を挙げた金融機関の割合。

（出典）金融情報システムセンター「平成 30 年度金融機関アンケート調査結果」（2018 年 11 月）

機械学習システムとその仕組み

機械学習システムは、主にデータの判定・予測の準備のための訓練フェーズと、新たなデータに対して判定・予測を行う判定・予測フェーズからなる（図表 2 参照）。

【図表 2】機械学習システムの概要



訓練フェーズでは、まず、(A)訓練データ提供者がシステムの利用目的に合う訓練データを収集し、訓練実行者に提供する[2]。(B)訓練実行者は、訓練データを一定のアルゴリズム（学習モデル）

に適用して訓練データに内在する関係を見つけ出し、新たな判定・予測対象のデータが与えられたときにその関係に基づいて判定・予測結果を出力する一種の関数（判定・予測エンジン）を生成する[3]。(C)訓練実行者は、判定・予測エンジンを判定・予測実行者に提供する。

判定･予測フェーズでは､(D)システム利用者が、判定・予測を行いたいデータ（判定・予測用データ）を判定・予測実行者に提示する。判定・予測実行者は､(E)それを判定・予測エンジンに入力して判定・予測を行い、(F)その出力（判定・予測結果）をシステム利用者に提示する。(G)判定・予測結果は、システム利用者から訓練データ提供者に還元される場合がある。例えば、判定・予測結果が誤っていた場合、訓練データ提供者は、還元データを分析し、正しい判定・予測結果に修正したうえで、これらのデータを訓練実行者に送信して再度訓練フェーズを実行することにより、判定・予測エンジンの精度の向上を図る。

セキュリティ目標と脆弱性

（セキュリティ目標）

情報システム一般における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その機能の機密性（confidentiality）・完全性（integrity）・可用性（availability）の確保が掲げられることが多い。機密性は、システムで取り扱われるデータやその機能が第三者に知られないことを、完全性は、それらのデータや機能が不正に偽造・改変されないことを、可用性は、システムが正常に稼動することをそれぞれ意味する。機械学習システムについても、これらがセキュリティ目標となる。

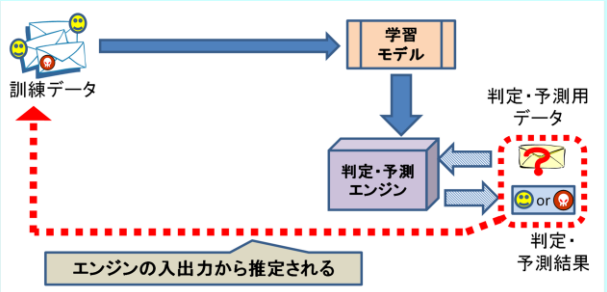
一般に、情報システムに対する攻撃者は、各エンティティとそれらの間の通信路に存在する脆弱性を狙い、システムのセキュリティ（機密性・完全性・可用性）を脅かす。例えば、各エンティティに対して大量のサービス要求を送ることにより、その機能や業務を妨害する攻撃や、エンティティ間の通信路上でデータを盗取・改変する攻撃が知られている。前者への対策としては、大量のアクセス要求を制御するサービスを利用したり、外部からのデータの受信を制御するゲートウェイを設置したりする方法が挙げられる[4]。後者への対策としては､TLS（Transport Layer Security）等の暗号通信プロトコルによる保護が知られている。

機械学習システムには、こうした情報システム一般に想定される脆弱性に加え、特有の脆弱性が存在する。特に、セキュリティ上のリスクとなりうるものとして、①訓練データや判定・予測エンジンにかかる情報の漏洩につながりうる脆弱性と、②判定・予測の精度低下につながりうる脆弱性がある[5]。

（情報漏洩につながりうる脆弱性）

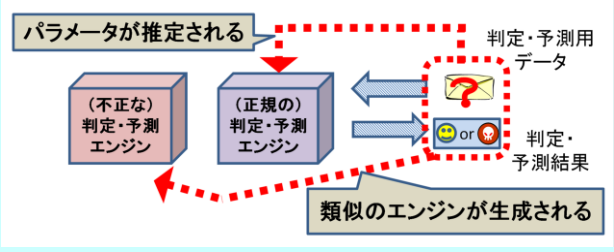
機械学習システムでは、判定・予測エンジンの入出力から、訓練データにかかる情報が漏洩しうる（図表 3 参照）。例えば、氏名等の個人を識別する情報と顔画像を訓練データとして使用した顔画像認識システムにおいて、判定・予測エンジンの入出力から、訓練データとして用いられた特定の個人の顔画像を高い確率で推定することに成功した研究事例が知られている[6]。

