基于机器学习的宋词风格识别

赵建明[1],李春晖[2],姚念民[2]
ZHAO Jianming[1], LI Chunhui[2], YAO Nianmin[2]

1.福建师范大学 福清分校 电子与信息工程学院,福建 福清 350300
2.大连理工大学,辽宁 大连 116024
1.School of Electronic and Information Engineering, Fuqing Branch, Fujian Normal University, Fuqing, Fujian 350300, China
2.Dalian University of Technology, Dalian, Liaoning 116024, China

ZHAO Jianming, LI Chunhui, YAO Nianmin. Classification of SonCi style using machine learning algorithms. Computer Engineering and Applications, 2018, 54(1):186-190.

Abstract: This paper presents a study of the style of SonCi using many machine learning algorithms whose parameters are optimized by the results of experiments. At the same time, the reverse analyses is performed to get which words have the most effects on the decision making. This method can be used in analyzing the writing style of some author.
Key words: machine learning; Natural Language Processing(NLP); SonCi style

摘 要:使用多种机器学习算法对宋词的风格进行了分类研究,通过比较测试结果选择了较优算法和较优的参数配置。同时,对实验的结果进行了回溯分析,定量分析了哪些单字对宋词风格的判定起到更大的作用。这种分析方法可以推广,用来作为作者写作风格的特征进行更进一步的研究分析。
关键词:机器学习;自然语言处理;宋词风格
文献标志码:A **中图分类号:**TP183 doi:10.3778/j.issn.1002-8331.1607-0016

1 引言

宋词历来有婉约派和豪放派之分,相应的,词人也大致分为这两派,如辛弃疾、苏东坡等被认为是豪放派的代表,秦观、柳永等被认为是婉约派的代表。在悠长的历史中,每首宋词都有学者对之进行了分析和分类,甚至学者们对一些词的分类一直争论不休,这些研究凝聚了无数学者的辛勤劳动[1]。本文尝试使用机器学习算法对宋词进行分类。基于机器学习的文本分类方法可以对文本进行客观、可量化的评估,并进一步加深对宋词风格的主观分类的理解。

2 相关工作

对文本进行情感分析的研究是近年来的热点,如文献[2]和[3]。但使用基于机器学习的文本分类方法对宋词进行分类的研究不多,国内查到的基本都是重庆大学易勇[4-6]的研究。易勇使用的是朴素贝叶斯方法,并且使用信息增益和遗传算法来选择特征。他们对188首豪放风格的宋词和210首婉约风格的宋词进行了实验,最高准确率达到88.5%。该项研究所使用的语料库较小,这影响了其模型的泛化能力。另外,该项研究在研究宋词分类上仅使用了一种机器学习方法,参数的设置也较为单一。本文在此基础上使用了较大的语料库,并应用了多种机器学习方法,在参数的设置上也进行了多重比较和筛选。另外,本文在获得分类结果的基础上,对分类决策的影响因素进行了回溯分析,得出了哪些字在分类决策中起到更大的作用。该项研究可以进一步扩展,加深对宋词分类的理解。

3 文本分类工具介绍

Scikit-learn是基于Python的机器学习开源工具包。基本功能主要分为6部分:分类、回归、聚类、数据降维、模型选择、数据转换,以下简单介绍本文用到的数

作者简介:赵建明(1976—),男,副教授,主要从事机器学习研究;李春晖(1977—),女,讲师;姚念民(1974—),通讯作者,男,教授,主要从事自然语言处理研究,E-mail:lucos@dlut.edu.cn。
收稿日期:2016-07-04 **修回日期:**2016-09-02 **文章编号:**1002-8331(2018)01-0186-05
CNKI **网络优先出版:**2016-12-28, http://www.cnki.net/kcms/detail/11.2127.TP.20161228.0955.012.html

据转换、模型选择和分类功能。

Scikit-learn提供了TfidfVectorizer类用于文档向量化，它支持把字符串形式的文档转换成词袋(Bag of Words)形式的向量。当文档以词袋形式向量化时，不再保有文档中词语之间的顺序信息。

