Poisoning Attacks to Graph-Based Recommender Systems

Minghong Fang, Guolei Yang, Neil Zhenqiang Gong, Jia Liu
Iowa State University; Facebook, Inc.

Shilong Zhao †

November 21, 2022

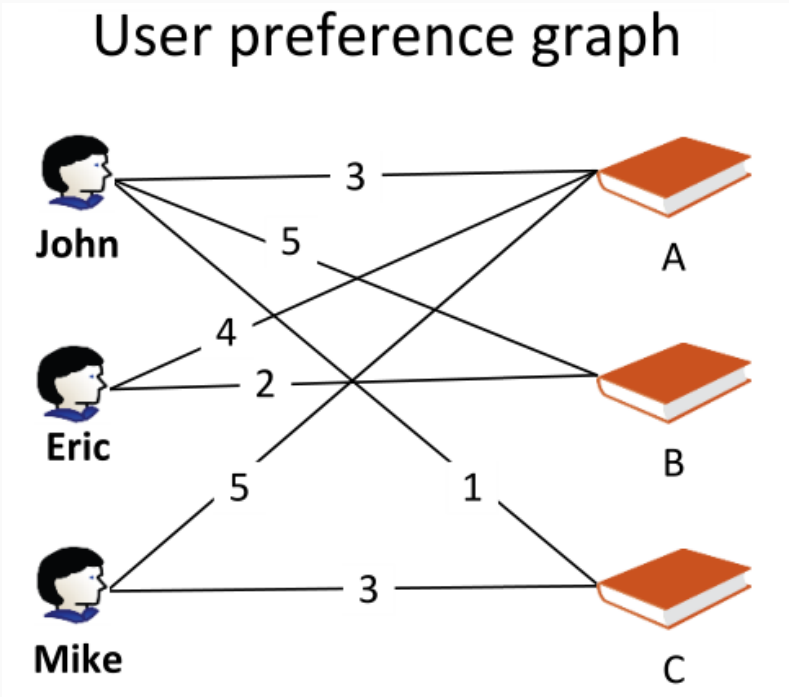
† Institute of Computing Technology
University of Chinese Academy of Sciences

Outline

Introduction

Introduction

Graph-Based Recommender Systems: **实际上是图上的** Random Walk **算法**。



Introduction

Contributions:

1. 将投毒攻击形式化为优化问题并提出近似求解的方法
2. 使用真实世界数据与其他攻击方法进行比较
3. 提出检测假用户的反制措施并验证效果

Problem Formulation

Problem Formulation

Attack goal: **最大化** Hit ratio, $\mathrm{h(t)}$ = **所有正常用户** top-N **推荐中包含了** Target Item t **的用户数量/正常用户数量**

Attack approach: **攻击者向** RS **注入假用户，控制假用户给目标物品高分，给一些其他** item **精心设计的评分（这部分** item **叫做** filler items，**每个用户有** n **个)。**

Attack knowledge: white-box setting. **攻击者知道给定推荐系统使用的推荐算法和用户-物品评分矩阵。**

$$\max h(t)$$
$$\text{subject to } |r_v|_0 \leq n+1, \forall v \in \{v_1, v_2, \cdots, v_m\}$$
$$r_{vi} \in \{0, 1, \cdots, r_{max}\}, \forall v \in \{v_1, v_2, \cdots, v_m\}$$

0 **范数：向量中非零元素的个数。**

$\mathrm{v_m}$: **第** m **个恶意用户**

Proposed Method

Proposed Method

很难得到上面提出的优化问题的精确解，于是对其进行近似：

1. 不同时优化 m 个假用户的评分，而是逐个优化他们的评分。具体来说，给定目前的正常用户和假用户，我们找到下一个假用户的评分分数，以优化 HR。（贪心）
2. 使用其他容易优化的函数近似 HR。
3. 将评分的取值范围 $\{0, 1, ..., r_{max}\}$ 放宽到实数域 $[0, r_{max}]$.

使用 PGD(projected gradient descent) 解决近似后的优化问题。
最后生成假用户对 Target item 和 filler items 的评分：

1. fake user 对 target 打最高分；
2. 根据近似优化问题中求出的 w_i 从高到低选择 n 个 item 作为 filler item；
3. 为每个 filler item 打分：从正常用户为这个 item 打分的分布中 sample 一个数，离散化后作为分数。

Experiment

Experiment

数据集:

1. MovieLens-100K(943 users, 1,682 movies, and 100,000 ratings)
2. Amazon Instant Video(5,073 users, 10,843 items, and 48,843 ratings)

Table 1: Dataset statistics.

Dataset	#Users	#Items	#Ratings	Sparsity
Movie	943	1,682	100,000	93.67%
Video	5,073	10,843	48,843	99.91%

Metric: HR@N

Baseline Attacks: On paper...

