第 43 卷 第 4 期 计 算 机 工 程 2017 年 4 月
Vol. 43 No. 4 Computer Engineering April 2017

·人工智能及识别技术· 文章编号：1000-3428(2017)04-0171-06 文献标志码：A 中图分类号：TP391

基于 AHP 与 SVM 的微博机器用户检测方法

张晓艺，路 燕，翟惠良

（山东科技大学 信息科学与工程学院，山东 青岛 266590）

摘 要： 以新浪微博中的用户为研究对象，分析并提取机器用户的特征，提出一种新的微博机器用户检测方法。通过层次分析法构建分类指标体系，对各指标特征进行量化评估，利用支持向量机（SVM）算法构建机器用户检测模型。测试 SVM 中不同核函数对各分类指标的重要性预测，并与量化评估结果进行比对，同时测试不同核函数模型的分类精度，对比两项结果综合选择出最优分类器。实验结果表明，该方法能够对微博中的机器用户进行较为精确的检测。

关键词： 机器用户检测；特征提取；量化评估；层析分析法；支持向量机；最优分类器

中文引用格式： 张晓艺，路 燕，翟惠良. 基于 AHP 与 SVM 的微博机器用户检测方法[J]. 计算机工程，2017，43(4)：171-176.

英文引用格式： Zhang Xiaoyi，Lu Yan，Zhai Huiliang. Microblog Bot-user Identification Method Based on Analytic Hierarchy Process and Support Vector Machine[J]. Computer Engineering，2017，43(4)：171-176.

Microblog Bot-user Identification Method Based on Analytic Hierarchy Process and Support Vector Machine

ZHANG Xiaoyi，LU Yan，ZHAI Huiliang

（College of Information Science and Engineering，Shandong University of Science and Technology，Qingdao，Shandong 266590，China）

【Abstract】 Taking sina Microblog bot users as the object of study，this paper analyses and extracts features of the bot user and proposes a new Microblog bot user identification method. Through the Analytic Hierarchy Process（AHP），it constructs an index system and makes quantitative evaluation of each index feature. It uses Support Vector Machine（SVM）to construct a bot-user identification model. It tests different kernel functions that the importance prediction of each classification index，compared with the result of quantitative evaluation. Meanwhile，using different kernel functions tests the classification accuracy. According to the two results，the optimal classifier is selected. Experimental result shows that the identification method can make an accurate detection to the bot user.

【Key words】 bot-user identification；feature extraction；quantitative evaluation；Analytic Hierarchy Process（AHP）；Support Vector Machine（SVM）；optimal classifier

DOI： 10.3969/j.issn.1000-3428.2017.04.029

0 概述

各大社交平台的开放性为每个人都提供了表达情感、记录生活、发表言论的机会。社交平台的社会媒体性使得用户不仅将其用作交友的工具，更加倾向于使用社交平台来接收有价值的信息[1]。但随着社交平台功能的完善和全民参与度的提高，社交平台也成为搜集隐私情报、舆论话题炒作、恶意营销的工具，大量依靠程序运作的“机器用户”也随之出现。

机器用户是垃圾用户其中的一种，其存在扰乱了社交平台的正常秩序，因此机器用户的识别问题受到了国内外学者的广泛关注。文献[2]提出了根据统计特征与双向投票算法的 AttriBiVote，利用用户信任的双向传播与其邻居节点的统计特征共同决定用户类别，区分垃圾用户与普通用户。文献[3]针对微博中的反垃圾处理问题，提出了基于重用检测模型的垃圾用户检测算法，该方法综合考虑了消息序列中文本相关性和时间相关性，对垃圾用户的发布行为进行建模。文献[4]基于蚁群优化算法，构建基于规则的自适应分类器来区分出邮件中的机器用户。文献[5]通过对用户发博时间分布情况的研究，分析出机器用户与普通用户之间的差别。文献[6]从图论和博文内容两方面进行分析，得到每个用户的信誉度值，根据信誉度值利用贝叶斯分类器区分出垃圾用户。本文基于国内外学者的研究成果，提出一种基于 AHP 与 SVM 的微博机器用户检测方法。

