レコメンドシステムに対する
ポイズニング攻撃のサーベイ論文報告

第 1 回リクルート MTG

August 12, 2025

1　概要

本報告では、レコメンドシステムに対するポイズニング攻撃に関する包括的なサーベイ論文「Manipulating Recommender Systems: A Survey of Poisoning Attacks and Countermeasures」の内容を整理し、主要な知見と今後の研究方向性について報告する。

2　ポイズニング攻撃の種類

2.1　攻撃手法の分類

1. **古典的なヒューリスティック攻撃**
    - あらかじめ定められたヒューリスティックなルールやアルゴリズムに基づいた攻撃
    - この攻撃自体は学習を伴わない
    - 比較的単純に見えるが効果的
2. **AI ベースの攻撃**
    - 機械学習モデルを利用して実行される攻撃
    - 敵対的生成ネットワークなどを用いて既存ユーザーの評価パターンを学習し、より検知されにくいデータを偽装
    - 学習によって進化する点が大きな特徴

2.2　市況の変化

- 過去 7 年間でレコメンドシステムに対する攻撃数が増加
- 古典的な攻撃から AI ベースの攻撃への明確な移行が見られる
- 2020 年から 2023 年の期間では、AI ベースのポイズニング攻撃がほとんどを占める

3　ポイズニング攻撃と敵対的攻撃の違い

3.1　敵対的攻撃

- モデル学習後の推薦段階で仕掛けられる
- 根底にあるモデルを変更することなく、推薦時にレコメンドシステムの出力を不正にできるサンプルを見つける
- 入力データを操作することで攻撃中にシステムを一時的に欺く
- 具体例：ログデータの改ざんによる入力値の変更

3.2 ポイズニング攻撃

- モデルのトレーニング段階で仕掛けられる
- モデルのトレーニングデータを汚染し、攻撃者に有益な結果をもたらす攻撃モデルへと変換させることを目的とする
- 具体例：悪意のあるログデータの注入による訓練妨害

4 レコメンドシステムの概要

主要なレコメンドシステムの種類：

- 行列分解ベース
- グラフベース
- アソシエーションルールベース
- 近傍ベース
- ディープラーニングベース

4.1 実用的なアーキテクチャ

リトリーバル（候補生成）フェーズ → ランキング（LightGBM）フェーズ

5 ポイズニング攻撃の分類

5.1 攻撃の目的による分類

1. **非ターゲット型**
    - エラーの最大化により役に立たなくする目的
    - 研究は比較的少ない
2. **ターゲット型**
    - 特定のターゲットアイテムの人気を増減させる目的
    - プロモーション攻撃 or デモーション攻撃
    - 攻撃のインセンティブが明確なので、研究が多い

5.2 敵対者の知識による分類

1. **ブラックボックス攻撃**
    - アーキテクチャ、予測関数、パラメータ、相互作用履歴 R を知らない
2. **グレーボックス攻撃**
    - 限定的な知識を持つ
    - ユーザー・アイテム間の相互作用行列 R に悪意のあるデータをインジェクトするのみ
    - 特に重要になる攻撃手法
3. **ホワイトボックス攻撃**
    - 徹底的な知識を持つ
    - 予測関数、パラメータ、相互作用の履歴 R を完全に把握

5.3 攻撃の影響による分類

1. **システムの可用性（Availability）への影響**
    - 主にホワイトボックス攻撃とグレーボックス攻撃が入力データを妨害
    - 研究の 3 分の 2 を占める
2. **複製（Replication）の試み**
    - ブラックボックス攻撃はシステムの可用性を損なう点では効率が劣る
    - シミュレーションとしてリバースエンジニアリングを試行
    - 研究の 3 分の 1 を占める
3. **不可知性（Unnoticeability）**
    - データ特性を維持することでアプローチが気付かれないようにする

6 攻撃者の能力

6.1 偽ユーザーの作成

- 偽ユーザーの最大数（攻撃サイズ）はシステム全体のユーザー数よりはるかに少ない
- 予算など運用上の観点から制限される

6.2 偽評価の作成

- 注入できる評価の数にはプロファイルサイズと呼ばれる上限 k が存在
- 現実のユーザーはシステム内のごく一部のアイテムとしかやり取りしないため
- 評価の値にも上界と下界がある
- 実際の評価行列の近傍である必要がある

6.3 その他の攻撃手法

- 偽の共起作成：ターゲットアイテムと関連付けられているユーザーに対して他のアイテムへの訪問を注入
- 偽リンクの追加：知識グラフの脆弱性を突く
- 偽画像の使用：主にコールドスタート問題の対処のための画像認識を悪用

