

# 基于混合特征值的托攻击检测算法

雷梦宁, 丁爱玲, 王新美, 韩佳倩, 曹 苗

(长安大学 信息工程学院, 陕西 西安 710061)

**摘 要:**传统的托攻击检测方法多采用基于评分值差异的算法, 其在小规模情况下易造成误判率过高的问题。通过分析真实用户和攻击用户评分项目选择方式的差异, 文中提出了一种基于混合特征值的托攻击检测算法。该算法在 Degsim、MeanVar、WDA 特征检测指标组成的特征模型基础上, 加入了流行项目卡方估计值(Chi-square of popular item, CHIP)、新颖项目卡方估计值(Chi-square of novel item, CHIN)两个特征检测指标, 构成一种新的特征模型。该特征模型在传统方法的基础上, 提出对项目与流行项目、项目与新颖项目之间的关联程度的考量, 依据特征属性选择 K-means 聚类与阈值判断相结合的分类方法, 可有效区分攻击用户和正常用户。实验对比表明, 该算法在小规模情况下可有效解决误判率高的问题, 具有更好的检测准确度。

**关键词:**推荐系统; 托攻击; 混合特征; 卡方估计值; 聚类算法

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2021)10-0087-06

doi:10.3969/j.issn.1673-629X.2021.10.015

## Shilling Attack Detection Algorithm Based on Hybrid Eigenvalue

LEI Meng-ning, DING Ai-ling, WANG Xin-mei, HAN Jia-qian, CAO Miao

(School of Information Engineering, Chang'an University, Xi'an 710061, China)

**Abstract:** Aiming at the problem of high misjudgment rate in shilling attack detection of traditional differential algorithm based on score value, we propose a shilling attack detection algorithm based on the hybrid eigenvalue by analyzing the selection mode difference of score item between real users and false users. The algorithm adds two feature detection indexes CHIP (Chi-square of popular item) and CHIN (Chi-square of novel item) to form a new feature model on the basis of the feature model with three feature detection indexes, Degsim, MeanVar and WDA. In the new feature model, we consider the correlation degrees between the item and popular item, and between the item and novel item based on the traditional differential algorithm, and according to characteristic attribute, adopt the classification method which combines K-means clustering and threshold judgment to distinguish the attacking user from the normal user effectively. Experiment shows that the proposed algorithm can effectively solve the problem of high misjudgment rate in small scale and has better accuracy.

**Key words:** recommendation system; shilling attack; hybrid feature; Chi-square; clustering algorithm

## 0 引言

互联网时代的迅速发展,使“信息过载”现象愈发严重,寻找一种可以辨别有效信息的手段至关重要。随着用户对信息筛选的需求,搜索引擎应运而生,其通过在特定位置输入一些简单的关键词寻找与该关键词相关的信息。但其提供的海量信息仍需用户消耗大量时间精力去筛选。

推荐系统(recommender systems)<sup>[1-4]</sup>的出现有效缓解了信息过多带来的影响,其能够在海量的搜索结果中,依据用户的浏览记录、行为习惯、兴趣爱好等记

录进行分析,为用户推荐最符合搜索预期的信息,从而缩短用户寻找有效信息的时间,为客户信息检索带来了极大的便利。其中,协同过滤(collaborative filtering, CF)作为推荐系统中最为有效的手段之一,广泛应用于生活中的各种领域,如 Facebook、YouTube 等。

推荐系统依靠其庞大的用户群体来为客户推荐较为准确的信息,一些商家利用该系统的开放性,通过注入大量攻击概貌<sup>[5]</sup>影响系统推荐结果,以此来提高或降低商品的系统推荐频率,从而谋取暴利。这种行为

收稿日期: 2020-07-03

修回日期: 2020-11-04

基金项目: 国家自然科学基金-青年科学基金项目(61806023)

作者简介: 雷梦宁(1997-),女,硕士研究生,CCF 会员(C3640G),研究方向为数字图像处理、托攻击检测;丁爱玲,博士,教授,研究方向为数字图像处理、智能信号与信息处理。