自然语言处理中，经常使用到n-gram模型，该模型基于这样一种假设，第n个词的出现只与前面$n-1$个词相关，而与其他任何词都不相关。在提取文档特征时，如果不考虑词之间的相关性($n=1$)，得到文档的unigram(1-gram)特征。如果提取相邻词语的特征($n=2$，当前词只和前一个词相关)，得到文档的bigram(2-gram)特征。unigram+bigram特征表示使用unigram和bigram的特征并集。

通常计算一个给定的词语在文档中出现的次数作为词袋向量中该词语的权重，称为词频(Term Frequency，TF)权重表示。如公式(1)所示，式中分子是该词在文件中的出现次数，而分母则是在文件中所有字词的出现次数之和。

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

词频通常会被归一化，以防止它偏向长的文件。归一化有两种方式，分别称为$L1$归一化和$L2$归一化。$L1$归一化是用词语出现次数除以文档中所有词语总数；$L2$归一化是用词语出现次数除以文档中各词语出现次数的平方和的平方根。

如果把词频向量中所有非零元素设置为1，即只考虑词语在该文档中是否出现，这里称为布尔权重表示。

在词频基础上乘以IDF反文档频率(Inverse Document Frequency)，称为TF-IDF权重表示，如公式(2)所示。某一特定词语的IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。其中$|D|$是语料库中的文件总数，分母是包含词条t_i的文档个数。IDF的主要思想是：如果包含词条t_i的文档越少，IDF越大，则说明词条t_i具有很好的类别区分能力。

$$idf_i=\lg\frac{|D|}{\left|\{j:t_i\in d_j\}\right|} \tag{2}$$

Scikit-learn提供了GridSearchCV类用于模型的参数搜索，GridSearchCV在数据集上产生K对训练/确认集，对给定参数，在K对训练集上分别训练并在确认集上进行检验，最后取K次实验的平均得分作为该模型在该参数上的得分估计。在完成参数空间搜索后，给出该模型在该数据集上的最优参数估计。

Scikit-learn提供了众多的分类模型供研究者用于解决问题，每个模型都有适合其应用的领域。可以划分为线性模型、支持向量机、最近邻分类、朴素贝叶斯、决策树、集成方法(ensemble methods)等几大类。

本文选择线性模型中随机梯度下降分类(SGDClassifier)[7-8]、逻辑回归分类(LogisticRegression)[9-11]，支持向量机中的LinearSVC和SVC[12-15]作为候选模型。

4　实验设计

本文实验由以下几个步骤组成：

(1)对下载到的《全宋词》(.txt格式)进行预处理，清除无关的内容，执行繁简转换，把每首词保存到一个单独的.txt文件中，以作者姓名、词牌名称、词ID(给每首词分配唯一ID，确保词的引用唯一)组合构成的文件名保存到数据处理目录下。经过处理共得到宋词18 987首。

(2)选择豪放派和婉约派的代表词人作品作为样本数据。为了简化数据处理，把公认为豪放派的作家，如苏轼、辛弃疾等的作品都作为豪放派词进行归类，实际上，这些作家的少量作品具有明显的婉约派风格，同样的，对于一些公认为婉约派的作家的作品，也做了统一分类处理。表1给出了样本数据的构成，从步骤(1)中得到的文件中按作者名称筛选出特定的词，共得到豪放派作品2 210首，婉约派作品1 812首。

表1　实验中使用样本的构成

豪放派	数量	婉约派	数量
辛弃疾	628	柳永	212
苏轼	349	晏殊	138
陈亮	74	晏几道	257
陆游	144	周邦彦	207
张孝祥	200	李清照	48
张元干	185	秦观	145
刘过	71	姜夔	87
王安石	27	吴文英	340
刘克庄	263	欧阳修	235
范仲淹	5	李之仪	143
陈与义	18		
叶梦得	103		
黄机	96		
戴复古	47		

(3)按照随机抽样方式从豪放派作品，婉约派作品中各抽取80%(共3 217首词)的文档作为训练集，剩余20%(共805首词)作为测试集。训练集用于训练及选择模型，测试集用于确认模型的有效性。

