

基于用户特征提取的新浪微博异常用户检测方法

袁得崙^{1,2} 章逸钊¹ 高见^{1,2} 孙海春^{1,2}

1 中国人民公安大学信息技术与网络安全学院 北京 102623

2 安全防范与风险评估公安部重点实验室 北京 102623

(yuandeyu@ppsuc.edu.cn)

摘要 随着互联网的发展,微博逐渐成为重要的社交媒体。然而,在在微博中异常用户通过传播有害信息、发送恶意链接,甚至发起恶意攻击等方式影响用户的行为,从而影响了社交网络的价值。因此,实现对异常用户的检测具有重要的意义。文中以多种途径获取的微博异常用户和正常用户数据集为基础,对其进行数据清洗后,提出综合提取并分析用户的多种属性。多种数据挖掘方法建立异常用户检测模型,从而进行异常用户账号的识别。对 C4.5 决策树、随机森林等算法的实验结果表明,所提方法选取的特征有效,检测异常用户的精度较高。

关键词: 微博;异常用户;特征提取;数据挖掘

中图法分类号 TP309

Abnormal User Detection Method in Sina Weibo Based on User Feature Extraction

YUAN De-yu^{1,2}, ZHANG Yi-fan¹, GAO Jian^{1,2} and SUN Hai-chun^{1,2}

1 Institute of Information Technology and Cyber Security, People's Public Security University of China, Beijing 102623, China

2 Key Laboratory of Safety Precautions and Risk Assessment, Ministry of Public Security, Beijing 102623, China

Abstract With the development of the Internet, Weibo has gradually become an important social media. However, in Weibo, abnormal users influence the behaviors of users by spreading harmful information, sending malicious links, and even launching malicious attacks, thus affecting the value of social networks. Therefore, it is important to realize the detection of abnormal users. Based on the Weibo abnormal users and normal user data sets obtained from multiple ways, this paper proposes to comprehensive extract and analyze various attributes of users. An abnormal user detection model is established through various data mining methods to identify abnormal user accounts. Experimental results of C4.5 decision tree and random forest algorithms show that by using the proposed method, the selected features are effective and the detection accuracy of abnormal users is high.

Keywords Weibo, Abnormal user, Feature extraction, Data mining

近年来,随着互联网的飞速发展,使用互联网获取信息的人越来越多。中国互联网络信息中心(CNNIC)发布的第43次《中国互联网络发展状况统计报告》显示,截至2018年12月,中国网民规模已达8.29亿,排名全球第一,互联网普及率相较于2017年提高了3.8个百分点,增至59.6%^[1]。互联网上的在线社交媒体(如Twitter、Facebook、微信、微博、LinkedIn等)成为网络用户在线参与的主要平台,社交媒体上的用户形成了在线社交网络(Online Social Network, OSN)。其中,新浪微博以强大的信息分享能力和便捷、快速的特性得到了广泛的应用。2018年第二季度发布的财报显示,新浪微博在2018年6月的微博月活跃用户数超过4.31亿,在微博日活跃用户也已突破2亿。人们通过微博分享知识、意见和经验,获取信息和资源,拓展人际关系,微博逐渐演变成人们生活中不可或缺的一部分。

在微博平台中大量用户活跃的同时,大量恶意信息和恶意用户也涌现了出来:广告用户通过恶意营销和广告诱导污染了社交媒体环境;僵尸用户通过恶意关注正常用户干扰了

社交网络平台的秩序,异常用户如女巫节点、网络水军和僵尸粉等蓄意散布谣言、恶意炒作、制造不良舆论导向,通过创建和传播虚假信息非法谋取利益,严重影响了用户的正常上网体验,浪费了网络资源,增加了不安全系数。微博每天都要处理海量的信息,但是如何鉴别恶意消息成为了亟待解决的问题。

异常用户检测是微博信息安全研究中的重要内容。在当前的微博平台上,存在大量的异常用户,有的异常用户利用微博关注度高、传播速度快的特点,在微博平台投放虚假广告,严重影响用户体验。还有一类异常用户是网络诈骗的幕后黑手,通常大量发布虚假中奖、色情、网络欺诈等有害信息。另外,微博上充斥着大量的网络水军和僵尸粉,通过创建大量假冒帐户,操纵异常用户进行恶意互粉、添加好友、点赞等,窃取普通用户账号,并且逃避公安机关和微博平台对异常用户的检测,严重危害了普通用户的信息安全和社交网络的信用系统。因此,如何准确地检测出新浪微博中的异常用户,对于完善微博网络环境、净化网络空间十分必要。

面对复杂的网络舆情环境,首先,要做到在海量的数据中

基金项目:国家自然科学基金(61771072);中国人民公安大学专项项目(2020JWCX01)

This work was supported by the National Natural Science Foundation of China (61771072) and Special Project of People's Public Security University of China (2020JWCX01).

