现当代文学作品的作者身份识别研究

年洪东,陈小荷,王东波
NIAN Hong-dong,CHEN Xiao-he,WANG Dong-bo

南京师范大学 文学院,南京 210097
School of Chinese Language and Literature,Nanjing Normal University,Nanjing 210097,China
E-mail:nianhongdong@163.com

NIAN Hong-dong,CHEN Xiao-he,WANG Dong-bo.Research on authorship attribution of contemporary literature.Computer Engineering and Applications,2010,46(4):226-229.

Abstract: This paper uses the statistical model (SVM) for the identification of the author of contemporary Chinese literature works to eight representatives.In the identification process to select a vocabulary based on a variety of statistics as identifying features,and to take training methods based on the low-density and more features,having achieved better result in cross-style works of the author identification.

Key words: authorship attribution;machine learning;computational stylistics;contemporary literature

摘 要:主要利用了SVM统计机器学习模型对中国现当代文学八位代表人物的作品进行了作者身份识别研究,在识别过程中选取了以词汇为基础的多种统计量作为识别特征,并且采取了基于低密度多特征的训练方法,在跨文体的作品的作者身份识别中取得了非常优异的识别性能。

关键词:作者身份识别;机器学习;计算风格学;现当代文学
DOI:10.3778/j.issn.1002-8331.2010.04.071 **文章编号**:1002-8331(2010)04-0226-04 **文献标识码**:A **中图分类号**:TP391.1

1 引言

我国是世界上文学遗产最丰富的国家之一,但由于各种各样的原因,许多传世之作的作者身份不能十分准确确定,而传统的文献考证手段既费时又费力,从而在信息时代需要一种新的方法和技术来迎接这一挑战。国外早在20世纪30年代就开始引入统计学中的定量分析方法来分析作家作品的文体风格[1],这也促使一个新兴的学科——计算风格学的诞生。最初,学者们是用简单的手工统计方式来完成对作品的定量统计的,直到发明了计算机,大规模的文学作品精确定量统计分析才成为可能[2],这时一些国外学者开始利用计算机统计方法去研究一些经典书籍的作者身份(著作权)和写作年代的确定[3]。近些年来由于统计机器学习方法的广泛应用,很多国外的学者将一些成熟的机器学习模型应用到了作者身份的识别领域当中,并且取得了比较好的识别效果[4-6]。将计算统计方法应用于汉语语言计算风格学研究最早始于20世纪七八十年代,1980年6月美国威斯康星大学陈炳藻在首届国际《红楼梦》研讨会上发表的《从词汇上的统计论<红楼梦>作者的问题》的论文。他在文中首次运用计算风格学的方法分析了该书前80回与后40回的用词特点,认为两者风格一致,从而得出了120回均系曹雪芹一人所作的结论[7]。日本学者金明哲在其论文《中文文章的作者识别》对中文短篇文章分别进行了以字符为单位和以单词词性为单位的n-gram数据分析。从聚类分析结果的对比中得出以下结论:(1)字符为单位的unigram可以以较高的正判别率判别作者。在没有筛选变量的前提下以大约98%的判别率判别1 000字的文章的作者。(2)词性为单位的n-gram也含有作者特征。其正判别率可达95%,略低于以字符为单位的unigram的正判别率[8]。武晓春等利用HowNet知识库,提出一种新的基于词汇语义分析的相似度评估方法,有效利用了功能词以外的其他词汇,对5位《人民日报》记者发表文章进行训练和测试,最好成绩宏平均F值达86.23%[9]。

作者身份识别的关键问题是从其已知作品中统计出能代表其独特风格的识别特征,如:词汇总量及其特色词汇构成的数量比例、标点符号的使用频率、词语频率、句子长度、句式的使用分布、辞格的运用、声调和韵律分布等[10]。其中标点符号的使用频率、特色词汇分布和句子长度受到了广泛的认同。但是作者的作品风格可能因为不同的文体和话题而产生相应的改变,也可能因为作者随着年龄的增长,其写作技巧变得日臻成熟,因此只利用上面所提到的识别特征很难对不同作家的作品进行有效的区分。总之,一个理想的作品风格识别模型应该能够排除这些"噪声"并且利用学习到的有效的特征来区分不同作者的风格。

