<http://www.sohu.com/a/274019860_129720>

南大庞明、周志华等

无组织恶意攻击，即攻击者在没有任何组织者的情况下单独使用少量的用户配置文件来攻击目标

解决方案——无组织恶意攻击检测算法（UMA）

背景：针对推荐系统的攻击。

常用于推荐的方法：协同过滤。

针对协同过滤推荐方法的某种攻击：排名操纵（协同过滤无力应对这种攻击），攻击者可能会向 user-item 评分矩阵中插入虚假的评分来使系统产生偏差。分为推攻击（增加他们内容的流行度），核攻击（减少竞品的流行度）。按XX可分为托攻击和无组织恶意攻击。

托攻击：目前大多数攻击检测研究聚焦于托攻击，并且在多种托攻击策略上表现出了较好的检测性能。他们认为所有攻击事件都是通过相同的策略来促进或降级特定项目。 例如，攻击组织者可以通过一个策略产生数百个假用户配置文件，每个假用户专业人员为最受欢迎的电影提供高分，并为目标电影提供低分数以降级。在这种策略中，每个假用户会给最流行的电影给出高分评价（推攻击），而给要降级的目标电影给出低分评价（核攻击）。各种各样切实可行的技术被开发用于控制托攻击，例如，网站注册需要实名和电话号码认证；验证码被用于确定某个响应是否由机器人生成；客户在购物网站上购买了某个产品之后才能进行评价。基于这些方法，传统的托攻击可能面临比较高昂的代价。例如，Amazon 等电商网站上的小型在线零售商可能不太愿意生成百上千个虚假客户来实现一次托攻击。

无组织恶意攻击：攻击者在没有任何组织者的情况下单独使用少量的用户资料来攻击目标。这种攻击类型在很多实际应用中都有发生，例如，Amazon 上的在线商店可能会制造一些虚假评价，降低其竞品高质量鞋子的评分；作家可能会雇佣几个读者给他们的低质量书籍打好评。实际证明这种少数的无组织恶意攻击会严重地影响系统，例如，首次恶意差评能够将卖家的销量降低13%。

|  |  |  |
| --- | --- | --- |
|  | 托攻击 | 无组织恶意攻击 |
| 参与的用户属性 | 虚假 | 真实? |
| 参与的用户数量 | 成百上千 | 少量 |
| 是否有组织 | 有组织 | 无组织 |

研究者先将无组织恶意攻击公式化为矩阵补全问题的变体。X 代表没有噪声和攻击的真实评价矩阵，该矩阵是低秩矩阵，因为用户的偏好会受多个因素的影响 (Salakhutdinov et al., 2007)。让 Y 代表稀疏攻击评分矩阵，Z 代表噪声矩阵。我们可以观察到一个（部分）矩阵 M = X + Y + Z。据我们所知，之前的研究工作没有对攻击检测做出类似的公式化阐述。本研究中的优化问题和鲁棒 PCA 之间的主要区别是：鲁棒 PCA 主要聚焦于从完全或者不完全矩阵中恢复低秩矩阵 X，本研究则更注重从微弱扰动的噪声项 Z 中区分稀疏攻击项 Y。（区分噪声还是攻击）（一个[矩阵](https://baike.baidu.com/item/%E7%9F%A9%E9%98%B5)*A*的**列秩**是*A*的[线性独立](https://baike.baidu.com/item/%E7%BA%BF%E6%80%A7%E7%8B%AC%E7%AB%8B/3209637" \t "_blank)的**纵列**的极大数。）

理论上，本研究证明了低秩评价矩阵 X 和稀疏矩阵 Y 可以在一些经典的矩阵补全假设下恢复。本研究提出了无组织恶意攻击检测算法（UMA），可以看作是一种近似交替分裂增广拉格朗日（proximal alternating splitting augmented Lagrangian）方法。研究者开发了一些新技术，证明该方法在全局收敛的最糟糕的情况下也具有 O(1/t) 的收敛速度。实验结果证明了该算法的有效性，且优于当前最优的攻击检测方法。