【図表 3】訓練データにかかる情報の漏洩



また、判定・予測フェーズの処理をクラウド上で提供するサービスのうち、判定・予測エンジンが判定・予測結果の確信度を出力するタイプの一部では、エンジンのパラメータを推定することにより、ほぼ同一の入出力関係を実現するエンジンを生成できるとの研究事例もある（図表 4 参照）[7]。

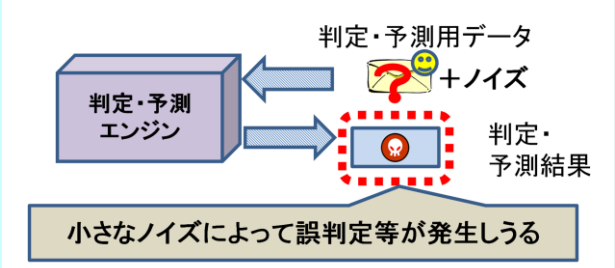
【図表 4】判定・予測エンジンにかかる情報の漏洩



（判定・予測の精度低下につながりうる脆弱性）

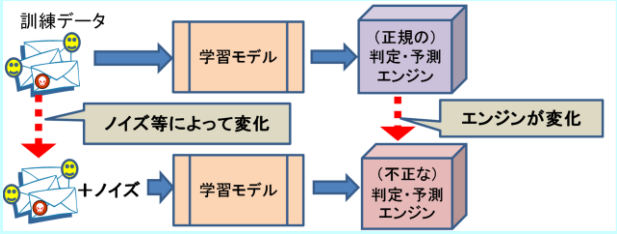
判定・予測用データにノイズ等が加えられると、判定・予測の精度が低下し、誤判定等が誘発されうる。例えば、画像認識や手書き文字認識のシステムでは、与えられた正規の画像に対して、微細なノイズを加えることにより誤った判定・予測結果を誘発させることができることを示した研究事例がある（図表 5 参照）[8]。正規の画像とノイズが付加された画像は、見た目には区別することが困難である。そのため、こうした画像が攻撃に用いられると、攻撃の検知が遅れることが懸念される。

【図表 5】入力へのノイズ付加による
判定・予測の精度低下



また、訓練データにノイズ等が加えられると、生成される判定・予測エンジンの精度が低下し、不特定多数の判定・予測用データに対して誤った判定・予測結果が誘発される可能性がある（図表 6 参照）。実際、こうした脆弱性を悪用して、判定・予測結果を不正に操作することができることを示す研究事例が発表されている[9]。

【図表 6】訓練データへのノイズ付加による
判定・予測の精度低下



金融分野における機械学習システムの活用事例とセキュリティ対策上のポイント

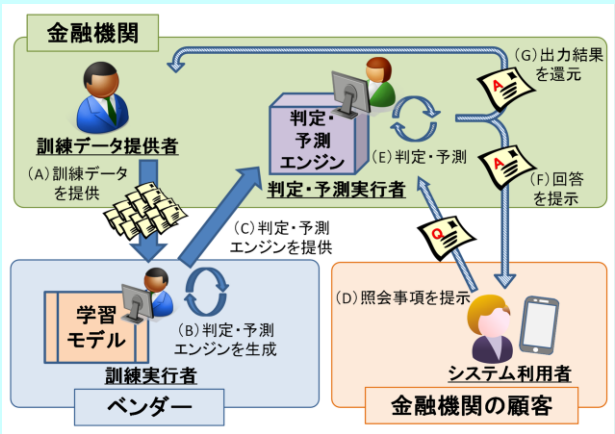
金融分野での活用が期待されている機械学習システムのうち、チャットボットによる照会対応システム、個人ローンの顧客向け信用度評価システム、異常検知システムを取り上げ、それらのセキュリティ対策上の留意点について考察する[10]。

（チャットボットによる照会対応システム）

近年、顧客とのコミュニケーションにおけるサービス品質向上の手段として、多くの金融機関でチャットボットを導入する動きがみられる[11]。チャットボットは、コールセンターや SNS、スマートフォン・アプリ、ウェブ上において、顧客からの照会への自動応答や、顧客の状況に合わせた金融商品の提案等を行う。チャットボットにより、顧客ニーズに合致した付加価値の高い回答の提示が可能になると期待されている。

