文章编号:1004-9037(2014)01-0071-05

一种新型朴素贝叶斯文本分类算法

邸 鹏 段利国

(太原理工大学计算机科学与技术学院,太原,030024)

摘要:针对在文本分类中先验概率的计算比较费时而且对分类效果影响不大、后验概率的精度损失影响分类准确率的现象,对经典朴素贝叶斯分类算法进行了改进,提出了一种"先抑后扬"(抑制先验概率的作用,扩大后验概率的影响)的文本分类算法。算法中去掉了对先验概率的计算,并在后验概率的计算中引入了一个放大系数。实验结果表明,分类时不计算先验概率对分类精度影响甚微但可以明显加快分类的速度,在后验概率的计算中引入放大系数减少了误差传播的影响,提高了分类精度。

关键词:文本分类;朴素贝叶斯;先验概率;后验概率

中图分类号:TP391.1 **文献标识码**:A

New Naive Bayes Text Classification Algorithm

Di Peng, *Duan Liguo*

(Department of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, 030024, China)

Abstract: According to the phenomena that the calculation of prior probability in text classification is time-consuming and has little effect on the classification result, and the accuracy loss of posterior probability affects the accuracy of classification, the classical naive Bayes algorithm is improved and a new text classification algorithm is proposed which restrains the effect of prior probability and amplifies the effect of posterior probability. In the new algorithm, the calculation of prior probability is removed and an amplification factor is added to the calculation of posterior probability. The experiments prove that removing the calculation of prior probability in text classification can accelerate the classification speed and has little effect on the classification accuracy, and adding an amplification factor in the calculation of posterior probability can reduce the effect of error propagation and improve the classification accuracy.

Key words: text categorization; naive Bayes; prior probability; posterior probability

引 言

自动文本分类是自然语言处理领域中的一个研究热点,其研究目的是借助自动分类技术判断文本的类别。数量急剧增长的网络文本成为人们获取信息的主要来源,借助文本分类技术,可以更加快捷、准确地获取用户需要的信息。此外,文本分类技术在电子政务、垃圾邮件过滤、文本情感分析、网络舆情监控等领域都有着广泛的应用。[1]

在英文文本分类方面,Dublin 大学 Finn 等人[2]研究主客观句分类,得出基于词性标注的特征选择方法比词袋效果好。Columbia 大学 Yu 等人[3]对新闻这类主要讲"事实"的文本进行主客观句子识别,利用 SimFinder 工具计算句子相似度,构造训练集,结合各类词性信息构建贝叶斯分类器,提出多分类器的构建以解决训练集构造的不确定性和训练集质量的问题。Cornell 大学 Pang 等人[4]利用属性相同的句子位置分布较近的特点,将候选句子构成一幅图,从而将主客观句分类转化为求图的最小割问题,实现 Cut-based 分类器,对主客观句进行分类识别。

基金项目:国家重点实验室开放课题(SKLSE2012-09-30)资助项目;山西省自然科学基金(2013-011015-2)资助项目;太原理工大学"语言信息处理学科建设和研究"专项项目资助。

收稿日期:2013-09-01;**修订日期**:2013-11-02

在中文文本分类方面，文献[5]提出了一种 K 近邻元分析分类算法，该算法利用 NCA 算法对训练集进行距离测度学习和降维，使用了 K 近邻方法，通过测试集的类条件概率进行类别判定，取得了很好的分类效果。文献[6]提出了一种基于属性频率的朴素贝叶斯算法，该方法放松了属性之间相互独立的假设条件，利用 RoughSet 可辨识矩阵，对不同属性赋予不同权值，在最后的实验中取得了良好的分类效果。文献[7]提出了一种归一化向量的分类算法，将单个类别中的词频及文档频率用到矩阵投影运算中去，将文本特征的三维空间投影到二维空间上，有效地降低了特征空间维数，提高了分类效率。

综合分析现有研究成果，没有文献对先验概率的计算进行改进，对贝叶斯算法的研究也普遍是为了提高分类的准确率，而很少考虑分类的时间，并且在后验概率的计算中存在的一个小误差没能被很好地改进。因此，本文对传统的贝叶斯算法进行改进，提出了一种新的文本分类算法。