作者简介： 张晓艺（1993—），女，硕士研究生，主研方向为社交网络分析、数据挖掘；路 燕，副教授；翟惠良，硕士研究生。

收稿日期： 2016-05-25 **修回日期：** 2016-08-08 **E-mail：** 275030843@qq.com

1 总体思路

目前对于网络中的各种机器用户的识别取得了很多成果,从不同的切入点分析机器用户的特征以建立区分模型。但在已有的识别方法中,均未考虑到机器用户特征的权重问题。分析并提取微博机器用户的特征,运用层次分析法构建分类指标体系,建立判断矩阵来量化各个分类指标的权重,利用支持向量机算法构建了机器用户检测模型,将分类指标的权重值考虑到分类模型中,其基本思路如图 1 所示。

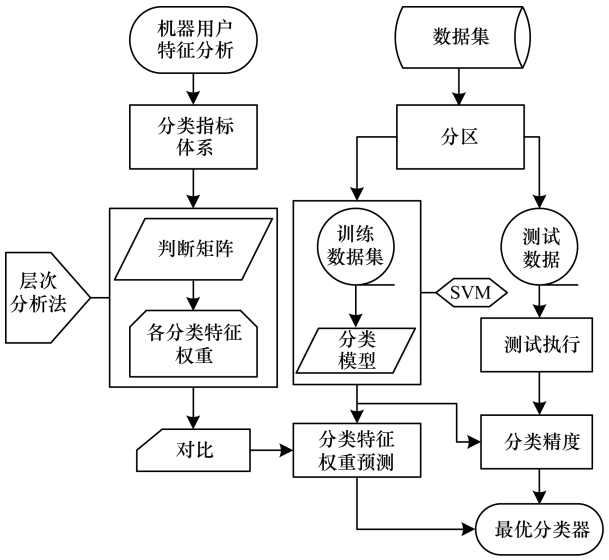

图 1 机器用户识别模型建立的基本框架

2 基于层次分析法的机器用户分类模型

2.1 机器用户特征分析

依靠程序自动地发布大量垃圾博文,其本身的用户行为以及发布的博文内容都与正常使用微博的正常用户存在很大差异。本文从 2 个角度对机器用户进行特征分析,即用户行为特征和博文内容特征。

2.1.1 用户行为特征

用户行为特征方法如下:

1)用户被关注度。使用微博时会因自己在现实生活中的社交圈子而获得一部分数量的“粉丝”。但机器用户则不同,他们不存在自己的“社交圈”,其发布的博文内容也大多是以营销、炒作为目的,其内容并不足以吸引普通用户去关注,也就是说机器用户会有很少的“粉丝”数量,并且他们为了让更多的人看到他们,会大量地关注其他普通用户。用户被关注度可以表示为:

$$\text{用户关注度}=\frac{\text{粉丝数}}{\text{关注人数}+\text{粉丝数}}$$

可以看出,机器用户的被关注度非常小。

2)互粉率。互粉率也是表现用户正常社交的重要指标。普通用户利用微博来进行日常的社交活动,互粉好友在其关注的人和粉丝数中一定会占有一定的比重。而机器用户的身份都是由是机器人程序创建的,其存在的目的并不是社交。因此,机器用户的互粉好友会非常少或几乎没有。互粉率可表示为:

$$\text{互粉率}=\frac{\text{互粉数}}{\text{关注人数}+\text{粉丝数}}$$

3)时间合理度。发博时间合理度表现了用户发博的时间规律。机器用户由程序自动发布微博,为吸引更多人的眼球,通常会在短时间内发布大量的博文,这与普通用户相比显然具有不合理性,为这一特征设置一定的阈值,统计出每分钟发布博文超过阈值的次数,即可统计出该用户发博的时间合理度[7]。

2.1.2 发布博文内容特征

发布博文内容特征方法如下:

1)关键词词频。机器用户发布博文的目的通常就是为了推广或炒作某些对象,因而其博文内容很多都会存在相同的关键词,或者大量发布内容相同的博文。设置一些特定的垃圾关键词,运用适用于中文的词频逆文档频率与向量空间模型的方法[8]统计博文的垃圾关键词词频。