(4)使用基于现代汉语语料进行训练得到的分词模型在宋词上的分词效果不理想。考虑到宋词基本由单字词和少量双字词构成的情况，采用了基于单字的切分方案，把一个汉字作为一个词以提取单字词特征，结合bigram方案用于提取双字词特征。

(5)采用常用的词袋表示法把每首词转换成向量形式。

(6)在训练集上使用5折交叉确认[16]的方式，使用网格搜索功能分别搜索四种分类模型在该样本上的最优参数配置。表2给出了各分类模型的搜索参数空间。

表2　各分类模型的搜索参数空间

模型	中文名称	参数搜索设置
SGD	随机梯度下降	{'alpha':(0.0001,0.00001,0.000001),'penalty':('l1','l2','elasticnet')}
LR	逻辑回归	{'C':(1,10,100,1000),'penalty':('l1','l2')}
LinearSVC	线性支持向量机	{'C':(1,10,100,1000),'loss':('hinge','squared_hinge')}
SVC	支持向量机	{'kernel':['rbf','linear'],'C':[1,10,100,1000]}

(7)在测试集上验证步骤6得到的分类模型的有效性,输出其在测试集上的报告。

(8)分析模型的权值矩阵,找出分类背后的依据。

在Ubuntu系统下使用Anaconda Python完成该实验,考虑到每首词的词汇数量较少,做了两组对比实验:第一组只提取unigram(1-gram)特征;第二组提取unigram(1-gram)和bigram(2-gram)特征,并分别使用BOOL(布尔)、TF with $L1$ Normalization(词频并进行$L1$归一化)、TF-IDF with $L2$ Normalization(词频-逆文档词频并进行$L2$归一化)3种权重表示方式进行文档向量化。使用scikit-learn的网格搜索功能在训练集上对表2给出的4个分类模型的不同参数进行五折交叉确认,以确定最优模型及相关参数。

图1、图2给出了两组实验的结果,从图中可以很清楚的看到,提取unigram+bigram特征对分类精确度提高具有比较明显的作用。图2中,逻辑回归模型和线性支持向量机模型都取得了比较好的结果,在训练集上的分类精确度都达到81.4%。选择unigram+bigram特征,使用TF-IDF with $L2$ Normalization权重表示方式进行文档向量化,逻辑回归模型通过网格搜索得到的最优参数设置为{'penalty':'l2','C':100}。

选择分类效果最好的LR模型作为分类模型,在测试集上对该模型进行了验证,表3给出了LR模型在测试集上的测试结果。其中*Precision*(精确度)=$TP/(TP+FP)$,反映了被分类器判定的正例中真正的正例样本的比重,*Recall*(召回率)=$TP/(TP+FN)$,反映了被正确判定的正例占总的正例的比重,$F1$($F1$值)=$2\times Precision\times Recall/(Precision+Recall)$;$TP$表示被系统识别为该类的样本中有$TP$个是正确的(即真正的样本数量),$FP$表示被系统识别为该类的样本中有$FP$个是错误的(即假正的样本数量),$FN$表示有$FN$个属于该类的样本没有被系统正确识别;support表示从多少个样本中统计得到的数据;avg为加权均值。

表3　逻辑回归模型在测试集上的结果

流派	*Precision*	*Recall*	*F1-score*	support
婉约派	0.78	0.8	0.79	363
豪放派	0.83	0.82	0.83	442
avg/total	0.81	0.81	0.81	805

从表3可以看出,本文模型在测试集上的平均精确度、召回率和$F1$值都为0.81。其中,精确度值接近于其在训练集上通过交叉确认得到的分类精确度,说明该模型具有很强的泛化能力。

5　回溯分析

为了对模型的分类机制有进一步了解,用训练好的模型对岳飞和李清照的词进行预测,下载到的全宋词中岳飞的词有3首,李清照的词48首。表4、表5分别给出了逻辑回归模型在两位词人作品集上实验结果,由于篇幅原因,表5只给出部分代表作及全部预测错误的实验数据,48首李清照词人作品有3首预测错误。

表4　逻辑回归模型在岳飞词集上的预测结果

词名	决策值	婉约派概率	豪放派概率	分类结果
小重山	0.865 2	0.296 3	0.703 7	豪放派
满江红·登黄鹤楼有感	3.309 6	0.035 2	0.964 8	豪放派
满江红·写怀	3.42 81	0.031 4	0.968 6	豪放派