1 贝叶斯相关理论

用于文本分类的机器学习算法主要有 SVM，Bayes，KNN，LLSF 和决策树等。其中，朴素贝叶斯算法是一种以贝叶斯相关理论为基础的最常用的方法，它以属性之间的相互独立性为前提，当该前提成立时，与其他分类算法相比，朴素贝叶斯算法的准确率往往最高[8]。而在朴素贝叶斯算法中经常会用到全概率公式以及贝叶斯公式。

1.1 全概率公式

设 A，B 是随机试验的两个随机事件，事件 B 发生的概率 $P(B)>0$，则称

$$P(A \mid B)=\frac{P(AB)}{P(B)} \tag{1}$$

为在事件 B 发生的条件下，事件 A 发生的条件概率[9]。

全概率公式的定义如下：设随机试验 E 的样本空间为 Ω，A 为 E 的一个事件，$B_1,B_2,\cdots,B_n$ 是对空间 Ω 的一个有限划分，且 $P(B_i)>0(i=1,2,\cdots,n)$，则

$$P(A)=P(A \mid B_1)P(B_1)+P(A \mid B_2)P(B_2)+\cdots+P(A \mid B_n)P(B_n)=\sum_{i=1}^{n}P(B_i)P(A \mid B_i) \tag{2}$$

称为全概率公式[9]。也可以通俗的理解为：事件 A 发生有 $B_1,B_2,\cdots,B_n$ 这 n 种方法，将每一种方法下事件 A 发生的概率相加，就可得到事件 A 发生的概率。

1.2 贝叶斯公式

设随机试验 E 的样本空间为 Ω，A 为 E 的事件，$B_1,B_2,\cdots,B_n$ 为 Ω 的一个划分，则

$$P(B_j \mid A)=\frac{P(A \mid B_j)P(B_j)}{\sum_{i=i}^{n}P(A \mid B_i)P(B_i)} \tag{3}$$

称为贝叶斯公式[9]。

1.3 朴素贝叶斯算法

在朴素贝叶斯算法中会用到先验概率，它是指一个假设能够成立的背景知识。在本文中，某个假设 h 在未进行训练之前的初始概率用 $P(h)$ 表示，即假设 h 的先验概率。在实际处理中，当没有先验概率时，可以为每一种假设赋予一个相同的先验概率。同理，用 $P(D)$ 表示训练样本数据 D 的先验概率。对于 $P(D/h)$，它表示当假设 h 成立时观察到数据 D 的条件概率。在机器学习中，通常需要研究的是 $P(h/D)$，即给定一个训练样本数据 D 之后，判断在数据 D 的基础上假设 h 成立的条件概率，也把它叫做后验概率，它表示训练样本数据 D 出现时假设 h 成立的置信度。

根据贝叶斯公式

$$P(h \mid D)=\frac{P(D \mid h)P(h)}{P(D)} \tag{4}$$

可以去掉公式中的 $P(D)$，因为 $P(D)$ 通常是不依赖于假设 h 的常量，则公式中后验概率 $P(h|D)$ 的值就取决于 $P(D|h)P(h)$ 这个乘积，这就是本文所要用到的朴素贝叶斯算法的核心思想。在给定了候选假设集合 H 以及训练数据 D 之后，从假设集合 H 中找出 D 出现时可能性最大的假设 h。简而言之，就是当给定一些训练样本数据集之后，如何让计算机去学习这个训练样本数据集，从而当碰到新的数据时，可以自动将新数据归类到已经定义好的某一个类别中去。

一个数据通常包含多个属性，这里假设数据 D 中包含 $a_1,a_2,\cdots,a_n$ 这 n 个属性，而朴素贝叶斯算法基于这样一个假设：给定数据的属性值之间相互独立。该假设说明当给定一个具体的目标值时，$a_1,a_2,\cdots,a_n$ 同时发生的联合概率等于每个属性单独发生的概率的乘积，用公式表达

$$P(D \mid h)=P(a_1,a_2,\cdots,a_n \mid h)=\prod_{i=1}^{n}P(a_i \mid h) \tag{5}$$

则朴素贝叶斯算法中的后验概率就可以表示成[10]

$$P(h \mid D) = P(h)\prod_{i=1}^{n} P(a_i \mid h) \tag{6}$$

2　朴素贝叶斯分类器设计

当给定一篇中文文本时,计算机只能将其识别成一个很长的字符串,如何让计算机更加细致地识别这篇文本是首先要面对的问题。因此,在利用朴素贝叶斯算法进行文本分类之前,需要对文本进行一些预处理。

