プレスリリース
2025 年 6 月 10 日

国立研究開発法人情報通信研究機構
国立大学法人神戸大学
EAGLYS 株式会社

プライバシー保護連合学習技術「DeepProtect」を活用した銀行の不正口座検知の実証実験を実施し、検知精度向上を確認

【ポイント】

- 銀行 4 行と連携し、プライバシー保護連合学習技術「DeepProtect」を活用した不正口座検知の実証実験を実施
- 個別学習モデルと連合学習モデルを組み合わせたアンサンブル学習を適用し、不正口座検知の精度向上を確認
- 不正取引モニタリング業務での実運用に向け、現行システム（AML システム）と並行運用の可能性を検討

国立研究開発法人情報通信研究機構（NICT、理事長: 徳田 英幸）は、国立大学法人神戸大学（学長: 藤澤 正人）及び EAGLYS 株式会社（代表取締役社長: 今林 広樹）に委託し、りそな銀行他 3 行と連携して、プライバシー保護連合学習技術「DeepProtect」[*1]を活用した不正口座検知の実証実験を実施しました。

本実験では、銀行間でデータを共有せずに学習モデルを構築し、より実運用に近い条件下で検知精度[*2]を向上させることを検証し、データ項目にばらつきがあっても情報を余すことなく利用できるアンサンブル学習[*3]という新しいアプローチを適用した結果、従来の個別の銀行ごとに構築する学習モデル（個別学習モデル）と比較して適合率[*4]が最大約 10 ポイント向上し、再現率[*5]が 95%を超える高精度な検知を達成、個別学習モデルよりも安定した検知精度を維持できることを確認しました。また、従来のルールベースの監視や AI を用いた個別学習では検知が困難だった潜在的な不正口座を特定できる可能性が示されました。

今後は金融機関の不正取引モニタリング業務における実運用に向け、現行の AML システム[*6]と並行運用できる形での導入可能性を検討します。

【背景】

複雑化・巧妙化する金融犯罪手法に対し、口座への入出金や顧客ごとの取引の監視、また AI を用いた検知システムの導入・検討など、不正取引モニタリングの取組が各金融機関で進められています。しかし、単独の金融機関では十分な AI 開発の肝となる学習データの確保が難しく、顧客の個人情報やプライバシーの保護を実現した上で、複数の金融機関が組織横断的に協調して AI を開発していくことが極めて重要となっています。

この課題を解決するため、NICT はデータを外部に開示することなく機密性を保ったまま深層学習を行う「DeepProtect」を活用することで、複数の金融機関と連携して不正取引を自動検知するシステムの開発に取り組んできました。NICT は、2023 年から、これまでの成果を基に、次のステップとして神戸大学及び EAGLYS に対し、それぞれの金融機関で日々蓄積される取引データを継続して学習に取り込めるなど、より実用性の高い不正取引モニタリング AI の研究開発及び実証実験を、高度通信・放送研究開発委託研究として委託していました。

【今回の成果】

本実証実験は、りそな銀行等の協力を得て、各銀行から正常口座と凍結となった不正口座の両方について、顧客情報が特定されない形で一定期間のデータを提供いただきました。時系列解析が可能な形に前処理を実施した上で個別学習のモデルを作成、続いて DeepProtect を用いて 4 銀行で連合学習を実施し、検知精度の評価を行いました。その結果、個別学習と比較し連合学習では全体的に精度の低下が起こりました。各銀行のデータのフォーマットや定義がカラムレベルで異なり、4 銀行で共通して使えるデータ項目が極端に少ないことが理由の一つと考えられました。

そこで、研究を受託した神戸大学及び EAGLYS は、DeepProtect にアンサンブル学習を組み合わせることで、データ項目にばらつきがあっても情報を余すことなく利用し、不正検知を高性能化できるアプローチを新たに適用しました。このアプローチでは、各銀行の個別学習モデル（4 個）と 4 銀行での連合学習モデル（1 個）に加え、3 銀行ずつの連合学習モデル（4 個）でも連合学習を行い、全部で 9 個のモデルを活用しデータ項目の標準化を実質的に実現しまし

た（図 1 参照）。

その結果、連合学習モデルでは、個別学習モデルと比較して適合率が最大約 10 ポイント向上するケースも見られ、95%の再現率を超える高精度な検知を達成するケースもありました。

また、再現率の高い箇所で連合学習モデルの方が良い検知精度を確認しました（図 2 参照）。アンサンブル学習により、検知の安定性も向上しました。

実証実験後の協力銀行とのワークショップにおいて、「個別学習では不正口座として検知されず、連合学習では不正口座として検知された取引口座」を確認した結果、一部において「グレーな口座」と評価され、既存のルールベースでの監視ではすり抜けていたことがわかり、従来のルールベースの監視や AI を用いた個別学習では検知が困難であった「潜在的な不正口座」を特定できる可能性が示されました。