被称为托攻击(shilling attacks)<sup>[6-7]</sup>。其不正当的商业竞争行为造成系统推荐信息虚假或精确度不高等影响,偏离客户搜索预期。因此对托攻击进行防范检测具有重大的意义。

现有的托攻击检测方法对基本托攻击模型检测效果明显,文献[8-9]提出了一种基于特征分析的托攻击检测算法,可以针对不同类型的托攻击选取有效的检测指标,通过托攻击检测指标识别出攻击用户。但该方法不适合用在复杂的攻击模型下。文献[10]对推荐系统中现有的托攻击检测技术和鲁棒性能进行了分析,发现现有的检测算法大多是基于评分值差异提取的特征,容易造成误判率过高的问题。

受此启发,文中针对用户选择评分项目方式的不同,提出了一种基于混合特征值的托攻击检测算法。该算法考虑到项目流行度和新颖度的特性,选择了五

项特征检测指标构建特征模型对托攻击进行检测。最后,通过在 MovieLens 数据集上的实验,验证该特征模型可以有效检测出攻击用户。

## 1 相关工作

### 1.1 攻击概貌

攻击概貌由攻击者的所有评分构成,包括四个部分<sup>[7]</sup>:填充项目集、选择填充项目集、未评分项目填充集、目标项目集。填充项目是攻击者选取其他评分项目进行填充,填充项往往是随机的,可以掩护目标项目躲避检测。选择填充项目是特定的,由攻击者精心挑选,进行有效攻击。即攻击用户除对目标项目进行评分外,还对其他项目进行评分,使得攻击用户与正常用户更加接近,增加检测难度。攻击概貌的结构如表 1 所示。

表 1 攻击概貌的结构

填充项目集 $I_f$			选择填充项目集 $I_s$			未评分项目填充集 $I_e$			目标项目集 $I_t$
项集成员	item <sub>fi</sub>	...	item <sub>fm</sub>	item <sub>si</sub>	...	item <sub>sm</sub>	item <sub>ei</sub>	...	item <sub>et</sub>
评分函数	$r_f( * )$			$r_s( * )$			$r_e( * )$		
							$r_t( * )$		

### 1.2 攻击类型

文献[11]提出了随机攻击和均值攻击,其为两种基本的标准攻击模型,文献[12-13]提出了流行攻击、分段攻击和 love/hate 攻击。Gunes 等<sup>[14]</sup>在流行攻击

基础上,讨论了逆流行攻击等混淆攻击。不同攻击模型对推荐系统评分集所需的先验知识不同。表 2 列出了 4 种常见攻击模型的生成策略,其中  $I_s$  代表选择填充项目集,  $I_f$  代表填充项目集,  $I_t$  代表目标项目集。

表 2 四种攻击模型

攻击模式	$I_s$		$I_f$		$I_t$
	项目	评分	项目	评分	
随机攻击	无	无	随机选择	$r_{random}$	$r_{max}/r_{min}$
平均攻击	无	无	随机选择	$r_{average}$	$r_{max}/r_{min}$
流行攻击	流行项目	$r_{max}$	随机选择	$r_{random}$	$r_{max}/r_{min}$
分段攻击	目标项目的近邻	$r_{max}/r_{min}$	随机选择	$r_{min}/r_{max}$	$r_{max}/r_{min}$

表中,  $r_{max}$  表示在评分时给予最高分,  $r_{min}$  表示给予最低分,  $r_{random}$  表示随机评分,  $r_{average}$  表示均值评分。由表 2 可以观察到,不同攻击模式的主要区别在于对装填项目的评分方式不同。

根据攻击用户信息的生成策略可知,攻击用户与真实用户不同之处主要体现在 3 个方面:①目标项目的评分;②填充项目的评分;③由于所有的攻击用户信息采用同样的生成策略,致使攻击用户信息之间具有高度的相似性。文中利用以上数据差异生成统计特征,提出基于混合特征的攻击检测算法,以此区分正常用户与攻击用户。

