中国科学: 信息科学 2018 年 第 48 卷 第 2 期: 177–186
SCIENTIA SINICA Informationis
论文

无组织恶意攻击检测问题的研究

庞明[1,2], 周志华[1,2*]

1. 南京大学计算机软件新技术国家重点实验室, 南京 210023
2. 南京大学软件新技术与产业化协同创新中心, 南京 210023
* 通信作者. E-mail: zhouzh@lamda.nju.edu.cn

收稿日期: 2017–05–16; 接受日期: 2017–05–31; 网络出版日期: 2018–01–15
国家自然科学基金 (批准号: 61333014) 资助项目

摘要 推荐系统在我们的生活中被广泛应用, 对人们的生活起着越来越重要的影响. 然而, 协同过滤作为一种常见的推荐技术, 很容易受到伪造虚假用户评分信息的恶意攻击的影响. 为了保证推荐的质量, 很多恶意攻击检测的方法被提出用于检测恶意攻击. 现有的攻击检测方法大多是针对有组织大规模攻击的检测, 即攻击者根据同一种策略, 伪造大量的虚假用户评分信息用于提升或贬低一个目标物品. 本文研究了一种不同的攻击类型: 无组织恶意攻击, 即攻击者们在没有组织的情况下, 分别伪造少量的虚假用户评分信息来提升或贬低同一个目标物品. 无组织恶意攻击出现在很多真实的应用中, 对推荐系统的鲁棒性造成严重影响, 而针对该攻击类型的研究还很初步. 实验结果表明现有攻击检测方法不能够有效地检测无组织恶意攻击. 本文分析了现有的多种攻击检测方法无效的原因, 进而通过分析无组织恶意攻击的特性, 总结出无组织恶意攻击检测的关键.

关键词 攻击检测, 推荐系统, 协同过滤, 无组织恶意攻击, 鲁棒性

1 引言

随着互联网的发展, 网上活动已经成为我们生活中很重要的一部分. 例如, 越来越多的顾客选择在亚马逊、易趣、淘宝等平台上购物, 人们开始习惯于在 YouTube、Netflix 等视频网站观看电视节目, 等等. 推荐系统面对数量庞大的用户和物品, 需要将合适的物品推荐给相应的用户. 为此, 多种协同过滤的技术被提出, 用于帮助用户选择到适合自己的物品 [1~3].

然而, 协同过滤算法容易受到恶意攻击的影响 [4, 5], 攻击者可以通过向用户 – 物品评分矩阵中插入虚假的用户评分信息来操纵系统的推荐. 一些攻击者会通过攻击来增加自己的物品的受喜爱程度 (推广攻击), 一些攻击者则通过攻击降低竞争对手的物品的喜爱程度 (贬低攻击). 现有攻击检测工作大多考虑有组织攻击, 并在各种有组织攻击上展现了良好的检测能力 [5~7]. 有组织攻击主要体现在攻击者通过同一种攻击策略生成大量虚假用户评分信息来攻击同一个目标物品. 例如, 一个攻击组织者

引用格式: 庞明, 周志华. 无组织恶意攻击检测问题的研究. 中国科学: 信息科学, 2018, 48: 177–186, doi: 10.1360/N112017-00112
Pang M, Zhou Z-H. Unorganized malicious attacks detection (in Chinese). Sci Sin Inform, 2018, 48: 177–186, doi: 10.1360/N112017-00112

根据同一种攻击策略生成数百个虚假用户评分信息来攻击一个目标电影, 其中攻击策略为每个虚假用户给最流行的电影最高的评分, 给目标电影最低的评分.

为了减少恶意攻击的发生, 多种机制被提出, 用于提高恶意攻击的成本. 例如, 很多网站的注册需要实名制或者电话验证; 验证码被用来判断用户的反馈是否源自机器; 顾客在购买相应商品后, 才能给其评分等等. 上述这些机制导致有组织的攻击的成本变得很高. 例如, 亚马逊网站上的店家需要购买大量的商品才能伪造出数百个虚假用户的评分信息来实施一次有组织的攻击.

本文研究了一种不同的攻击模式: 无组织恶意攻击, 即攻击者们在没有组织的情况下, 分别伪造少量的虚假用户评分信息来攻击同一个目标物品. 这类攻击模式出现在很多真实的应用中. 例如, 亚马逊网站上的店家可能会在没有组织的情况下, 分别伪造少量的虚假用户评分信息来贬低同一个热门商家的商品; 作家的支持者可能会在没有组织的情况下, 分别伪造少量的虚假用户来推广其作品. 多个现实应用指出无组织恶意攻击对推荐结果产生了严重的影响[1)2)]. 我们通过实验进一步证实了这一点.

