doi:10.11959/j.issn.1000-436x.2016152

新浪微博反垃圾中特征选择的重要性分析

张宇翔[1,2,3]，孙菀[1]，杨家海[2,3]，周达磊[4]，孟祥飞[5]，肖春景[1]

（1. 中国民航大学计算机科学与技术学院，天津 300300；2. 清华大学网络科学与网络空间研究院，北京 100084；
3. 清华信息科学与技术国家实验室，北京 100084；4. 北京邮电大学网络技术研究院，北京 100876；
5. 北京航空航天大学虚拟现实技术与系统国家重点实验室，北京 100876）

摘 要：微博中的垃圾用户非常普遍，其异常行为及生产的垃圾信息显著降低了用户体验。为了提高识别准确率，已有研究或是尽可能多地定义特征，或是不断尝试提出新的分类检测方法；那么，微博反垃圾问题的突破点优先置于寻找分类特征还是改进分类检测方法，是否特征越多检测效果越好，新的方法是否可以显著提高检测效果。以新浪微博为例，试图通过不同的特征选择方法与不同的分类器组合实验回答以上问题，实验结果表明特征组的选择较分类器的改进更为重要，需从内容信息、用户行为和社会关系多侧面生成特征，且特征并非越多检测效果越好，这些结论将有助于未来微博反垃圾工作的突破。

关键词：新浪微博；特征生成；特征选择；垃圾用户检测

中图分类号：TP391 **文献标识码**：A

Feature importance analysis for spammer detection in Sina Weibo

ZHANG Yu-xiang[1,2,3], SUN Yu[1], YANG Jia-hai[2,3], ZHOU Da-lei[4], MENG Xiang-fei[5], XIAO Chun-jing[1]

(1.College of Computer Science, Civil Aviation University of China, Tianjin 300300, China;
2.Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China;
3.Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China;
4.Institue of Network Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
5. State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100876, China)

Abstract: Microblog has drawn attention of not only legitimate users but also spammers. The garbage information provided by spammers handicaps users' experience significantly. In order to improve the detection accuracy of spammers, most existing studies on spam focus on generating more classification features or putting forward new classifiers. Which kind of issues would be put the high priority of an enormous amount of research effort into? Are extensive features or novel classifiers better for the detection accuracy of spammers? It is tried to address these questions through combining different feature selection methods with different classifiers on a real Sina Weibo dataset. Experimental results show that selected features are more important than novel classifiers for spammer detection. In addition, features should be derived from a wide range, such as text contents, user behaviors, and social relationship, and the dimension of features should not be too high. These results will be useful in finding the breakpoint of Microblog anti-spam works in the future.

Key words: Sina Weibo, feature definition, feature selection, spammer detection

收稿日期：2015-11-23；**修回日期**：2016-04-24

基金项目：国家重点基础研究发展计划(“973”计划)基金资助项目(No.2009CB320505)；国家科技支撑计划基金资助项目(No.2008BAH37B05)；国家自然科学基金资助项目(No.61170211, No.U1533104, No.61301245)；教育部博士点基金资助项目(No.20110002110056)

Foundation Items: The National Basic Research Program of China (973 Program) (No.2009CB320505), The National Key Technology R&D Program of China(No.2008BAH37B05), The National Natural Science Foundation of China (No.61170211, No.U1533104, No.61301245), Ph.D. Programs Foundation of Ministry of Education of China (No.20110002110056)

1　引言

微博是一种近年来新兴的在线社交网络(online social network)，用户可通过 Web、WAP 等各种客户端在其上组建个人社区，并允许发布 140 字左右的文字更新信息，用户之间通过建立单向或双向的友好关系实现信息即时分享。

在微博成为人们日常交流的重要方式之时，同时也成为垃圾用户(spammer)发布非法广告和垃圾消息的平台。2013 年 7 月，新华网报道[1]“新浪微博社区公约体系上线运行约一年时间，微博管理中心共接到超过 1 500 万次的用户举报，其中垃圾广告达到 1 200 多万次，淫秽色情危害信息达到 100 多万次”。根据人民网报道[2]，大量虚假粉丝严重侵害用户利益并影响微博生态，2015 年 1 月起新浪微博根据用户举报和数据分析清除垃圾粉丝。微博中垃圾问题非常严重，垃圾用户的异常行为及生产的垃圾信息显著降低了用户体验，增添了社会风险。

学术界开展较早的反垃圾研究包括垃圾网页检测[3]、垃圾邮件过滤、虚假在线评论过滤[4]、网络众包(crowdsourcing)中的欺骗检测[5]、传统社交网络(如人人网[6])中的垃圾过滤等，研究中用到的反垃圾方法对于微博中垃圾用户检测有一定的借鉴意义，但因为微博的构成要件及其功能均不同于前述应用，故不能将其直接应用于微博反垃圾。

微博反垃圾问题的解决非常困难，其原因主要有以下几个方面：1) 微博文字信息非常短，并且带有大量的不规范用语，因此微博文字内容具有噪声多、特征词少等特点；2) 简短的文字信息中可包含页面、图片、音频、视频的链接，非法用户将链接指向与文字信息不一致的垃圾内容，加之目前广泛使用的 URL 缩短服务，很难做到机器自动鉴别链接指向内容；3) 垃圾制造者不断创新，以更高明的方式躲避检测，且更新周期越来越短[7]。

微博反垃圾研究起步较晚，研究成果不多，而且已有研究绝大多数均是针对 Twitter，少部分针对 Myspace[8,9]、Facebook[8,10]、Foursquare[11]等。目前，鲜有针对新浪微博的反垃圾研究工作，尽管新浪微博与 Twitter 在基本功能上较为相似，但在网民构成、传播内容、转发模式、开放性、好友管理、扩展功能等方面均存在较大差异[12~15]，加之针对 Twitter 的反垃圾研究也处于研究初期，因此不能将 Twitter 中的反垃圾方法直接用于新浪微博反垃圾中。

在微博垃圾用户检测研究中，先要确定待检测垃圾用户所指的具体对象，是发布垃圾内容(如虚假信息、垃圾链接等)者，还是僵尸、水军等；接着通过微博提供的 API 接口等方式采集检测所需的数据；最后选取有利于垃圾检测的特征，利用机器学习方法对所有用户进行分类检测，确定垃圾用户。

微博反垃圾研究大都围绕上述步骤展开，方法上的差异主要体现在最后一步。研究初期采用的方法是试图找到能较好地区分正常用户与垃圾用户的少数几个特征，通过设定恰当的阈值来区分，如 2011 年，文献[16]针对 Twitter 选取用户发布连续消息的时间间隔(Timestamp gap<10 s)和文本内容相似性(Levenshtein<5 或 Jaccard>0.6)2 个特征，通过设置阈值来识别自动程序垃圾用户，检测结果的检测精度为 81.48%，召回率为 82.07%。这种方法简单易操作，但检测效果不理想。