### 1.3 托攻击检测指标

特征指标用于捕捉攻击用户与正常用户在评分方式上的差异。文献[15-16]中定义的 9 个统计量从不同角度反映了攻击用户概貌有别于真实用户概貌的

特征。

文献[8]针对流行攻击对统计量进行了研究,给出了有效检测指标排行,文中选择其前三项作为检测指标,如下所示:

(1) K 近邻用户相似度(DegSim)。

在进行托攻击时,大量注入系统的攻击概貌往往具有相同的攻击模型,具有数量大,相似度高的特点,故攻击用户的此项特征值比真实用户高。DegSim 的计算公式如下:

$$\text{DegSim}_u = \frac{\sum_{v=1}^k \text{sim}(u, v)}{k} \quad (1)$$

其中,  $\sum_{v=1}^k \text{sim}(u, v)$  是皮尔逊相似度,  $u, v$  表示数据集 DATE 中两个不同的用户,  $r$  表示用户  $u$  对项目的评分,  $k$  指所选取的最近的用户数目;此项特征值越高,

表示该用户是攻击用户的可能性就越大。

### (2) 均值方差 (MeanVar)。

对用户评分项目进行均值方差运算,体现用户模型评分项目与所有评分项目平均值之间的二阶矩关系,第  $u$  个用户的 MeanVar 的计算公式如下:

$$\text{MeanVar}_u = \frac{\sum_{j \in p_{u,f}} (r_{u,j} - \bar{r}_u)^2}{|p_{u,f}|} \quad (2)$$

其中,  $|p_{u,f}|$  是指用户  $u$  的所有评分项目中除去最高评分之后的集合项目总数,  $r_{u,j}$  为用户  $u$  对填充项目  $j$  的评分值,  $\bar{r}_u$  为用户  $u$  对项目的平均评分。

### (3) 加权评分一致度 (WDA)。

此特征值通过计算相应项目评分数目的逆向权重,以此衡量用户对项目的评分背离该项目评分均值的程度。第  $u$  个用户的加权评分一致度的计算公式如下:

$$\text{WDA}_u = \sum_{i=0}^{N_i} \frac{|r_{u,i} - \bar{r}_i|}{\text{NR}_i} \quad (3)$$

其中,  $N_u$  表示用户  $u$  评价过的项目个数,  $\text{NR}_i$  表示项目  $i$  被评价过的次数,  $\bar{r}_i$  表示项目  $i$  的评分均值,  $r_{u,i}$  表示用户  $u$  对项目  $i$  的评分。

目前很多检测器通过计算出各个特征指标值,形成用户评分矩阵构建特征模型,以此作为属性对分类器进行训练,最终能够将真实用户和攻击用户进行分类。文中结合以上三个特征指标得到特征模型,绘制三维图,如图 1 所示,图中“+”代表攻击用户,即圆圈中的数据,“o”代表正常用户。由图可知该特征模型可以较好地区分真实用户和攻击用户,但部分攻击用户与正常用户数据重叠,存在一定误判率。

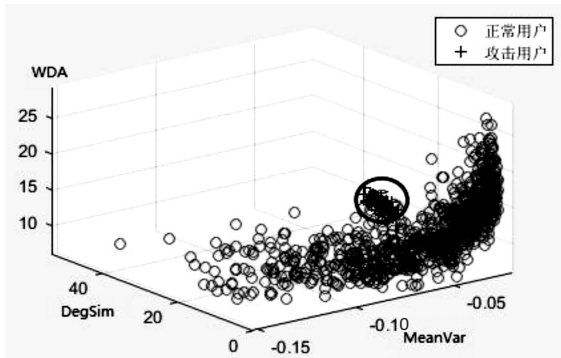


图 1 Degsim, MeanVar, WDA 三维图

该特征模型在实际应用中准确率和召回率不够高,为进一步提高检测准确率,文中加入对项目流行度和新颖度的考量。考虑到项目流行度、项目新颖度以及攻击用户装填项目服从不同的概率分布,其所得到的用户平均流行度以及新颖度数值与正常用户的平均流行度以及新颖度数值始终具有差异,因此文中提出了两个新的特征检测指标,分别是检测项目与流行项