通信作者:高见(gaojian@ppsuc.edu.cn)

分析用户行为特征并检测出异常用户,这也是对异常用户进行管控、采取措施、予以制止和打击的前提和基础。其次,要通过微博舆情的采集和分析,清除异常用户。目前,海量的社交网络信息为我们的研究提供了基础,但现有的微博异常用户检测领域仍然缺乏科学的理论指导和技术支持,且学术界现有方法普遍选取的用户特征维度较低,往往仅选择一类或几类特征明显的用户行为等单一特征进行分析,特征之间的结合程度不高,识别率和召回率较低。如何将多种特征有机结合,并选取合适的特征、应用高效的算法进行分析处理,是异常用户检测工作的关键环节。

本文以多途径获取的微博异常用户和正常用户数据集为基础,对其进行数据清洗后,创新性地提出综合提取分析用户的多种属性,通过多种数据挖掘方法建立异常用户检测模型,从而进行异常用户账号的识别。最终结果表明,本文构建的特征提取方法可以有效提取用户特征,运用 C4.5 决策树和随机森林等分类算法,对网络舆情预警和谣言控制提供了有力的理论和技术支持,为公安机关识别异常用户提供相应意见,具有重要的理论及实际研究价值。

1 相关工作

随着移动互联网的发展,异常用户检测的问题困扰着各大社交平台,成为了微博安全研究的热点。近年来, Twitter 等各大社交网络平台开始研究识别异常用户的技术,并取得了一定成效。异常用户检测方法主要从两个方面展开研究:一是分析并提取恶意用户的特征,包括文本特征、用户属性、行为特征等;二是基于异常用户之间的关系网络图进行异常用户的检测。

1.1 异常用户检测研究现状

对于微博异常用户的检测研究主要是对微博用户进行分类,以正确判断出正常用户或者异常用户。

Fabrico 等^[2]选取 Youtube 作为研究对象,在分析用户特征的基础上,使用支持向量机的分类方法分析了在视频网络中识别垃圾用户和正常用户的有效方法,并从分类效果上评估了属性集合数量的价值。Benevenuto 等^[3]对大量的用户特征数据和文本内容数据进行分析 and 特征提取,并借鉴统计学中的相关知识在特征排序上进行实验,最终保留排名前十位的特征做进一步研究。Stringhini 等^[4]以 Twitter 为对象,通过抓取粉丝数与关注数的比例、微博所含链接数的比例、信息相似度等属性,建立用户行为模型,并选取随机森林算法在社交网络中有效地预测垃圾用户。彭希羨等^[5]在新浪微博上展开实验,通过网络爬虫与计量学相结合的方式对用户特征进行了提取和分析,在微博用户的分类上取得了优异的效果。刘勘等^[6]以新浪微博为例,对微博用户的行为、文本、关系、属性特征进行提取,并使用随机森林分类算法,对机器用户的识别方法针对性地进行实验和预测,取得了良好的效果。Aphinyanaphongs 等^[7]在微博文本分类方面,结合相关数据挖掘的算法,对 Twitter 中饮酒的话题进行了研究,通过对所获取到的数据进行分析,得出人们饮酒习惯在地域、时间等方面的差别。

在国内,微博的地位和国外的 Twitter 一样,但它们在许多方面存在区别。此外,国内对微博用户的研判并不完善,在许多方面存在不足:1)从用户分类的角度来说,定性研究的数量要远多于定量研究,通常所遇到的定量分析往往是指定一