6.4 研究の分布

- 偽ユーザー × 偽評価の研究が大部分
- 偽の共起は一部
- 偽リンク、偽画像はごくわずか

7　攻撃のアプローチ

7.1　インジェクション

- 限定的な知識しか持たない場合に使用
- うまく設計された少数のユーザーの注入
- 研究のすべての手法が何らかの形で偽ユーザーと評価を訓練データに注入している

7.2　シミュレーション

- ブラックボックスの状況で使用
- 通常、標的のレコメンドシステムをシミュレートし代替モデルを訓練
- インジェクションに加えてシミュレーションも行う研究は少数

8　ポイズニング攻撃のグループ分け

8.1　モデル非依存攻撃（Model-agnostic attacks）

- あらゆるレコメンドシステムを攻撃
- 特定のモデルのアルゴリズムに依存しない

8.2　モデル内在攻撃（Model-intrinsic attacks）

- 特定の種類のレコメンドシステムを標的とする攻撃
- 例えば CF や行列分解など

8.3　攻撃手法の分布

- 行列分解ベースに対する攻撃（Bo Li など）が最多
- グラフベースに対する攻撃が次に多い
- 近隣ベース、ディープラーニングベースは各 1 件
- FDRS ベースは 2 件が注目に値する

9　ドメインとインタラクションの種類

9.1　対象ドメイン

- 映画
- POI（Point Of Interest）：観光地、商業施設、公共施設など
- 位置情報サービス
- 引用ネットワーク
- ニュース
- その他多数

結論：ポイズニング攻撃の影響は広範囲にわたる。オンラインデータの完全性と信頼性を守るため、堅牢な対策を開発する必要がある。

9.2 インタラクションの種類

- 明示的なインタラクションの研究が 8〜9 割
- 暗黙的なインタラクションの研究はわずか

10 ポイズニング攻撃の評価指標

10.1 主要な評価指標

1. **HR@K（The hit rate ratio）**
    - 最も優勢で、研究の 3 分の 2 で採用
2. **nDCG@K（Normalized discounted cumulative gain）**
    - 次いで研究の 4 分の 1 で採用

上記 2 つの指標以外は一般的ではない。

11 データセット

11.1 使用頻度の高いデータセット

- **MovieLens** が研究の 3 分の 2 以上で使用される
- **Amazon** の様々な製品カテゴリのデータセットが研究の半分で使用される
- Netflix は 3 件
- Yelp は 2 件

12 ポイズニング攻撃への対策

12.1 対策の分類

1. **検知手法**
    - 攻撃で作成されたプロファイルを特定することを目的とする
2. **防御手法**
    - 悪意のあるプロファイルを明示的に特定しようとせず、攻撃に対してよりロバストにすることを目指す

12.2 検知手法の特徴量

12.2.1 Model-agnostic な特徴量

評価行動の絶対的な異常さに基づく：

- **RDMA**：特定のユーザーが多数の特定のアイテムに与えた評価が他のユーザーの評価と比べてどれだけ平均的に乖離しているか
- **WDMA**：RDMA の重み付き分散で、疎なアイテムに対するユーザー評価の累積的な違いを捉える
- **長さ分散（LengthVar）**：プロファイルの長さの違いを測る指標

- **トップ近隣ユーザーとの類似度（DegSim）**：あるユーザーがトップ k 人の近隣ユーザーとどれだけ似ているか

特徴：本物のユーザーが異常な行動を示した場合に弱い

12.2.2 Model-intrinsic な特徴量

攻撃者プロファイル内の評価間の相対的な関係性の歪みに基づく：

- **平均分散（MeanVar）**：悪意のあるプロファイルを構成する 3 つの部分に分けて測定
  1. ターゲットアイテムへの極端な評価
  2. プロファイルを埋めるためのフィラーアイテムへの評価
  3. 未評価のアイテム
- **フィラー平均ターゲット差分モデル（FMTS）**：ターゲットとなる部分の評価とフィラー部分の評価との差の度合いを評価
- **フィラー平均相関（FAC）**：プロファイル内の評価と全アイテムの平均評価との間の相関を反映
- **フィラー平均差分（FMD）**：プロファイルの評価と全アイテムの平均評価の絶対差の平均値を測定

12.3 検知特性のフレームワーク

1. ユーザープロファイル：類似性、サイズ、グループ行動、属性
2. ターゲット評価：集合性、歪んだ評価
3. フィラー評価：評価、長さ
4. サイド情報：共起グラフ、ユーザー・ユーザーグラフ