目之间的卡方估计值 (Chi-square of popular item, CHIP) 和与新颖项目之间的卡方估计值 (Chi-square of novel item, CHIN), 通过这两个指标统计检测项目与所选的流行项目或新颖项目之间的相关程度。其中流行项目的选择依据项目流行度 (item popularity, IPop), 新颖项目的选择依据项目新颖度 (item novelty, INov), 其计算公式分别如下所示:

$$\text{IPop} = \sum_{u \in D_i} \varphi(r_{ui}) \quad (4)$$

其中,  $D_i$  表示所给数据库中所有真实用户的合集,  $r_{ui}$  表示用户  $u$  对任意一个项目  $i$  的评分。若  $r_{ui} = \emptyset$ , 则  $\varphi(r_{ui}) = 0$ , 若  $r_{ui} \neq \emptyset$ , 则  $\varphi(r_{ui}) = 1$ 。

$$\begin{cases} \text{Nov}_i = \frac{\sum_{u \in D_i, r_{ui} \neq \emptyset} \text{Nov}_{ui}}{|D_g|} \\ \text{s. t. } \text{Nov}_{u,i} = \frac{\sum_{r_{u,j} \neq \emptyset} (1 - w(i,j))}{N_u} \end{cases} \quad (5)$$

其中,  $|D_g|$  表示现在集合中的所有用户数目,  $\text{Nov}_{u,i}$  表示该用户对其任意一个项目的新颖程度,  $N_u$  表示用户  $u$  的项目评分数, 也就是相似度。

流行项目以及新颖项目的卡方估计值通用公式如下:

$$\text{CHI} = |I| \times \frac{(A \times D - B \times C)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (6)$$

其中,  $I$  表示数据集中所有的项目,  $A$  表示既属于有评分项目又属于流行项目/新颖项目的个数,  $B$  表示属于有评分的项目但是不属于流行项目/新颖项目的个数,  $C$  表示虽然不属于有评分项目却属于流行项目/新颖项目的个数,  $D$  表示既不属于有评分项目的也不属于流行项目/新颖项目的个数。通过计算用户评分项目与新颖项目/流行项目之间的关联程度, 得到特征矩阵。

## 2 K-means 聚类算法

聚类作为统计数据分析中的一项重要技术, 目前各个领域得到广泛应用, 如数据挖掘、机器学习等。它通过静态分类的方法, 将更为相似的对象分到相同的组别, 即该组别中的对象拥有较多相似的属性。

K-means 聚类算法源于信号处理中的一种向量化方法, 其主要目的是将所给的样本数据聚类。其算法流程为:

(1) 随机创建  $K$  个对象作为起始聚类中心;

(2) 计算每个对象与  $K$  个聚类中心点之间的欧氏距离, 将每个对象分到距聚类中心距离最短的类别中;

(3) 重新对每一类中的对象进行计算, 找到新的

聚类中心点,重复(2)过程;

(4)直到聚类中心点的位置不再改变,样本聚类完成。

文中使用 K-means 聚类算法对攻击用户与真实用户集合进行初步分类。

### 3 基于混合特征值的托攻击检测算法

文中构建了一种新的特征模型,该特征模型由两部分组成:①由特征指标 Degsim、MeanVar、WDA 组成特征模型的第一层;②由特征指标 CHIP 和 CHIN 组成特征模型的第二层。在该特征模型的基础上提出了一种基于混合特征值的托攻击检测算法,将其命名为 T-Kmeans 算法。该算法的具体步骤如下:

步骤一:向用户评分矩阵注入攻击概貌,得到混合

数据集;对其提取特征 Degsim、MeanVar、WDA、CHIP、CHIN,并按列排序,得到用户特征向量矩阵  $V$ 。

步骤二:提取特征向量矩阵  $V$  的前三列,即 DegSim、MeanVar、WDA 三个特征值,通过 K-means 聚类算法将用户初步聚成两类,称为第一真实用户集合和第一攻击用户集合。

步骤三:对特征矩阵  $V$  的后两列进行阈值判断操作;将大于阈值的标记为真实用户,小于阈值的标记为攻击用户,称其为第二真实用户集合和第二攻击用户集合;其中阈值的选择根据经验选择<sup>[17]</sup>。