我们评测了现有的多种恶意攻击检测方法在无组织恶意攻击上的检测效果. 实验结果表明现有攻击检测方法不能够有效地检测无组织恶意攻击. 本文分析了现有多种攻击检测方法可以成功的检测有组织攻击, 但无法有效地检测无组织恶意攻击的原因. 进而通过分析无组织恶意攻击的特性, 本文总结出无组织恶意攻击检测的关键.

2 相关工作

协同过滤被广泛地应用在推荐系统中, 其基本假设是原来表现出类似的兴趣爱好的用户, 在以后也应该有类似的兴趣爱好 [8]. 由此, 大量的关于协同过滤的工作被提出 [1~3]. 协同过滤有两种主要的类别, 即基于存储的协同过滤方法和基于模型的协同过滤方法. 基于存储的协同过滤方法直接利用用户给物品的评分信息来预测用户感兴趣的物品. 这类方法又分为两大类, 即基于用户的协同过滤方法和基于物品的协同过滤方法. 其中基于用户的协同过滤方法会首先找到一个用户的相似用户, 再将相似用户喜欢的物品推荐给该用户. 用户之间的相似度被一种相似度度量定义, 通常是余弦相似度或者 Pearson 相关系数 [9]. 关于相似度度量的多种变种工作也被提出 [10, 11]. 基于物品的协同过滤方法会把一个用户喜欢的物品的相似物品推荐给该用户 [12].

基于模型的协同过滤方法首先利用用户给物品的评分信息训练得到一个预测模型, 再利用该预测模型来生成对每个用户的推荐 [13]. 例如, 混合模型 [14] 在每个聚类中学习物品的概率分布; 根据用户给物品的评分矩阵, 矩阵分解得到用户和物品对应的隐变量, 再利用该低秩近似矩阵来预测评分矩阵中的未评分项 [15]. 考虑用户 – 物品评分矩阵外的边际信息, 有很多拓展协同过滤的工作被提出 [2, 16].

上述两大类协同过滤的方法一般都假设用户给物品的评分如实地反映了用户的喜好. 但在我们实际生活中, 攻击者会通过伪造用户的方式, 来操控推荐系统, 增加自己的利益. 现有的研究工作表明基于协同过滤的推荐方法容易受到恶意攻击的影响 [4, 5].

针对这种问题, 很多恶意攻击检测的方法被提出. 现有的恶意攻击检测的方法主要包含统计的方法、分类的方法和聚类的方法 [4]. 统计的方法通过检测可疑的评分来查找恶意用户. 文献 [6] 提出一种基于 Neyman-Pearson 准则的攻击检测方法来区分正常用户和恶意用户. 分类的方法首先根据每个

1) http://www.forbes.com/sites/suwcharmananderson/.
2) http://how-to-post-fake-reviews-on-amazon.blogspot.com/.

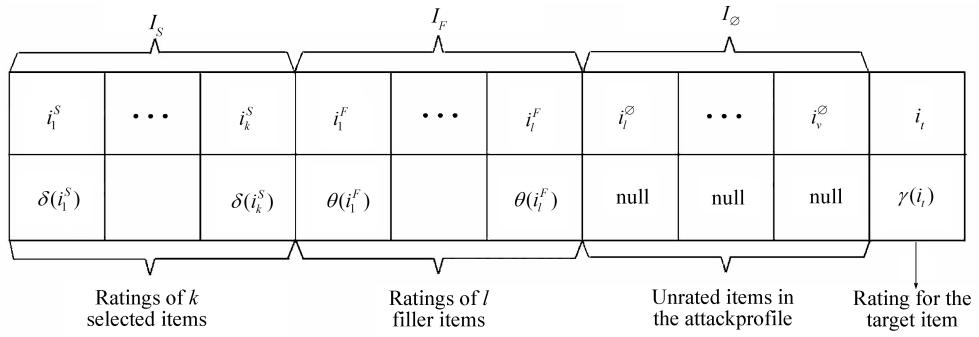


图 1 恶意用户的评分信息的一般形式

Figure 1 General form of an attack profile

用户给物品的评分信息提取出该用户的特征, 再根据用户的特征和标记 (即是否是恶意攻击者) 训练得到一个检测恶意攻击者的分类模型 [17].