种具体的分类方法,而忽略了与其他方法进行比较与结合;2)研究重点往往放在某一类异常用户的识别上,而没有从总体上去把握;3)缺少对微博用户提取后的特征进行评估和选择,并且对于选择的特征没有属性约简^[8],从而影响预测效果;4)理论层面的研究较多,实践层面的研究较少,在一定程度上会对研究成果的应用价值产生影响。

1.2 微博异常用户检测方法

微博异常用户的检测方法主要分为两类:1)基于用户属性和行为特征;2)基于微博的文本内容。

基于用户属性和行为特征的方法主要是对微博用户账号的属性进行研究,提取能区分正常用户与异常用户的特征。早期的研究通常采用统计学研究方法构建评价指标,如夏崇欢^[9]基于主成分分析法对微博用户行为特征进行提取,并对特征进行降维分析,进而构建分类模型,获得对异常用户检测的良好表现。当前学术界一般采用监督学习的分类方法检测节点异常性。

基于文本内容的检测方法通过提取用户发布信息内容、关键字、URL 指向网页等判断账号异常与否及情感倾向。郝亚洲等^[10]用 NLP 方法,通过分类器和触发词从大量网络舆情文档集中过滤出包含“攻击”“受伤”“死亡”“拘捕”4 种异常行为的句子,提取有效命名实体,构建出异常行为共现网络并体现关键实体和社交关系。

2 微博异常用户特征分析

本节介绍新浪微博异常用户的特征,给出本文采用的异常用户检测框架,并介绍 C4.5 决策树和随机森林两种数据挖掘算法。

2.1 微博异常用户分类

微博平台上的异常用户账号有创建、发展和应用 3 个阶段,具有动态化的特点,因此在创建、维护到发挥作用的过程中,异常用户的行为状态存在很大区别。通过对异常用户账号不同阶段的特点进行分析,可以将异常用户分为以下几类。

僵尸用户(social bot):由攻击者创建的用来模拟正常用户发布消息、关注或点赞好友等的虚假账号^[11]。通常僵尸用户在批量被创建的过程中有相似的命名规则。由于一部分僵尸粉的存在,异常用户和正常用户在微博发布时间间隔、关注数、粉丝数等属性上的差异较大。

女巫节点:攻击者制造虚假账号频繁与正常用户进行互动,从而提高自身的影响力^[12],进而更容易在微博上对用户进行欺骗或者控制用户进行选择。大部分女巫节点会经常更换用户名并且与同类节点频繁交互。

spam 用户:由攻击者创建的用于发布广告、钓鱼、色情等恶意消息或通过恶意互粉、添加好友、点赞等行为改变社交网络中信誉等恶意行为的虚假用户,其文本内容中大多会出现链接,而且该类用户发表的信息一般不会有人去评论和关注,因此在评论数、收藏数、转发数等属性上的差异较大。

被劫持用户:该类用户通常原本是社交网络中的正常用户,在被攻击者盗取账号之后实施恶意行为。对其检测主要通过研究用户行为是否产生突变。

2.2 异常用户行为特点

本文研究的微博异常用户的行为特征主要有重复关注、大量发布广告、传播恶意消息、过度转发。

重复关注的特点是:用户频繁关注普通用户,导致其关注

数上升,但他们的动态并不多,因此其微博数量以及长度都小于普通用户。

大量发布广告的特点是:用户发布微博的数量和频率均高于普通用户,而且为了加入热门话题提高影响力,发布的微博文本内容所提及的链接数、“@”个数、“#”个数等也高于普通用户。

