

文章编号: 2096-1618(2023)03-0306-08

# 基于极小极大博弈的水军识别算法研究

穆云翔, 盛志伟, 卢嘉中

(成都信息工程大学网络空间安全学院, 四川 成都 610225)

**摘要:**随着互联网的发展,用户越来越多地在线上完成购物、订餐,并倾向于先参考线上评论。评论对用户决策的重要导向作用催生了网络水军。网络水军会为了自身利益或其他不良动机,发布与实际体验不相符的评价,且会随时调整自己的策略来逃避平台的识别。现提出一个基于行为特征的水军识别算法(FBS),并将FBS加入到极小极大博弈,在这个博弈中,水军与识别器相互竞争,将博弈转换为两个相互依赖的马尔可夫决策过程,不断优化各自的策略,最终得到一个当前场景下最优的识别器。与当前先进的水军识别算法对比,性能有了明显提升,在公开数据集 YelpChi 上实际效应可以达到3.69。

**关键词:**网络水军;水军识别;极小极大博弈;马尔可夫决策过程

**中图分类号:**TP393

**文献标志码:**A

**doi:**10.16836/j.cnki.jcui.2023.03.009

## 0 引言

网络水军是指在网络中针对特定内容发布特定信息的、被雇佣的网络写手,简称水军,又名网络枪手。他们通常活跃在电子商务网站平台中,通过发布虚假的商品评价来影响正常消费者的决策。

网络的快速发展为水军提供了滋生条件。一方面,网络环境提供的便利:网络开放性、即时性、自由性、交互性等特点为网络水军的发展壮大提供了环境支撑。网络的开放性为网络水军提供了自由出入的媒体门户,不需要提供任何真实信息即可徜徉于各大论坛、贴吧,在网络上任意发表言论;网络的即时性则有助于网络水军引导网民对舆情的推动,让受众在缺乏理性分析的前提下引爆预设议题。网络的自由性塑造了无中心的狂欢广场,任何网民都可自由发布信息,也为网络水军的“灌水”行为提供了便利。网络的交互性为政府、媒体、网民之间的交流互动提供了便利,从而形成“围观”的力量。另一方面,网络受众的媒介素养偏低。中国网民群体庞大,但媒介素养参差不齐,整体不容乐观,在一定程度上纵容了网络水军的发展壮大。一些年轻网民由于心态不成熟,往往不能客观、全面、辩证地看待社会问题,容易受网络负面情绪影响。面对网络水军故意炒作的热点事件,极易非理性地跟风发帖,成为网络水军的“帮凶”。一些在经济上比较失意的网民则容易产生“仇富”心态,当网络上曝出诸如“房妹”等新闻热点时,他们会不自觉地与网络水军

站在同一战壕;一些网民对娱乐化的追求不断削弱主流媒体的权威性与影响力,为更能把握网民心理的网络水军提供了抢占先机的机会。

电子商务平台提供在线评论系统作为商家与消费者的桥梁。消费者通过浏览评论细节来决定是否购买产品,产品评论成为影响消费者购买欲望的重要因素。由于消费者倾向于购买好评产品,而放弃购买负面评论产品,因此在竞争激烈的电商市场中,许多商家通过雇佣“水军”,在自己的店铺下用图片伪造好评,操纵评论。并在竞争对手的门店下进行恶意评论,误导消费者。水军虚假评论的存在干扰了产品描述的真实性,对电商平台和消费者产生了显著的负面影响。因此,识别网络水军评论并保护消费者权益非常重要。总而言之,通过大量同质行为或内容形成回声室效应,从而影响其他用户的观点和决策,是网络水军的最终目的。

在电商平台中,网络水军采取发布不实信息来混淆视听。对于消费者而言,水军的存在影响购买意愿;对于商家而言,如何请水军刷好评变成了影响销售的最大因素;对于市场而言,水军扰乱了原有的市场秩序。这对于行业的发展具有很大的影响力,因此网络水军的监管成为一个亟待解决的难题。

本文主要关注 Yelp 系统中的水军。关于这类水军识别器主要有以下缺点:(1)大多数识别器都假设水军有相同的特征,并且可以根据这个特征来识别水军。但是在现实世界里有很多种水军,他们有不同的目标、对象和策略。如一个水军可能想要推广某一件商品,而另一个水军想要贬低竞争对手。(2)专业水