聚类的方法根据用户的评分信息将用户聚成若干表现相近的簇, 其中最小的簇中的用户被视为恶意攻击者. 文献 [18] 提出了一种基于若干种分类属性 [19] 的无监督聚类算法. 他们根据这些属性进行 k-means 聚类, 将最小簇中的用户判定为恶意用户. 文献 [7] 提出一种基于 PLSA 的聚类方法, 以替代传统的最近邻方法. 文献 [20] 提出一种变量选择方法, 将用户当作变量计算用户间的协变量矩阵, 对该矩阵进行主成分分析, 选出前 l 个主成分最小的一批用户, 判定为恶意用户.

上述方法大多是针对某一种攻击策略的特性, 通过考虑用户的某些共有特征, 判定其中一类用户为恶意用户. 即现有恶意攻击检测的应用场景均是针对有组织的大规模攻击, 而非无组织的小规模的攻击. 文献 [5] 通过低秩矩阵分解方法预测用户的评分, 从而计算出每个用户的置信度, 将置信度最低的一批用户判定为恶意用户. 这种做法考虑了用户的真实评分和实际评分的差距作为评判用户置信度的指标, 而不再是针对某一种特定的攻击策略. 但由于没有考虑到不完全评分矩阵中存在恶意攻击项, 其低秩矩阵分解并不能很好地还原出用户的真实评分.

3 无组织恶意攻击

3.1 小节介绍了恶意攻击中虚假用户评分信息的一般形式, 以及多种常见的攻击策略. 在此基础上, 3.2 小节对无组织恶意攻击和有组织恶意攻击进行了区别, 给出了无组织恶意攻击的正式定义. 3.3 小节探究了无组织恶意攻击对推荐结果的影响.

3.1 恶意攻击的一般形式

文献 [21] 第一次提出了恶意攻击中恶意用户的评分信息的一般形式, 见图 1. 选择评分项 I_S 由评分函数 δ 赋值; 填充项 I_F 由评分函数 θ 赋值; 目标评分函数决定了目标项 i_t 的评分. 剩余的物品项不给予评分. 恶意攻击者利用上述形式, 生成虚假的用户评分信息. 具体过程为, 确定选择评分项 I_S 和填充项 I_F, 在 I_S 和 I_F 的评分上尽可能模仿正常用户的评分, 从而伪装成正常用户, 再通过给予目标项 i_t 偏离真实分数的评分, 达到其恶意攻击的目的.

常见的攻击策略包括随机攻击策略、平均攻击策略和从众攻击策略等 [22].

• 随机攻击策略: 攻击者从候选物品中随机选择 l 个物品组成填充项 I_F, 将随机的分数赋予填充项 I_F;

• 平均攻击策略: 攻击者从候选物品中随机选择 l 个物品组成填充项 I_F, 将每个物品的系统平均分赋予填充项 I_F;

• 从众攻击策略: 攻击者从候选物品中随机选择 k 个物品组成选择评分项 I_S, 将最高评分赋予选择评分项 I_S.

3.2 无组织恶意攻击的定义

在有组织恶意攻击中, 攻击者通过同一种策略生成大量虚假用户评分信息来攻击同一个目标物品, 即在同一种策略中, 目标项 i_t 为特定的同一个目标; 选择评分项 I_S 和填充项 I_F 的个数 k, l 在确定后也固定不变; 各个评分函数确定后也保持不变. 举例来说, 在平均攻击策略下, 攻击者从除目标 i_t 外的所有物品中随机选择 l 个物品组成填充项 I_F, 评分函数 δ 将每个物品的系统平均分赋予填充项 I_F; 在从众攻击策略下, 攻击者从除目标项 i_t 外的所有物品中随机选择 k 个物品组成选择评分项 I_S, 评分函数 θ 将最高评分赋予选择评分项 I_S. 除此之外, 在有组织恶意攻击中, 虚假用户的数目通常很大 [4].

然而在无组织恶意攻击中, 选择评分项 I_S 和填充项 I_F 的个数 k, l, 以及各个评分函数都不被限定为同一种. 除此之外, 我们还假设存在多个攻击者, 每个攻击者分别只伪造少量的虚假用户评分信息来提升或贬低他们自己的目标物品. 在下文中, 我们给出无组织恶意攻击的正式定义.

让 $U_{[m]} = \{U_1, U_2, \ldots, U_m\}$, $I_{[n]} = \{I_1, I_2, \ldots, I_n\}$ 分别表示 m 个用户和 n 个物品. 让 M 表示用户给物品的实际评分矩阵, M_{ij} 表示用户 U_i 给物品 I_j 的实际评分, 即用户在实际给出的评分. 让 X 表示用户给物品的真实评分矩阵, 即不遭受攻击的真实评分. X_{ij} 表示用户 U_i 给物品 I_j 的真实评分, 即 X_{ij} 反映出用户 U_i 给物品 I_j 的真实喜恶. 假设可以给出的最高分数为 $R_{\max}$, 则 $0 < X_{ij} \leqslant R_{\max}$, $0 < M_{ij} \leqslant R_{\max}$.