传播恶意消息的特点是:伪装能力强,会通过发布一些类似正常用户发布的微博内容以及频繁更换登录地址等行为来伪装自己。

过度转发的特点是:大量且频繁地转发同一条微博,其微博文本内容通常会出现大量相同的链接。

2.3 微博异常用户检测框架

本文采取的微博异常用户检测框架如图 1 所示。

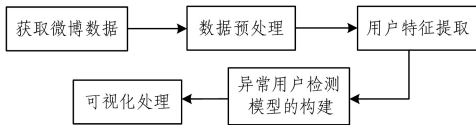


图 1 微博异常用户检测框架图

Fig. 1 Framework of abnormal user detection on Weibo

通过获取的微博的结构化数据,我们对异常用户检测进行建模,其核心在于对分类算法的选择。本文异常用户检测框架包含以下几个部分。

(1)获取微博数据:通过微博爬虫、蜜罐账户、下载共享数据集等方式获取所需微博结构化数据集。

(2)数据预处理:分析所获取数据的结构,对异常用户和正常用户进行人工标记,为之后的研究提供数据支撑。

(3)用户特征提取:在对异常用户行为特征的分析的基础上,利用分类工具进行可视化处理,最后对数据的有效特征进行提取和融合。

(4)异常用户检测模型的构建:基于提取和融合过的用户数据,选择高效的分类算法,通过 weka 构造分类模型,进行异常用户检测的训练与测试。

(5)可视化处理:通过可视化处理,评估各分类方式的优劣与劣势,从而建立更加完善的检测机制。

2.4 基于数据挖掘的分类算法

2.4.1 C4.5 决策树算法

C4.5 决策树算法为了选取当前最优决策属性的评估系数,采用信息增益率的作为衡量标准。信息增益率的计算公式如下:

$$IGR(attr) = \frac{IG(attr)}{SplitInfo(attr)} \quad (1)$$

其中, $IG(attr)$ 是某属性对于集合的信息增益, $SplitInfo(attr)$ 为分离信息度量。

C4.5 决策树算法基于从上到下的递归分治策略,根节点存放某个属性,并产生下一个分治用来存放可能的属性值,将数据分成多个部分,每个部分对应一个根节点的分支;然后在每个分支上应用递归算法并重复,当所有数据有相同的分类时停止。该算法的优势在于可信度高、原理简单易懂,不仅能够处理连续型和离散型的数据,还能够处理有缺失值的数据,并且分类结果比较容易接受。

2.4.2 随机森林算法

随机森林作为一种简易的分类算法,在大多数情况下不需要超参数调优也可以获得良好的结果,因此被大量用于分

类和回归的数据挖掘研究中。随机森林算法的原理是将几个决策树组合起来形成森林,并将它们组合起来以产生更准确和稳定的预测。不同于决策树的是,随机森林算法通过随机选择的函数来创建最佳分割,每个节点被分割成最佳函数以最小化错误。基于此过程形成的分类结果可以更好地进行建模。

3 基于新浪微博数据的用户特征提取

本节研究新浪微博异常用户的检测方法,通过对大量微博用户数据的分析,对实验所需的用户行为特征进行提取,完成数据预处理的目标,从而达到后期通过机器学习进行异常用户检测的数据要求。在分析的过程中,我们无法每时每刻监测用户账号的行为数据,但是,可以通过对大量用户数据的特征分析和提取来预测账号的状态。本节从用户基本属性、行为模式和文本内容 3 个方面对微博用户特征进行分析,基于 Python 语言对微博用户的特征进行提取和挖掘。

3.1 数据来源

本文以基于 HTTP 协议发送请求模拟网页访问的微博爬虫^[13]、公安蜜罐账户、公安共享数据集等方式共获取了 14000 条微博数据。

新浪微博作为国内最大的社交网络平台,对外开放可供第三方进行插件开发、非隐私数据抓取的 API 接口。API 接口提供了 30 多个用户属性,在本文的微博爬虫中,我们选取所需部分进行研究,选取的字段说明如表 1 所列。

表 1 微博用户接口字段说明

Table 1 Description of Weibo user interface

字段名	字段中文名	字段类型
uid	账号编号	varchar(20)
screen_name	账号昵称	varchar(50)
province	省份编号	int
city	城市编号	int
location	所在地	varchar(20)
url	个人简介中嵌入的链接	varchar(80)
gender	性别	varchar(10)
followersnum	粉丝数	int
friendsnum	关注数	int
statesnum	全部微博数	int
favouritesnum	收藏数	int
created_at	账号创建时间	datetime
mid	消息编号	varchar(20)
data	消息发布时间	datetime
text	消息内容	varchar(50)
source	消息来源	varchar(100)
repostsnum	转发数	int
commentsnum	评论数	int
attitudesnum	点赞数	int
topic	话题	varchar(20)

3.2 微博用户特征分析

为了尽量全面地描绘微博用户的特征,基于调用新浪微博 API 所获得的数据,我们从用户基本属性、行为模式、文本内容 3 方面对用户特征进行研究,共提取 10 个用户特征。