该文使用了SVM统计机器学习模型,将作者身份的识别

基金项目:国家社会科学基金项目(the National Social Science Fundation of China under Grant No.07BYY050)。
作者简介:年洪东(1978-),硕士研究生,研究方向为计算语言学;陈小荷(1952-),教授,博导,研究方向为计算语言学;王东波(1982-),硕士,研究方向为计算语言学。
收稿日期:2008-09-18 修回日期:2008-12-05

问题看成了不同领域的文本分类问题，这样不同作者的风格包括词汇、标点、句子、句式和修辞方面等都可以转化为以词汇为统计基础的概率值融入到机器学习的模型当中，并且通过机器学习实现对不同作者身份的识别。

2　SVM 模型简介

SVM[1]（Support Vector Machine）即支持向量机是在统计学习理论（Statistical Learning Theory）的基础上发展而来的一种机器学习方法，它基于结构风险最小化原理，将原始数据集合压缩到支持向量集合（通常为前者的 3%~5%），学习到分类决策函数。其基本思想是构造一个超平面作为决策平面，使正负模式之间的空白最大。支持向量机在解决小样本、非线性及高维模式识别问题中表现出了许多优势，并在很多领域得到了成功的应用，如：人脸识别、手写体识别、文本分类等领域。在文本分类方面 SVM 的表现尤为突出，其分类的查全率几乎超过了现有的所有方法。

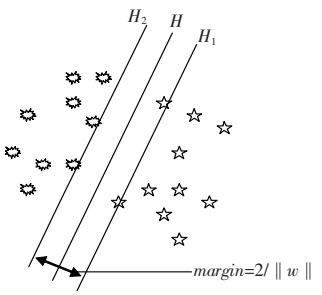

图 1　SVM 分类原理图

SVM 的基本思想可用图 1 的两维情况说明，图 1 中，五角形和多边形代表两类样本，H 为分类线，H_1、H_2 分别为过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫分类间隔。所谓最优分类线就是要求分类线不但要将两类正确分开（训练错误率为 0），而且使分类间隔最大。分类线方程为 $x*w+b=0$，可以对它进行归一化，使得对线性可分的样本集 (x_i,y)，$i=1,2,\cdots,n$，$x\in R_d$，$y\in\{+1,-1\}$，满足

$$y_i[(w*x)+b]-1\geqslant 0,\ i=1,2,\cdots,n \qquad (1)$$

此时分类间隔等于 $2/\|w\|$，使间隔最大等价于使 $\|w\|^2$ 最小。满足公式（1）且使 $1/2\|w\|$ 最大的分类面就叫做最优分类面，H_1 和 H_2 上的训练样本点就称作支持向量。

基本的 SVM 是针对两类分类问题的，为了实现对多个类别的识别，需要对 SVM 进行扩展。常用的 SVM 多类分类方法有 One-vs-Rest、One-vs-One、ECOC（Error-Correcting Output Coding）和二叉树等方法[11]。实验主要利用一对多（One-vs-Rest）的多分类方法扩展了 SVM 机器学习模型，根据文中的任务对每次识别实验分别训练 8 个不同的 SVM 识别模型，在识别结果中选取识别概率最大的类别作为最终识别结果。

3　特征选择及识别实验

在实验中，选取了现当代文学的 8 位代表作家的作品作为实验语料（经过分词处理的纯文本格式），这 8 位作家分别是鲁迅、沈从文、老舍、余秋雨、贾平凹、路遥、余华、王小波，他们的作品具体信息见表 1。