在实际的评分系统中, 真实评分矩阵 X 可能会被系统噪声干扰. 例如, 如果对于 $i \in [m]$, $j \in [n]$, $X_{ij} = 4.5$, 那么用户给出的实际评分为 5 或者 4 都应该被视为正常. 如果 $|M_{ij} - X_{ij}| > \epsilon$, 则将该评分视为恶意评分, 其对应的用户被视为恶意用户. 例如在五分制评分中, 实际评分和真实评分的差距大于 3, 那么可以将其视为恶意评分. 我们定义无组织恶意攻击为 $\forall j \in [n]$, 满足 $|M_{ij} - X_{ij}| > \epsilon$ 的恶意用户来自不同策略, 且每种策略生成的用户数小于 η, 则该评分矩阵中存在的攻击为无组织恶意攻击.

3.3 无组织恶意攻击对推荐结果的影响

本小节探究了无组织恶意攻击对推荐结果的影响. 假设推荐算法为标准的基于用户的 k 近邻推荐算法, 其中 $k = 20$, 使用 Pearson 相关系数度量距离. 评测使用数据集 MovieLens 100 K, 其评分范围是 1 到 5. 从所有平均分小于 2 的物品中随机选出一个物品作为攻击目标. 攻击者们随机选取平均攻击策略、随机攻击策略和从众攻击策略中的一种生成伪造用户, 并给攻击目标打高分. 我们对不同评分比率 (评分物品数占全部物品数的百分比) 设置的无组织恶意攻击进行了探究, 其中评分比率分别为 0.01, 0.03, 0.06, 以及这 3 种评分比率的混合. 实验重复 20 次. 预测偏移表示遭受攻击后的推荐预测结果较遭受攻击前发生的改变. 预测操纵率表示遭受攻击后的推荐预测中评分超过 4 的比率. 图 2(a) 展示了平均预测偏移, 其中横坐标表示恶意用户占全体用户的比例, 从 0.01 到 0.10. 图 2(b)

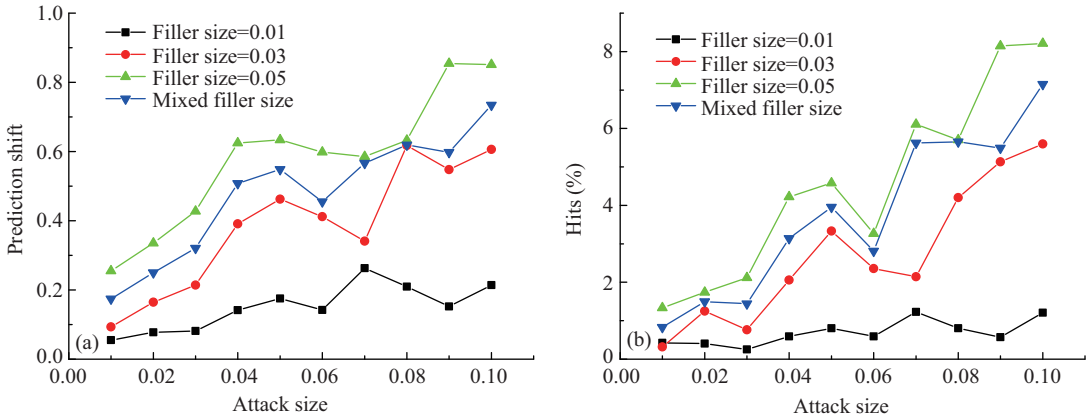


图 2 (网络版彩图) 无组织恶意攻击对推荐系统的影响

Figure 2 (Color online) Effectiveness of unorganized malicious attacks. (a) Prediction shift; (b) hits

展示了平均预测操纵率. 由图 2 可以看出, 无组织恶意攻击在不同的评分比率设置下, 都对目标物品的推荐结果产生了明显的提升作用, 评分比率越大, 提升作用越明显.

4 实验测试

本节评测了现有的多种恶意攻击检测方法在无组织恶意攻击上的检测效果.

为了评估这些方法的检测效果, 我们采用了攻击检测的查准率 P, 查全率 R 和 $F1$ 作为评价指标. 这些评价指标可由如下方式计算 [4]:

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\ R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\ F1=\frac{2\times P\times R}{P+R},$$

其中, TP 表示判定为恶意用户的集合中真实的恶意用户的比例, FP 表示判定为恶意用户的集合中真实的正常用户的比例, FN 表示判定为正常用户的集合中真实的恶意用户的比例.