チャットボットの基本的な機能は、想定される照会内容に対し、自動的に回答を提示する質疑応答機能である。近年、こうした機能を実現する汎用的な学習モデルが各ベンダーから提供されている。それらを活用したチャットボットによる照会対応システムでは、例えば、次のような処理の流れが想定される（図表 7 参照）。

【図表 7】チャットボットによる
照会対応システムのモデル



まず、(A)金融機関は、過去に蓄積した照会ノウハウや、顧客の特性に合致した金融商品に関するデータを訓練データとしてベンダーに提供する。ベンダーは、(B)訓練データをチャットボット用の学習モデルに適用して判定・予測エンジンを生成し、(C)金融機関にそれを提供する。(D)金融機関の顧客は、スマートフォン・アプリや SNS を用いて照会事項を金融機関に提示する。金融機関は、(E)照会事項を判定・予測エンジンに適用して判定・予測を行い、(F)それに基づく回答を顧客に提示する。(G)金融機関は、必要に応じて、出力結果を還元データとして活用する。

攻撃者は、まず、金融機関の顧客になりすますなどして、チャットボットの入出力を用いて訓練

データや判定・予測エンジンに関する情報を推定する可能性がある。また、チャットボットへの入力を改変し、誤った回答（判定・予測結果）を誘発させる可能性もある。

金融機関のセキュリティ対策は、各攻撃が成功した場合に生じうる影響や経済的損失を踏まえて決定していくことになる。

チャットボットが担う機能が一般的な照会事項への回答や金融商品の説明に限られる場合、それらのデータの機密性は低いことが想定される。そのため、訓練データや判定・予測エンジンの推定が成功したとしても、顧客の個人情報が漏洩したり、金融機関の収益に悪影響を及ぼしたりするような、致命的な脅威にはなりにくいと考えられる。

もっとも、不正な判定・予測用データによって誘発された判定・予測結果が還元データとして用いられると、不正な判定・予測エンジンが生成されることになる。その結果、不適切な回答が繰り返し生じるような場合には、金融機関に対する信頼低下を招く可能性がある。こうした観点から、金融機関は、還元データを用いてチャットボットの回答内容を確認することが求められる。

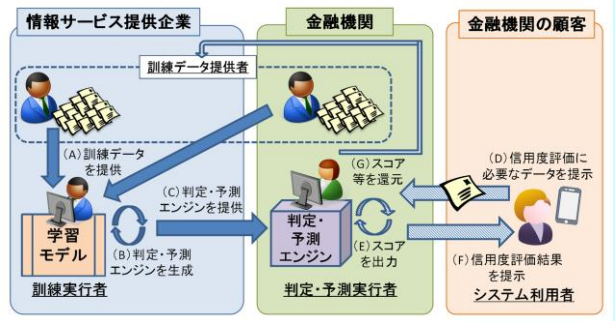
（個人ローンの顧客向け信用度評価システム）

個人ローンの融資審査では、顧客の年齢や年収、勤務先といった従来の審査項目のほか、顧客の性格や趣味、ライフスタイル、ネットショッピングの実績といった多種多様な情報を基に、機械学習システムを用いて顧客の信用度を数値化し、信用度評価に活用する事例がみられる。

このような信用度評価システムを構築する際には、金融機関自身が有する情報のほか、顧客に関するさまざまなデータ（ビッグデータ）を有する企業から情報提供を受けるケースがある。例えば、ビッグデータを有し、機械学習システムを構築するノウハウも有する企業（情報サービス提供企業）と連携して信用度評価システムを構築する場合、次のような処理の流れが想定される（図表 8 参照）[12]。

(A)金融機関は、顧客に関するデータを訓練データとして情報サービス提供企業に提供する。情報サービス提供企業は、(B)金融機関から提供されたデータに加え、自社が有するデータも訓練データ

【図表 8】個人ローンの顧客向け信用度評価システムのモデル



として用いて判定・予測エンジンを生成し、(C)金融機関に提供する。(D)金融機関の顧客は、スマートフォン・アプリや SNS を用いて、信用度評価に必要となるデータを判定・予測用データとして金融機関に提示する。金融機関は、(E)それを判定・予測エンジンに適用し、判定・予測結果としてスコアを出力するとともに、(F)スコアに基づく信用度評価結果を顧客に提示する。(G)金融機関および情報サービス提供企業は、必要に応じて、スコア等を還元データとして活用する。