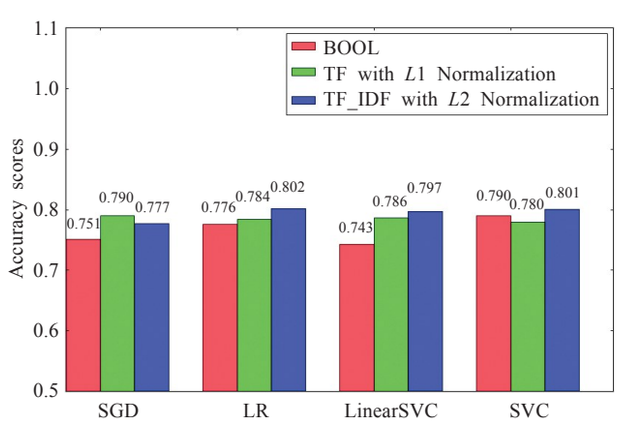

图1　Unigram特征下各分类模型交叉确认最高精确度得分

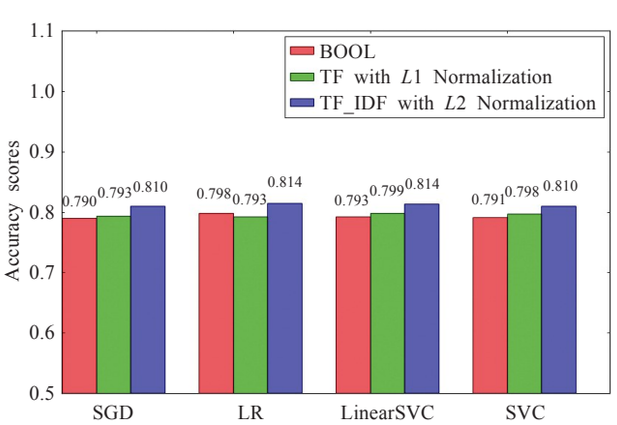

图2　Unigram+bigram特征下各分类模型交叉确认最高精确度得分

表5　逻辑回归模型在李清照词集上的部分预测结果

词名	决策值	婉约派概率	豪放派概率	分类结果
一剪梅	−2.375 1	0.914 9	0.085 1	婉约派
醉花阴	−2.197 3	0.900 0	0.100 0	婉约派
点绛唇·闺思	−2.143 5	0.895 1	0.104 9	婉约派
浣溪沙·小院闲窗春色深	−1.581 0	0.829 4	0.170 6	婉约派
凤凰台上忆吹箫	−1.449 1	0.809 9	0.190 1	婉约派
永遇乐·落日熔金	−1.023 4	0.735 6	0.264 4	婉约派
减字木兰花	0.187 1	0.453 4	0.546 6	豪放派
行香子	0.242 8	0.439 6	0.560 4	豪放派
渔家傲·天接云涛连晓雾	2.455 6	0.079 0	0.921 0	豪放派

这个结果符合主观预期。比如岳飞《满江红写怀》一词,充满了豪情壮志,读来让人热血沸腾,而《小重山》一词虽然大体上属于豪放风格,但明显比其他两首词内敛,又如李清照的《渔家傲-天接云涛连晓雾》实际属于豪放词作品。那么哪些字或词对宋词的风格起到很大作用呢?下面进行回溯分析。

对于惩罚项为*L*2('penalty':'l2')的二类逻辑回归模型,其目标函数为 $\min_{w}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+C\sum_{i=1}^{l}\ln(1+\mathrm{e}^{-y_i\boldsymbol{w}^{\mathrm{T}}x_i})$,其中 $\min_{w}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ 为 *L*2 规则化项,用于防止模型过拟合和加速模型收敛速度,$\sum_{i=1}^{l}\ln(1+\mathrm{e}^{-y_i\boldsymbol{w}^{\mathrm{T}}x_i})$ 为损失项,$C>0$ 为权重系数。