4.1 数据集

我们首先在攻击检测常用的数据集 MovieLens 100 K 和 MovieLens 1 M 上构造人工数据进行了攻击检测的实验, 这两个数据集来自 GroupLens[3]. 评分的范围是 1 到 5, 其中 1 为最低评分, 5 为最高评分. 数据集 MovieLens 100 K 包含 943 个用户在 1682 个电影上的 100000 个评分, 数据集 MovieLens 1 M 包含 6040 个用户在 3706 个电影上的 1000209 个评分. 这些数据本身并不包含攻击, 我们通过人工生成的方式, 将恶意用户加入其中, 具体细节会在 4.3 小节详细给出. 重复实验 10 次, 最终实验结果取均值.

除此之外, 我们还从 Douban 网[4]上搜集得到了一个真实的包含恶意用户的数据集 Douban 10 K. Douban 10 K 包含 213 个用户在 155 个物品上的 12095 个评分, 评分的范围是 1 到 5, 其中 1 为最低评分, 5 为最高评分. 在 213 个用户中有 35 个用户为恶意用户.

3) http://grouplens.org/datasets/movielens/.
4) http://www.douban.com/.

4.2 比较方法

本小节列举了实验中对比的 4 种具有代表性的恶意攻击检测方法, 以及也可以被用于恶意攻击检测的鲁棒主成分分析 (RPCA) 方法. 各方法的实验设置皆使用相应引文的推荐设置.

• N-P 方法: 该方法为基于统计的攻击检测方法, 其基于 Neyman-Pearson 准则来区分正常用户和恶意用户 [6].

• k-means 方法: 该方法为基于聚类的攻击检测方法, 其基于若干种分类属性进行 k-means 聚类, 将最小簇中的用户判定为恶意用户 [18].

• PCAVarSel 方法: 该方法将用户当作变量计算用户间的协变量矩阵, 对该矩阵进行主成分分析, 选出前 l 个主成分最小的一批用户, 判定为恶意用户 [20].

• MF-based 方法: 该方法通过低秩矩阵分解预测用户的评分, 从而计算出每个用户的置信度, 将置信度最低的一批用户判定为恶意用户 [5].

• RPCA 方法: 该方法是考虑到稀疏噪声和微小扰动噪声的低秩矩阵恢复方法 [23].

4.3 实验结果

数据集 MovieLens 100 K 和 MovieLens 1 M 上并不包含攻击, 我们通过人工生成的方式, 将恶意用户加入其中. 在第 1 组实验中, 我们通过多种传统攻击策略的组合构造无组织恶意攻击, 将生成的恶意用户评分信息加入到原始的数据集中, 得到遭受无组织恶意攻击的数据集.

我们采取的传统攻击策略包括随机攻击策略、平均攻击策略和从众攻击策略 [22].

从平均分少于 2 的物品中随机选择一个作为目标物品, 每个攻击者随机选取一种策略生成一个虚假的用户评分信息, 并给该物品高分, 以推广该物品. 为了与之前的攻击检测工作保持一致, 我们设置评分比率 (评分物品数占全部物品数的百分比) 为 0.01, 评分物品从评分最多的前 10% 物品中随机挑选. 我们设置恶意用户比率 (恶意用户占全部用户的比率) 为 0.2. 表 1 总结了比较方法在数据集 MovieLens 100 K 和 MovieLens 1 M 上构造的人工数据上的实验结果.

在第 2 组实验中, 我们除了考虑传统攻击策略的组合外, 还考虑攻击者可以雇佣系统中已存在的用户进行恶意攻击. 我们按照如下方式为数据集 MovieLens 100 K 和 MovieLens 1 M 添加无组织恶意攻击. 我们设置评分比率 (评分物品占全部物品的百分比) 为 0.01, 评分物品从评分最多的前 10% 物品中随机挑选. 此处, 无组织恶意攻击的生成策略包含两种, 一小部分恶意用户 (25%) 的生成策略与第 1 组实验中相同; 剩余恶意用户 (75%) 的生成按照如下策略:

(1) 从平均分少于 2 的物品中随机选择一个作为目标物品 i_t.

(2) 从系统中给 i_t 的评分小于 2 的用户中, 随机挑选一个用户 U_C.

(3) 将 U_C 给 i_t 的评分更改为 5.

