基于微博文本的用户人格分析模型研究

舒晓敏,马晓宁

(中国民航大学 计算机科学与技术学院,天津 300300)

摘　要:传统的微博用户人格分析将人格分为五类,但未考虑人格类别之间潜在的关联性。为此基于多标签集成分类方法(RAkEL)进行改进,构建RAkEL-PA模型。RAkEL-PA模型使用标签集合中不同的随机子集训练相应的Label Powerset(LP)分类器,然后集成所有分类结果作为最终分类结果。在微博用户文本消息数据上进行实验,结果表明,RAkEL-PA模型的两个不同策略对用户人格分类准确率较高。RAkEL-PA模型充分考虑多个人格之间的相关性,以提高用户人格分类鲁棒性。

关键词:大五人格;人格分析;多标签学习;RAkEL-PA;微博文本

DOI:10.11907/rjdk.201356　　开放科学(资源服务)标识码(OSID):

中图分类号:TP303　　**文献标识码:**A　　**文章编号:**1672-7800(2020)011-0025-04

Research on User Personality Analysis Model Based on Weibo Text

SHU Xiao-min, MA Xiao-ning

(*College of Computer Science and Technology*, *Civil Aviation University of China*, *Tianjin* 300300, *China*)

Abstract: Traditional personality analysis of Weibo users divides personality into five categories without considering the potential correlation among personality categories. The multi-label ensemble classification method (RAkEL) is improved to construct the RAkEL-PA model. The RAkEL-PA model uses different random subsets in the label set to train the corresponding Label Powerset (LP) classifier, and then ensembles all the classification results as the final classification result. The effectiveness of RAkEL-PA in personality analysis has been verified experimentally on Weibo users' text messages. The experimental results show that the accuracies of the two different strategies of RAkEL-PA are higher for user personality classification. RAkEL-PA fully considers the correlation between multiple personalities and improves the robustness of user personality classification.

Key Words: big-five personality; personality analysis; multi-label learning; RAkEL-PA; Weibo text

0　引言

心理学把个体人格研究与社交网络结合,用社交网络中用户行为数据对用户人格进行分析与预测[1],如工作绩效预测[2]、青少年网络成瘾诱因分析[3]、抑郁症预测[4]、人格与情绪表达关系[5]等,价值巨大。

文献[6]统计地理位置、发布频率等移动互联网用户特征,将人格分类看作三分类和五分类问题实验;文献[6,7,8]分别采用新浪微博、Facebook、Twitter和YouTube数据集进行人格识别;文献[9,10]采用二进制粒子群算法和半监督算法建立社交网络用户人格分析模型;文献[11]将人格分类问题转化为二分类问题;Rosen等[12]针对用户个体网站内容分析用户人格;Ross等[13]通过研究用户数据得出外向型与组成成员个数关系密切。

以上方法都是将五维人格看作不相干任务执行,而事实上五个维度之间有一定关联[1,6-8,11,14]。本文通过对多标签集成方法—随机k标签集(Random k-LabELsets, RAkEL)[10]进行改进,构建基于微博文本的RAkEL-PA(RAkEL-Personality Analysis)模型,综合考虑五维人格相关性,弥补前人工作的空白。

1　研究流程

人格模型泛指大五人格模型(Big-Five Model),包括外向性(Extraversion, E)、神经质(Neuroticism, N)、宜人性

收稿日期:2020-04-11

基金项目:中央高校基本科研业务费专项资金项目(3122014C018);中国民航大学科研启动基金项目(09QD02X)

作者简介:舒晓敏(1992-),女,中国民航大学计算机科学与技术学院硕士研究生,研究方向为舆情分析、文本分析、机器学习;马晓宁(1979-),男,博士,中国民航大学计算机科学与技术学院副教授、硕士生导师,研究方向为信息安全、网络舆情分析、机器学习、文本分析。本文通讯作者:舒晓敏。

(Agreeableness,A)、责任型(Conscientiousness,C)和开放性(Openness,O)五个维度[1]。

本文研究流程:①获取数据:在微博上发放大五人格量表问卷,志愿者填写问卷以及微博 userID,采用 userID 通过爬虫获取志愿者微博文本数据;②特征提取:从微博文本中提取与人格相关度高的特征,创建人格分析模型的特征属性;③建立模型:构建 RAkEL-PA 模型;④评估模型:采用分类准确率 Accuracy 和损失函数 Hammingloss 两个指标进行评估。