对于两类(C_1,C_2)问题,给定两个类的样本$\chi=\{x_i,y_i\}$,$i=1,2,\cdots,l,x_i\in R^n,y_i\in\{-1,+1\}$,其中$x_i\in C_1$则$y_i=1$,如果 $x_i\in C_2$ 则 $y_i=-1$。使用样本训练得到矩阵参数 $\boldsymbol{w}$ 后,即可用于分类,$\boldsymbol{w}^{\mathrm{T}}x_i$ 即决策值,当 $\boldsymbol{w}^{\mathrm{T}}x_i>0$ 时 $x_i\in C_1$,否则 $x_i\in C_2$,$\boldsymbol{w}^{\mathrm{T}}x_i$ 越大,$x_i\in C_1$ 的概率就越大,$\boldsymbol{w}^{\mathrm{T}}x_i$ 越小,$x_i\in C_2$ 的概率越大。

对于两类问题,$\boldsymbol{w}$ 是 $1\times n$ 维的矩阵,n 是文档向量的维度,x_i 是输入的文档向量,把决策值展开,有公式(3):

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i=\sum_{j=1}^{n}\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_{i,j}\tag{3}$$

即决策值是由文档向量化后的 x_i 第 j 个分量和权值矩阵 $\boldsymbol{w}$ 的第 j 个分量的乘积累加和组成。根据上述讨论,分别计算 x_i 每个分量 $x_{i,j}$ 对决策值的贡献 $\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_{i,j}$,并按贡献大小进行排序。训练后的权值矩阵 $\boldsymbol{w}$ 最大的10个分量对应的词为山、君、公、却、然、发、子、作、老、笑,最小的10个分量对应的词为红、金、情、渐、相、心、远、深、筵、杳。

表6、表7分别给出了判定为豪放派、婉约派贡献最

表6　判定为豪放派贡献最大的前10个词

词名	小重山(岳飞)									
特征词	山	白	明	老	三	自	功名	松	行	已
贡献	0.265 8	0.216 7	0.201 6	0.191 7	0.173 6	0.161 1	0.157 4	0.149 8	0.145 0	0.143 9
词名	满江红·登黄鹤楼有感(岳飞)									
特征词	山	骑	龙	却	洛	作	万	钱	里	续
贡献	0.432 8	0.268 1	0.184 1	0.141 2	0.131 7	0.126 6	0.104 0	0.103 4	0.100 0	0.093 6
词名	满江红·写怀(岳飞)									
特征词	山	发	壮	白	笑	栏	子	从	头	三
贡献	0.429 3	0.185 4	0.176 3	0.175 0	0.165 5	0.159 9	0.149 4	0.147 8	0.142 6	0.140 2

表7　判定为婉约派贡献最大的前10个词

词名	一剪梅(李清照)									
特征词	红	情	心	雁	相	罗裳	残	秋	心头	轻
贡献	−0.208 28	−0.175 6	−0.155 9	−0.145 22	−0.141 51	−0.128 16	−0.109 59	−0.106 05	−0.104 18	−0.098 68
词名	醉花阴(李清照)									
特征词	金	暗	魂	把酒	黄昏	消	袖	夜	帘	消魂
贡献	−0.201 32	−0.184 05	−0.143 02	−0.128−61	−0.123 13	−0.109 74	−0.107 17	−0.105 01	−0.103 45	−0.098 44
词名	点绛唇·闺思(李清照)									
特征词	寸	情	阑	深	一寸	惜	绪	情绪	倚	闺
贡献	−0.257 32	−0.224 87	−0.199 27	−0.192 68	−0.187 90	−0.172 11	−0.170 96	−0.162 87	−0.127 45	−0.126 47
词名	浣溪沙·小院闲窗春色深(李清照)									
特征词	沈	远	深	暮	恐	倚	轻	谢	帘	闲
贡献	−0.439 53	−0.244 24	−0.218 84	−0.193 49	−0.153 15	−0.144 76	−0.143 53	−0.141 42	−0.140 72	−0.139 07
词名	凤凰台上忆吹箫(李清照)									
特征词	念	眸	红	闲	金	暗	楼	锁	掩	凝眸
贡献	−0.197 50	−0.185 70	−0.158 62	−0.145 64	−0.143 40	−0.131 09	−0.120 26	−0.109 16	−0.104 22	−0.102 93
词名	永遇乐·落日熔金(李清照)									
特征词	金	相	柳	暮	侣	意	染	闺	夜	谢
贡献	−0.271 66	−0.102 08	−0.101 37	−0.095 97	−0.084 53	−0.082 63	−0.080 43	−0.071 25	−0.070 85	−0.070 14