随后的研究从 2 个方面展开，一方面选取尽可能多的特征，然后利用分类算法对微博用户进行分类检测。如文献[17]针对 Twitter 定义 39 个文本内容和 23 个用户行为特征，采用支持向量机(SVM, support vector machine)对用户进行分类检测，大约有 70%的垃圾用户和 96%的正常用户被正确检测。文献[9]针对 Twitter 选出 1.2 万个文本内容特征，分别采用 5 种分类器进行检测，决策树(C4.5)的准确率为 99.4%，检测效果最佳。另一方面，针对所定义的特征不断尝试各种分类方法，除上述之外，包括朴素贝叶斯(Naïve Bayes)[9]、*k*-最近邻(*k*-NN, *k*-nearest neighbor)[18]、AdaBoost[19]、神经网络(neural network)[20]、随机森林(random forest)[21]、数据流聚类方法(StreamKM++、Den-Stream)[22]等、混合马尔可夫模型(mixture of Markov models)[23]等。

研究结果存在如下 2 种现象：1) 针对同一个社交网络，同一个分类算法在不同的文献中分类效果迥异，如针对 Twitter 反垃圾，贝叶斯方法在文献[24]效果最佳，而在文献[9]效果最差；2) 针对不同的社交网络，同一个分类算法的分类效果均为最佳，如针对 Twitter 和 MySpace，Decorate 分类器的检测效果均最佳[25]。于是会有如下问题，解决微博反垃圾问题的突破点优先置于寻找分类特征还是改进分类检测方法，特征越多检测效果是否越好，新颖的方法是否可以显著提高检测效果。

本研究将以新浪微博为实例对该问题进行深入探讨。首先，通过调用新浪微博开放的 API 接口

收集新浪微博中山大学社区用户的个人页面信息，包括用户个人资料、粉丝数、关注数、微博创建时间、微博内容、微博数量，共计获取了 9 万个微博用户的信息。接着，结合已有研究提出的区分度大的特征，从内容信息、用户行为和社会关系 3 个方面生成 17 个极具代表性的特征。最后，将 7 个特征选择算法(其中 6 个为经典算法, 另外一个为本文提出的综合特征选择算法)与 10 个典型的分类识别学习算法组合进行实验，从而回答上述问题。

2 相关工作

如前所述，早期的研究试图找到少数几个有利于分类检测的关键特征(如文献[16]), 然而检测效果非常不理想。鉴于此，研究者将特征的选取扩大到某个单一方面，常常会伴随提出一些新颖的检测方法。如文献[26]检测 Twitter 中热门话题中的不相关内容，仅选取文本特征，发现在 5 个典型的分类器中 SVM 的检测准确率最高。文献[27]检测人人网中的垃圾用户，仅考虑用户的社会活动行为，引入活动数量矩阵(user-activity count matrix)，矩阵的行向量表示一个特定用户的活动数量，列向量对应不同类型的社会活动，采用矩阵分解和支持向量机相结合方法对用户进行了分类检测。文献[28]仅局限于 Twitter 中推文内容的情感特征，利用结合矩阵分解的优化模型来识别垃圾。文献[29]针对商业网站在线评论，根据文本内容中 URL 连接关系的变化使用无监督方法识别垃圾。文献[30]限于 Twitter 中推文中包括的 URL, 共定义 9 个相关特征, 利用 SVM 进行分类检测。2015 年，SIGKDD 中的文献[23]针对 Tagged.com 中用户在时序上的相关关系特征, 利用混合马尔可夫模型来识别垃圾用户。2015 年，SIGIR 中的文献[31]专注于 Twitter 中的 Hashtag 特征，先选用 *k*-NN 算法过滤掉明显的垃圾信息，后利用最大期望算法(EM，expectation-maximization)识别剩余的难于识别的垃圾信息。

仅使用少数几个特征或某一特定方面的特征的反垃圾检测不够准确，因为垃圾用户易于推断反垃圾检测的主要依据特征，进而有针对性地伪装为合法用户，从而避免检测[29, 32, 33]。

另一条研究主线是从多个侧面定义尽可能多的特征，然后借助机器学习分类方法来检测垃圾用户。文献[34]针对 Twitter 从文本内容、用户跟随关系 2 个方面定义特征, 利用优化模型检测垃圾用户。文献[35]对 Twitter 中新闻的可信度展开检测，从文本内容、用户行为、主题和传播 4 类方面生成 74 个特征，采用决策树分类器对每条新闻的可信度进行检测。文献[36]从用户拓扑、文本内容和众包 3 方面生成 18 个特征，采用 AdaBoost 和支持向量机对垃圾信息进行分类检测。文献[25]针对 MySpace 和 Twitter 中的垃圾用户进行检测, 针对前者从个人注册信息和私信文本内容生成特征，针对后者从用户行为和信息文本内容生成特征，然后采用标准分类器对进行分类检测，实验表明不管是前者还是后者，Decorate 分类器的检测效果最佳。文献[37]针对新浪微博垃圾用户检测, 除了使用常用的特征(如 URL 链接比、关注粉丝比等)之外，还关注社交网络传播的有向特性，在此基础上提出了基于统计特征与双向投票的垃圾用户检测算法。2014 年 SIGIR 中的文献[38]利用 Web、邮件等中的垃圾信息进行特征映射迁移学习，采用矩阵分解与优化模型结合的方式检测 Twitter 等社交网络中的垃圾。

特征选取的范围不同，采用的分类方法也不同。当特征类型单一时，有利于提出新颖的、高技术难度的分类方法，但受特征源单一的局限使容易被垃圾用户识破，从而避免被检测到。当特征类型多时，垃圾用户不易于躲避检测，同时也不利于提出新的分类方法。总之，已有研究通常以用户分类检测效果好坏为唯一目标，并没有深入探讨将研究重点优先置于寻找分类特征还是改进分类方法会更有利于提高分类效果，本文以新浪微博为实例，对上述问题进行详细讨论。

3 问题形式化描述

3.1 问题形式化

设微博用户集为 $U=\{u_1, u_2, \cdots, u_N\}$，其中，$N$ 为用户数目。用户 u_i 拥有个人页面 P_i，包括个人资料、微博、关注/粉丝等信息。垃圾用户检测定义为根据事先抓取的用户个人页面 P_i 和分类器 *Classifier* 预测用户 u_i 是正常用户还是垃圾用户, 形式化为：*Classifier*: $u_i \rightarrow \{spammer, legitimate\ user\}$。