2 RAkEL-PA 模型构建

2.1 数据获取

2.1.1 获取用户五维人格得分

在问卷星网站上制作大五人格量表[1]作为调查问卷。制作 5 个分量表,每个分量表包括 5 个选项(非常不符合、不太符合、不确定、比较符合、非常符合)12 个题目,分别记 2、4、6、8 和 10 分,其中有题目反向计分,满分为 100 分。将问卷发放到微博,志愿者填写问卷,根据得分标注用户五维人格标签。

2.1.2 微博用户数据获取及数据预处理

利用 userID 使用 Python 语言编写微博爬虫程序,爬取用户 3 个月微博文本数据。删除仅含图片、表情等无用数据。

2.2 特征提取

本文使用 CCPL 开发的中文心理分析系统 TextMind[14],产生已验证的 76 个微博文本特征[14],如表 1 所示。另外,表情符号更能反应用户情绪,所以本文统计微博消息中含有的表情符号,并统计每条消息的影响力,如表 2 所示。

表 1 用户微博文本内容特征

基于内容的特征	特征表示	个数
基于词性特征	冠词、动词、助动词、副词、介词、连接词、否定词、量词	8
基于人称代词特征	代名词、特定人称代名词、第一人称单数代名词、第一人称复数代名词、第二人称代名词、第三人称单数代名词、第三人称复数代名词、非特定人称代名词、第二人称复数代名词	9
基于标点符号特征	句号、逗号、冒号、分号、问号、感叹号、引号、括号、其它标点符号	9
基于时态特征	过去、现在、将来	3
基于情感特征	正向情绪词、负向情绪词、情感词数量	3
基于词类型特征	社会历程词、家庭词、朋友词、人类词、情感历程词、焦虑词、生气词、悲伤词、认知历程词、洞察词、因果词、差距词、确切词、限制词、包含词、排除词、感知历程词、视觉词、听觉词、感觉词、生理历程词、身体词、健康词、性词、摄食词、相对词、移动词、空间词、时间词、工作词、成就词、休闲词、金钱词、宗教词、死亡词、应和词、停顿赘词、填充赘词、心理词、关爱词	40
基于词数量特征	词数、数字比率、词长大于等于 6 的比率、数量	4

表 2 用户微博文本其它特征

其它特征	特征表示	个数
基于影响力特征	转发数、评论数、点赞数	3
基于表情符号特征	@、给你小心心、比心、击掌、爱你、微笑、难过、流泪	8

由于特征量化为数值后差异巨大,必须对其先归一化[11]。将每个特征进行[0,1]区间归一化,如公式(1)所示。

$$f^* = \frac{f - f_{\min}}{f_{\max} - f_{\min}} \tag{1}$$

其中,f和f^*分别为文本特征的原始值和归一化值,$f_{\min}$和$f_{\max}$分别为所有用户相应特征的最小值和最大值。

2.3 RAkEL-PA 模型构建

2.3.1 多标签分类方法

多标签学习方法主要有算法自适应和问题转换方法两种[15]。前者主要包括支持向量机[7]和多标签k近邻算法(ML-kNN)[16];后者主要有 Binary Relevance(BR)[5]和 Label Powerset(LP)。

2.3.2 基于微博文本的 RAkEL-PA 模型构建

LP 方法优点是考虑标签相关性,但也存在不足[17],因此将大量标签的集合随机分成很多小的标签集,采用 LP 为每个小标签集训练多标签分类器,将所有 LP 分类器决策集成得到 RAkEL 的最后结果。本文基于微博文本的用户人格分析模型,构建基于人格分析的不相交子集策略 RAkEL_d-PA 和基于人格分析的重叠子集策略 RAkEL_o-PA。

确定 RAkEL_d-PA 标签集k的大小,将标签集合L随机分成$m=\lceil M/k \rceil$个不相交的k标签集$R_j, j=1,2\cdots m$。用LP学习m个多标签分类器$h_j, j=1,2\cdots m$。每个分类器h_j学习一个单标签分类任务,包含训练集中所有R_j的子集类值。该策略中不同标签集中的标签不相交,所以标签数越多性能越好[18]。