军会研究最新的识别技术,并从中发掘新的策略来逃过识别器<sup>[1-2]</sup>。(3) 现有的识别器大都基于精确率和召回率作为识别目标。而根据 Luca<sup>[3]</sup>利用市场研究的现有成果,从商品评分变化的角度定义了网络水军的实际效应。实际效应可以用来表示水军和平台在对抗过程中的实际目标,假设水军的目标是推广目标商品,那么平台的目标就是尽可能地减少推广的程度。文献[3]也通过实验证明,即使在召回率很高的情况下,水军依然可以达到很高的推广效果。

本文提出一种利用用户行为特征的水军识别算法(FBS),将 FBS 应用到一个水军对抗模型,最终训练出的识别器在公开数据集 YelpChi 上取得了不错的性能表现。并利用网络水军账号的一些特点,提出几种新的用户特征和一个基于用户行为特征分析的水军识别算法(FBS)。将 FBS 应用到一个水军对抗模型中,最终训练出的识别算法性能相较于以前有明显提升。

## 1 相关工作

### 1.1 AP 算法

AP(affinity propagation)算法<sup>[4]</sup>无需事先指定聚类数目,且没有明确的质心(聚类中心点),样本中的所有数据点都可能成为 AP 算法中的质心。根据现实环境中网络水军种类繁多,且同类型水军之间较为类似的特点,再基于 AP 算法的上述特征,将其应用到水军对抗模型。

### 1.2 分类器的选择

目前最常见的分类器有人工神经网络、K-近邻(K-NN)、朴素贝叶斯和决策树。尽管这些分类算法在几十年的发展中衍生出很多的改进算法,但是仍然没有一种完美的分类算法能适应所有的环境问题。根据 Kotsiantis<sup>[5]</sup>在主流分类算法性能比较,人工神经网络和支持向量机对大规模数据训练比较困难,且对缺失数据敏感;K-NN 时间复杂度和空间复杂度高,可解释性差;朴素贝叶斯通常准确率较低,且只能用于处理二分类问题;相对于其他几种分类算法,决策树计算量简单,可解释性强,比较适合处理有缺失属性值的样本,能够处理不相关的特征。此外,决策树能够很好地处理同时具有离散和连续属性的分类问题。因此,使用决策树算法进行分类识别最为合适,详情如表 1 所示。

表 1 机器学习常用分类算法比较

分类算法	优势	劣势
人工神经网络	准确率高	训练数据大、学习时间长
支持向量机	与特征维度无关,其适用于特征多、样本少的分类任务	对数据缺失敏感、计算复杂度与样本个数有关
K-NN	无需训练	时间效率低、K 的选择不固定
朴素贝叶斯	对数据缺失不敏感	特征之间需要相互独立
决策树	综合性能均衡、可解释性强	信息增益偏向于有多数值的特征

### 1.3 网络水军对抗模型

2020 年 Dou 等<sup>[6]</sup>利用强化学习建立的水军对抗模型,提出一种全新的水军识别性能的评测指标 PE (practical effect)。该模型利用多个水军识别算法和多种水军攻击策略进行博弈,将博弈过程转换为两条相互依赖的马尔可夫决策过程。利用双方博弈直至达到纳什均衡,此时的水军识别算法即可认为是当前环境下的最优算法。本文将提出 FBS 并将其应用到对抗模型中,以此训练出的水军识别算法达到的效果相较于以前 PE 提升了 8%。

### 1.4 研究现状

现有识别网络水军的方法主要有 3 种:基于文本与情感分析法、基于行为特征分析法和基于图结构法。其中,基于文本与情感分析需要花费较长时间进行训练,且随着 NLP 等人工智能算法的发展,水军文本内容已经和普通用户的评论文本差别越来越小。因此,基于文本分析法的性能相较于其他两种普遍偏低。

早期的水军由于发布的评论信息都很类似,因此研究方法大多基于语言学特征。其中,词袋特征是大部分研究者的首选语言特征。M Mccord 等<sup>[7]</sup>提取重复评论的 bigram 特征,在推特数据集训练回归模型,利用随机森林分类器识别只关注品牌的评论和评论文本无关的两类垃圾评论,精确率高达 95.7%。

Li 等<sup>[8]</sup>基于新扩展的黄金标准数据集识别网络水军,该数据集由来自 3 个不同领域(酒店、餐厅、医生)的数据组成,每个领域都包含 3 种类型的评论,即客户生成的真实评论、网络水军生成的欺骗性评论和员工(领域专家)生成的欺瞒性评论。该文试图捕捉欺骗性评论和真实评论之间语言特征的一般差异和水军检测的领域迁移性问题。实验表明该模型在餐厅数据集上分类准确率都能达到 75% 左右,而在医生数据集上准确率只有 50% 左右。实验表明该特征用于水军的虚假评论检测的领域迁移性差。