3.2 垃圾用户

垃圾用户通常是指在微博中展示、发表和传播垃圾信息的用户。通常不同的研究会从不同的角度赋予垃圾用户不同的内涵。

本文根据微博的实情将垃圾用户分为内容垃圾、僵尸垃圾、封号垃圾 3 类。内容垃圾主要传播

黄色信息、虚假中奖信息、不良网站链接。僵尸垃圾可分为文本僵尸和异常转发用户，主要以兜售粉丝为目的。封号垃圾是指被官方关停的垃圾用户，多数是由自动程序产生。

4　特征生成与分析

对于微博，人们在其上浏览、发布、转发和评论信息，而信息传播主要依赖于用户间社会性的交往与互动，在微博中这种社会关系是由用户间的“关注—被关注”体现出来，它是现实世界中社会关系在社交网络中的复制和重构。基于此，垃圾应该产生于内容信息、用户行为和社会关系 3 方面，故本文从这 3 个方面定义特征。

选取特征的基本原则是：根据统计指标，挑拣区分度大的特征；使特征之间的相关性最小；保留中性特征。借鉴相关文献中有代表性的特征，结合新浪微博的实际情况，经过反复计算与分析，最终选取了 17 个特征，其统计指标如表 1 所示。

4.1　社会特征

关注数(F_1)为相关微博关注其他微博总数。垃圾用户的关注数远远高于正常用户，且垃圾用户的关注数的离散度较正常用户的小，其原因很可能是正常用户常会根据自己的兴趣有选择地关注其他用户，而以获得更多粉丝为目的垃圾用户，必然会大量关注其他用户，期待所关注用户回粉。

粉丝数(F_2)为相关微博的粉丝总数，表明垃圾用户的粉丝数要明显少于正常用户，且它的粉丝数的离散度较正常用户的小很多，很可能因为垃圾用户没有正常的社会关系，导致很少有人会关注它。

互粉数(F_3)为互为粉丝的数量，反映用户的真实好友数量。正常用户的互粉数远多于垃圾用户的，且它的离散度较垃圾用户的小很多。很可能因为真实的社会关系会给正常用户带来许多互粉数，而垃圾用户即使主动关注了其他用户，因其不在真实的社交关系中，故其他用户回粉的概率很小。

关注粉丝比(F_4)为关注数与粉丝数的比值。正常用户的关注数少、粉丝多，而垃圾用户的恰与其相反，故二者的比值更大，有利于提升检测效果。

关注互粉比(F_5)为关注数与互粉数的比值，较 F_4 更能放大正常用户与垃圾用户之间的差距，其原因可根据 F_1 和 F_3 的定义简单推得。

4.2　用户行为特征

用户名复杂度(F_6)为用户名字的复杂度。部分垃圾用户有着极其相似的命名特征，名字长度较长并且较复杂，定义如下(该特征和特征 F_{15}~F_{17} 均需先经过分词处理，采用了 NLPIR(natural language processing & information retrieval)中文分词工具[39])。

$$complex = n + \sum_{i=1}^{k} \mathrm{ceil}\left(\frac{length_i}{3}\right) \tag{1}$$

其中，n 表示词的数量，k 表示数词的个数，$length_i$ 表示第 i 个数词的长度。正常用户与垃圾用户的名字复杂度的统计特征差距并不明显，说明许多垃圾用户常会起与正常用户相近的名字，但该特征能够较准确地检测出少部分垃圾用户。

微博数(F_7)为发布的微博总数。较正常用户，内容垃圾用户通常会发布较多的博文，僵尸为了避免检测也会发布适当数量的微博。

月均微博(F_8)为数据采集期间用户每月所发的微博数，可衡量用户所发微博的活跃频度。垃圾用户的活跃度要高于正常用户的，特别是内容垃圾用户最为活跃，而僵尸最不活跃。

时间间隔(F_9)为用户最近一次发布微博距数据采集结束时刻的时间间隔(单位为天)。垃圾用户的时间间隔均比较大，其原因是它会为某种利益目的进行短暂非法活动，当活动结束就不再发送博文，如有些广告用户可能卖完某一商品就停用。

转发比(F_{10})为转发的微博与微博总数之比。较正常用户，垃圾用户常会转发其他用户的微博，以达到其扩散某些非法信息的目的。

4.3　内容特征

URL 链接比(F_{11})为含有 URL 的微博数量与微博总数之比。内容垃圾用户的 URL 链接比最大且变异系数最小，其原因很可能是在微博中放置链接诱使用户进入，从而达到某些恶意目的。

微博评论比(F_{12})为收到的评论数与微博总数之比。正常用户会与好友就所发微博进行交流，而垃圾用户由于其微博的信息价值低且没有真正的“好友”，故所发的微博一般不会有用户去评论。

原创微博评论比(F_{13})为收到的评论数与原创微博总数之比。较 F_{12} 更具区别性。

微博平均长度(F_{14})为博文的平均长度。从统计指标上看，正常用户与垃圾用户的差别不大，其原因可能是因内容垃圾的平均长度较大造成的，但在特征选择算法中该特征的排名相对靠前。

表 1 特征的统计指标

特征	中值		均值					变异系数				
	合法用户	垃圾用户	合法用户	垃圾用户	内容垃圾	僵尸垃圾	封号垃圾	合法用户	垃圾用户	内容垃圾	僵尸垃圾	封号垃圾
F_1	317.00	1 107.00	492.92	1 110.95	1 106.02	1 396.29	1 004.48	1.01	0.48	0.60	0.41	0.33
F_2	276.00	141.50	1 137.95	428.13	772.52	239.67	241.19	28.20	7.12	6.52	5.53	1.05
F_3	119.00	4.00	184.42	67.66	146.75	46.22	15.07	1.28	2.88	1.93	3.30	4.41
F_4	1.11	6.33	3.00	50.31	46.11	140.50	16.95	5.23	3.21	2.41	1.99	6.24
F_5	2.50	249.71	18.85	419.37	316.20	560.24	443.94	5.94	1.19	1.73	1.13	0.81
F_6	3.00	3.00	3.56	4.27	3.77	3.30	5.05	0.81	1.13	1.10	0.90	1.14
F_7	555.00	349.00	1 271.76	541.86	818.45	372.96	397.47	1.73	2.22	2.37	1.59	0.65
F_8	3.33	4.94	5.27	7.84	10.56	4.13	7.31	1.63	4.15	5.14	1.12	0.71
F_9	62.00	62.00	86.51	135.35	165.86	247.50	66.56	1.00	1.09	0.96	0.81	0.55
F_{10}	0.58	1.00	0.55	0.73	0.51	0.57	0.96	0.49	0.51	0.74	0.70	0.17
F_{11}	0.09	0.01	0.17	0.23	0.47	0.29	0.03	1.22	1.50	0.81	1.21	5.15
F_{12}	0.82	0.38	1.70	0.36	0.35	0.16	0.42	1.56	2.09	3.22	1.87	0.33
F_{13}	1.50	0.00	2.75	0.20	0.41	0.15	0.03	1.37	4.73	3.35	3.26	6.36
F_{14}	97.06	124.11	97.63	111.64	100.37	97.93	126.25	0.30	0.24	0.30	0.28	0.11
F_{15}	0.04	0.04	0.07	0.09	0.16	0.07	0.04	1.18	1.51	1.15	1.54	0.89
F_{16}	5.12	7.53	6.37	8.62	10.44	7.02	7.92	1.26	0.99	1.26	0.76	0.41
F_{17}	0.05	0.05	0.05	0.07	0.09	0.07	0.05	0.50	0.71	0.60	0.41	0.33