このモデルにおいても、攻撃者は、金融機関の顧客になりすますなどして、判定・予測エンジンの入出力を悪用することが想定される。もっとも、攻撃が成功した場合に生じうる影響や経済的損失は、一般的な照会への回答を行うチャットボットのシステムと、個人情報をもとに顧客の信用度を評価する信用度評価システムでは異なる。

信用度評価システムでは、システム利用者として不特定多数の個人が利用することを想定しており、認証等によってなりすましを防ぐことは困難である。そのため、判定・予測エンジンの入出力を手掛かりとして訓練データが推定されたとしても、顧客の個人情報や個人の特定につながる情報が漏洩しないようにすることが求められる。例えば、年齢や収入等の個人情報を訓練データとして用いる場合には、そのままの数値ではなく、幅を持ったカテゴリー（年齢は 10 歳ごと、年収は 100 万円ごと等）に分類したうえで使用することなどが考えられる。

また、判定・予測エンジンを推定する攻撃や、不正な判定・予測用データにより誤ったスコアを誘発する攻撃の結果、本来よりも緩い条件での不適切な融資が実行されたり貸倒れに至るリスクが高まったりする可能性がある。さらに、誘発さ

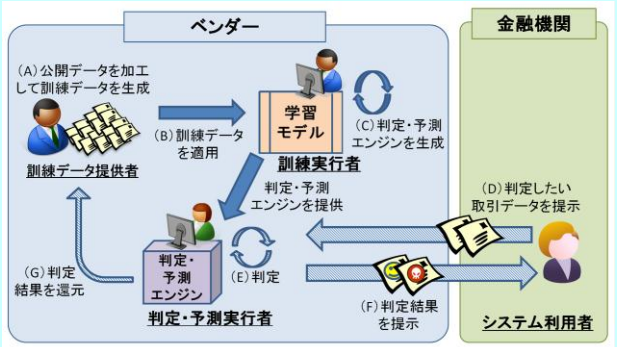
れたスコアが訓練データに還元されることによって判定・予測エンジンが不正に改変されると、他の顧客の信用度も正しく評価できなくなる可能性がある。そのため、不正な判定・予測用データおよび訓練データを検知・排除する機能や、それらがスコアに与える影響を低減する工夫を、判定・予測エンジンに組み込むなどの対応を検討することが求められる。

（異常検知システム）

金融機関の経営リスクを低減する手段として、近年、機械学習システムを用いた金融市場の異常検知やクレジットカード取引等の不正取引検知が注目されている。金融市場の異常検知では、過去の注文、市場流動性、価格変動等の情報から、金融市場の正常状態を学習して異常を検知する。また、クレジットカード取引等の不正取引検知では、過去の不正取引のデータのパターンや特徴を学習することで、類似の不正取引を検知する。

ベンダーが公開データを基にこうした機械学習システムを構築し、金融機関向けに提供している事例もある。例えば、処理の流れとして以下のケースが想定される（図表 9 参照）。

【図表 9】異常検知システムのモデル



ベンダーは、(A)公開データを加工して訓練データを生成した後、(B)それを学習モデルに適用し、(C)判定・予測エンジンを生成する。(D)金融機関は、判定したい取引データをベンダーに提示する。ベンダーは、(E)それを判定・予測エンジンに適用して判定結果を出力し、(F)それを金融機関に提示する。(G)ベンダーは、必要に応じて判定結果を還元データとして活用する。

このモデルでは、攻撃者は、金融機関になりすましたり、金融機関内部の者と結託したりすることで、判定・予測エンジンの入出力を悪用することが想定される。

攻撃が成功した場合に生じる影響や経済的損失について考えると、まず、訓練データが推定されたとしても公開データを用いるため問題はない。また、判定・予測エンジンが推定されたとしても、ベンダーにとってそれは重要な資産であるが、金融機関にとっては特段の影響は生じない。

一方、判定・予測用データを改変する攻撃による影響は、システムの利用目的により異なる。例えば、金融市場の異常を検知するシステムの場合、異常を検知できない、または、正常時に異常と判断するといった誤判定が誘発される可能性がある。金融機関が、こうした判定結果を基に金融取引を行うとすれば、経済的損失を被ったり不正な取引を実行したりするなどのリスクが生じる。