Conclusion

Conclusion

文章提出了一种针对图模型投毒攻击，目的是让目标商品推荐给尽可能多的用户。

实验结果表明：

1. 针对图模型进行攻击效果很好。
2. 也可以攻击其他类型的 RS，但效果不太好。
3. 通过使用监督学习可以检测出很大一部分虚假用户，但也可以错误地预测一小部分正常用户是虚假用户。

Q&A

Q & A

Supplement

Supplement: 近似 HR

在具有新加入的假用户 v 的图中，为了对普通用户 u 进行推荐，我们首先从 u 执行 Random Walk 算法并计算其平稳概率分布 P_u，其中 P_{ui} 是项目 i 的平稳概率。我们根据平稳概率对用户 u 未评分的项目进行排名，选择平稳概率最大的前 n 项被推荐给用户 u, 生成推荐列表 L_u。P_u 是从用户 u 开始的随机游走的平稳概率分布，即下式的解：

$$p_u = (1-\alpha)\cdot Q\cdot p_u + \alpha\cdot e_u,$$

$$Q_{xy} = \begin{cases} \dfrac{r_{xy}}{\sum_{z\in\Gamma_x} r_{xz}} & \text{if } (x,y)\in E \\ 0 & \text{otherwise,} \end{cases}$$

对于用户节点 x，Γx 是 x 评分的项目集; 对于项目节点 x，Γx 是评价 x 的用户集。转移矩阵 Q 是边权值 w_v 的函数。

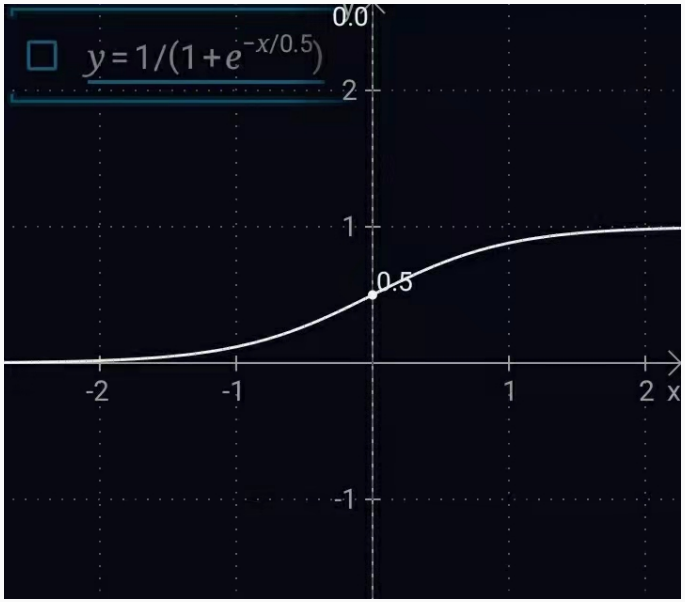
Supplement: 近似 HR

设计一个损失函数满足：

1. Filler items 排在 Target item 前时 Loss 大，反之小
2. Target 排的越高 Loss 越小

$$l_u = \sum_{i \in L_u} g(p_{ui} - p_{ut}),$$

$$g(x) = \frac{1}{1+\exp(-x/b)}$$



Supplement: 近似 HR

对所有用户求和：

$$l = \sum_{u \in S} l_u,$$

最终的优化目标：

$$\min F(w_v) = \|w_v\|_2^2 + \lambda \cdot l$$
$$\text{subject to } w_{vi} \in [0, r_{max}],$$

其中 $||\mathrm{w_v}||_2^2$ 用来限制每个假用户只能对少量商品进行评分。
这样我们就可以计算 $\mathrm{F(w_v)}$ 对 $\mathrm{w_v}$ 的导数。(On paper...)

假用户评分的生成

Algorithm 1 *Our Poisoning Attacks*

Input: Rating matrix R, parameters t, m, n, λ, b.
Output: m fake users $v_1, v_2, \cdots, v_m$.

1: //Add fake users one by one.
2: **for** $v = v_1, v_2, \cdots, v_m$ **do**
3: Solve the optimization problem in Equation 6 with the current rating matrix R to get w_v.
4: //Assign the maximum rating score to the target item.
5: $r_{vt} = r_{max}$.
6: //Find the filler items
7: The n items with the largest weights are filler items.
8: //Generate rating scores for the filler items.
9: $r_{vj} \sim \mathcal{N}(\mu_j, \sigma_j^2)$, for each filler item j.
10: //Inject the fake user with rating scores r_v to the system.
11: $R \leftarrow R \cup r_v$.
12: **end for**
13: **return** $r_{v_1}, r_{v_2}, \cdots, r_{v_m}$.