因垃圾用户发布的微博具有很强的相关性，故文本内容相似性是非常重要的反垃圾检测特征，本文采用基于词语级别的博文之间的余弦相似度(F_{15})、模相似度(F_{16})和词语共享率(F_{17})3 个特征，分别来从不同的时间粒度来度量微博之间的相似性。其中，F_{15} 计算用户在相邻两天所发微博的相似程度，F_{16} 计算用户在一天内所发微博的相似程度，而 F_{17} 没有强调时间上的相似度。

5 特征选择算法与分类器

5.1 特征选择算法

特征选择(feature selection)是指从原始特征集中选出与任务最相关的特征子集，使任务达到和特征选择前近似甚至更好的效果。通过特征选择，一些与任务无关和相互冗余的特征被删除，无关和冗余特征不仅增加特征空间的维数，降低学习的效率，而且还增加噪声数据的可能，从而干扰学习算法的学习过程，并最终影响分类模型的构造。

特征选择通常选择与类别相关性强、且特征彼此间相关性弱的特征子集，由特征子集生成、子集评价、终止条件判断和子集验证 4 个步骤组成[40]。根据特征子集评价与分类学习算法的结合方式，特征选择算法可主要分为 Filter、Wrapper 2 大类，前者使用独立于学习算法的评估准则来滤去任务无关特征和冗余特征，后者使用后续的分类准确率作为评价函数。总体来说，前者识别精度较低，但识别效率高；后者与其相反。

本文选用的特征选择算法如表 2 所示，FS_1~FS_5 是有代表性的输出特征权重的有监督特征选择算法，FS_6 是本文提出的综合特征排名算法，FS_7 是以选择最小特征组为输出的有监督特征选择算法。

FS_1~FS_5 特征选择算法，因其评价标准的专一性分别有其最佳适用范围，由于事先并不能预知哪个算法适合本文所涉及的应用环境，为此本文提出了综合特征排名算法 FS_6，基本思想是综合考虑每个特征在不同的特征选择算法中的贡献，将在各个选择算法结果中排名靠前的特征的权值加大，这样既克服了每个利用了特征选择算法因专一性而带来的缺点，又利用了其优点，其计算过程如下。

已知特征集 $F=\{F_1, F_2, \cdots, F_M\}$，设第 i 个特征选择算法的特征排名 $F^R{}_i=(F_{i,1}, F_{i,2}, \cdots, F_{i,M})(1\leq i\leq L, L$ 为特征选择算法数目)，在 L 个特征选择算法的结果排名中，将前 $k(1\leq k\leq M)$名的所有特征组成特

表 2　特征选择算法

编号	名称	分类	评价标准
FS_1	CHI(chi-squared)[41]	Filter	CHI-square
FS_2	IG(information gain)[42]	Filter	信息熵
FS_3	ReliefF[43]	Filter	欧拉距离
FS_4	SVM-RFE(recursive feature elimination for SVM)[44]	Wrapper	预测分析
FS_5	SU(symmetrical uncertainty)[45]	Filter	不确定分析
FS_6	CR(comprehensive ranking)	—	—
FS_7	CFS(correlation-based feature selection)[46]	Filter	相关分析

征集 $Top_k=\{F_{i,j}|1\le i\le L, 1\le j\le k\}$。算法如下。

1) 计算特征 F_j 在 Top_k 中的出现概率，公式为 $P_{Top_k}(F_j)=\dfrac{num(F_j)}{sizeof(Top_k)}$，$num(F_j)$为特征 F_j 在 Top_k 中出现次数，$sizeof(\cdot)$为 Top_k 中特征数目。

2) 计算特征 F_j 在 $Top_k(F)$中的出现概率的平均值，计算公式为 $\overline{P}(F_j)=\sum_{k=1}^{M}P_{Top_k}\dfrac{(F_j)}{M}$。

5.2　分类器

基于机器学习的分类检测是通过学习训练出一个分类模型，其将数据集中的样本映射到给定类别中的某一个类别。由于分类器对样本数量的敏感度、特征之间相关度的敏感度等均不相同，故选择不同的分类器得到的分类效果往往不同。本文使用了 10 个经典的分类器（如表 3 所示），包括了相关文献已验证识别效果最好的分类器。

6　实验与评估

6.1　数据标注

基于对多次实验结果的分析, 共随机抽取 4 300 个用户，进行人工标注，正常用户为 3 710 个(包括 199 个新浪微博身份实名认证的用户)，约占总数的 86.3%，垃圾用户为 590 个(约 13.7%)。在垃圾用户中，内容垃圾为 208 个(约 4.8%)；僵尸垃圾为 111 个(约 2.6%)；封号垃圾为 271 个(约 6.3%)。

6.2　实验设置

为了得到可信的结果，实验采用 10 折交叉验证方法[54]来验证分类性能, 将原来样本随机分成 10 等份互不相交的样本子集，每等份样本的类别比例近似等于总样本的，其中用 9 份样本子集作为训练集建立分类检测模型，而用剩下的 1 份样本子集作为验证集，然后交叉验证重复 10 次，使得每份样