3.2.1 基于用户基本属性的特征

本文以新浪微博数据的用户账号为中心,针对所获取的全部微博数据,对基于用户基本属性的特征进行研究,分析用户 ID、账号昵称、省份、城市、所在地、个人简介中嵌入的链接、性别、粉丝数、关注数、账号创建时间等特征。为了更好地对正常用户和异常用户进行识别,提取表 2 所列的用户基本属性的字段。

表2 用户基本属性的相关字段说明

Table 2 Description of related fields of user basic attributes

字段	说明
uid	用户 id,账号的唯一标识
followersnum	该账号的粉丝数
friendsnum	该账号的关注数
created_at	账号的创建时间

3.2.2 基于用户行为模式和文本内容的特征

用户在微博上主要进行如下行为:浏览、发布、转发、评论微博以及对微博进行点赞。虽然我们并不能抓取用户浏览微博的相关数据,但是通过对用户行为模式的分析,可以提取相关特征来反映和判断用户的行为方式与特征。本文提取的基于用户行为模式的特征如表3所列。

表3 用户行为模式的相关字段

Table 3 Related fields of user behavior patterns

字段	说明
text	该条微博的文本内容
created_at	微博的发布时间
statesnum	该账号的微博总数
reportsnum	微博被转发的次数
commentsnum	微博被评论的次数
attitudesnum	微博被点赞的次数

3.3 数据预处理

异常用户通常具有明显的行为特征,为了更好地了解影响异常用户检测效率的因素,需深入探索相关特征对检测技术的贡献度。因此,除了对用户所发布的微博内容、微博发布时间、粉丝数、关注数等基于基本属性和行为模式的特征进行提取外,本文还对表4所列行为特征进行提取。

表4 微博用户行为特征选取

Table 4 Selection of Weibo user behavior features

字段	说明
tNum	计算发布微博的长度
tNum_at	计算发布微博中@的个数
tNum_A	计算发布微博中大写字母的个数
tNum_a	计算发布微博中小写字母的个数
tNum_num	计算发布微博中数字的个数
tNum_space	计算发布微博中词的个数
tNum_h	计算发布微博中标点符号的个数
http_num	计算发布微博中http个数
tTime	计算账号注册时长
tFrq	计算发布频率(条/天)

通过运行Python的pandas库,对用户发布微博的长度、发布微博中“@”字符、大写字母、小写字母、数字、词(也就是空格)、标点符号、http的个数等特征进行提取,然后通过字符串匹配的方法,沿微博数据的账户id进行聚合并计算各特征的平均数,接着通过计算账号创建时间(created_at)、微博收集时间(collect_at)和发布微博总数(statesnum)等属性,提取出账号注册时长(tTime)和发布频率(tFrq)等特征。其核心步骤如下:

第一步 对微博数据的相关行为特征和文本特征进行提取。

第二步 根据微博数据的uid进行聚合,计算各用户全部特征的平均值,最后对每个用户id只保留一行数据,形成新列表,并计算注册时长和发布频率。

利用优化的分析工具对丰富特征以后的数据进行分类,将每个行为特征转变为影响异常用户检测的因素,并进行有监督的离散化处理(Discretize)。图2展示了离散化后的各类特征对异常用户的预测效果。

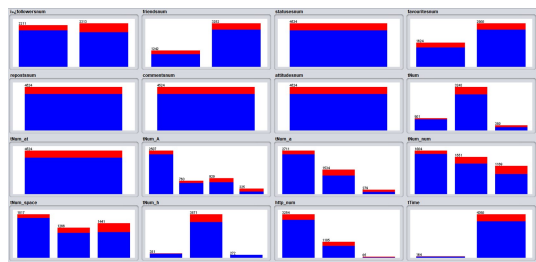


图2 各类特征对异常用户的预测效果(电子版为彩色)

Fig. 2 Prediction effect of various features on abnormal users

图2中,柱状图的同一种颜色的比例越大,其所代表的用户特征对异常用户的预测效果越好。根据计算出的各类特征对异常用户检测结果的预测效果,可以发现关注数、发布微博中@的个数、数字的个数、http个数、账号注册时长、发布频率(条/天)对检测结果影响较大,公安部门在开展舆情管控等相关工作时应着重关注。