表 1　作家作品实验语料信息表

作家	作品	语料大小/MB	体裁
鲁迅	《坟》、《热风》、《呐喊》、《彷徨》、《野草》、《朝花夕拾》、《故事新编》、《华盖集》、《华盖集续编》、《而已集》、《三闲集》、《二心集》、《南腔北调集》	2.36	杂文，小说，散文，散文诗
沈从文	《沈从文文集：第一卷》[2]，《沈从文文集：第三卷》、《沈从文文集：第五卷》、《沈从文文集：第九卷》	2.84	小说，散文
老舍	《骆驼祥子》、《蜕》、《二马》、《火葬》、《离婚》、《茶馆》、《龙须沟》、《残雾》、《春华秋实》、《方珍珠》、《柳树井》、《女店员》、《全家福》、《西望长安》	2.94	小说，戏剧
余秋雨	《文化苦旅》、《霜冷长河》《山居笔记》、《借我一生》《行者无疆》、《千年一叹》	1.82	散文
贾平凹	《废都》、《秦腔》、《怀念狼》、《浮躁》	2.36	小说
路遥	《平凡的世界》、《人生》、《你怎么也想不到》、《黄叶在秋风中飘落》、《惊心动魄的一幕》、《在困难的日子里》、《我和五叔的六次相遇》	2.28	小说
余华	《余华散文随笔集》[3]、《高潮》、《河边错误》、《现实一种》、《夏季台风》、《活着》、《许三观卖血记》、《在细雨中呼喊》、《兄弟上，下》	1.75	散文，小说
王小波	《白银时代》、《黄金时代》、《革命时期》、《红拂夜奔》、《万寿寺》、《王小波-2015》、《未来世界》、《我的阴阳两界》、《寻找无双》	1.20	小说

要根据作品来识别出其作者首先要确定所选取的识别特征。文学是语言的艺术，语言是文学的第一要素，是构成文学作品的“物质材料”，人们常说某某作家的作品“犀利”、“睿智”，其实正是通过其作品的字里行间所表现出来的独特风格来确定的，这也就是文学理论上常说的“文如其人”和“风格即人”的具体表现。具体地来说，某种特定的风格的物质承载者就是作家所运用的语言（包括词汇、句式、辞格等），因而，在实验里选取的基本特征单位是词而不是字，虽然写作时用字也可能区别不同的作家，但一般来说，字的数量是有限的，而作家使用的词汇数量有明显的不同，从修辞学角度来说，文学作品所展现的“雄壮”、“柔婉”等风格正是由于作品中包含不同概率的特定词汇造成的。其他影响作品风格的语言特点如：句式、音律、辞格等等，也都能够体现出作者写作的风格特色，这些都可以通过转化为基于词语一些常用的统计量来作为识别作者作品风格的量的标准。

实验中主要使用了 4 个常用的统计量：

（1）互信息

互信息是用来统计两个变量间的相关性，其值越大，越说明其两个变量的相关程度越高。在文学作品中，互信息体现在作家使用的词语上，即作者习惯使用的词语和作家的作品风格具有一定的相关性。

[1] 该文使用的是李荣陆博士提供的基于 SVM-light 的文本分类工具包，在此表示感谢。

[2] 具体篇目见《沈从文文集》丛书，花城出版社，1982。

[3] 具体篇目见《余华散文随笔集》，人民日报出版社，1998.12。

$$MI(c,f)=\log\left(\frac{P(c,f)}{P(c)P(f)}\right)$$

其中，f 是作者使用的词语，c 是属于某作家的类别，当 f 和 c 相关性越高，$MI(c,f)$ 的值也越大。当 f 独立于 c 时，$MI(c,f)$ 为 0。在应用时，一般取平均值：

$$MI_{avg}(f)=\sum_{c\in C}P(c)MI(c,f)$$

(2) 卡方统计 (χ^2)

χ^2 统计也是用来衡量两个变量的相关性，但它比互信息更强，因为它同时考虑了特征存在与不存在时的情况。即作家在自己作品中使用某些词语和不使用某些词语 χ^2 统计都会考虑到。

$$\chi^2(c,f)=\frac{(P(c,f)P(\bar{c},\bar{f})-P(c,\bar{f})P(\bar{c},f))}{P(c)P(f)P(\bar{c})P(\bar{f})}$$