表 3　分类器

编号	名称	说明
$Classfier_1$	Naive Bayes (NB)[47]	基于贝叶斯定理与特征之间独立假设基础之上，根据某对象的先验概率利用贝叶斯公式计算出其后验概率，选择具有最大后验概率的类作为该对象所属的类
$Classfier_2$	logistic regression (LR)[48]	使用逻辑回归 sigmod 函数来计算后验概率，根据后验概率对所给对象进行分类识别
$Classfier_3$	support vector machine (SVM)[49]	建立在统计学理论中的结构风险最小化准则基础上，原理是将低维空间的点映射到高维空间，使它们成为线性可分，再使用线性划分的原理来判断分类边界
$Classfier_4$	radial basis function network (RBFN)[50]	该方法是一种前馈神经网络，采用径向基函数作为激活函数
$Classfier_5$	*k*-nearest neighbor (IBk/kNN)[18]	一种基于实例学习的非参数估计的分类方法，计算新样本与训练样本之间的距离，找到距离最近的 *k* 个邻居，如果邻居的大多数属于某一个类别，则该样本也属于这个类别
$Classfier_6$	AdaBoost.M1 (ABM1)[19]	一种提高给定学习算法精度的方法，使用同一个训练集训练不同的弱分类器，然后把这些弱分类器集合起来，构成一个强的最终分类器
$Classfier_7$	bootstrap aggregating (BA)[51]	与 AdaBoost 一样，也是一种集成学习分类方法，但在训练集的选取和预测函数的生成方面存在明显差异，通常 AdaBoost 的分类准确度较 BA 的高，不过 BA 可以有效避免过拟合
$Classfier_8$	decision trees (J48/C4.5)[52]	一种简单且快速的非参数树状分类方法，利用信息增益率来选择特征，将信息增益率最大的特征作为决策树的分裂节点，每个分支均重复这一过程
$Classfier_9$	random forest (RF)[21]	以决策树为基本分类器的一个集成学习分类方法，它包含多个由 BA 集成学习技术训练得到的决策树，当输入待分类的样本时，最终的分类结果由单个决策树的输出结果投票决定
$Classfier_{10}$	logistic model trees (LMT)[53]	在决策树中引入了线性逻辑回归，节点包含逻辑回归函数

本都被验证一次。最终模型的预测分类性能评估指标就是这 10 次分类评估指标的平均值。

6.3 特征选择实验

设包含 M 个特征的集合为 F，C 为类别特征，数据集中正样本(正常用户)与负样本(垃圾用户)比例为δ，数据集记为 $D_\delta(F, C)$。

实验包括特征选择、用户分类检测和实验结果评估 3 部分。特征选择是采用不同的特征选择算法(FS)对数据集 $D_\delta(F, C)$进行计算，按照特征对分类的贡献计算出特征排名 F^R，或从 M 个特征中选出 $m(1\leq m\leq M)$个最佳特征子集 F^{best}。

分别将δ=1(共 1 184 条)和δ=5.9(共 4 101 条)样本数据输入不同的特征选择算法中，分别得到每个样本比例对应的特征选择结果。其中，δ取不同的值是为了考察不平衡数据集对特征选择结果的影响。

表 4 分别给出了 6 个经典特征选择算法的不同结果, 其中 CFS 方法计算出最小数目的特征子集(用来与第 6.4.2 节的实验结果对比分析)，而其他特征选择算法均给出了特征排名。表 5 给出了综合特征排名算法(CR)的结果。

6.4 分类检测实验

用户分类检测是将不同的分类器(*Classifier*)与不同的特征选择算法(FS)进行组合<*FS*, *Classifier*>对用户进行识别，也即将特征选择的结果作为分类器的输入，然后根据度量指标对分类结果进行评估分析，包括特征选择算法对分类器的影响、特征数目对分类效果的影响、样本数量对分类器的影响。

6.4.1 特征选择对分类器影响分析

分别将δ=1 和δ=5.9 的 6 个不同特征选择算法的结果与新浪微博中正常用户与垃圾用户真实比例δ=5.9 的 4 101 条样本数据输入至 10 个经典的分类器中，共计 120 组实验，然后记录每个实验的 6 个分类结果评价指标。本节使用准确率(*Acc*)来衡量分类器对整个样本的识别能力。由于不同分类检测实验结果的准确率之间的绝对差距不是很大，为了在图上将其显著区分开，引入了准确率之间的比率 $Ratio(Acc_i)=\dfrac{Acc_i}{\min(Acc_k)}$ $(k=1,\cdots,10)$，表示每组实验中每个实验结果的准确率与最小者的比值。

图 1 和图 2 分别给出δ=1 和δ=5.9 的同一个特征选择算法组合不同分类器的检测结果的准确率，从图中可知，无论是δ=1 还是δ=5.9，就单个特征选择算法而言，其与不同分类器组合后的分类效果之间存在一定差异，但差异非常微小，如前所述，为了使差异显著，图中纵轴采用了准确率之间的比率 *Ratio*；就所有的特征选择算法而言，分类器的性能较为稳定，一些分类器无论与哪个特征选择算法结合，其分类效果均表现出色。

表 4 特征选择实验结果

δ	方法	Top_1	Top_2	Top_3	Top_4	Top_5	Top_6	Top_7	Top_8	Top_9	Top_{10}	Top_{11}	Top_{12}	Top_{13}	Top_{14}	Top_{15}	Top_{16}	Top_{17}
5.9	ChiSq	F_5	F_{10}	F_1	F_{14}	F_{17}	F_{11}	F_4	F_9	F_3	F_{13}	F_{15}	F_6	F_7	F_{12}	F_{16}	F_8	F_2
	IG	F_5	F_{10}	F_1	F_{14}	F_{17}	F_{11}	F_4	F_9	F_3	F_{13}	F_7	F_{15}	F_{12}	F_6	F_{16}	F_8	F_2
	ReliefF	F_{10}	F_1	F_{14}	F_{11}	F_{17}	F_3	F_5	F_{13}	F_6	F_{15}	F_7	F_{12}	F_9	F_4	$F16$	F_8	F_2
	SVM-RFE(c=0.05)	F_5	F_4	F_3	F_1	F_7	F_{17}	F_{10}	F_{11}	F_9	F_6	F_{13}	F_{15}	F_{12}	F_8	F_{14}	F_{16}	F_2
	SU	F_5	F_1	F_{10}	F_4	F_{17}	F_{14}	F_9	F_{11}	F_{13}	F_3	F_7	F_{12}	F_{15}	F_{16}	F_6	F_8	F_2
	CFS	$\{F_5, F_{10}, F_{13}\}$																
1	ChiSq	F_{10}	F_1	F_5	F_{17}	F_{12}	F_{14}	F_{11}	F_{13}	F_9	F_3	F_4	F_7	F_{15}	F_{16}	F_6	F_8	F_2
	IG	F_5	F_{10}	F_1	F_{17}	F_{12}	F_{14}	F_{11}	F_{13}	F_9	F_3	F_4	F_7	F_{15}	F_6	F_{16}	F_8	F_2
	ReliefF	F_1	F_{10}	F_5	F_{14}	F_{11}	F_{17}	F_{12}	F_9	F_3	F_{13}	F_6	F_{16}	F_{15}	F_4	F_7	F_2	F_8
	SVM-RFE(c=0.01)	F_{10}	F_1	F_5	F_{11}	F_{17}	F_{12}	F_{14}	F_9	F_3	F_{13}	F_7	F_{15}	F_6	F_{16}	F_4	F_8	F_2
	SU	F_5	F_1	F_{10}	F_{12}	F_{17}	F_{13}	F_{11}	F_{14}	F_4	F_9	F_3	F_7	F_{15}	F_6	F_{16}	F_8	F_2
	CFS	$\{F_1, F_5, F_{10}, F_{12}, F_{17}\}$																