4 基于用户特征提取的异常用户检测方法

为了提高对微博异常用户的检测效率,本节基于C4.5决策树和随机森林算法进行实验,通过对比与总结,结合第3节的用户特征提取方法提高准确率和召回率,从而实现较好的检测效果。Weka是数据挖掘领域的常用工具,利用Weka分别基于C4.5决策树算法和随机森林分类算法进行异常用户检测,实验结果表明实现了不错的分类效果。将数据挖掘算法与提出的微博用户特征提取方法有机结合,取长补短,对于微博异常用户的检测具有长远的研究价值。

4.1 实验方法和评估标准

首先,我们将获取到的微博数据集进行人工标注,基于所提出的特征提取方法,将正常用户标为“No”,异常用户标为“Yes”,用于训练和测试模型对微博异常用户的检测效果。

其次,调用Weka的相关接口进行数据挖掘,采用十折交叉验证的方式得出结果。Weka(Waikato Environment for Knowledge Analysis)^[14]是由新西兰学者参与研究和开发的一款成熟的数据挖掘工具,拥有丰富的数据挖掘的算法库,并能对分类结果进行快速详尽的分析。

Weka虽然可以直接进行分类器的可视化实现,但是它对于数据样本的格式有着严格要求。一般情况下,Weka要求使用ARFF格式的文件进行训练,于是我们调用Weka中的Simple CLI接口,将本地的数据集在Weka命令行中进行格式转化,转化后数据如下。

```
@relation test1_data_all
```

```
@attribute followersnum numeric
```

```
@attribute friendsnum numeric
```

```
@attribute statusesnum numeric
```

```
@attribute favouritesnum numeric
```

```
@attribute repostsnum numeric
```

```
@attribute commentsnum numeric
```

```
@attribute attitudesnum numeric
```

```
@attribute tNum numeric
```

```
@attribute tNum_at numeric
```

```
@attribute tNum_A numeric
```

```
@attribute tNum_a numeric
```

```
@attribute tNum_num numeric
```

```
@attribute tNum_space numeric
```

```
@attribute tNum_h numeric
@attribute tTime numeric
@attribute tFrq numeric
@attribute class {No, Yes}

@data
9310,665,20745,4,0,0,150,5,3,0.5,12.5,2,5,130,5,1,
797.265972,26.020175,No
501,699,714,13,0,0,180,0,0,0,6,139,35,0,867.524306,
0.823032,No3185,2425,26191,335,0,0,0,75,5,0,0.5,8,0,1,5,
65.5,0.5,188.965278,138.602183,No
```

可以看出,ARFF文件第一部分给出了头信息(Head information)。关系声明在第一行,格式为@relation <relation-name>。由于数据集中每个属性都有一个“@attribute”语句来对应,因此属性声明由“@attribute”开头的语句表示,后面的字符表示属性名称和数据类型。“@data”标记之后的就是数据信息,最后一列的class属性表示数据集类别。

获取到ARFF的文件后,为了对模型进行校验,我们调用Weka中的Explorer接口,点击Classify,再点击Choose,选

择实验所需的分类算法,如朴素贝叶斯、随机森林等,接着选择K-fold交叉验证算法,设置 $k=10$,最后点击Start,便可以进行实验。

本实验采用的评估标准为准确率(Precision)、召回率(Recall)、F值(F-Measure)等,通过对这些系数的综合分析,判断各特征对微博异常用户检测的效果。一般情况下,检测结果的准确程度和准确率成正比,且根据F值来综合衡量准确率和召回率。此外,还可以通过对检测结果的混淆矩阵来评估实验效果,即,混淆矩阵的对角线数值之和越大,实验效果越好。

4.2 基于C4.5分类算法的分类实验

本实验中采用C4.5分类算法构建的剪枝决策树。使用4524个微博用户数据,其中正常用户3801个,异常用户723个,异常用户分类实验的准确度可以达到84.0186%。结合微博用户分类实验结果的评估标准,我们对准确率、召回率、F-Measure值等评估值进行平均加权处理,得到表5所列结果。分析可知,各项指标在加权之后都能达到80%以上,具有比较好的检测效果。

表5 基于C4.5决策树的微博异常用户检测结果评估

Table 5 Evaluation of Weibo abnormal user detection results based on C4.5 decision tree