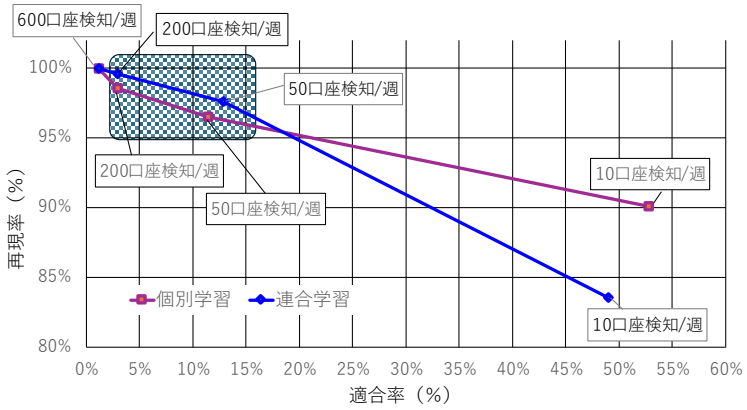
さらに、高度通信・放送研究開発委託研究の取組である DeepProtect の高度化に関して、(1)不正取引データの合成手法の提案、(2)訓練データの不均衡データ問題[*7] の緩和及び敵対的サンプル攻撃[*8] に対する防御手法への応用とその有効性の提示、(3)破滅的忘却[*9] を抑制しながら継続的な連合学習を可能にするアルゴリズムの検証、(4)銀行における不正取引監視者を支援する不正取引モニタリングシステムのプロトタイプの開発、を行いました。

今回の成果を不正取引モニタリングシステムに実装することで、より精度の高い不正取引の検知や不正口座の早期検知が可能となるとともに、金融機関の監視業務の効率化やコスト削減効果が期待されます。

9 個の学習モデルを活用
個別学習モデル 4 個＋ 4 銀行での連合学習モデル 1 個＋ 3 銀行ずつの連合学習モデル 4 個

※実際には各銀行が持つ学習モデルのセットは異なります。

図 1 アンサンブル学習の実施イメージ



再現率の高い箇所（▨）で連合学習モデルの方が良い検知精度を確認

図 2 個別学習及び連合学習モデルのシミュレーション結果

本実証実験において、適合率、再現率は以下を意味する。
・適合率：モデルが不正口座と予測したもののうち、実際に不正口座であるものの割合
・再現率：実際に不正口座であるもののうち、モデルが正しく不正口座と判定した割合

【今後の展望】

本実証実験で得られた成果を踏まえ、NICT は DeepProtect の基盤技術の更なる高度化を目指します。また、神戸大学と EAGLYS は検知精度の更なる向上を図るとともに、不正取引検知業務への実装に向けた取組を進めます。そのため、まずは現行の AML システムと並行運用する簡易的なシステムの導入を検討し、金融機関での実用化の可能性を検討していきます。

＜各機関の役割分担＞

・NICT: 全体マネジメント、DeepProtect 技術の提供、実証実験結果報告及びヒアリング調査のためのワークショップ開催
・神戸大学: 実証実験環境の開発と性能評価、DeepProtect の敵対的サンプル攻撃等への耐性向上、不正送金検知特徴量と不正判定基準の標準化
・EAGLYS: DeepProtect の継続学習化、モジュールの開発、人間系フィードバックを容易とする支援ツールの開発

＜関連する過去の NICT のプレスリリース＞

・2022 年 3 月 10 日 プライバシー保護連合学習技術を活用した不正送金検知の実証実験を実施
https://www.nict.go.jp/press/2022/03/10-1.html
・2020 年 5 月 19 日 プライバシー保護深層学習技術を活用した不正送金検知の実証実験において金融機関 5 行との連携を開始
https://www.nict.go.jp/press/2020/05/19-1.html
・2019 年 2 月 1 日 プライバシー保護深層学習技術で 不正送金の検知精度向上に向けた実証実験を開始
https://www.nict.go.jp/press/2019/02/01-2.html

なお、本研究の一部は、NICT の令和 4 年度高度通信・放送研究開発委託研究における委託研究課題「プライバシー保護連合学習の高度化に関する研究開発（課題番号 229）」を、神戸大学及び EAGLYS に委託し実施されました。

＜ 本件に関する問合せ先 ＞
国立研究開発法人情報通信研究機構
サイバーセキュリティ研究所
セキュリティ基盤研究室
小川 一人
E-mail: security@ml.nict.go.jp