表 5 综合特征算法实验结果及特征在 Top_k 中出现的平均概率

δ	特征/概率	Top_1	Top_2	Top_3	Top_4	Top_5	Top_6	Top_7	Top_8	Top_9	Top_{10}	Top_{11}	Top_{12}	Top_{13}	Top_{14}	Top_{15}	Top_{16}	Top_{17}
5.9	CR	F_5	F_{10}	F_1	F_{17}	F_{14}	F_4	F_{11}	F_3	F_9	F_{13}	F_7	F_{15}	F_6	F_{12}	F_{16}	F_8	F_2
	概率	0.59	0.45	0.41	0.26	0.26	0.25	0.22	0.21	0.15	0.13	0.13	0.09	0.09	0.07	0.04	0.03	0.01
1	CR	F_{10}	F_1	F_5	F_{17}	F_{12}	F_{11}	F_{14}	F_{13}	F_9	F_3	F_4	F_7	F_{15}	F_6	F_{16}	F_8	F_2
	概率	0.55	0.51	0.51	0.28	0.26	0.24	0.23	0.16	0.15	0.13	0.09	0.08	0.07	0.06	0.05	0.02	0.01

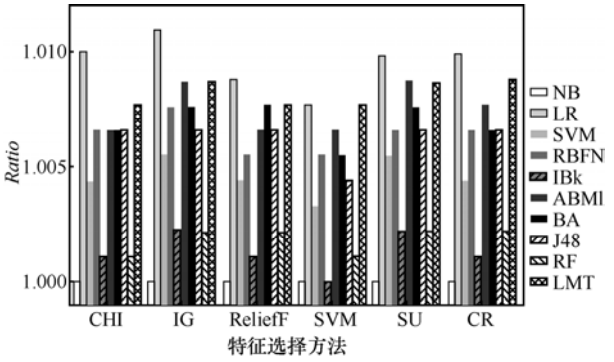

图 1　特征选择方法与分类器组合的检测性能(δ=1)

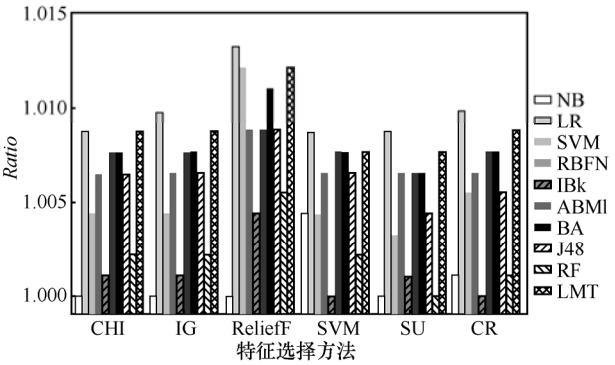

图 2　特征选择方法与分类器组合的检测性能(δ=5.9)

此外，就所有的特征选择算法而言，特征选择算法对分类器的支持在很大程度上具有稳定性。具体来说，任意给定一个特征选择算法 FS_x 和一个分类器 $Classifier_y$, 其组合的分类结果准确率 $Acc<FS_x, Classifier_y>$在 $Acc<FS_i, Classifier_y>(i=1,\cdots,6)$中的排名与对于某一分类器 $Classifier_j(j=1,\cdots,10$ 且 $j\neq y)Acc<FS_x, Classifier_j>$在 $Acc<FS_i, Classifier_j>(i=1,\cdots,6)$的排名基本一致。也即对于某个特征选择算法，其与某个分类器组合的准确率在该分类器与所有特征选择算法组合的准确率中的排名，大致可以代表该特征选择算法与其他任一分类器组合在该分类器与所有特征选择算法组合的准确率中的排名。直观而言，在图 2 中，ReliefF 与每个分类器组合的分类结果的准确率排名均靠前。这一现象表明，新浪微博中反垃圾分类检测效果在一定程度上依赖于特征组的选择。

在特征选择实验中δ=5.9, 其实验结果较δ=1 有明显差别，在δ=1 中，同一个分类器与不同特征选择算法组合的分类准确率较为接近，而在δ=5.9 中，差距较为明显，特别是与 ReliefF 特征选择算法组合的分类器的分类效果更为突出, 这表明在用户比例接近真实的环境下新浪微博中特征选择对分类器的影响较为明显。此外，在δ=1 中，分类效果整体上较好的排名前 3 位的特征选择算法分别是 IG、CR 和 SU，而在δ=5.9 中，排名前 3 位的特征选择算法分别是 ReliefF、CR 和 IG；无论δ=1 还是δ=5.9，分类效果整体上较好的排名前 2 位的均是 LR 和 LMT，排名第 3 位的分别为 ABM1(δ=1)和 BA(δ= 5.9)。该现象说明在不同的用户比例下特征的选择对分类器的影响不完全相同，另外，本文提出的 CR 方法排名第 2，虽不是最好的方法，但起到了均衡其他特征选择方法的效果。

总之，整体而言，对于新浪微博的垃圾用户检测，特征组的选择较分类器的选择更为重要，也即特征组的选取较分类器的改进更为重要。

6.4.2　特征数目对分类效果影响分析

探寻特征数目对分类效果的影响，是否存在最小特征数目。实验针对δ=5.9，选取排名前 3 的特征选择算法(分别为 ReliefF、CR 和 IG)与排名前 3 的分类器(LR、LMT 和 BA)进行组合，从而得到特征重要性排名及数目与准确率指标之间的关系，如图 3 所示，横坐标为特征的数目且根据特征对分类结果的贡献程度由大至小排列(在图中, 因在所有特征选择方法中，第 17 个特征相同，为了降低计算量，该特征没有参与分类计算), 纵坐标为 3 个不同分类器与不同的特征选择算法组合的准确率的平均值，及图中<ReliefF, .>表示 ReliefF 特征选择算法分别与分类器 LR、LMT 和 BA 组合的准确率的平均值。

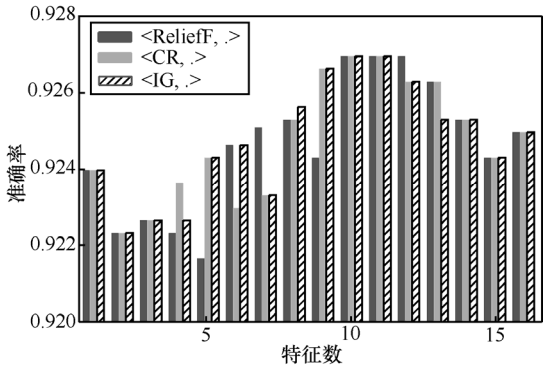

图 3　特征数目对准确率的影响(δ=5.9)