Noekhah S 等<sup>[9]</sup>通过提取分析词频、信息丰富度、内容定罪等特征,实现了基于欺骗性语言的评论文本在线欺骗识别系统。将这些特征集应用到之前使用的3个分类器(支持向量机、朴素贝叶斯和C4.5决策树),并使用5倍交叉验证。最终的实验结果表明,识别欺骗性评论的准确率达到80%,但是该识别方法时间复杂度很高,并不适用于一些较大的数据集。

Wang 等<sup>[10]</sup>首先提出虚假评价检测中的冷启动问题,在Yelp评价数据上提取一部分“新的评价”,即该用户只发布了一条评价。作者将之前研究中提到的文本和行为特征检测模型应用到这些新评价上,发现检测的效果并不好。为解决这种冷启动问题,一种直观的想法是从历史数据中去寻找和这个新评价发布者特征相似的评论者,然后把最相似的评价者或者评价的标签作为该新评价的标签。总而言之,虽然新评价信息很少,但可以通过深度学习,在历史数据上学习到有效的关系嵌入(embedding),然后利用该模型得到新的数据嵌入,这样就可以结合历史嵌入和其标签来预测新数据的标签。

Hooi B 等<sup>[11]</sup>利用二部图提出了Fraudar算法。Fraudar定义了一个可以表达结点平均可疑度的全局度量 $G(\cdot)$ ,在逐步贪心移除可疑度最小结点的迭代过程中,使 $G(\cdot)$ 达到最大的留存结点组成可疑度最高的致密子图。在此算法中,由于无法模仿每个节点与其他节点的联系,因此准确率大幅度提升。但是Fraudar的一个缺点是它的串行运算特性导致在大规模二部图上运算缓慢,其每次迭代只动态地删除一个结点并更新剩余结点状态。

Wang 等<sup>[12]</sup>除了利用用户本身的一些信息,还利用用户在社交网络中的好友关系对一些可疑用户进行识别。作者基于图结构的方法将水军和水军的虚假评论识别看作联合分类或排序问题,再采用马尔科夫随机场模型和LBP<sup>[13]</sup>(loopy belief propagation)计算每个节点的可疑程度。还对LBP算法进行优化,提高了算法的效率并且有收敛性的保证。实验表明,该模型在新浪微博数据集上的分类Accuracy都能达到75%。

Shah N 等<sup>[14]</sup>提出利用网络结构特征来识别在亚马逊上通过众包发送水军虚假评论的用户,提出TwoFace算法,更多关注召回率,该算法有的召回率能够达到91%。该算法也有缺点,ground truth的可信度不是很高。

S Rayana<sup>[2]</sup>提出SPEAGLE框架来做网络水军识别,利用关联数据和元数据,结合了图、行为和文本进行水军识别,该方法中图由user-review-product图构成,3种类型的结点都有标签,user:水军与否,review:

虚假与否,product:为被攻击目标与否。论文用图来做分类,用metadata来估计有关节点类分布的先验知识。该算法在数据集YelpZip上的准确率可以达到79.4%。

尽管研究者们针对不同情况下的用户特征进行深入研究,但其往往集中在某几个方面。随着平台和水军的发展,上述方法大多只能识别出某一类水军,适用性并不够广泛。

## 2 FBS-基于行为特征的水军识别算法

### 2.1 问题描述

由于现实环境中电商平台的网络水军复杂且多样,因此水军检测主要面临的困难在于水军检测涉及的特征难以规范化表达。一方面,水军覆盖范围广,在不同平台其特征不尽相同,导致没有一个标准的水军特征集;另一方面,水军经过多年发展,不断通过模仿正常用户来伪装自己,导致识别模型准确率不高。因此特征的选择变得尤为重要,应选用水军无法模仿的一些特征加入特征集。

### 2.2 特征定义

在Yelp系统中,刻画用户的特征有很多,如MNR(一天内写的最大评论数)、PR(积极评论比例)、NR(负面评论比例)等。结合Mukherjee等<sup>[1]</sup>的研究,选取了4个原始特征,如表2所示。

表2 原始特征描述	
特征名称	特征含义
MNR	用户一天内写的最大评论数量。用户的评论一般都较为均匀地分布在账号的存活期中,而水军用户就更有可能在一段时间内爆发式地评论。
PR	负面评论比例。计算方法如下公式所示: $PR = \frac{\text{Number}(\text{Review\_Negative})}{\text{Review\_Number}}$
NR	正面评论比例。计算方法如下公式所示: $NR = \frac{\text{Number}(\text{Review\_Positive})}{\text{Review\_Number}}$
RD	与产品平均评级的绝对评级偏差。