RAkEL_d-PA 模型训练过程和分类过程分别如图 1 和图 2 所示。

RAkEL_o-PA 中L^k表示L中所有不同k标签集的集合。L^k大小由二项式系数$\left|L^k\right| = \binom{M}{k}$决定。与 RAkEL_d-PA 不同的是,已知标签集k的大小以及期望的分类器数量$m \leqslant \left|L^k\right|$,$\text{RAkEL}_\text{o}$-PA 通过从$L^k$随机采样选择$m$个$k$标签集$R_i, i=1,2\cdots m$。当$mk > M$时标签集会重叠。

在 RAkEL_o-PA 模型上训练过程和分类过程分别如图 3 和图 4 所示。

3 实验

3.1 实验数据集和特征提取

本文共收到 258 份问卷,经过筛选(如:每个问题答案相同)得到有效问卷 169 份。使用爬虫得到用户在微博上的文本消息。利用文心软件提取文本特征,如表 1 和表 2

所示,并进行归一化处理。标签数 M 为人格的五个维度。因此标签集界限是 2^5=32,而实际标签集数量范围为此边界的 5%~44%[17]。本文标注的标签集中有 8 种标签集出现次数最多,将集中 60% 的数据作为训练集,其余作为测试集。

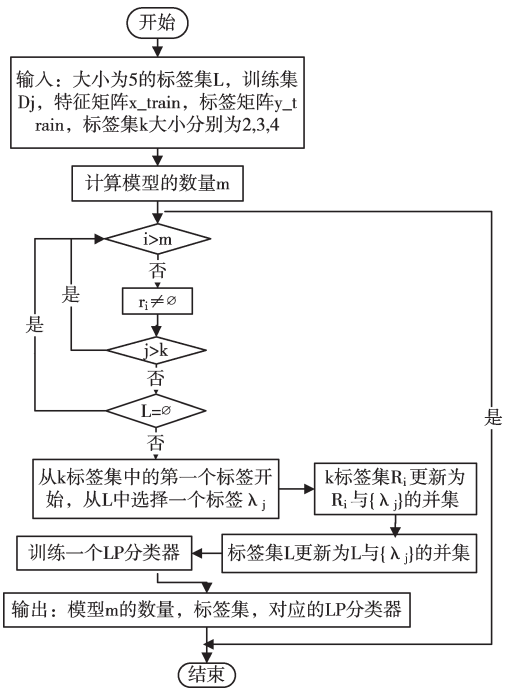

图 1　RAkEL$_d$-PA 模型训练流程

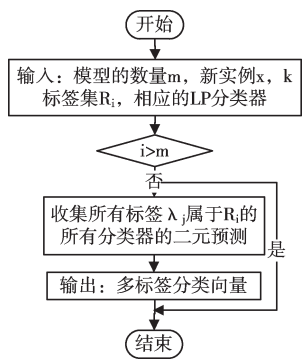

图 2　RAkEL$_d$-PA 模型分类流程

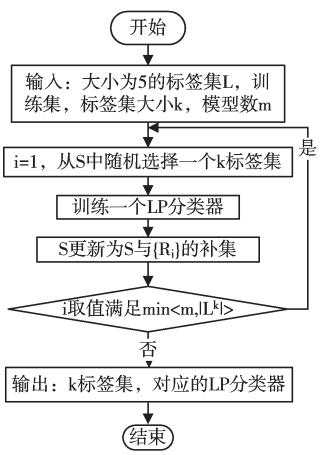

图 3　RAkEL$_o$-PA 模型训练流程

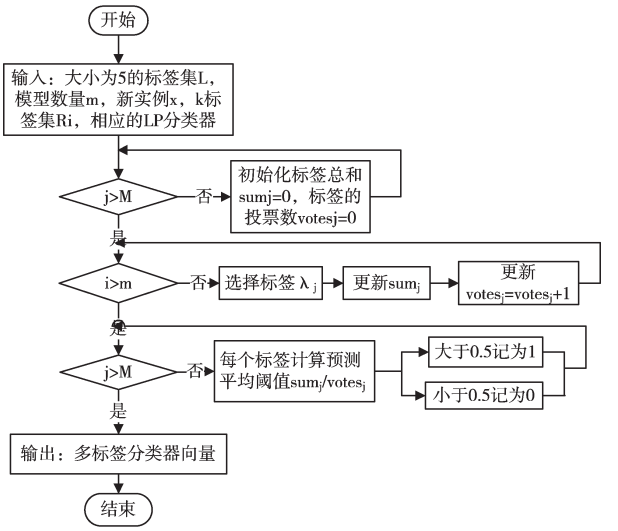

图 4　RAkEL$_o$-PA 模型分类流程

3.2　模型评价指标

本文使用分类准确度 Accuracy(A)和 Hammingloss(H)评估多标签分类效果。

用 D 表示一个多标签数据集,$|D|$ 表示样本个数,x_i 表示第 i 个样本,$y_i \subseteq L$ 表示 x_i 的标签集,i=1,2…$|D|$。本文通过学习一个多标签分类器 h 预测实例 x_i 的标签集 z_i,即 $z_i=h(x_i)$。