如果忽略局部的波动，总体来说，准确率随着特征数目的逐渐增加呈现抛物线形状，随着特征数目的增加准确率会逐渐升高, 达到峰值(从图 3 知特征数目为 10 个时准确率达到峰值)，然后下降。该结果表明在分类检测中仅有少数几个关键特征是不够的，只有特征数目达到一定的数量，准确率才能达到峰值；当然，过多的冗余特征又会导致准确

率的降低。此外，值得一提的是，此处的最小特征数目 10 个与第 6.3 节采用 CFS 算法选出的最小特征数目 5 个(如表 4 所示)所示不尽一致，需进一步讨论。

6.4.3 最佳特征来源分布分析

旨在分析最佳特征的来源分布。在 6.4.2 节实验中，对于δ=5.9 样本，取排名第 1 的特征选择算法(ReliefF)中的最佳特征子集 F^{best}={F_5, F_1, F_{10}, F_4, F_{17}, F_{14}, F_9, F_{11}, F_{13}, F_3}。其中{F_5, F_1, F_4, F_3}属于社会特征，{F_{10}, F_9}属于用户行为特征，{F_{17}, F_{14}, F_{11}, F_{13}}属于内容特征，也即最佳特征来源于内容信息、用户行为和社会关系 3 个方面。这一结果表明需要从多侧面生成特征，这将有助于提高识别准确率。

6.4.4 样本数量对分类效果影响分析

旨在分析样本数量对分类器性能的影响，掌握分类器性能收敛与样本数量之间的关系，并探寻实验中所需的最佳训练样本数量，进一步说明之前的实验对样本数量的假设是合理的。

根据 6.4.2 节实验，从所有样本集中随机抽取不同数量的样本，以 200 为步长使样本数量从 400 逐渐增加到 4 000。将不同的样本数目输入至不同的分类器中，进行分类检测实验，观察样本数量与准确率之间的变化关系。图 4 给出了 10 个分类检测算法准确率的统计曲线，从图中可以看出，虽然不同分类器的准确率各有差异，但是总体的趋势都是特征数目在 1 000 到 2 000 之间时准确率的变化发生由快到慢的转折。

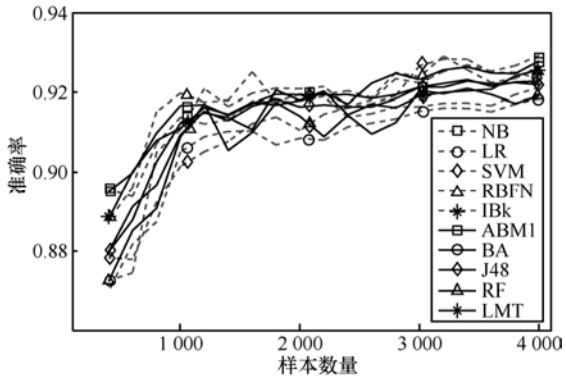

图 4 10 个分类检测算法的准确率

由于图中所有曲线的变化趋势相似，因此计算每个样本数目下 10 个分类检测算法准确率的平均值，如图 5 所示的 avgAcc 曲线，其拟合结果为 fitCurve 曲线。对于拟合曲线，当样本数目达到 3 000 以上时，只增大样本数目已经很难使分类器的准确率得到提高。这说明有关样本数量假设合理可行。

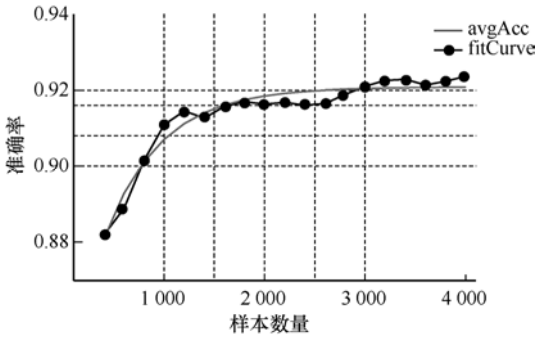

图 5 样本数量对准确率的影响

7 结束语

本文旨在回答在微博反垃圾中优先将研究重点投入到寻找分类特征还是改进分类方法。以新浪微博为例，实验结果表明特征组的选择较分类器的改进更为重要，需从内容信息、用户行为和社会关系多侧面定义特征，且特征并非越多检测效果越好。鉴于此，希望未来在特征的选取方面投入更多的工作，以便在反垃圾研究中有进一步的突破。尽管实验是以新浪微博为例展开，但其结果同样适用于腾讯微博、搜狐微博等微博的反垃圾。

参考文献：

[1] Available online[EB/OL]. http://news.xinhuanet.com/2013-07/04/c_116410610.htm.
[2] Available online[EB/OL]. http://it.people.com.cn/n/2015/0212/c1009-26552746.html.
[3] SPIRIN N, HAN J W. Survey on web spam detection: principles and algorithms[J]. ACM SIGKDD Explorations Newsletter, 2012, 13(2): 50-64.
[4] MUKHERJEE A, LIU B, GLANCE N S. Spotting fake reviewer groups in consumer reviews[C]//The WWW. c2012: 191-200.
[5] WANG T Y, WANG G, LI X. Characterizing and detecting malicious crowdsourcing[C]//The ACM SIGCOMM. c2013: 537-538.
[6] WANG G, WILSON C, ZHAO X H. Serf and turf: crowdturfing for fun and profit[C]//The WWW. c2012: 679-688.
[7] SRIDHARAN V, SHANKAR V, GUPTA M. Twitter games: how successful spammers pick targets[C]//The ACSAC. c2012: 389-398.
[8] STRINGHINI G, KRUEGEL C, VIGNA G. Detecting spammers on social networks[C]//The ACSAC. c2010: 1-9.
[9] IRANI D, WEBB S, PU C. Study of static classification of social spam profiles in MySpace[C]//The ICWSM. c2010: 82-89.
[10] GAO H Y, HU J, WILSON C. Detecting and characterizing social spam campaigns[C]//The CCS. c2010: 681-683.
[11] AGGARWAL A, ALMEIDA J M, KUMARAGURU P. Detection of spam tipping behaviour on foursquare[C]//The WWW. c2013: 641-648.
[12] GAO Q, ABEL F, HOUBEN G J. A comparative study of user's microblogging behavior on Sina weibo and Twitter[C]//The 20th International Conference on User Modeling. c2012: 88-101.
[13] YU L, ASUR S, HUBERMAN BA. What trends in Chinese social media[C]//SNA-KDD Workshop. c2011: 1-10.
[14] YU LL, ASUR S, HUBERMAN B A. Artificial inflation: the real story of trends and trend-setters in Sina weibo[C]//The International Con-