2)链接比例。在博文中直接插入链接,能够更直接地推广自己的产品,因此很多机器用户会在自己的博文内容中添加链接,吸引看到的用户点入自己的电商店铺等,以实现推广的目的,因此其博文中含有链接的比例会大大高于普通用户。博文中的链接比例指标可表示为:

$$\text{链接比例}=\frac{\text{含有链接的博文数量}}{\text{博文总数}}$$

可以看出,机器用户的链接比例会明显高于普通用户。

3)提及比例。在微博中,如果想要提醒用户关注到自己的博文,就可以采用提及的方式,即“@”其他用户,机器用户为了使自己的博文被别人关注到,以实现扩大影响力的作用,往往会在博文中大量地“@”普通用户,博文中的提及比例这一指标可表示为:

$$\text{提及比例}=\frac{\text{提及数量}}{\text{博文总数}}$$

可以看出,机器用户的提及比例明显高于普通用户。

4)标签比例。微博的热门标签会随着当天的热点话题不断更新,为自己的博文加上形式为#热门话题#的标签,会使自己的博文更容易被普通用户搜索到,机器用户常常采用这一方式发布博文,其博文中含有热门标签的比例会比普通的用户大。这一指标可表示为:

$$\text{标签比例}=\frac{\text{含有标签的博文数量}}{\text{博文总数}}$$

2.2 分类指标体系的构建

本文运用层次分析法构建分类指标体系。层析分析法是美国运筹学家在 20 世纪 70 年代初提出的,将与决策相关的各个特征元素分解成目标、准则、方案等层次,在解决多层多因素的问题时能科学地体现出个指标的权重问题。本文在对机器用户的特征进行分析后,构建分类指标体系,在此基础之上进行定性和定量分析。

2.2.1 分类指标体系

经过上文的分析,本文得到的如图 2 所示的机器用

户分类体系,将分类特征进行分层,以得到机器用户分类体系模型。模型分为 3 层,即目标层 A、准则层 B 和指标层 C。机器用户分类指标体系如图 2 所示。

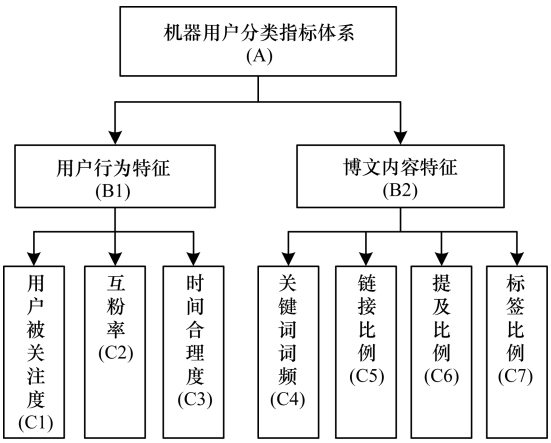

图 2　机器用户分类指标体系

2.2.2　各分类特征权重的确定

分类特征的权重确定步骤如下:

步骤 1　判断矩阵的建立。

本次实验选取 10 位经常使用新浪微博的学生,对层次体系结构中的各项指标进行两两比较,采用"1 ~9"的评判方法进行赋值,建立两两判断矩阵[9]。判断矩阵是进行相对重要程度计算的重要依据[10],可表示为:

$$A=(a_{ij})_{n\times n}$$

其中,a_{ij}代表矩阵元素 U_i与 U_j相对于其上一层元素重要性的比例标度,比值越大,则 U_i的重要度就越高。

各层的比较标度如表 1、表 2 所示。

表 1　机器用户分类体系用户行为特征评价要素的比较标度

分类要素	用户被关注度	互粉率	时间合理度
用户被关注度	1	3	6
互粉率	1/3	1	5
时间合理度	1/6	1/5	1

得到判断矩阵为:

$$\boldsymbol{U}_1=\begin{bmatrix}1 & 3 & 6\\ 1/3 & 1 & 5\\ 1/6 & 1/5 & 1\end{bmatrix}$$

表 2　机器用户分类体系博文内容特征评价要素的比较标度

分类要素	关键词词频	链接比例	提及比例	标签比例
关键词词频	1	4	8	6
链接比例	1/4	1	5	3
提及比例	1/8	1/5	1	1/3
标签比例	1/6	1/3	3	1

得到判断矩阵为:

$$\boldsymbol{U}_2=\begin{bmatrix}1 & 4 & 8 & 6\\ 1/4 & 1 & 5 & 3\\ 1/8 & 1/5 & 1 & 1/3\\ 1/6 & 1/3 & 3 & 1\end{bmatrix}$$