步骤四:将步骤二、步骤三中得到的第一攻击用户集合和第二攻击用户集合做交集,得到最终检测结果,即攻击用户集合,剩余的用户则为真实用户集合。

算法流程如图 2 所示。

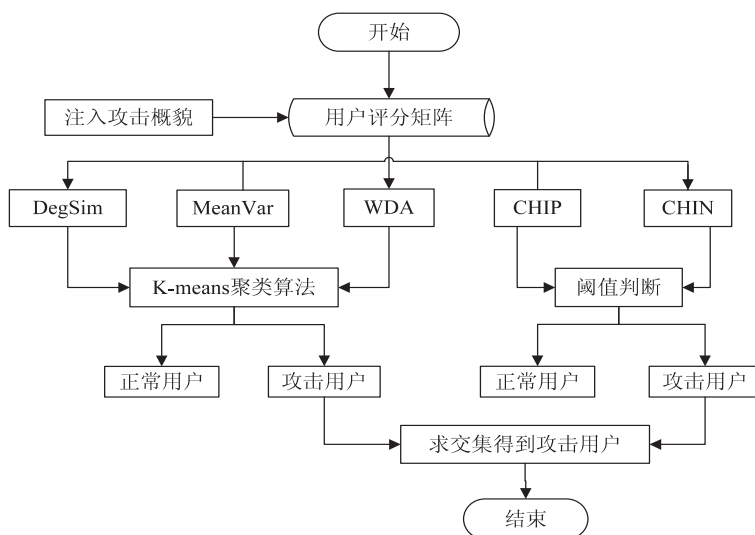


图 2 T-Kmeans 算法流程

### 4 仿真实验

文中实验采用 Movielens 数据集,包括 943 个观众对 1 682 部电影的随机评价,共计 100 000 条评价,采取 5 分制,即最高分记 5 分,最低分记 1 分,未评分的记为 0。

实验选取的攻击模型为流行攻击,攻击目的为推攻击。分别在攻击规模为 3%、5%、8%、10%、12%,填充规模为 3%、5%、8%、10% 的条件下进行实验。

#### 4.1 算法评估标准

文中通过计算准确率 (precision) 和召回率 (recall),与主成分分析 (principal components analysis, PCA) 检测方法进行对比,以此评估 T-Kmeans 检测方法的有效性与准确率。其计算公式如下:

$$\text{precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (8)$$

其中,TP 表示被正确识别的攻击用户的数目,FP 表示被误判的真实用户的数目,FN 表示未被识别出来的攻击用户的数目。

#### 4.2 实验结果与对比

##### 4.2.1 准确率对比

将文中提出的检测方法的准确率与 PCA 检测算法的准确率进行对比,得到的实验结果如图 3(a)~(d)所示。

如图 3 所示,在填充规模分别为 3%、5%、8%、10% 的情况下,随着攻击规模的增大,PCA 检测算法和 T-Kmeans 检测算法的准确率都在持续增加,但 T-Kmeans 检测算法准确率一直比 PCA 检测算法准确率高,且最高时候可达到 98%,这说明在小规模攻击情况下 T-Kmeans 检测算法在准确率方面比 PCA 检测算法效果好。