分类准确度(A)[18]定义如下:

$$Accuracy=\frac{1}{|D|}\sum_{i=1}^{|D|}I(Z_i=y_i) \tag{2}$$

当 $z_i=y_i$ 时,$I(z_i=y_i)=1$,否则 $I(z_i=y_i)=0$。

Zhang M L & Zhou Z H[16]提出采用 Hammingloss(H)评估实例标签被错误分类的次数,公式如下:

$$Hammingloss(h,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|y_i\triangle Z_i|}{|L|} \tag{3}$$

其中,△ 代表两集合的对称差异,$|L|$ 表示标签数量。

3.3　实验结果与分析

3.3.1　RAkEL$_d$-PA 模型实验结果分析

在 RAkEL$_d$-PA 实验中,标签集 k 取 2、3 和 4。k 值不同模型数 m 也不同。

如图 5 所示:k=2 时,模型的 A 值最高;k=3 和 k=4 时,A 值略低于 k=2 时,而 LP 的 A 值保持不变。原因是同时具有两种人格特质的人较多。随着 k 值增大,m 变小,参与训练的分类器个数变少,导致 RAkEL$_d$-PA 性能变差。

如图 6 所示:k=2 时,模型 H 值最小;k=3 和 k=4 时,H 值略高,可见随着 k 值增大,H 值也在变大,而 LP 的 H 值不变。该模型的 H 最大值和 LP 的 H 值接近,说明随着 k 值接近 M,模型性能与 LP 性能相当。

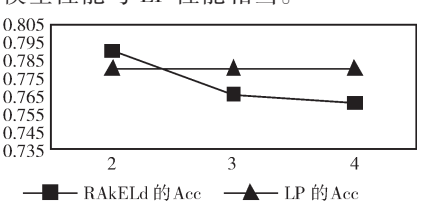

图 5　RAkEL$_d$-PA 模型中不同 k 值的 A 值与 LP 的 A 值对比

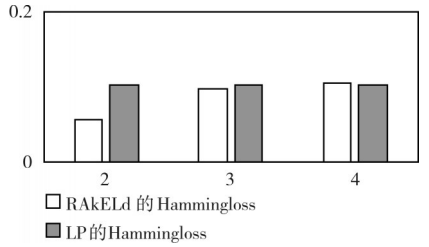

图 6 RAkEL$_d$-PA 模型中不同 k 值的 H 值与 LP 的 H 值对比

3.3.2 RAkEL$_o$-PA 模型实验结果分析

RAkEL$_o$-PA 模型使用 k(2~4)的所有有意义值进行实验。在 k=2 和 k=3 时,m 范围为 1~10,k=4 时,m 范围为 1~5。RAkEL$_o$-PA 模型的分类决策计算方式采用多数投票规则。

如图 7 所示:①k=2(同时具有两种人格特质)时,A 值在 m=8 时最高,与文献[9]得出的结论一致,即 A 和 C、C 和 E、C 和 O、O 和 E 分别具有很强的相关性;②k=3 时,A 值在 m=8 时最高,文献[11]也表明,C、A、E,E、C、O,O、A、C 分别有强相关性;③k=4 时,A 值在 m=4 时最高,与 k=2 和 k=3 相比,同时具有 4 种人格特质的人相对较少,所以 A 值略低于 k=2 和 k=3 时的 A 值,而 LP 的 A 值不随 m 和 k 的改变而改变。

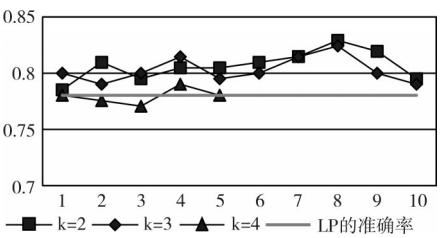

图 7 RAkEL$_o$-PA 模型中不同 m 和 k 的 A 值与 LP 的 A 值对比

如图 8 所示:随着 m 值增大,模型的 H 值在减小。k=2,m=7、8、9 时,H 值最小;k=3,m=8 时,H 值最小;k=4,m=3 时,H 值最小。LP 分类器的 H 值不随 m 和 k 的改变而改变。可以看出,模型的 H 值均比 LP 小,说明该模型性能比 LP 好。