步骤 2　权重的确定。

对判断矩阵中的每一列进行归一化处理之后,对归一化的值进行求和,再将求和结果进行归一化,即可得到特征向量的值。经过计算[11],最终得到机器用户分类体系中目标层、准则层和指标层的各分类指标的权重,如表 3 所示。

表 3　各分类指标体系权重

目标层(A)	准则层(B)	指标权重(B)	指标层(C)	指标权重(C)	指标总权重
机器用户分类指标体系	用户行为特征	1/2	用户被关注度	0.63	0.315
			互粉率	0.29	0.150
			时间合理度	0.08	0.040
	博文内容特征	1/2	关键词词频	0.60	0.300
			链接比例	0.23	0.120
			提及比例	0.05	0.030
			标签比例	0.11	0.060

3　对机器用户进行的分类检测

3.1　支持向量机算法描述

支持向量机是由 Vapnik 等人于 1995 年首先提出的,它是基于 VC 维理论和结构风险最小化原则的学习机器。在解决小样本、非线性和高维模式识别问题中表现出许多特有的优势,即在有限的训练集样本得到小的误差仍然能够保证对独立的测试集保持小的误差[12]。

本文从 2 个方面考量分类模型是否最优:一方面是分类模型对各个分类特征的重要性预测;另一方面是分类模型的分类精度。

1)在支持向量机中,分类超平面将数据分为 2 类。从数学上来看,分类线性方程为:

$$<w\cdot x>+b=0$$

其中,$<w,x>$为 2 个向量的内积,w 为权值;b 为一个常数,当两类样本线性可分时,满足条件:

$$y_i[<w,x>+b]\geqslant 1,i=0,1,\cdots,l$$

其中,y_i为样本点 x_i的类别。$<w,x_i>$越大,第 i 个变量对分类超平面的贡献就越大,这个变量在分类中所占的权重也就越大。

2)在支持向量机中,核函数决定了支持向量机的非线性处理能力。对于非线性的情况,SVM 的处理方法是选择一个核函数 $K(x)$。概括地说,支持向量机就是首先通过用内积函数定义的非线性变换将输入空间变换到一个高维空间,因此允许在高维空间中构造分类超平面。通过构造核函数,将运算放于样本空间进行,减少了计算量,以线性的代价获得非线性的优良特性[13],这就是核方法的思想。

根据核函数 K 的不同,可生成不同的支持向量机,进而产生不同的分类结果,常用的有:

(1)高斯径向基函数(RBF Kernel):

$K(x,x_i)=\exp(-\|x-x_i\|^2/\sigma^2)$

(2)多项式核函数(Poly Kernel):

$K(x,x_i)=(1+x\cdot x_i)^d$

(3)Sigmoid 核函数(Sigmoid Kernel):

$K(x,x_i)=\tanh[k_1(x\cdot x_i)+k_2]$

3.2 实验数据准备

本文实验中所用的数据来自新浪微博,参照文献[1]中对 Twitter 数据进行挖掘的方式,采用"滚雪球"式的数据获取方法。首先人工从微博中选取粉丝数大于 100 的用户作为种子用户,对于每个种子用户,获取最近 200 条博文以及该种子用户所有的粉丝和关注,同时将粉丝列表中的用户再加入到种子用户中,再获取他们的博文内容以及粉丝和关注,如此反复迭代,最终选取 700 个用户的信息作为原始数据集。

在获取原始数据之后,将数据按照机器用户分类体系中所需的特征进行统计,按如表 4 所示的属性获取数据表,并对数据做归一化处理。

表 4 用户数据属性表

特征	属性	说明
用户行为特征	Attention rates	用户被关注度
	Follows rates	互粉率
	Time rates	时间合理度
博文内容特征	Key words	关键词词频
	Link rates	链接比例
	Mention rates	提及比例
	Tag rates	标签比例

本文将原始数据集分为训练数据和测试数据,抽取一半的数据集作为训练数据集,另一半作为测试数据集。进行预先的标记分类时,为保证对训练结果的正确分类,由 10 位经常使用新浪微博的学生,进入用户的微博主页,由他们人工判别抽取的用户是否是机器用户,判别标准参照文献[14]。首先进入微博用户的主页,如果其个人资料较为完整,并且博文内容中有自己通过客户端发布的原创内容,博文下有与其他好友进行的互动,相册中还包含一些相关照片,一般就判别该用户为普通用户。反之,就是判别为机器用户。将分类结果用标号 -1 和 +1 来表示,其中,-1 表示机器用户;+1 表示普通用户。