针对当前网络水军特征覆盖面不足导致识别率不高的问题,综合Yelp系统中水军用户与正常用户的差异性,结合水军用户的一些突出特征,本文提出AW、ISR、ERD、ETG、SQD等5个新特征,以此来扩充现有特征集对网络水军特征的覆盖面。

定义1 AW 是用户发表的第一条评论和最后一



条评论的时间差。通常情况下,水军的时间差较短。因为部分水军可能只是为了某几次完成任务而注册,且完成刷分之后便弃用账号。

$$AW = \text{Time\_last} - \text{Time\_first}$$

定义2 ISR 是用户是否仅有唯一评论。水军账号的注册可能仅是为了某一次的刷分行为,在之后便不再使用,而正常用户的使用大多是长期的。

定义3 ERD 是用户评论的时间分布熵。分布熵是对不同概率分布的刻画,它是概率分布的不确定性的期望值。值越大,表示时间分布的不确定性越大。正常用户的评论时间一般是长期且稳定的,不会在短时间内出现大量的评论,因此时间分布的不确定性较大,分布熵较大。而水军大多情况下,在任务集中时存在爆发式的评论。因此,时间分布的不确定性较低,分布熵较小。

$$\text{ETG} = - \sum_{i=1}^m p_i \lg_2(P_i)$$

其中  $p_i$  表示第  $i$  个类别出现的概率,一般可以通过用属于此类别的样本数量除以样本容量来估计该值。

定义4 ETG 是用户评论的评分分布熵。一般来说,正常用户的评分分布不稳定,而水军的评分分布大多分布在最高分和最低分。因此,正常用户的评分分布熵较高,而水军的评分分布熵较低。

定义5 SQD 是用户评分中最高评分和最低评分在全部评论中的占比。因水军是为了提高或降低某一商品的评分,故水军的最高评分和最低评分在全部评论中的占比较高,正常用户占比较低。

$$\text{NR} = \frac{\text{Number}(\text{Review\_maximim} \cup \text{Review\_minimum})}{\text{Review\_Number}}$$

为验证所选取特征与构建的新特征的有效性,本文利用卡方检验算法<sup>[13]</sup>对上述特征进行相关性验证,得到的特征  $P$  值排序结果,如表3所示。

表3 特征  $P$  值排序

序号	特征描述	$P$ 值
1	MNR	3.05E-20
2	SQD	4.57E-17
3	RD	1.13E-15
4	ISR	2.88E-13
5	AW	4.56E-10
6	PR	8.69E-8
7	ERD	1.56E-5
8	ETG	1.05E-4
9	NR	3.79E-2

从表3可以看出,所提出的 SQD、ISR、AW、ERD、ETG 的特征  $P$  值分别排在第2名、第4名、第5名、第7名、第8名。因此,新构造的5个特征和数据任务相关性较强,具有一定的有效性。因此,将采用这9个特征作为水军特征集。

2.3 FBS 算法模型

结合网络水军与正常用户之间的差异,给出一个能准确反映水军和正常用户之间差异的特征集合,采用 AP 聚类算法的特征集,引入 AP 聚类算法,通过刻画用户和用户之间的相似性,结合同一类别水军高度相似的特点,解决多类别水军适应性问题,再通过引入一个合适的欧氏距离阈值 Radius,将阈值之内所有未标注用户标注为其所属质心的标签,再将扩充后的标注集通过 C4.5 决策树算法进行分类模型训练,其流程如图1所示。

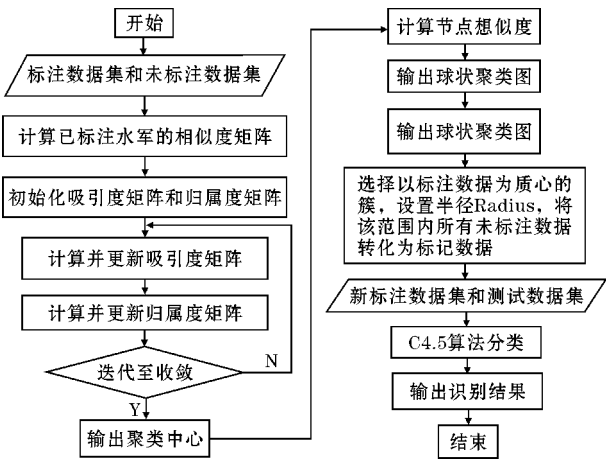


图1 FBS 流程

2.4 算法描述和分析

具体算法描述如下。

输入:Yelp 用户信息集合  $U\{u_1, u_2, u_3, \dots, u_n\}$   
输出:每个Yelp用户  $u$ , 为水军的可能性大小  $y$ ,  
方法:按以下步骤执行