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
正常用户评估值	0.952	0.719	0.869	0.952	0.909	0.704	No
异常用户评估值	0.281	0.048	0.537	0.281	0.369	0.704	Yes
加权平均值	0.840	0.608	0.814	0.840	0.819	0.704	

在基于C4.5分类算法的实验后,随机选取一组微博数据集,验证此类模型的检测效果。

```
==== Confusion Matrix ====
a   b   <-- classified as
882 8|   a=No
71  1|   b=Yes
```

图3 C4.5决策树分类实验的混淆矩阵

Fig. 3 Confusion matrix of C4.5 decision tree classification experiment

如图3的混淆矩阵所示,一共962位训练用户,其中包含883位正常用户和79位异常用户,最终实验结果显示,有882

位正常用户和1位异常用户被正确分类,准确率达91.7879%,有71位正常用户和8位异常用户被错误分类。总体来说,异常用户的分类效果比较不错,召回率达91.8%。

4.3 基于随机森林算法的分类实验

使用4524个微博用户数据,其中正常用户3801个,异常用户723个,异常用户分类实验的准确度可以达到85.4332%。结合微博用户分类实验结果的评估标准,我们对准确率、召回率、F-Measure值等评估值进行平均加权处理,得到表6所列结果。分析可知,各项指标在加权之后都能达到80%以上,具有比较好的检测效果。

表6 基于随机森林的微博异常用户检测结果评估

Table 6 Evaluation of Weibo anomaly user detection results based on random forest

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
正常用户评估值	0.976	0.758	0.866	0.976	0.918	0.795	No
异常用户评估值	0.242	0.024	0.672	0.242	0.356	0.795	Yes
加权平均值	0.854	0.636	0.834	0.854	0.824	0.795	

在基于随机森林分类算法的实验后,随机选取一组微博数据集,验证此类模型的检测效果。

```
==== Confusion Matrix ====
a   b   <-- classified as
431 7|   a=No
73  2|   b=Yes
```

图4 随机森林分类实验的混淆矩阵

Fig. 4 Confusion matrix of random forest classification experiment

如图4的混淆矩阵所示,一共513位训练用户,其中有433位正常用户和80位异常用户,最终实验结果显示,有431位正常用户和2位异常用户被正确分类,准确率达84.4055%,有73位正常用户和7位异常用户被错误分类。总体来说,异常用户的识别效果较为不错,召回率达84.4%。

4.4 检测结果比较

针对上述分类算法的检测效果,我们将微博数据集随机

划分为4份,分别作为训练集和测试集,对C4.5决策树和随机森林分类算法对异常用户检测效果做进一步评估。加权之后的评估系数如表7所列。

表7 不同分类算法对异常用户预测效果的对比

Table 7 Comparison of different classification algorithms on prediction effect of abnormal users

分类算法	评估指标					
	C4.5决策树			随机森林		
Precision	88.7	78.8	80.1	89.7	81.6	81.5
Recall	86.2	81.5	81.5	90.2	83.3	78.2
F-Measure	87.3	79.4	80.7	90.0	79.5	79.6

由表7可知,经过对不同分类算法对微博异常用户的预

(下转第385页)

- tional Publishing, 2014;143-163.
- [5] KORCZYNSKI M A. Classifying Service Flows in the Encrypted Skype Traffic [C] // 2012 IEEE International Conference on Communications (ICC). 2012:1064-1068.
- [6] WANG T, CAI X, NITHYANAND R, et al. Effective attacks and provable defenses for website fingerprint [C] // 23rd (USENIX) Security Symposium (USENIX). 2014;143-157.
- [7] CHENG G, CHEN Y X. Encrypted Traffic Identification Method Based on Support Vector Machine [J]. Journal of Southeast University (Natural Science Edition), 2017(4):655-659.
- [8] CHEN W, HU L, YANG L. Fast Identification Method of Encrypted Traffic Based on Load Characteristics [J]. Computer Engineering, 2012(12):22-25.
- [9] ZHANG B Y. Analysis of the Principle and Application of HTTPS Protocol [J]. Network Security Technology and Application, 2016(7):36-37.
- [10] XU P, LIN S. Traffic Classification Method Based on C4.5 Decision Tree [J]. Journal of Software, 2009(10):2692-2704.