表 2 总结了比较方法在数据集 MovieLens 100 K 和 MovieLens 1 M 上按照上述方式构造的人工数据上的实验结果. 表 3 总结了比较方法在数据集 Douban 10 K 上的实验结果. 由表 1~3 可以看出传统的恶意攻击检测方法无法有效地检测出无组织恶意攻击. 除 MF-based 方法在第 1 组实验中的结果以外, 其他实验结果的 $F1$ 都远低于 0.8, 无法在真实的无组织恶意攻击检测中使用. 而且可以观察到, 当数据规模变大数据变稀疏 (从 MovieLens 100 K 到 MovieLens 1 M), 这些方法的表现明显变差.

传统的恶意攻击检测方法主要是针对某一种攻击策略的特性, 通过考虑用户的某些共有特征, 判定其中一类用户为恶意用户. 当面对无组织攻击时, 因为攻击策略不单一, 多种多样, 每种策略生成的恶意用户数目较小, 导致传统的恶意攻击检测方法无法准确地找到具有共同特征的一批用户. 例如,

表 1 恶意攻击检测的查准率、查全率和 $F1$, 其中 MovieLens 受到传统攻击策略组合的无组织恶意攻击

Table 1 Detection precision, recall and $F1$ on MovieLens under unorganized malicious attacks based on traditional strategies

	MovieLens 100 K			MovieLens 1 M		
	P	R	$F1$	P	R	$F1$
RPCA	0.908±0.010	0.422±0.048	0.575±0.047	0.342±0.003	0.558±0.028	0.424±0.009
N-P	0.774±0.015	0.641±0.046	0.701±0.032	0.711±0.007	0.478±0.018	0.572±0.014
k-means	0.723±0.171	0.224±0.067	0.341±0.092	0.000±0.000	0.000±0.000	0.000±0.000
PCAVarSel	0.774±0.009	0.587±0.024	0.668±0.019	0.278±0.007	0.622±0.022	0.384±0.011
MF-based	0.911±0.009	0.814±0.008	0.860±0.009	0.407±0.005	0.365±0.004	0.385±0.005

表 2 恶意攻击检测的查准率、查全率和 $F1$, 其中 MovieLens 受到一般形式的无组织恶意攻击

Table 2 Detection precision, recall and $F1$ on MovieLens which are under general unorganized malicious attacks

	MovieLens 100 K			MovieLens 1 M		
	P	R	$F1$	P	R	$F1$
RPCA	0.797±0.046	0.659±0.097	0.721±0.097	0.635±0.012	0.391±0.022	0.484±0.015
N-P	0.244±0.124	0.145±0.089	0.172±0.084	0.273±0.020	0.099±0.031	0.144±0.035
k-means	0.767±0.029	0.234±0.042	0.357±0.051	0.396±0.026	0.300±0.039	0.341±0.035
PCAVarSel	0.481±0.027	0.168±0.017	0.248±0.023	0.120±0.006	0.225±0.012	0.157±0.008
MF-based	0.556±0.023	0.496±0.021	0.524±0.022	0.294±0.012	0.264±0.010	0.278±0.011

表 3 恶意攻击检测在 Douban 10K 上的查准率、查全率和 $F1$

Table 3 Detection precision, recall and $F1$ compared with other algorithms on dataset Douban 10 K

	RPCA	N-P	k-means	PCAVarSel	MF-based
P	0.535	0.250	0.321	0.240	0.767
R	0.472	0.200	0.514	0.343	0.657
$F1$	0.502	0.222	0.396	0.282	0.708

k-means 方法和 N-P 方法需要在恶意用户的分类属性或潜在类别相似的情况下, 才可以检测出恶意用户; PCAVarSel 方法只有当恶意用户之间相同的未评分项比正常用户之间更多的时候, 才能检测出恶意用户. 然而, 这些条件在无组织恶意攻击中是不成立的. RPCA 方法和 MF-based 方法试图根据观测到的评分矩阵, 恢复出真实的评分矩阵. 然而这两种方法很难将由攻击造成的稀疏矩阵和由环境导致的噪声矩阵区分开, 因此很难取得良好的查准率和查全率. 这种现象在更大更稀疏的数据上更加明显.

考虑到不同的推荐系统中可能存在不同的恶意用户比率, 因此我们变化恶意用户比率从 2% 到 20%, 比较不同方法的效果. 实验结果见图 3. 随着恶意用户比率降低, 除 RPCA 方法以外的其他方法的表现变差, 即恶意用户比率变低会导致传统恶意攻击检测方法的效果变差. 随着恶意用户比率增高, 虽然 k-means 方法的查准率有所上升, 但其查全率显著下降, 导致 $F1$ 并没有明显变化. MF-based 方法的查准率和查全率虽然有所上升, 但仍然很低, 达不到可以实际使用的效果. 随着恶意用户比率的变化, 虽然 RPCA 表现很稳定, 但根据表 1 和 2 可知, 当数据规模变大、数据变稀疏, RPCA 方法的表现显著下降, 无法应对现实应用中的无组织恶意攻击者检测.