また、クレジットカード取引等の不正取引検知システムの場合においても、不正取引に関するデータが改変され、不正取引を検知できなくなる可能性が考えられる。さらに、誤った判定・予測結果が還元データとして用いられると、判定・予測エンジンが改変され、攻撃に用いられたデータ以外の取引データも正しく判定できなくなる可能性がある。

こうした観点から、不正な判定・予測用データや還元データを検知・排除したり、それらが判定結果に与える影響を低減したりする機能を判定・予測エンジンに組み込むなどの対応を検討することが求められる。

おわりに

金融分野における機械学習システムの利活用は始まったばかりであり、執筆者たちが知る限り、深刻なセキュリティ上の被害は報告されていないようである。もっとも、機械学習システムには、従来の情報システムが有する脆弱性に加えて特有の脆弱性が存在する。機械学習システムを導入する際には、こうした脆弱性を踏まえ、想定しうる攻撃を洗い出したうえで、実際に攻撃を受けた際の影響の多寡を見極めて対策を検討することが重要である。

AI や機械学習システムの技術は日々進化している。一方で、それらを狙った攻撃手法も巧妙化しており、これまで対処可能であった攻撃に対し

て、さらなる対策が必要となる場合もある。機械学習システムのセキュリティ対策を考える際には、最新の攻撃手法とそれらへの対策について、技術の進展を踏まえつつ検討していくことが求められる。

[1] 金融情報システムセンター「平成 30 年度金融機関アンケート調査結果」（2018 年 11 月）

[2] ここでは、判定・予測の対象となりうるデータ（の一部）とそれらの判定・予測の結果を示すデータを訓練データとするケースを取り上げる。こうした機械学習の手法は「教師あり学習」と呼ばれる。

[3] 文献によっては、学習モデルを学習アルゴリズム、判定・予測エンジンを学習モデルと呼ぶこともある。

[4] 大量のアクセスを制御するサービスとして、例えば、CDN（contents delivery network）が知られている。CDN は、インターネット・ユーザーへのコンテンツ配信を、効率的かつ高速に実行するとともに、大量のアクセスを制御する機能を有している。

[5] 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」（金融研究所ディスカッション・ペーパー No. 2018-J-16）などに詳しい。

[6] このシステムでは、判定・予測用データとして顔画像が与えられ、判定・予測結果として個人の識別情報（氏名等）と、判定・予測の確からしさを示す値（確信度）が出力される。詳細については、Fredrikson, M. *et al.*, “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures”（Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS) 2015, Association for Computing Machinery, 2015）を参照されたい。

[7] 詳細については、Tramèr, F. *et al.*, “Stealing Machine Learning Models via Prediction APIs”（Proceedings of USENIX Security Symposium, Advanced Computing Systems Association, 2016）を参照されたい。

[8] 詳細については、Szegedy, C. *et al.*, “Intriguing Properties of Neural Networks”（Proceedings of International Conference on Learning Representations (ICLR) 2014, Cornell University Library, 2014）を参照されたい。なお、近年では、こうした攻撃への対策の 1 つとして、敵対的生成ネットワーク（GAN : generative adversarial networks）と呼ばれる技術を用いて判定・予測の精度低下を防ぐ手法が注目されている。もっとも、こうした手法を用いた場合でも、誤った判定・予測が高い確率で発生する事例が報告されており、未だ十分な対策手法は確立されていない。

[9] 詳細については、Barreno, M. *et al.*, “The Security of Machine Learning”（Machine Learning, 2010）を参照されたい。

[10] 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」（金融研究所ディスカッション・ペーパー No. 2019-J-1）などに詳しい。

[11] 最近では、顧客の口座残高に関する照会対応のほか、支出状況等の情報を分析して、消費動向に関するアドバイスを提供したり、不正出金や二重払いの可能性を警告したりするスマートフォン・アプリ・サービスが提供されている。例えば、バンク・オブ・アメリカの「Erica」やキャピタル・ワンの「Eno」等がある。また、SNS 上におけるチャットボットとのやり取りを通じて、顧客に適した保険商品や保険料金の見積りを提示するものとして、ライフネット生命のサービス等が知られている。

[12] 金融機関と情報サービス提供企業が一体となって合弁会社を設立した事例も存在する。例えば、みずほ銀行とソフトバンクが設立した J.Score の事例が該当しうる。