3.3 模型建立

创建如图 3 所示的基本流,用不同的核函数建模。将数据集通过分区节点在源数据表中添加一个字段,根据字段取值的不同,将数据随机分为训练数据和测试数据,并将数据转化为 SVM 分类器可识别的数据[15];一部分用于建立和训练模型;另一部分用于测试模型。类型节点用来选择数据表中各字段的角色设置。例如,将 Class 字段的角色设置为"目标",表示数据的分类结果;ID 字段是各行源数据的标号,其值不会对模型的分类和预测产生影响,将其角色设置为"无";其他字段的角色设置为"输入",作为分类特征字段用作预测变量。然后选择不同的核函数创建 SVM 模型,对比生成结果。本文选用 3 种最常用的核函数,即 RBF 高斯径向基核函数、poly 多项式核函数和 Sigmoid 核函数。

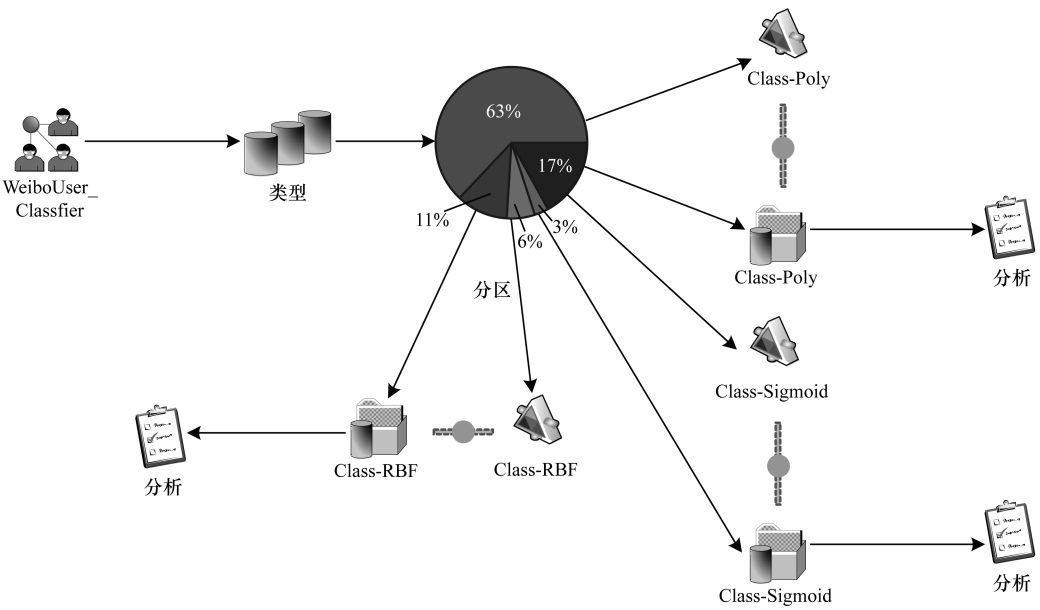

图 3 基本流建模图

4　实验结果与分析

4.1　各分类特征预测的重要性

不同的核函数对各分类特征的权重会有不同的判断。将上文得出的机器用户分类指标体系中所得到的各分类特征的权重与 3 种核函数对各分类特征的权重所作出的预测相比较,判断模型的可靠性。