步骤1:对于Yelp用户  $u_1$ ,按照表2的内容提取原始特征;

步骤2:利用上一步提取的基础特征,分别计算 AW、ISR、ERD、ETG、SQD;

步骤3:将按上述步骤处理好特征后的数据集输入到 AP 聚类模型中进行聚类,对以标注数据为质心的簇,引入 Radius 阈值,将 Radius 范围内所有未标注数据标注未与其所属质心同一标签;

步骤4:将步骤3中得到的新标注集和原始标注集一起输入到 C4.5 决策树中进行模型训练;

步骤5:将测试集输入到步骤4中训练好的分类



模型中,输出识别结果每个 Yelp 用户  $u$ , 为水军的可能性大小  $y$ ;

步骤 6: 计算模型 PE 值, 算法结束。

## 2.5 使用的识别器和水军攻击策略

将提出的 FBS 算法引入到 Dou 等<sup>[6]</sup>提出的水军对抗模型中, 训练出在 YelpChi 上性能更好的识别器。具体使用到的识别算法和水军攻击策略如下:

识别器: 采用如下几种识别算法作为博弈中的水军识别方。

(1) GANG<sup>[14]</sup>: 基于马尔科夫随机场的识别器, 利用有向图模型识别水军用户的方法。

(2) SpEagle<sup>[2]</sup>: 对用户、评论与商品组成的马尔科夫随机场进行概率推算的识别器。

(3) fBox<sup>[14]</sup>: 基于 SVD 的识别器, 利用子图密度寻找小规模的可疑用户。

(4) Fraudar<sup>[11]</sup>: 找出所有用户中最善于伪装的水军簇的识别器。

(5) FBS: 基于用户行为特征的水军识别器。

水军攻击策略: 采用 Dou 等<sup>[6]</sup>提出的 IncBP、IncDS、IncPR、Singleton 攻击方法, 具体如下:

(1) IncBP: 利用 VIP 用户, 尽量避免利用用户行为特征的 FBS 识别器和利用图形信息的 GANG 和 SpEagle 识别器。具体是利用在用户节点组成的马尔科夫随机场上进行置信传播, 用可疑度最低的用户节点发布水军的虚假评论。

(2) IncDS: 每轮先计算用户节点组成的子图密度, 用子图密度最小的用户节点发布水军的虚假评论。

(3) IncPR: 每轮攻击前计算用户行为特征的可疑程度, 用可疑程度最小的用户节点发布水军的虚假评论。

(4) Singleton: 创建新用户, 用新用户发布水军的虚假评论。

训练过程如图 2 所示。

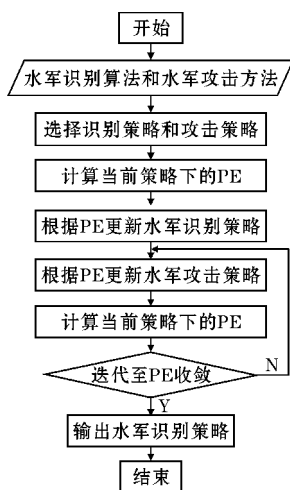


图2 训练过程

水军攻击的实际效应: 假设  $p$  为水军的攻击策略,  $q$  为识别器的策略,  $R(p, q)$  则为水军策略为  $p$ 、识别器策略为  $q$  时一轮攻击之后未被识别出的水军用户。这轮水军攻击的实际效应算法如下:

$$PE(v; R, p, q) = f(v; R(p, q)) - f(v; R)$$

水军的目标就是优化策略  $p$ , 使  $PE(v; R, p, q)$  的值最大化。

识别器的实际目标: 优化识别策略  $q$ , 使  $PE(v; R, p, q)$  的值最小化。

水军和识别器的目标已经确定: 水军要使目标商品的销售额提升, 而识别器要抑制这种提高。双方将在实际的场景中进行零和博弈, 假设水军攻击策略为  $p$ , 识别器策略为  $q$ , 则双方的博弈目标:

$$\min_q \max_p \sum_{v \in \gamma_T} \max \{0, PE(v; R, p, q)\}$$

利用上述的几种水军识别算法和水军攻击算法进行训练, 最终训练出 Hybrid detect。如此训练出的水军识别算法为当前环境下性能最优的水军识别算法。

## 3 实验

实验运行环境为: Windows10 操作系统, 2.90GHz 6 核处理器, 16 GB 内存, 算法的性能实验利用 PyCharm 软件实现。

### 3.1 衡量指标

平均评分的提升或降低会实际影响产品的销售额, 在 M Luca 的研究中发现, 平均评分每提升 1 分 (满分 5 分), 其销售额会增长 5% ~ 9%。并且普通用户与 VIP 用户的评论对商品的销售额的影响也有所不同, VIP 用户的评论对产品收益<sup>[3]</sup>的影响更大。因为他们在评论系统中对产品的评价更重要, 比普通用户的评论更频繁地呈现给客户。利用这项研究的结果, 设计了一种更能表现水军实际影响力的指标: 实际效应 PE。计算方法:

$$f(v; R) = \beta_0 \times RI(v; R) + \beta_1 \times ERI(v; R_E(v))$$

其中, ERI 用来计算 VIP 用户评论的影响, RI 用来计算所有用户评论的影响。  $\beta_0$  和  $\beta_1$  是两种影响的系数,  $\beta_0 = 0.035$ ,  $\beta_1 = 0.036$ 。这两个系数的值是用 Yelp 的数据估算的。

### 3.2 数据集

为准确地验证本文所提算法的效果, 准备了 2 套数据集。第一套为公开数据集 YelpChi<sup>[1]</sup>, 数据集中包含标记的水军的虚假评论和正常评论。第二套数据集 (YMX) 通过购买的方式获得, 其中有 300 条水军的虚



假评论为新用户所发,正常评论是公开数据集 YelpNYC<sup>[1]</sup>中被标记的正常评论。两套数据集均是在 Yelp 平台获得的数据。YMX 将作为训练集使用,YelpChi 将作为测试集使用。

两个数据集的详细情况如表 4 所示。

表 4 数据集详细信息

	YMX	YelpChi
用户数量	35642	38063
评论数量	66348	67395
商品总数	105	201
水军评论数量	800	450
正常评论数量	65548	37613

水军攻击策略设置:在测试集中共添加 600 条水军的虚假评论,包含 450 个用户和 100 件商品。

识别策略设置:每一次识别器进行节点可疑度计算之后,将可疑度排名靠前的 1% 节点作为水军删除。

VIP 用户选择:由于无法从 Yelp 抓取这些用户的 VIP 信息,Yelp 将用户评论的数量作为是否为 VIP 的关键因素,因此把每个数据集中单个用户评论超过 10 条的作为 VIP 用户,分别占 YelpChi 的 1.4%,YMX 的 1.2%。

3.3 FBS 算法的验证

为验证 FBS 算法在 Hybrid detect 中的有效性和其在识别模型中的影响,在本文所构建的水军识别模型的基础上,将利用加入 FBS 的 Hybrid detect 和未加入 FBS 的 Hybrid detect 在数据集 YelpChi 的实际效应进行对比,实验结果如图 3 所示。

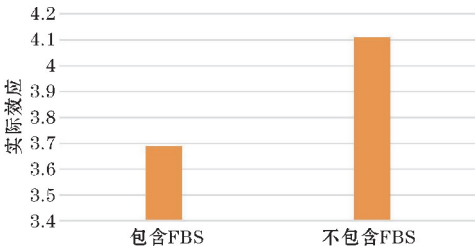


图 3 不同的 Hybrid detect 的实际效应

由图 3 可以看出,不包含 FBS 的 Hybrid detect 的实际效应高,能达到 4.11。而加入 FBS 之后,实际效应能够降到 3.69,说明 FBS 能够有效地识别出原本识别不出的网络水军,使模型性能提高。

3.4 对比算法

为验证本文算法的有效性,实验选择的对比算法有 GANG、SpEagle、fBox、Fraudar、Nash detect 算法。首先验证了以 GANG、SpEagle、fBox、Fraudar、Prior 算法为基础训练出的 Nash detect 在数据集 (YelpChi) 上的效

果,再验证以 GANG、SpEagle、fBox、Fraudar、FBS 算法为基础在训练出的 Hybrid detect 在数据集 (YelpChi) 上的效果,以此验证引入 FBS 的有效性。再将 Hybrid detect 与当前先进的水军识别模型 Tow Face 做对比。

3.5 Hybrid detect 性能测试

表 5 显示了单个识别器面对单个攻击时的实际效果。可以看到,每个识别器只有识别特定类型的水军效果较好。如果识别器选用了单个的识别器,那么水军可以采用相应策略使其效果大大减弱。