##### 4.2.2 召回率对比

将文中检测方法的召回率与 PCA 检测算法的召回率进行对比,得到的实验结果如图 4(a)~(d)所示。



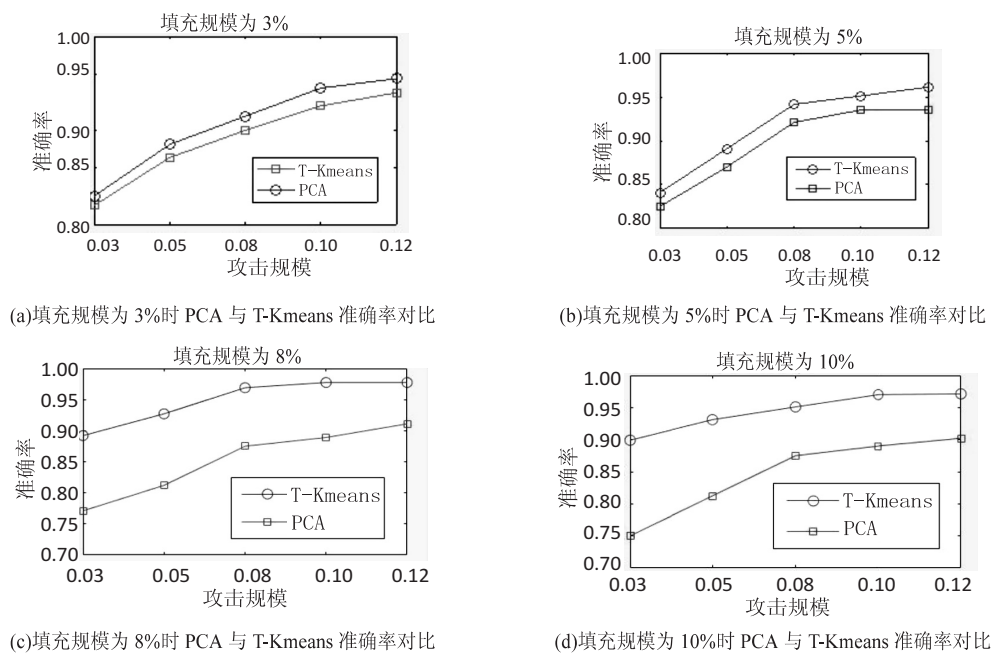


图 3 T-Kmeans 与 PCA 准确率对比

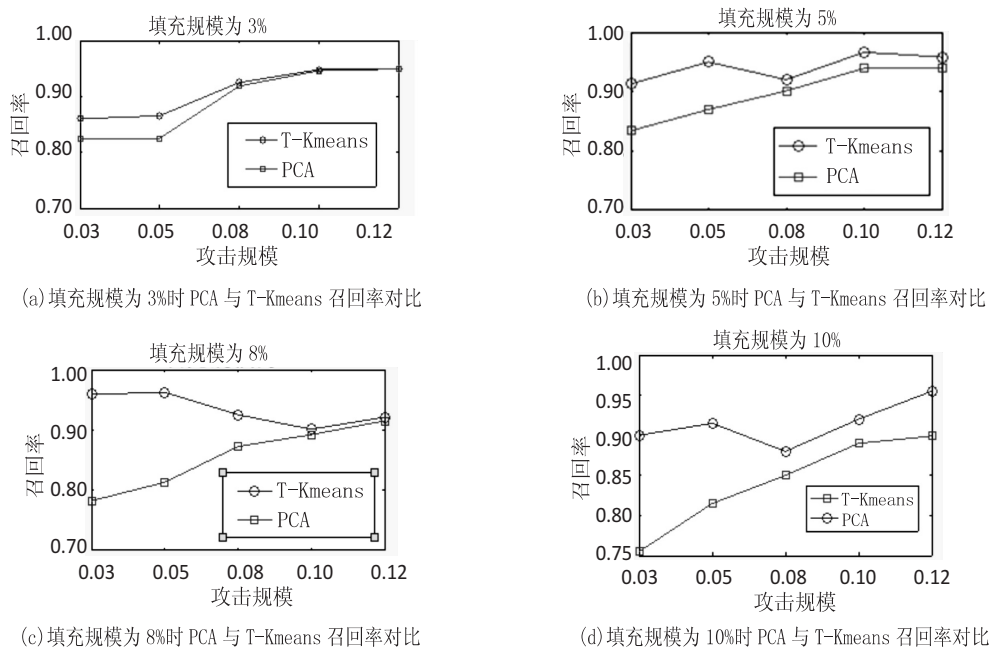


图 4 T-Kmeans 与 PCA 召回率对比

如图 4 所示,在填充规模分别为 3%、5%、8%、10% 的情况下,随着攻击规模的增大,T-Kmeans 检测算法的召回率变动较大,但其一直比 PCA 检测算法的召回率高,且最高时候可达到 97%,这说明 T-Kmeans 检测算法在召回率方面比 PCA 检测算法的检测效果好。