当 c 与 f 相互独立时，$\chi^2(c,f)$ 为 0。平均值为：

$$\chi^2_{avg}(f)=\sum_{c\in C}P(c)\chi^2(c,f)$$

(3) 交叉熵 (Cross Entropy)

熵和互信息正好相反，体现在两变量间的分离成度，熵越大，相关性越差。交叉熵只考虑特征在文本中发生的的情况。

$$CE(f)=\sum_{c\in C}P(c,f)\log\left(\frac{P(c,f)}{P(c)P(f)}\right)$$

(4) 证据权值 (Weight of Evidence)

证据权值反映的是类概率 (文中为确定为某一作家类别的概率) 在给定某一特征值下的类概率的差。对于特征 f，其证据权值记为 $WE(f)$，计算公式如下：

$$WE(f)=\sum_{c\in C}P(c)P(f)\left|\log\left(\frac{odds(c|f)}{odds(c)}\right)\right|$$

当 n 是训练文档实例数目，且有：

$$odds(X_i)=\begin{cases}\dfrac{1/n^2}{1-1/n^2} & P(X_i)=0\\ \dfrac{1-1/n^2}{1/n^2} & P(X_i)=1\\ \dfrac{P(X_i)}{1-P(X_i)} & P(X_i)\neq 0\wedge P(X_i)\neq 1\end{cases}$$

为了尽可能客观准确地得出识别结果，采取了如下的实验步骤：

(1) 采用随机抽取的方法来进行实验，首先以 10 KB 大小为单位把每个作家的作品分为若干份 (该实验 8 位作家共分为 1 632 个文件)，再按比例分别从每个作家作品语料中随机抽取训练语料和测试语料。

(2) 采用低密度多特征方法训练 SVM 模型，所谓低密度就是训练语料属于某类的特征点覆盖度，通过对训练语料进行进一步的分割来降低特征点的密度，在选取特征维度时要尽量往高取，具体数值在实验中确定，测试语料也采取相应策略。

(3) 识别性能评价，使用宏平均查全率 ($\bar{r}$)、宏平均查准率 ($\bar{p}$) 和宏平均调和平均值 ($\bar{f}$)。宏平均是先对每一个类统计 r、p 值，然后对所有的类 r、p 的平均值，即：

$$\bar{r}=\frac{\sum_{c\in C}r_c}{|C|}\qquad \bar{p}=\frac{\sum_{c\in C}p_c}{|C|}\qquad \bar{f}=\frac{2\bar{r}\bar{p}}{\bar{r}+\bar{p}}$$

具体的作者识别流程图如图 2。

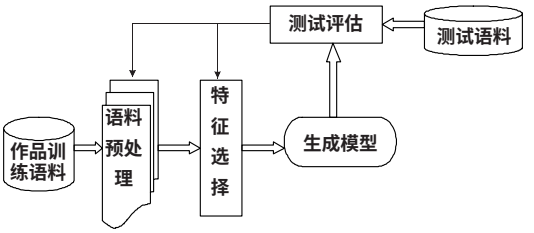

图 2 识别流程图

4 实验结果及分析

为了找出对现当代文学作品作者身份识别最有效的统计量，进行了三组对比实验。

(1) 特征空间维数选取实验

在实验中，在切分出的 1 632 个文件中按 6:4 比例随机抽取训练和测试语料，并且使用互信息作为识别统计特征。从图 3 可以看到随着特征空间维数的增加查全率和准确率也随之增加，在特征维数达到 9 000 时，宏准确率已达到 90.561%，再次增加特征维数，查全率和准确率仍继续增加，在特征维数为 50 000 左右时查全率和准确率达到了最大值为 100%。因此在基于 SVM 的作者身份识别实验中，如果训练时间允许的话，为达到较高的查全率和准确率，应该在训练语料时尽量多选取词语的空间特征维数进行训练。