二、相关工作

协同过滤（CF）是构建推荐系统最成功的技术之一。 CF的核心假设是，如果用户过去表达类似的兴趣，他们将在未来分享共同的兴趣（Goldberg等，1992）。自那时以来已经取得了关于CF的重大进展（Salakhutdinov等人，2007; Li等人，2009; Bresler等人，2014; Rao等人，2015）。传统CF有两种主要类别（基于用户项目评级矩阵），它们是基于存储器和基于模型的CF算法。基于存储器的CF基于用户项目矩阵的全部或部分来预测用户对项目的评级。它可以细分为基于用户和基于项目的CF.典型的基于用户的CF方法通过聚合一些类似用户的评级来预测用户的评级。用户相似性由相似性度量来定义，通常是余弦相似度或Pearson相关性（Singhal，2001）。已经提出了许多关于相似性度量的修改和调整（Adomavicius＆Singhal，2005; Zhang＆Pu，2007）。基于项目的CF方法根据用户给出的项目的评级来预测用户的项目评级（Deshpande＆Karypis，2004）。

基于模型的CF方法使用用户项目矩阵来训练预测模型，并且从预测模型生成推荐（Ekstrand等，2011）。 例如，混合模型学习每个集群中项目的概率分布（Kleinberg＆Sandler，2008）; 矩阵分解技术从用户项目矩阵中学习用户和项目的潜在因素，然后使用低秩近似矩阵来预测未评级项目的分数; 从概率的角度来看，Salakhutdinov＆Mnih（2008a）提出了概率矩阵分解框架。 考虑到除用户项目矩阵之外的辅助信息，许多工作扩展了CF范例（Basilico＆Hofmann，2004; Salakhutdinov等，2007）。

然而，CF方案的两个主要类别都易受攻击（Gunes等，2014; Aggarwal，2016）。 人们越来越关注攻击检测。 研究人员已经提出了几种方法，这些方法主要是基于统计，聚类，分类和基于数据简化的方法（Gunes等，2014）。 这些方法主要集中在托攻击，托攻击用相同的策略生成一大部分用户去升级或降级一个特定的项目。 统计方法用于检测给出可疑评级的异常。 Hurley等人（2009）提出了一个Neyman-Pearson统计分析检测方法，区分了正常用户的攻击者。 同样，概率贝叶斯网络模型用于Li＆Luo（2011）。 基于来自用户配置文件的属性，分类方法通过kNN，SVM，粗糙集理论等检测攻击（Mobasher等，2009; He等，2010）。（检测托攻击的方法：统计、聚类、分类、数据简化）

基于几种分类属性的无监督聚类算法（Bryanetal。，2008）在Bhaumik等人的论文中提出。（2011年）。 他们应用基于这些属性的k-means聚类，并将最小群集中的用户分类为攻击者。 Mehta（2007）提出了一种基于PLSA的聚类方法，而不是使用传统的最近邻方法。 Mehta＆Nejdl（2009）提出了变量选择方法，该方法将用户视为变量并计算其协方差矩阵。 通过分析协方差矩阵的主成分，在前l个主成分中有最小协方差的用户在最终的变量选择中被选择。 凌等人（2013）尝试使用低秩矩阵分解方法来预测用户的评级。 用户的声誉根据预测的评级计算，名誉低的用户被分类为恶意用户。

这些方法基于由相同攻击策略产生的攻击特征的共同特征来实现检测。 当推荐系统处于无组织的恶意攻击时，不同的攻击者使用不同的策略来产生攻击专业？或雇用现有用户来攻击不同的目标。 在这种情况下，传统的攻击检测方法可能不适合。

从破坏的矩阵中恢复低维度结构与鲁棒PCA相关（Cand'esetal。，2011; Yietal。，2016; Bouwmans等，2017）。 他们专注于从完整或不完整矩阵中恢复低级别部分X，它与攻击检测（这是我们的任务）不同。 我们更注重区分稀疏攻击项Y和小扰动噪声项Z.

三、公式化

在本节中，我们将介绍攻击配置文件的一般形式，然后我们详细比较无组织的恶意攻击和托攻击。还介绍了无组织恶意攻击的正式定义和相应的检测问题。