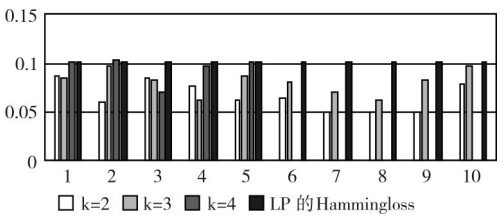

图 8 RAkEL$_o$-PA 模型中不同 k 值的 H 值与 LP 的 H 值对比

4 结语

针对传统人格分析方法未考虑五个人格维度之间的潜在相关性导致个体人格分类准确率较低问题,提出 RAkEL 的改进模型 RAkEL-PA 实现个体人格分类。实验结果表明,具有双重人格特质和三重人格特质的人较多,说明五维人格之间存在依赖性。该模型考虑了五维人格之间的相关性,提高了微博用户人格分类的准确率,从而验证了 RAkEL-PA 模型对人格分类的有效性。后续考虑获取更多微博用户数据,在更大数据集上进行实验,以进一步验证该模型的有效性。

参考文献:

[1] 张磊,陈贞翔,杨波.社交网络用户的人格分析与预测[J].计算机学报,2014,37(8):1877-1894.

[2] JUDGE T A, ZAPATA C P. The person - situation debate revisited: effect of situation strength and trait activation on the validity of the big five personality traits in predicting job performance[J]. Academy of Management Journal, 2015, 58(4): 1149-1179.

[3] ZHOU Y, LI D, LI X, et al. Big five personality and adolescent internet addiction: the mediating role of coping style[J]. Addictive behaviors, 2017, 64(8): 42-48.

[4] ALLEN T A, CAREY B E, MCBRIDE C, et al. Big five aspects of personality interact to predict depression[J]. Journal of personality, 2018, 86(4): 714-725.

[5] 刘真亦. 不同人格倾向微博用户的情绪表达分析[D].杭州:浙江大学,2019.

[6] 孙启翔.基于移动互联网社交行为的用户性格分析和预测[D].北京:北京理工大学,2016.

[7] FARNADI G, SITARAMAN G, SUSHMITA S, et al. Computational personality recognition in social media[J]. User Modeling and User-Adapted Interaction, 2016, 26(2-3): 109-142.

[8] 杨洁.基于用户情感和网络关系分析的人格预测模型[D].上海:东华大学,2016.

[9] 毛雨.基于社交网络的用户人格分析研究与实现[D].北京:北京邮电大学,2019.

[10] 郑赫慈.网络空间中人格分析的研究与实现[D].北京:北京邮电大学,2019.

[11] XUE D, HONG Z, GUO S, et al. Personality recognition on social media with label distribution learning[J]. IEEE Access, 2017, 5(142): 13478-13488.

[12] ROSEN P A, KLUEMEPER D H. The impact of the big five personality traits on the acceptance of social networking website[C]. AMCIS 2008 proceedings: AMCIS, 2008: 223-229.

[13] ROSS C, ORR E S, SISIC M, et al. Personality and motivations associated with facebook use [J]. Computers in Human Behavior, 2009, 25(2): 578-586.

[14] LIMA A C E S, DE CASTRO L N. A multi-label, semi-supervised classification approach applied to personality prediction in social media[J]. Neural Networks, 2014, 58(12): 122-130.

[15] BAI S, HAO B, LI A, et al. Predicting big five personality traits of microblog users[C]. Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01. IEEE Computer Society, 2013: 501-508.

[16] ZHANG M L, ZHOU Z H. ML-KNN: a lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[17] TSOUMAKAS G, KATAKIS I, VLAHAVAS I. Random k-labelsets for multilabel classification[J]. IEEE Transactions on Knowledge & Data Engineering, 2011, 23(7): 1079-1089.

[18] TSOUMAKAS G, VLAHAVAS I. Random k-labelsets: an ensemble method for multilabel classification[C]. European conference on machine learning, Springer, Berlin, Heidelberg, 2007: 406-417.

(责任编辑:杜能钢)