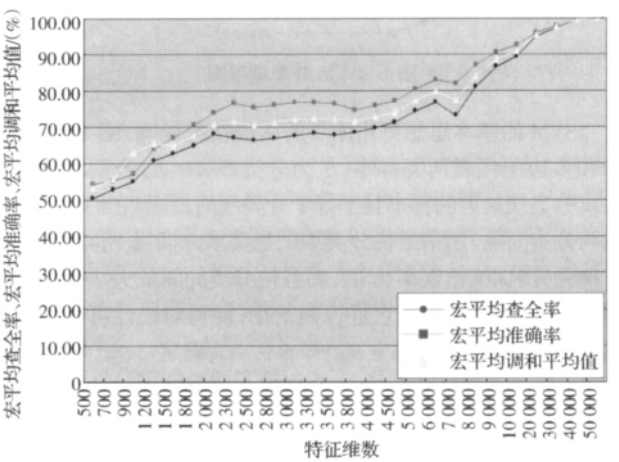

图 3 特征空间维数选取实验结果

(2) 不同特征统计量的对比实验

该组实验主要目的是测试 4 个不同的统计特征量在作者身份识别中的效果。和上组实验一样，重新在 1 632 个文件中按 6:4 比例随机抽取训练和测试语料，空间特征维数设定为 4 000，实验环境为 系统 XP、处理器 Intel Core2 Duo、内存 2 GB。从表 2 的实验结果可以看出使用证据权值作为统计特征的识别效果最好，而且训练时间只比采用互信息统计特征量多 125 s，但宏平均 $\bar{f}$ 值却高出 17.527%。

表 2 不同特征统计量的对比实验结果

实验次数	统计特征	特征维数	宏查全率/(%)	宏准确率/(%)	宏调和平均值/(%)	训练时间/s
第一次	互信息	4 000	80.719	83.113	81.899	491
第二次	卡方统计	4 000	99.348	99.417	99.382	667
第三次	交叉熵	4 000	98.902	98.861	98.881	566
第四次	证据权值	4 000	99.493	99.359	99.426	616

(3) 小语料训练不同统计量对比实验

该组实验的训练语料和测试语料是在 1 632 个文件中按 3:7 随机抽取训练语料和测试语料的，目的是为了测试 4 个不同的统计特征量在较小训练语料(30%)的识别效果。在实验中设定空间特征维数为 5 000，从表 3 的识别结果可以看出不同的统计特征识别效果从高到低的顺序为：证据权值>卡方统计>交叉熵>互信息。这一结果和上一组实验结果从高到底的排序是一样的，但证据权值特征的识别宏平均 $\bar{f}$ 值比互信息的 $\bar{f}$ 值要高出 21.54%。

表 3 小语料训练不同统计量对比实验结果

实验次数	统计特征	特征维数	宏查全率/(%)	宏准确率/(%)	宏调和平均值/(%)
第一次	互信息	5 000	71.524	84.254	77.369
第二次	卡方统计	5 000	98.589	98.920	98.754
第三次	交叉熵	5 000	98.474	98.711	98.592
第四次	证据权值	5 000	98.815	99.004	98.909

通过以上三组实验，可以得出：以词语为基础的统计特征量的多特征 SVM 识别模型在现当代文学作品作者身份识别中的效果是非常优异，特别是在使用较小的训练语料时(文中第三组实验训练语料为 489 个，测试语料为 1 143 个)，和采用大语料的识别效果相差无几(第二组最高识别结果和第三组最高识别结果 $\bar{f}$ 值相差仅为 0.517%)。另外文中使用的现当代文学作品语料体裁丰富，包括小说、戏剧、散文、杂文、散文诗，但在基于 SVM 的多特征识别中仍就保持非常好的识别效果，这也说明了基于词语的统计特征量能够量化作者作品的某些风格特征，这对计算风格学的研究有着十分重要的启示作用。另外随着古代汉语分词技术的成熟，基于词语的多特征 SVM 识别模型必将会在古文献作者识别领域中发挥出更大的作用。

参考文献：

[1] Yule G U.On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship[J].Biometrika，1938，30：363-390.

