レコメンド分野の著名な国際会議RecSys2025に連合レコメンドシステムのプライバシー検証に関する論文が採択

トヨタ自動車株式会社 InfoTech
篠田 謙司、笹井 健行、福島 真太朗

トヨタ自動車株式会社は、連合レコメンドシステムにおける新たな脆弱性を明らかにしました。レコメンド分野の著名な国際会議、RecSys2025（theNineteenth ACM Conference on Recommender Systems）に研究論文 "Popularity-Bias Vulnerability: Semi-Supervised Label Inference Attack on Federated Recommender Systems" が採択されました。また2025年9月の本会議にて発表を行いました。

背景

人に寄り添うAIエージェントの実現に向けて、ユーザ理解に基づくレコメンド技術は極めて重要です。特に、各事業者が保有データを直接共有することなく、プライバシーを保護しながら機械学習モデルを学習できる垂直連合学習（Vertical Federated Learning; VFL）が注目されています。VFLでは、各事業者は勾配情報等を準同型暗号といった強固な暗号技術で保護したうえで交換し、各自のローカルモデルを更新します。これにより、自組織だけでなく他組織のデータも、プライバシーを保ったまま学習に活用できることが期待されます。

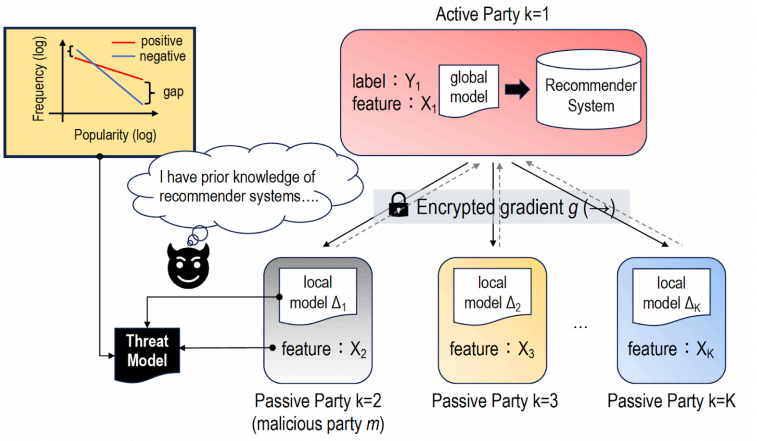
一方で本技術は発展途上にあり、悪意のあるVFL参加組織が、他組織の本来保護されるべきデータ（ラベル等）を漏えいさせてしまう攻撃も知られており、実用的かつセキュアなアルゴリズムの構築に向け、研究が進められています。我々は産業応用上のセキュリティ懸念を事前に調査する目的で、VFLをレコメンドシステムに適用した場合にドメイン固有の脆弱性が現れるかを検証しました。その結果、レコメンドにおける典型的な性質を悪用することで、事業者が保持するラベル（正解データ）を推論により漏えいさせる攻撃が成立することを明らかにしました。

技術概要

多くのVFLでは、参加組織は次の2つの役割に分かれます。1つは学習のラベルを保持するアクティブパーティ、もう1つは特徴量のみを提供するパッシブパーティです。アクティブパーティはラベルに基づく推論可能なモデルを保持し得ますが、パッシブパーティは特徴量のみを保持し、単独では推論能力を持ちません。

既存研究では、パッシブパーティがアクティブパーティのラベルを推定・漏えいさせる攻撃が報告されています。例として、少量の漏えいラベルでローカルモデルを半教師あり学習し推論能力を持たせる手法や、木ベースVFLで得られたローカルモデルの木構造を解析して各ID（レコメンドの場合はユーザとアイテムの組）の葉ノード割り当てを識別し、推論に用いる方法などがあります。VFLの産業応用を進めるには、既存研究から一歩進めてドメイン特有の脆弱性を明らかにすることが重要です。そこで本研究では、レコメンドシステムを対象としてその脆弱性を調査しました。

レコメンドシステムでは、ユーザの購買履歴や訪問履歴をラベルとして、特徴量からそれを当てるモデルが使われます。ここで、購買・訪問履歴は「ユーザが明確に興味を示したアイテム」は分かる一方で、それ以外のアイテムに興味がないかは明確ではありません。この状況を暗黙的フィードバックと呼びます。機械学習の訓練には通常、ユーザが興味を持つアイテム（正例）と、興味を持たないアイテム（負例）のラベルが必要なため、暗黙的フィードバック下では負例を人工的に生成する必要があります。



垂直連合レコメンドシステムと悪意のあるパッシブパーティに関する模式図

レコメンドシステムには、次の性質がしばしば見られます。

（1）人気度の分布がべき乗則に従う：人気度を横軸、該当人気度のアイテム数を縦軸とした分布が、しばしばべき乗則に従います。
（2）ランダムサンプリングによる負例生成：暗黙的フィードバック下で負例を作成する際、各ユーザが購買・訪問していないアイテムからランダムにサンプリングする方法（ランダムネガティブサンプリング）が一般的です。

（1）の下で（2）が実行されると、正例と負例の分布に大きな歪みが生じます。攻撃者がこの歪みを事前知識として保有している場合、それを悪用した攻撃モデルを構築できるのではないか、というのが我々の着眼点です。

我々は次の3ステージからなる手法を提案しました（攻撃対象のVFLアルゴリズムは木ベース（ブースティング木）であるSecureBoost）。

ステージ1）木の解析によるグラフ構築：悪意のあるパッシブパーティが得たローカルモデルから、同一の葉ノードに割り当てられたと推定されるユーザ-アイテムID間にエッジを張り、各IDをノードとするグラフを構築します。パッシブパーティが保有する特徴量はノード特徴量として付与します。
ステージ2）事前知識の活用：人気度のべき乗則とランダムネガティブサンプリングの組合せによる正例・負例の分布の歪みを用いて、擬似ラベルの作成や、損失関数のための事前確率を定義します。
ステージ3）半教師あり学習による攻撃モデル作成：ステージ1のグラフをもとに、ステージ2の擬似ラベルや確率分布を用いてグラフベースの半教師あり学習を行い、攻撃モデルを学習します。

以上の手続きにより、悪意のあるパッシブパーティはアクティブパーティのラベルを推論によって漏えいさせる攻撃が可能となります。3つのオープンデータセットで本攻撃モデルを検証し、ベースラインを上回る推論性能を確認しました。

まとめ

本研究により、ドメイン特有の性質（アイテム人気度と暗黙的フィードバック下の負例生成）を悪用したVFLの新たな脆弱性を明らかにしました。本攻撃の存在を事前に認識することで、有効な防御手法の構築につなげ、データを保護しながらユーザに寄り添ったAIエージェントの開発に貢献します。