RBF 核函数测得的各分类特征的权重如图 4 所示。

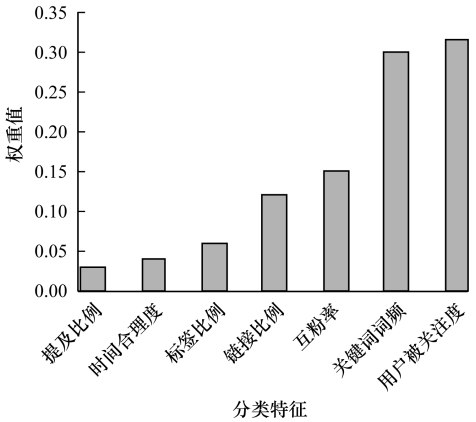

图 4　RBF 核函数的分类特征权重

多项式核函数测得的各分类特征权重如图 5 所示。

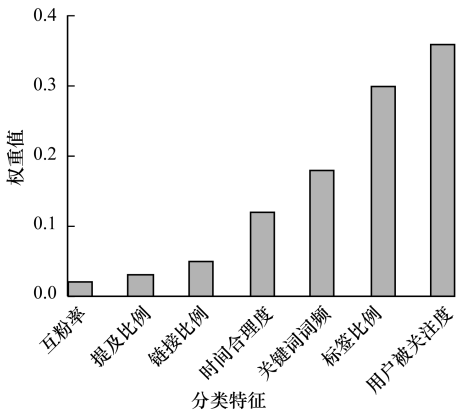

图 5　多项式核函数的分类特征权重

Sigmoid 核函数测得的各分类特征权重如图 6 所示。将不同核函数测得的分类特征权重与机器用户分类指标体系中的权重进行比对,比对结果如图 7 所示。

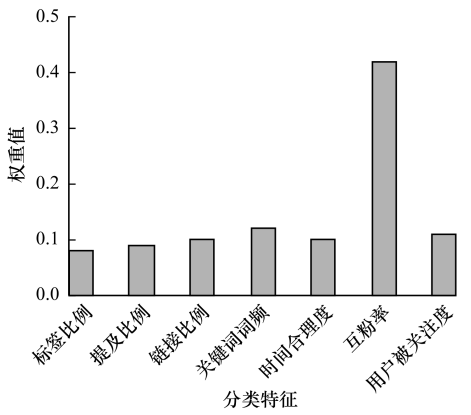

图 6　Sigmoid 核函数的分类特征权重

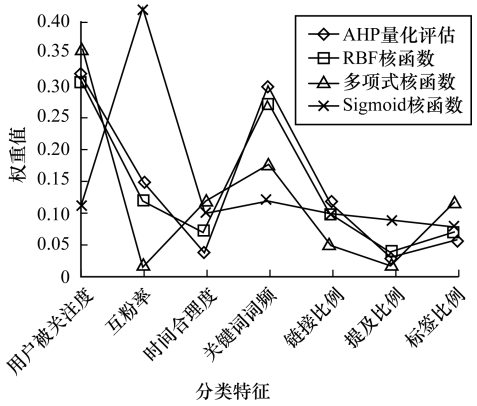

图 7　各分类指标特征权重比对结果

4.2　训练测试模型的分类精度

支持向量机分类中的核函数对分类结果的精度影响很大,确定了使用哪种核函数后,在参数估计和预测时,只需计算相应的核函数。但选择怎样的核函数并没有确定的准则,需要对数据集进行反复测试来选择最适合的核函数。RBF 核函数、多项式核函数、Sigmoid 核函数的分类精度结果如表 5 ~ 表 7 所示。

表 5　RBF 核函数分类精度

分类	训练规模		测试规模	
	训练数据集	分类精度/%	测试数据集	分类精度/%
正确	344	97.45	336	97.11
错误	9	2.55	10	2.89
总计	353		346	

表 6　多项式核函数分类精度

分类	训练规模		测试规模	
	训练数据集	分类精度/%	测试数据集	分类精度/%
正确	353	100	327	94.51
错误	0	0	19	5.49
总计	353		346	

表 7　Sigmoid 核函数分类精度

分类	训练规模		测试规模	
	训练数据集	分类精度/%	测试数据集	分类精度/%
正确	333	94.33	324	93.64
错误	20	5.67	22	6.36
总计	353		346	

4.3　分析结果

分析结果如下:

1)在分类指标重要性预测这一测试中,从图 7 的比对结果中可以看出,RBF 核函数所得到的各分类指标权重的结果与机器用户分类指标体系中的结

果最为相近,多项式核函数与 Sigmoid 核函数的预测结果存在很大差异。因此,在预测分类特征重要性的测试中,采用 RBF 这一核函数的分类模型具有较高的可靠性。

2)在分类精度这一测试中,对于训练数据集中数据的分类,3 个模型的正确率都比较高;对于测试数据集中的分类,RBF 核函数的正确率稍高。总体而言,3 个模型预测的准确率差不多,采用 RBF 核函数分类模型的预测能力稍好。