- [11] LIU K. Research on feature selection in network flow classification [D]. Yangzhou: Yangzhou University, 2013;18-19.
- [12] ZHOU Z H. Machine Learning [M]. Beijing: Tsinghua University Press, 2016:73-79.



ZOU Jie, born in 1996, postgraduate. Her main research interests include machine learning and network traffic analysis.



ZHU Guo-sheng, born in 1972, Ph. D., professor. His main research interests include next-generation Internet and software-defined networks.

(上接第 368 页)

测效果的比较,随机森林算法的分类效果要好于 C4.5 决策树,但这并不否定决策树的分类作用,因为不同的数据挖掘算法都有其优势和劣势,我们应当从微博用户的各类特征出发,在提取有效特征的基础上与传统数据挖掘算法结合使用,以起到仅使用其中一种方法所不能达到的良好效果。

结束语 随着对微博用户行为研究的不断深入,对微博上的大量异常用户的检测与识别也引起了学术界和产业界的重视。本文基于用户特征对微博异常用户分类方法进行分析,在获取到的微博用户数据的基础上,通过对用户基本属性、行为模式和文本内容的分析,提取相关特征并进行聚合,完成对数据的预处理。随后选取 Weka 作为分析工具,以微博用户数据集为样本,对微博的传播特点以及用户属性、文本内容等特征进行综合考量,将数据挖掘算法融入对新浪微博异常用户的检测方法当中,建立起微博异常用户检测模型,先后选取 C4.5 决策树和随机森林分类算法完成了模型训练和实验预测,并对其准确率进行了评估。本文结果可以为公安机关开展舆情管控工作提供参考。

参 考 文 献

- [1] 中国互联网信息中心. 第 43 次中国互联网络发展状况统计报告 [R]. 北京: CNNIC, 2019.
- [2] FABRICO B, MAGNO G, RODRIGUES T, et al. Detecting spammers and content promoters in online video social networks [C] // Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009:620-627.
- [3] BENEVENUTO F, MAGNO G, RODRIGUES T, et al. Detecting spammers on twitter [C] // Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference. Washington, 2010:6-12.
- [4] STRINGHINI G, KRUEGEL C, VIGNA G. Detecting spammers on social networks [C] // Proceedings of the 26th Annual Computer Security Applications Conference. ACM, 2010:1-9.
- [5] 彭希羨,朱庆华,刘璇. 微博客用户特征分析及分类研究——以

“新浪微博”为例 [J]. 情报科学, 2015, 33(1):69-75.

- [6] 刘勘,袁蕴英,刘萍. 基于随机森林分类的微博机器用户识别研究 [J]. 北京大学学报, 2015, 51(2):289-300.
- [7] APHINYANAPHONGS Y, RAY B, STATNIKOV A, et al. Text classification for automatic detection of alcohol use-related tweets: A feasibility study [C] // 2014 IEEE 15th International Conference on Information Reuse and Integration (IRI). IEEE, 2014:93-97.
- [8] 蒋鑫. 基于属性约简的社交网络异常用户识别系统的设计与实现 [D]. 北京: 北京邮电大学, 2016:2-3.
- [9] 夏崇欢. 基于行为特征分析的微博恶意用户检测方法 [D]. 南京: 南京邮电大学, 2018:5-6.
- [10] 郝亚洲,郑庆华,陈艳,等. 面向网络舆情数据的异常行为识别 [J]. 计算机研究与发展, 2016, 53(3):611-620.
- [11] 张玉清,吕少卿,范丹. 在线社交网络中异常帐号检测方法研究 [J]. 计算机学报, 2015, 38(10):2011-2027.
- [12] 吴大鹏,司书山,闫俊杰,等. 基于行为特征分析的社交网络女巫节点检测机制 [J]. 电子与信息学报, 2017, 39(9):2089-2096.
- [13] 刘琛. 基于行为分析的社交网络异常账号的检测 [D]. 北京: 北京交通大学, 2017.
- [14] 孙洋. LBSN 中基于好友聚类的社交搜索系统设计与实现 [D]. 南京: 东南大学, 2017.



YUAN De-yu, born in 1986, Ph. D., lecturer. His main research interests include cyber security, and complex networks.



GAO Jian, born in 1982, Ph. D., lecturer. His main research interests include botnet, malware analysis and cyber crime.