表 5 识别器对抗各种攻击的实际效果

	IncBP	IncDS	IncPR	Singleton
FBS	4.887	2.897	4.897	0.559
GANG	4.892	4.901	4.901	0.564
SpEagle	4.887	4.897	4.897	0.559
fBox	4.877	4.885	4.885	0.532
Farudar	2.01	4.885	3.134	0.532

在训练阶段,每种识别器的权重每一轮都在改变,变化如图 4 所示,每种识别算法的遗漏水军数量如图 5 所示。

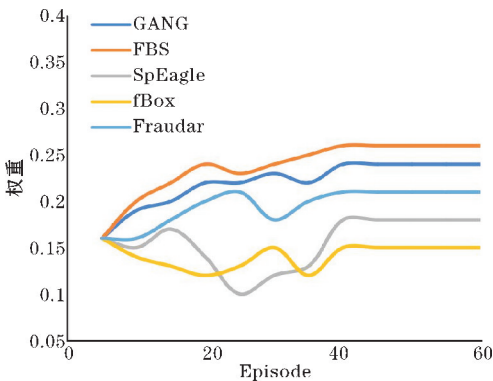


图 4 每种识别器在多轮博弈后的权重

从图 4 可以看出,Hybrid detect 在前 40 轮训练中,各水军识别算法的权重都在平稳地向最优配置移动。并且在 40 轮训练之后,各个权重都已趋于平稳,说明此时 Hybrid detect 已经收敛到最优配置。

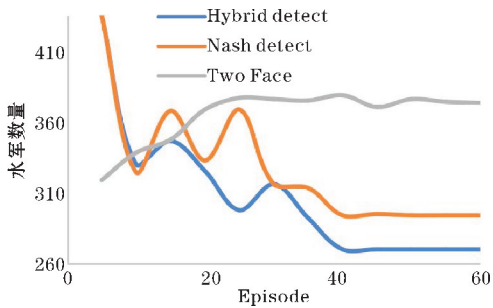


图 5 每种识别器每轮未检测到的水军数量



从图5可以看出,Hybrid detect在40轮训练后漏检的水军数量是最少的,并且在40轮后漏检水军数量也不会发生明显变化,这一结果也和图4的结论相同。

在训练阶段,在训练集上进行了60轮训练,混合识别器在最终训练完成后PE值明显低于Nash detect和Two Face。混合识别器在测试集上也取得了最优的成绩,其PE值达到3.69,而其余识别器的PE值都高于3.9,性能提升约8%。结果如图6和图7所示。

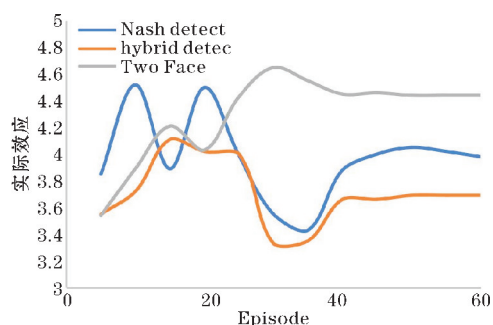


图6 每种检测器在多轮博弈后的实际效应

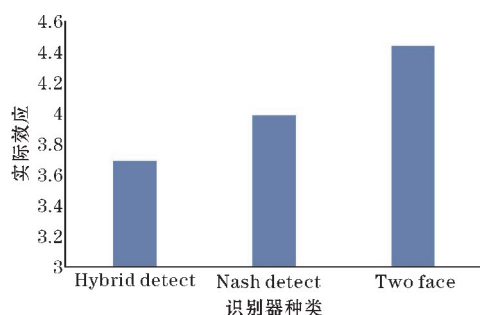


图7 几种识别器在测试集的实际效应

## 4 结束语

分析了现有的水军识别器存在的一些问题,利用博弈论的知识对现有识别器进行改进。针对不同场景的水军,只要设计好目标函数、攻击方法和识别器,通过运行该算法,就可以在线下找到最优的识别器配置。为解决传统水军识别方法中存在的一些问题,提出一种新的水军识别方法,该方法能充分利用用户信息的特征。最后,将Hybrid detect与Nash detect、Two Face这两种识别器进行对比,实验结果表明本文的方法具有更好的性能。

## 参考文献:

[1] Mukherjee A, Venkataraman V, Liu B, et al. What yelp fake review filter might be doing? [C]. Proceedings of the International AAAI Conference on

Web and Social Media. 2013, 7(1): 409-418.