国立大学法人神戸大学
数理・データサイエンスセンター
教授 小澤 誠一
E-mail: ozawasei@kobe-u.ac.jp

EAGLYS 株式会社
取締役/CSO 丸山 祐丞
E-mail: maruyama@eaglys.co.jp

＜ 広報 （取材受付）＞
国立研究開発法人情報通信研究機構
広報部 報道室
E-mail: publicity@nict.go.jp

国立大学法人神戸大学
総務部 広報課
E-mail: ppr-kouhoushitsu@office.kobe-u.ac.jp

EAGLYS 株式会社
E-mail: pr@eaglys.co.jp

＜用語解説＞

*1 プライバシー保護連合学習技術「DeepProtect」

連合学習技術に暗号技術を融合することによって、NICT が独自に開発したプライバシー保護連合学習技術のこと。まず、各組織で持つデータを基に深層学習を行う際に、学習中のパラメータ（勾配情報）を暗号化して中央サーバに送り、中央サーバでは、暗号化したまま学習モデルのパラメータ（重み）の更新を行う。次に、更新されたこの学習モデルのパラメータを各組織においてダウンロードすることで、より精度の高い分析が可能になる。DeepProtect は、各組織から中央サーバにデータそのものを送ることなく、学習中のパラメータのみを暗号化して送信するが、このパラメータは、複数のデータを集計した統計情報とすることによって個人を識別できない状態にすることが可能であり、さらに、暗号化を施すため、データの外部への漏えいを防ぐことができる。
本技術により、パーソナルデータのような機密性の高いデータを外部に開示することなく、複数組織で連携して多くのデータを基にした深層学習が可能となる。
本技術は、下記ジャーナルに採択・掲載されている。

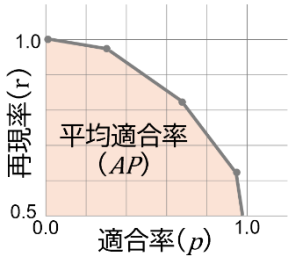
L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-Preserving Deep Learning via Additively Homomorphic Encryption", IEEE Transactions on Information Forensics and Security, Vol.13, No.5, pp.1333-1345, 2018.

L. T. Phong and T. T. Phuong, "Privacy-Preserving Deep Learning via Weight Transmission", IEEE Transactions on Information Forensics and Security, Vol.14, No.11, pp 3003-3015, 2019.

本技術は、下記動画でも紹介している。
『NICT ステーション　～DeepProtect～』
https://youtu.be/CpA9OD5vUIM

***2 検知精度**

今回の実証実験ではモデル評価の指標として、適合率（p）[*4]、再現率（r）[*5]、平均適合率（AP）を使用した。平均適合率はモデルの全体的な性能を表す指標であり、Precision-Recall 曲線の下側の面積を計算することで求められる。



参考：Precision-Recall 曲線の例

***3 アンサンブル学習**

アンサンブル学習は、一般的には複数の機械学習モデルを組み合わせて、精度の高い推論を導出する機械学習手法。当実証実験では各銀行の個別学習モデルと複数の連合学習モデルを組み合わせることで、データ項目のばらつきがあっても情報量を余すことなく利用するアプローチを適用した。

***4 適合率**

モデルが「不正」と予測したデータのうち、実際に「不正」であった割合を示す指標。適合率が高いほど、誤検出が少なく、正確に予想できていることを意味する。適合率は、スパムメールの検出など「誤検出を減らすことが重要なケース」で重視される。

***5 再現率**

モデルが実際に「不正」であるデータのうち、どれだけ正しく「不正」と予想できたかを示す指標。再現率が高いほど、取りこぼしなく正しく検知できていることを意味する。再現率は、不正検出など「取りこぼしを減らすことが重要なケース」で重視される。

***6 AML システム**

AML（アンチ・マネー・ローンダリング）システムとは、金融機関における資金の流れを常に監視し、不正が疑われる取引があれば自動で検出するソフトウェアのことである。近年、日本におけるマネー・ローンダリング及びテロ資金供与対策では、取引時確認や疑わしい取引の検知・届出といった様々な局面で、AI、ブロックチェーン、RPA（ロボティック・プロセス・オートメーション：人工知能等を活用した定型的作業の自動化）といった新技術が導入され、実効性向上に活用されている。

***7 銀行の不正取引検知における不均衡データ問題**

銀行の顧客口座の大多数は通常取引を行う正常口座であり、サンプルとして扱える不正取引を行う口座は極端に少数である。本研究ではデータ生成源モデルの推定とそれに基づくデータ生成を行い、合成データを訓練データに追加して学習することで性能改善を確認した。

***8 敵対的サンプル攻撃**

AI の入力に摂動を与えて意図的に AI に誤判定させる攻撃。不正取引検知においては銀行側が設置した不正検知 AI が顧客の口座取引を監視する中で、犯罪者による不正取引を見逃すように仕向けること。本研究では不正取引データの合成と敵対的攻撃アルゴリズムを組み合わせて、AI に検知されない不正取引を探索した。

***9 破滅的忘却**

継続学習を続けると、新しいデータの学習によりモデルパラメータが修正され、それにより過去の記憶に関連したパラメータも変更され、重要な過去の情報を忘却してしまう。本研究では過去の代表的なサンプルを残して学習を進める経験リプレイを実施することで性能が向上することを確認した。