综合分类模型对分类指标重要性的预测和对于分类精率这两项测试结果,在机器用户分类中,采用 RBF 核函数的支持向量机分类模型具有较为理想的检测结果。

5 结束语

本文针对机器用户自身的特点,从用户行为特征和博文内容特征 2 个方面入手,构建机器用户的分类指标体系,运用层次分析法确定各指标权重,利用支持向量机算法构建模型,通过比对各分类特征重要性预测和分类精度这 2 项结果来选择出最优分类器。实验结果表明,该方法能够对机器用户做出正确率较高的识别。但是目前机器用户的伪装手段越来越多样化,应结合其新的隐藏手段以及博文具体的语义特征做进一步的深度分析,强化分类体系的分类能力。下一步将研究机器用户的新特点以及伪装手段,结合其他分类算法选择最优分类器,使机器用户识别检测方法更具智能性。

参考文献

[1] 丁兆云,贾 焰,周 斌. 微博数据挖掘研究综述[J]. 计算机研究与发展,2014,51(4):691-706.

[2] 丁兆云,周 斌,贾 焰,等. 微博中基于统计特征与双向投票的垃圾用户发现[J]. 计算机研究与发展,2013,50(11):2336-2348.

[3] 赵 斌,吉根林,曲维光,等. 基于重用检测的微博垃圾用户过滤算法[J]. 南京大学学报(自然科学版),2013,49(4):456-464.

[4] Manaskasemsak B,Rungsawang A. Web Spam Detection Using Trust and Distrust-based Ant Colony Optimization Learning[J]. International Journal of Web Information Systems,2015,11(2):142-161.

[5] Gabriela T,Aldo F. Scaling-laws of Human Broadcast Communication Enable Distinction Between Hhuman,Corporate and Robot Twitter Users[J]. PloS ONE,2013,8(7):78-86

[6] Wang A H. Don't Follow Me:Spam Detection in Twitter[C]// Proceedings of IEEE International Conference on Security and Cryptography. Washington D. C.,USA:IEEE Press,2010:1-10.

[7] 胡建华. 微博用户行为与影响力分析系统的研究[M]. 北京:北京交通大学,2013.

[8] 陈 欣,郑 啸,焦媛媛,等. 一种基于支持向量机的垃圾微博识别方法[J]. 安徽工业大学学报(自然科学版):2013,30(10):440-445.

[9] 梁冬莹,周庆梅,王克奇. 基于层次分析法的数字资源服务绩效评价体系构建[J]. 情报科学,2013,31(1):78-81

[10] 张 瑜. 铁路客运站旅客安全预警管理系统研究[D]. 北京:北京交通大学,2008.

[11] 申志东. 运用层次分析法构建国有企业绩效评价体系[J]. 审计研究,2013(2):106-112.

[12] 顾亚祥,丁世飞. 支持向量机研究进展[J]. 计算机科学,2011,38(2):14-17.

[13] 梁礼明,钟 震,陈召阳. 支持向量机核函数选择研究与仿真[J]. 计算机工程与科学,2015,37(6):1135-1141.

[14] 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(自然科学版):2015,51(2):289-300.

[15] 董雨辰,刘 淡,罗军勇,等. 基于支持向量机的炒作微博识别方法[J]. 计算机工程,2015,41(3):7-14.

编辑 索书志

(上接第 170 页)

[11] Wu Wei,Yi M,Willy S,et al. Server-aided Verification Signatures:Definitions and New Constructions[C]// Proceeding of ProvSec'08. Shanghai,China:[s. n.],2008:141-155.

[12] Wu Wei,Yi M,Willy S,et al. Provably Secure Server-aided Verification Signatures[J]. Computers & Mathematics with Applications,2011,61(7):1705-1723.

[13] 杨小东,杨苗苗,高国娟,等. 强不可伪造的基于身份服务器辅助验证签名方案[J]. 通信学报,2016,37(6):49-55.

[14] 徐晓琴. 可证明安全数字签名的研究[D]. 秦皇岛:燕山大学,2009.

[15] Hu Xiaoming,Zhang Zhe,Yang Yinchun. Identity Based Proxy Re-signature Schemes Without Random Oracle[C]// Proceedings of CIS'09. Washington D. C.,USA:IEEE Press,2009:256-259.

编辑 索书志