[2] Gani J.Literature and statistics[M]//Kotz S，Jonhnson N L.Encyclopedia of Statistics.[S.l.]：Wiley，1985：90-95.

[3] Valenza R J.Are the Thisted-Efron authorship tests valid？[J].Computer and the Humanities，1991，25：27-46.

[4] Khmelev D，Tweedy F J.Using Markov chains for identification of Writers[J].Literary and Linguistic Computing，2001，16(4)：299-307.

[5] De Vel O，Anderson A，Corney M，et al.Multi-topic E-mail authorship attribution forensics[C]//Proc Workshop on Data Mining for Security Applications，8th ACM Conference on Computer Security，CCS'2001，2001.

[6] Short text authorship attribution via sequence kernels，Markov chains and author unmasking：An investigation[C]//Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP)，Sydney，2006：482-491.

[7] 曾毅平，朱晓文.计算方法在汉语风格学研究中的应用[J].福建师范大学学报：哲学社会科学版，2006(1).

[8] 金明哲.中文文章的作者识别[C]//第二届中国社会语言学国际学术研讨会暨中国社会语言学会成立大会，澳门，2003.

[9] 武晓春，黄萱菁，吴立德.基于语义分析的作者身份识别方法研究[J].中文信息学报，2006(6).

[10] 钱锋，陈光磊.关于发展汉语计算风格学的献议[M]//修辞学发凡与中国修辞学.上海：复旦大学出版社，1983.

[11] 李荣陆.文本分类及其相关技术研究[D].上海：复旦大学，2005.

(上接 126 页)

图 4 是在最小支持度为 0.9%，节点数量从 2 个增加到 10 个，在不考虑 I/O 操作时间和网络时间的情况下，计算耗时最长节点所用时间。实验结果显示，提出的算法有效地减少了总的计算时间，同样可以从图 5 中看出。

4 结束语

实验证明该算法有较好的均衡负载能力，算法可以将主要的计算量合理地分配到各个计算节点，有效地分解了计算量到各个计算节点，并且减少了挖掘频繁模式的时间。所提出的算法在构造多棵 LFP-tree 的时候会造成部分 LFP-tree 节点的冗余，但 LFP-tree 远比原 FP-tree 小，而各个计算节点只需要挖掘 LFP-tree，这样就降低了频繁模式挖掘的硬件需求，本来只能用一台高性能计算机才能完成的频繁模式挖掘任务，使用该文提出的算法就可以用多台计算性能一般的计算机来完成。

参考文献：

[1] Agrawal R，Imielinski T，Swami A.Mining association rules between sets of items in large databases[C]//Proceedings of the ACM SIGMOD International Conference Management of Data，Washington，1993：207-216.

[2] Han J，Kamber M.Data mining：Concepts and techniques[M].Beijing：High Education Press，2001.

[3] Agrawal R，Imielinski T，Swami A.Mining association rules between sets of items in large databases[C]//Proc 1993 ACM-SIGMOD Int Conf Management of Data，Washington，DC，May 1993：207-216.

[4] Savasere A，Omiecinski E，Navathe S.An efficient algorithm for mining association rules in large databases[C]//Proc of the 21st VLDB Conference，Zurich，Switzerland，1995：432-443.

[5] Agrawal R C，Agarwal C，Prasad V V V.A tree projection algorithm for generation of frequent itemsets[J].Journal of Parallel and Distributed Computing：Special Issue on High Performance Data Mining，2000.

[6] Han J W，Pei J，Yin Y W，et al.Mining frequent patterns without candidate generation[C]//Proc ACM-SIGMOD Int Conf on Management of Data (SIGMOD'00)，Dallas，TX，2000：1-12.

[7] Han J，Pei J，Yin Y，et al.Mining frequent patterns without candidate generation：A frequent-pattern tree approach[J].Data Mining and Knowledge Discovery，2004，8：53-87.

[8] 谈克林，孙志挥.一种 FP 树的并行挖掘算法[J].计算机工程与应用，2006，42(13)：155-157.