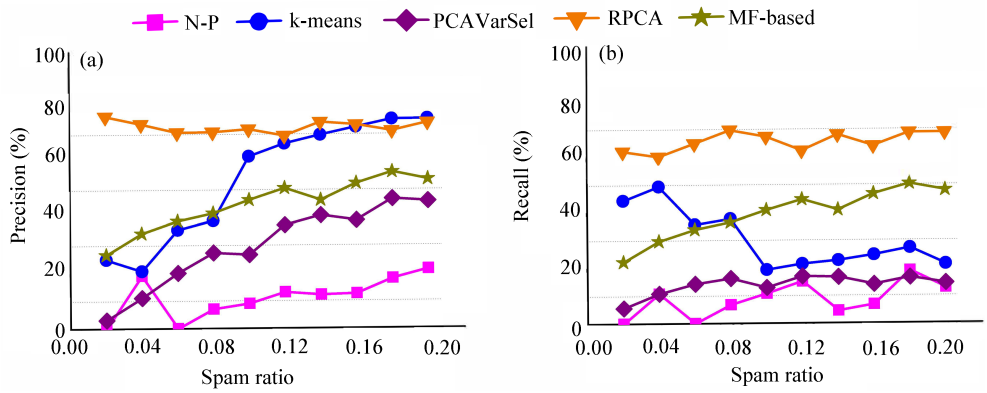


图 3 (网络版彩图) 无组织恶意攻击检测在 MovieLens 100K 上的查准率和查全率, 其中恶意用户比率的变化范围从 0.02 到 0.2

Figure 3 (Color online) Detection (a) precision and (b) recall on MovieLens 100K under unorganized malicious attacks. The spam ratio varies from 0.02 to 0.2

5 结束语

恶意攻击检测在保证推荐系统的推荐质量上起着十分重要的作用. 现有的攻击检测方法大多是针对有组织大规模攻击的检测. 本文研究了一种不同的攻击类型: 无组织恶意攻击, 即攻击者们在无组织的情况下, 分别伪造少量的虚假用户评分信息来提升或贬低同一个目标物品. 无组织恶意攻击出现在很多真实的应用中, 而针对该攻击类型的研究还很初步. 通过无组织恶意攻击检测的实验, 我们发现现有的攻击检测方法不能够有效地检测无组织恶意攻击, 进而分析了现有多种攻击检测方法无效的原因. 根据无组织恶意攻击的特性, 本文总结出在检测无组织恶意攻击时, 不能将检测局限在某一种攻击策略下, 而是需要根据实际评分矩阵恢复出真实的评分矩阵, 根据实际评分和真实评分的差距, 检测出恶意用户. 最近已经开始有这方面的探索 [24].

参考文献

1 Bresler G, Chen G, Shah D. A latent source model for online collaborative filtering. In: Proceedings of the 28th Advances in Neural Information Processing Systems, Montréal, 2014. 3347–3355

2 Li B, Yang Q, Xue G R. Transfer learning for collaborative filtering via a rating-matrix generative model. In: Proceedings of the 26th International Conference on Machine Learning, Montréal, 2009. 617–624

3 Rao N, Yu H F, Ravikumar P, et al. Collaborative filtering with graph information: consistency and scalable methods. In: Proceedings of the 29th Advances in Neural Information Processing Systems, Montréal, 2015. 2098–2106

4 Gunes I, Kaleli C, Bilge A, et al. Shilling attacks against recommender systems: a comprehensive survey. Artif Intell Rev, 2014, 42: 767–799

5 Ling G, King I, Lyu M R. A unified framework for reputation estimation in online rating systems. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, 2013. 2670–2676

6 Hurley N J, Cheng Z P, Zhang M. Statistical attack detection. In: Proceedings of the 3rd ACM Conference on Recommender Systems, New York, 2009. 149–156

7 Mehta B. Unsupervised shilling detection for collaborative filtering. In: Proceedings of the 22nd International Conference on Artificial Intelligence, Vancouver, 2007. 1402–1407

8 Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry. Commun ACM, 1992, 35: 61–70