fernece on Social Computing. c2012: 514-519.
[15] 樊鹏翼, 王晖, 姜志宏,等. 微博网络测量研究[J]. 计算机研究与发展, 2012,49(4):691-699.
FAN P Y, WANG H, JIANG Z H, et al. Measurement of microblogging network[J]. Journal of Computer Research Development, 2012, 49(4):691-699.
[16] SHARMA P, BISWAS S. Identifying spam in Twitter trending topics. technical report[R]. USC(University of Southern California) Information Sciences Institute, 2011.1-4.
[17] BENEVENUTO F, MAGNO G, RODRIGUES T. Detecting spammers on Twitter[C]//The 7th Collaboration, Electronic messaging, Anti-Abuse and Spam Conference. c2010: 1-9.
[18] HASTIE T, TIBSHIRANI R. DISCRIMINANT adaptive nearest neighbor classification[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence. 1996, 18(6):607-616.
[19] FREUND Y, SCHAPIRE RE. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1):119-139.
[20] ORR M J L. Regularization in the selection of radial basis function centres[J]. Neural Computation, 1995, 7(3):606-623.
[21] HO T K. The random subspace method for constructing decision forests[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998, 20(8):832-844.
[22] MILLER Z, DICKINSON B, DEITRICK W, et al. Twitter spammer detection using data stream clustering[J]. Information Sciences, 2014, 260(1): 64-73.
[23] SHOBEIR F, JAMES F, MADHUSHDANA S, et al. Collective spammer detection in evolving multi-relation social networks[C]//The KDD. c2015: 1769-1778.
[24] WANG A H. Detecting spam bots in online social networking sites: a machine learning approach[C]//DBSec. c2010: 335-342.
[25] LEE K, CAVERLEE J, WEBB S. Uncovering social spammers: social honeypots+machine learning[C]//The SIGIR. c2010: 435-442.
[26] MARTINEZ R J, ARAUJO L. Detecting malicious tweets in trending topics using a statistical analysis of language[J]. Expert Systems with Applications, 2013 40(8): 2992-3000.
[27] ZHU Y, WANG X, ZHONG E H. Discovering spammers in social networks[C]//The AAAI. c2012: 1-7.
[28] HU X, TANG J L, GAO HJ, et al. Social spammer detection with sentiment information[C]//The ICDM. c2014: 180-189.
[29] TAN E, GUO L, CHEN S, et al. Unik: unsupervised social network spam detection[C]//The CIKM. c2013: 479-488.
[30] ZHANG X, ZHU S, LIANG W. Detecting spam and promoting campaigns in the twitter social network[C]//The ICDM. c2012: 1194-1199.
[31] SURENDRA S, AIXIN S. HSpam14: a collection of 14 million tweets for hashtag-oriented spam research[C]//The SIGIR. c2015: 9-13.
[32] YANG C, HARKREADER R C, ZHANG J. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter[C]//The WWW. c2012: 71-80.
[33] HU X, TANG J L, LIU H. Online social spammer detection[C]//The AAAI. c2014: 1-7.
[34] HU X, TANG J L, ZHANG Y C, et al. Social spammer detection in microblogging[C]//The IJCAI. c2013: 177-188.
[35] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[C]//The WWW. c2011: 675-684.
[36] RATKIEWICZ J, CONOVER M, MEISS M. Detecting and tracking political abuse in social media[C]//The ICWSM. c2011: 1-8.
[37] 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.
DING Z Y, ZHOU B, JIA Y, et al. Detecting spammers with a bidirectional vote algorithm based on statistical features in microblogs[J]. Journal of Computer Research and Development, 2013, 50(11): 2336-2348.
[38] HU X, TANG J L, ZHANG Y C, LIU H. Leveraging knowledge across media for spammer detection in microblogging[C]//The ACM SIGIR. c2014: 547-556.
[39] Available online[EB/OL]. http://ictclas.nlpir.org/.
[40] DASH M, LIU H. Feature selection for classifications[J]. Intelligent Data Analysis, 1997, 16(21):131-156.
[41] LIU H, SETIONO R. CHI2: feature selection and discretization of numeric attributes[C]//The ICTAI. c1995: 338-391.
[42] NOWOZIN S. Improved information gain estimates for decision tree induction[C]//ICML. c2012: 1-8.
[43] KONONENKO I. Estimating attributes: analysis and extensions of RELIEF[C]//The ECML-PKDD. c1994: 171-182.
[44] GUYON I,WESTON J, BARNHILL SMD. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1-3):389-422.
[45] STECK J B. Netpix: a method of feature selection leading to accurate sentiment-based classification models[D]. Central Connecticut State University, 2005.
[46] HALL M A. Correlation-based feature selection for discrete and numeric class machine learning[C]//The ICML. c2000: 359-366.
[47] JOHN GH, EDU S, LANGLEY P. Estimating continuous distributions in Bayesian classifiers[C]//The UAI. c1995: 338-345.
[48] KEERTHI S S, DUAN K, SHEVADE S K. A fast dual algorithm for kernel logistic regression[J]. Machine Learning, 2005, 61(1):151-165.
[49] CORTES C, VAPNIK V N. Support-vector networks[J]. Machine Learning, 1995, 20(3):273-297.
[50] ORR M J L. Regularization in the selection of radial basis function centres[J]. Neural Computation, 1995, 7(3):606-623.
[51] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140.
[52] QUINLAN J R. C4.5: programs for machine learning[M]. Morgan Kaufmann Publishers, San Mateo, California, 1993.
[53] LANDWEHR N, HALL M, FRANK E. Logistic model trees[J]. Machine Learning, 2005, 59(1):161-205.
[54] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//The IJCAI. c1995: 1137-1143.

作者简介：

张宇翔（1975-），男，山西五寨人，博士，中国民航大学副教授，主要研究方向为社会网络分析、推荐技术。

孙菀（1991-），女，山东烟台人，中国民航大学硕士生，主要研究方向为社会网络分析与推荐技术。

杨家海（1966-），男，浙江云和人，清华大学教授、博士生导师，主要研究方向为计算机网络管理与测量、云计算与大数据等。

周达磊（1992-），男，江苏连云港人，北京邮电大学硕士生，主要研究方向为网络分析。

孟祥飞（1993-），男，山西太原人，北京航空航天大学硕士生，主要研究方向为数据分析技术。

肖春景（1978-），女，河北唐山人，中国民航大学讲师，主要研究方向为数据挖掘与推荐系统。