12.4 検知手法の概要

- 教師あり学習の手法
- 半教師あり学習の手法
- 教師なし学習の手法

12.5 防御手法

効果的な防御策の一つがロバスト最適化：

- オープン性の課題：外れ値検知、データサニタイゼーション、前処理技術
- コンセプトドリフト：動的学習アルゴリズム
- データの不均衡：アンサンブル手法

13 対策の効果性

13.1 効果的な対策

- ユーザーと評価に関する攻撃の研究が多く、それに伴い**ユーザーと評価に関する攻撃に効果的な対策が多い**

13.2 弱い対策

- 全体として、特定の攻撃に対して弱い対策はごく少数

14 今後の研究方向性

14.1 ポイズニング攻撃のギャップ、限界、そして今後の方向性

1. **攻撃者への知識：ホワイトボックス、グレーボックスからブラックボックス攻撃へ**
    - 既存のポイズニング攻撃はホワイトボックスおよびグレーボックス攻撃が多い
    - 通常はプライバシーとセキュリティの懸念からターゲットとなるレコメンドシステムへのアクセスは通常制限される
2. **攻撃手法：特定モデル向けからモデルに依存しない攻撃へ**
    - 現在は、特定モデルに限定された攻撃が主流だが、これは現実世界にあまり適用できない
3. **サイド情報の統合**
    - 最近ではドメインハイパーグラフ、ソーシャルプロパティ、空間情報など推薦の質を向上させるためにサイド情報を活用するレコメンドシステムが増えている
    - サイド情報は役立つ一方、複雑さと潜在的なセキュリティリスクをもたらす

14.2 攻撃対策のギャップ、限界、そして今後の方向性

1. **攻撃前の検出**
    - 攻撃前の振る舞いを検出するアプローチがほとんどない
    - 防御、検出に加えて予測という新しいカテゴリー
2. **主要な特性を超える**
    - ユーザーの属性のような補足情報を取り入れて検知
3. **精度以外も超える**
    - 公平性と説明可能性を担保することが同様の攻撃に対する予防策を開発する上で重要
4. **オーバーヘッドの最適化**
    - 検知・防御手法の追加にかかるレコメンドシステムのパフォーマンスへの影響を最小化

15 ポイズニング攻撃と対策の関連付け

- 新しいタイプのポイズニング攻撃に関する研究の量は多いが、対策の開発に関する研究は近年あまり進んでいない
- 悪用される可能性のある様々な脆弱性を抱えており、さらに攻撃を阻止するための効果的な対策が存在しない可能性がある

16 レコメンドシステムの一般的な脆弱性の統合

一般的な脆弱性がレコメンドシステムに与える影響は整っていない：

16.1 ソフトウェアの脆弱性

- プログラムのバグと設計ミス
- ゼロデイ攻撃
- SQL インジェクション
- マルウェア感染

16.2 運用上の脆弱性

- 不適切な認証・権限管理
- 設定不備
- パッチ適用遅延
- ヒューマンエラー

16.3 物理的な脆弱性

- 不法侵入
- 自然災害

17 サーベイ論文の貢献

1. 攻撃者の課題について議論し、レコメンドシステムの攻撃に焦点を当てている
2. 古典的なヒューリスティック攻撃と AI ベースの攻撃の両方を網羅した初の包括的なレビューを行い、攻撃を 5 つの次元で分類
3. ポイズニング攻撃に対する 41 の対策について広範なレビューを提示
4. この分野における未解決の問題について記述し、将来の研究の方向性を示している
5. 新規参入の研究者に向けてプログラムコード、データセットを含む公開リポジトリを構築

18 結論

本サーベイ論文は、レコメンドシステムに対するポイズニング攻撃の現状と対策について包括的に分析している。特に以下の点が重要である：

- 古典的なヒューリスティック攻撃から AI ベースの攻撃への明確な移行が確認されている
- 攻撃手法は多様化しており、モデル内在攻撃とモデル非依存攻撃の両方が存在する
- 検知・防御手法は発展しているが、新しい攻撃手法に対する対策の開発は追いついていない
- 今後の研究では、ブラックボックス攻撃やモデル非依存攻撃への対応、サイド情報の統合、攻撃前検出の実現が重要である

レコメンドシステムのセキュリティを向上させるためには、継続的な研究と実装が必要であり、本サーベイ論文が示す方向性に沿った取り組みが求められる。