大的前10个词。

从表6中可以看出,排在前面的单字确实从主观上带来豪放派的风格,3首词排在最前面的都是“山”字,由此可以得出结论,岳飞在写词的时候比较喜欢用“山”字来衬托其豪迈的性格。从表7中可以看出,李清照在写词的时候比较喜欢用“红”、“情”、“金”这些词。通过回溯分析,加深了对这些词的理解,并且可以做量化分析,目前还未发现相同的研究工作。

6 结论与展望

本文使用多种机器学习算法对宋词进行了分类研究,得到了较为满意的结果。与同类工作相比,使用的算法较多,并且进行了对比,通过比较测试结果选择了较优算法和较优的参数配置;另外,本文的训练集和测试集范围更大,因此所得的结果也更可信。

同时,对实验的结果进行了回溯分析,定量的分析了哪些单字对宋词风格的判定起到更大的作用。这种分析方法可以推广,用来作为作者写作风格的特征进行更进一步的研究分析。

在实验中,为了简化数据处理,按作者对数据进行了分类处理,这样的处理必然会在训练中引入一些噪声,如果对训练集进行人工筛选,相信本文模型的准确率会有进一步的提升

参考文献:

[1] 廖霞.词学之“婉约”与“豪放”范畴论[D].长沙:中南大学,2009-05-22.

[2] 黄伟,范磊.基于多分类器投票集成的半监督情感分类方法研[J].中文信息学报,2016,30(2):41-49.

[3] 黄仁,张卫.基于word2vec的互联网商品评论情感倾向研究[J].计算机科学,2016,43(S1):387-389.

[4] 易勇,何中市,李良炎,等.基于遗传算法改进诗词风格判别的研究[J].计算机科学,2005,32(7):156-158.

[5] 易勇,何中市,李良炎,等.单字词与古典诗词风格关系的研究[C]//第6届汉语词汇语义学研讨会论文集,2005.

[6] 易勇.计算机辅助诗词创作中的风格辨析及联语应对研究[D].重庆:重庆大学,2005-06-08.

[7] Zhang T.Solving large scale linear prediction problems using stochastic gradient descent algorithms[C]//Proceedings of International Conference on Machine Learning,2004:116-224.

[8] Keuper J,Pfreundt F.Asynchronous parallel stochastic gradient descent:A numeric core for scalable distributed machine learning algorithms[C]//Proceedings of Workshop on Machine Learning in High Performance Computing Environments,2015:1-11.

[9] Fan R,Chang K,Hsieh C.LIBLINEAR:A library for large linear classification[J].Journal of Machine Learning Research,2008(9):1871-1874.

[10] Chen Y,Lu B,Zhao H,et al.Parallel learning of large-scale multi-label classification problems with min-max modular LIBLINEAR[C]//Proceedings of International Symposium on Neural Networks,2012:1-7.

[11] Schmidt M,Roux N L,Bach F,et al.Minimizing finite sums with the stochastic average gradient[J].Mathematical Programming,2013,162(5):1-30.

[12] Cortes C,Vapnik V.Support-vector networks[J].Machine Learning,1995,20(3):273-297.

[13] Smola A,Scholkopf B.A tutorial on support vector regression[J].Statistics and Computing,2004,14(3):199-222.

[14] Shalevshwartz S,Singer Y,Srebro N,et al.Pegasos:Primal estimated sub-gradient solver for SVM[J].Mathematical Programming,2011,127(1):3-30.

[15] Shao Y,Chen W,Wang Z,et al.Weighted linear loss twin support vector machine for large-scale classification[J].Knowledge Based Systems,2015,73(1):276-288.

[16] Kohavi R.A study of cross-validation and bootstrap for accuracy estimation and model selection[J].International Joint Conference on Artificial Intelligence,1995,14:1137-1143.