9 Singhal A. Modern information retrieval: a brief overview. IEEE Data Eng Bull, 2001, 24: 35–43

10 Adomavicius G, Singhal A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng, 2005, 17: 734–749
11 Zhang J Y, Pu P. A recursive prediction algorithm for collaborative filtering recommender systems. In: Proceedings of the 1st ACM Conference on Recommender Systems, Minneapolis, 2007. 57–64
12 Deshpande M, Karypis G. Item-based top-n recommendation algorithms. ACM Trans Inf Syst, 2004, 22: 143–177
13 Ekstrand M D, Riedl J T, Konstan J A. Collaborative filtering recommender systems. Found Trends Human-Comput Interact, 2011, 4: 81–173
14 Kleinberg J, Sandler M. Using mixture models for collaborative filtering. J Comput Syst Sci, 2008, 74: 49–69
15 Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using markov chain monte carlo. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, 2008. 880–887
16 Salakhutdinov R, Mnih A, Hinton G. Restricted boltzmann machines for collaborative filtering. In: Proceedings of the 24th International Conference on Machine Learning, Corvallis, 2007. 791–798
17 Mobasher B, Burke R, Bhaumik R, et al. Attacks and remedies in collaborative recommendation. Intell Syst, 2009, 22: 56–63
18 Bhaumik R, Mobasher B, Burke R D. A clustering approach to unsupervised attack detection in collaborative recommender systems. In: Proceedings of the 7th International Conference on Data Mining, Vancouver, 2011. 181–187
19 Bryan K, O'Mahony M, Cunningham P. Unsupervised retrieval of attack profiles in collaborative recommender systems. In: Proceedings of the 2nd ACM Conference on Recommender Systems, Lausanne, 2008. 155–162
20 Mehta B, Nejdl W. Unsupervised strategies for shilling detection and robust collaborative filtering. User Model User-Adapted Interact, 2009, 19: 65–97
21 Bhaumik R, Williams C, Mobasher B, et al. Securing collaborative filtering against malicious attacks through anomaly detection. In: Proceedings of the 4th Workshop on Intelligent Techniques for Web Personalization, Boston, 2006
22 Lam S K, Riedl J. Shilling recommender systems for fun and profit. In: Proceedings of the 13th International Conference on World Wide Web, New York, 2004. 393–402
23 Candes E J, Li X D, Ma Y, et al. Robust principal component analysis? J ACM, 2011, 58: 1–37
24 Pang M, Gao W, Tao M, et al. Unorganized malicious attacks detection. arXiv:1610.04086, 2016

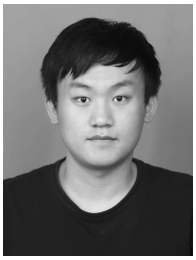
Unorganized malicious attacks detection

Ming PANG[1,2] & Zhi-Hua ZHOU[1,2*]

1. *National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023*, China*;
2. *Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University, Nanjing* 210023*, China*
* Corresponding author. E-mail: zhouzh@lamda.nju.edu.cn

Abstract Recommender system has attracted much attention during the past decade. However, collaborative filtering as a usual technique is vulnerable to malicious attacks that generate fake profiles to manipulate the system. Prior research has shown that attacks can significantly affect the robustness of the systems. Thus, many attack detection algorithms have been developed for better recommendation. Most previous approaches focus on organized malicious attacks, where the attack organizer fakes many user profiles using the same strategy to promote or demote an item. In this study, we analyze a different attack style: unorganized malicious attacks, where attackers fake a small number of user profiles to attack the same target item without any organizer. This attack style occurs in many real applications, which can significantly affect the robustness of a recommender system, yet relevant studies are inadequate. We conduct extensive experiments to study the performance of state-of-the-art attack detection approaches in unorganized malicious attack detection and discuss different approaches regarding their performance. Experimental results show that existing attack detection approaches cannot detect unorganized malicious attacks efficiently. By explaining the inefficiency of these attack detection approaches and the characteristics of unorganized malicious attacks in detail, we provide a possible research direction to develop new detection schemes for unorganized malicious attack detection.

Keywords attack detection, recommender systems, collaborative filtering, unorganized malicious attacks, robustness



Ming PANG was born in 1992. He received his B.S. degree in computer science and technology from Nanjing University, Nanjing, China, in 2014. Now, he is a Ph.D. candidate in Nanjing University. His research interests mainly include machine learning and data mining.

Zhi-Hua ZHOU was born in 1973. He received his B.S., M.S. and Ph.D. degrees in computer science from Nanjing University, China, in 1996, 1998 and 2000, respectively, all with the highest honors. Currently, he is a professor at Nanjing University. His main research interests include artificial intelligence, machine learning, and data mining.