[2] Rayana S, Akoglu L. Collective opinion spam detection: Bridging review networks and metadata [C]. Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining. 2015: 985-994.

[3] Luca M. Reviews, reputation, and revenue: The case of Yelp. com[J]. Com (March 15, 2016). Harvard Business School NOM Unit Working Paper. 2016, 12(16): 175-216.

[4] Leone, Michele, Sum ed ha, et al. Clustering by soft-constraint affinity propagation: applications to gene-expression data. [J]. Bioinformatics, 2007, 23(20): 2708-2715.

[5] Kotsiantis S B. Supervised Machine Learning: A Review of Classification Techniques[J]. Informatica, 2007, 31: 249-268.

[6] Dou Y, Ma G, Yu P S, et al. Robust spammer detection by nash reinforcement learning[C]. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 924-933.

[7] Mccord M, Chuah M. Spam Detection on Twitter Using Traditional Classifiers [C]. Autonomic & Trusted Computing-international Conference. DBLP, 2011: 175-186.

[8] Li J, Ott M, Cardie C, et al. Towards a general rule for identifying deceptive opinion spam [C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 1566-1576.

[9] Noekhah S, binti Salim N, Zakaria N H. Opinion spam detection: Using multi-iterative graph-based model [J]. Information Processing & Management, 2020, 57(1): 102140.

[10] Wang X, Kang L, Zhao J. Handling Cold-Start Problem in Review Spam Detection by Jointly Embedding Texts and Behaviors [C]. Meeting of the Association for Computational Linguistics. 2017: 366-376.

[11] Hooi B, Song H A, Beutel A, et al. Fraudar: Bounding graph fraud in the face of camouflage [C]. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 895-904.



- [12] Wang B, Gong N Z, Fu H. GANG: Detecting fraudulent users in online social networks via guilt-by-association on directed graphs[C]. 2017 IEEE International Conference on Data Mining (ICDM). IEEE, 2017: 465–474.
- [13] Si-Cai H U, Sun J P, Sheng-Gen J U, et al. Chinese emotion feature selection method based on the extended emotion dictionary and the chi-square model[J]. Journal of Sichuan University(Natural Science Edition), 2019, 56(1): 37–44.
- [14] Shah N, Beutel A, Gallagher B, et al. Spotting suspicious link behavior with fbox: An adversarial perspective[C]. 2014 IEEE International conference on data mining. IEEE, 2014: 959–964.
- [15] 孙文. 网络新闻评论用户行为分析及水军识别方法研究[D]. 杭州:杭州电子科技大学,2019.
- [16] 任亚峰,姬东鸿,张红斌,等. 基于PU学习算法的水军的虚假评论识别研究[J]. 计算机研究与发展, 2015, 52(3): 639–648.
- [17] Gatterbauer W, S Günnemann, Koutra D, et al. Linearized and Single-Pass Belief Propagation[J]. Proceedings of the Vldb Endowment, 2014, 8(5): 581–592.
- [18] Kaghazgaran P, Caverlee J, Squicciarini A. Combating crowdsourced review manipulators: A neighborhood-based approach[C]. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. 2018: 306–314.
- [19] Miller F P, A F Vandome, J Mcbrewster. Amazon Mechanical Turk. [C]. Alphascript Publishing. 2021:308–331.
- [20] Mukherjee A, Kumar A, Liu B, et al. Spotting opinion spammers using behavioral footprints[C]. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 632–640.
- [21] Y Chen, C Lou. Research on the formation path of fake reviews of online goods[J], Modern Intelligence. 2015, 8(10): 49–53.

## Research on the Algorithm of Online Water Army Recognition based on Minimax Game

MU Yunxiang, SHENG Zhiwei, LU Jiazhong

(College of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China)

**Abstract:** With the Internet's development, more and more users complete shopping and dining online. At the same time, the public will also tend to refer to online comments first. The important guiding role of comments in user decision-making gave birth to the network Navy. For its interests or other bad motives, the online Navy will release evaluations that are inconsistent with the experience. And the Navy will adjust its strategy at any time to avoid the platform's recognition. This paper proposes a behavior based Navy recognition algorithm (FBS), and adds FBS to the minimax game. In this game, the Navy and the recognizer compete, convert the game into two interdependent Markov decision-making processes, constantly optimize their strategies, and finally get an optimal recognizer in the current scene. Compared with the current advanced navy recognition algorithm, the neutral energy has been significantly improved, the actual effect based on the public dataset yelpchi can reach 3.69.

**Keywords:** network navy; navy identification; minimax game; Markov decision process