## 5 结束语

文中提出了一种基于混合特征值的托攻击检测算法。该算法构建了一种新的特征模型,在传统 Degsim、MeanVar、WDA 这三个特征检测指标基础上,考虑到项目与流行项目、项目与新颖项目之间的关联

程度,引入 CHIP、CHIN 检测指标,构成特征模型。通过对 Degsim、MeanVar、WDA 形成的特征矩阵进行 K-means 聚类,以及对 CHIP、CHIN 形成的特征矩阵进行阈值判断,并进行求交集操作,得到最终检测出的攻击用户集合。实验结果表明,该算法提高了检测准确度,具有一定的优越性。

## 参考文献:

- [1] GAO M, WU Z F, JIANG F. UserRank for item-based collaborative filtering recommendation[J]. Information Processing Letters, 2011, 111(9): 440-446.
- [2] LI C, LUO Z G. A metadata-enhanced variational bayesian

- matrix factorization model for robust collaborative recommendation[J]. *Acta Automatica Sinica*, 2011, 37(9): 1067–1076.
- [3] 李晓瑜. 协同过滤推荐算法研究[J]. *计算机与数字工程*, 2019, 47(9): 2118–2122.
- [4] 陆航, 师智斌, 刘忠宝. 融合用户兴趣和评分差异的协同过滤推荐算法[J]. *计算机工程与应用*, 2020, 56(7): 24–29.
- [5] 司明丹. 推荐系统中虚假评价用户识别方法研究[D]. 西安: 西安电子科技大学, 2019.
- [6] LI C, LUO Z. Detection of shilling attacks in collaborative filtering recommender systems[C]//*Proceedings of the international conference of soft computing and pattern recognition*. Dalian, China: IEEE, 2011: 190–193.
- [7] MOBASHER B, BURKE R, WILLIAMS C, et al. Analysis and detection of segment-focused attacks against collaborative recommendation[M]//*Advances in web mining and web usage analysis*. Berlin, Heidelberg: Springer, 2006: 96–118.
- [8] 伍之昂, 庄毅, 王有权, 等. 基于特征选择的推荐系统托攻击检测算法[J]. *电子学报*, 2012, 40(8): 1687–1693.
- [9] 胡德敏, 朱德福. 基于特征分析的推荐系统托攻击检测算法研究[J]. *软件导刊*, 2017, 16(2): 42–47.
- [10] 田俊峰, 蔡红云. 托攻击与推荐系统安全[J]. *河北大学学报: 自然科学版*, 2018, 38(6): 640–647.
- [11] LAM S K, RIEDI J. Shilling recommender systems for fun and profit[C]//*Proceedings of the 13th international conference on world wide*. New York: ACM, 2004: 393–402.
- [12] WILLIAMS C A, MOBASHER B, BURKE R. Defending recommender systems: detection of profile injection attacks[J]. *Service Oriented Computing&Applications*, 2007, 1(3): 157–170.
- [13] MOBASHER B, BURKE R, BHAUMIK R, et al. Effective attack models for shilling item-based Collaborative filtering system[C]//*Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery and data mining*. Chicago, USA: ACM, 2005: 13–23.
- [14] GUNES I, KALELI C, BILGE A, et al. Shilling attacks against recommender systems: a comprehensive survey[J]. *Artificial Intelligence Review*, 2014, 42(4): 767–799.
- [15] BURKE R, MOBASHER B, WILLIAMS C, et al. Research track poster ABSTRACT classification features for attack detection in collaborative recommender systems \* [C]//*Kdd 06 ACM SIGKDD international conference on knowledge discovery & data*. [s. l.]: [s. n.], 2008: 542–547.
- [16] WILLIAMS C A. Profile injection attack detection for securing collaborative recommender systems[D]. Chicago: DePaul University, 2006.
- [17] 周倩楠. 协同过滤系统中托攻击检测算法研究[D]. 秦皇岛: 燕山大学, 2017.