2.1　预处理

中文不像英文那样词与词之间都有空格分开,因此中文文本需要进行分词才能够进行下一步的处理。在本次朴素贝叶斯分类器的设计过程中,直接调用了中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS[11]。

在分词过后,作者利用哈尔滨工业大学信息检索研究中心提供的中文停用词表,对分词结果进行特征提取[12],以去除那些对分类结果无影响的词,比如"的"、"在"这一类词,降低文本向量的维数,提高运算的效率[13]。

2.2　朴素贝叶斯分类器设计

朴素贝叶斯分类器就是将朴素贝叶斯算法应用在分类器的设计上。根据贝叶斯

$$P(C \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid C)P(C)}{P(\boldsymbol{X})} \tag{7}$$

在特征提取之后,待分类的文本就会被表示成一个特征向量 $\boldsymbol{X}(x_1, x_2, \cdots, x_n)$,下一步的任务就是将该向量 $\boldsymbol{X}(x_1, x_2, \cdots, x_n)$归类到它最可能属于的类别 $C(C_1, C_2, \cdots, C_j)$中去。其中,$\boldsymbol{X}(x_1, x_2, \cdots, x_n)$表示文本的特征向量,$C_1, C_2, \cdots, C_j$ 为给定的 j 种类别。换句话说,就是求解向量 $\boldsymbol{X}(x_1, x_2, \cdots, x_n)$分别属于 $C_1, C_2, \cdots, C_j$ 的概率值$(P_1, P_2, \cdots, P_j)$,其中,P_j 表示将 $\boldsymbol{X}(x_1, x_2, \cdots, x_n)$归类到 C_j 的概率,那么 $\max(P_1, P_2, \cdots, P_j)$所对应的结果就是文本 X 所属的类别[14]。

根据朴素贝叶斯算法可以得到

$$P(C_j \mid x_1, x_2, \cdots, x_n) = P(x_1, x_2, \cdots, x_n \mid C_j)P(C_j) \tag{8}$$

式中:$P(C_j)$为待分类文本属于 C_j 的先验概率;$P(x_1, x_2, \cdots, x_n \mid C_j)$表示待分类文本属于 C_j 时;类别 C_j 中包含文本特征向量$(x_1, x_2, \cdots, x_n)$的后验概率。所以求解 $\max(P_1, P_2, \cdots, P_j)$就可以转化为求解式(9)的最大值

$$\arg\max_{C_j \in C} P(x_1, x_2, \cdots, x_n \mid C_j)P(C_j) \tag{9}$$

根据朴素贝叶斯算法的假设,文本的特征属性 $x_1, x_2, \cdots, x_n$ 之间相互独立,则其联合概率就等于各个属性的概率的乘积,所以,最后用来进行分类的函数就是

$$\arg\max_{C_j \in C} P(C_j)\prod_{i=1}^{n} P(x_i \mid C_j) \tag{10}$$

式中:$P(C_j) = \dfrac{N(C = C_j)}{N}$,$N(C = C_j)$为训练集中属于类别 C_j 的文本数量,N 为训练集中文本的总数量。$P(x_i \mid C_j) = \dfrac{N(X_i = x_i, C = C_j) + 1}{N(C = C_j) + M}$为 C_j 类的文本中包含 x_i 的文本数,$N(C = C_j)$为 C_j 类中所包含的总文本数,M 为文本特征向量的维数。

3　算法改进

在研究的过程中,主要从以下两个方面进行了改进。

(1)通常在给定一篇中文文本时,人们可能会有一个背景知识,根据这个背景知识可以对该篇文本进行大概的分类,这便是传统的朴素贝叶斯算法中的先验概率所起的作用。但在某些实际情况中,事先并不知道它属于哪一类,也就是没有相关的背景知识,因此本着公平的原则,该篇文档属于某一类的先验概率应该相同,不能因为训练文本集中某一类的文本数量多,就认为属于这一类的概率大,这是不公平、不科学的。因此本文就只针对先验概率相等的情况进行后续研究。根据以上观点,就可以去除分类函数中的先验概率的计算,则分类函数为

$$\arg\max_{C_j \in C} \prod_{i=1}^{n} P(x_i \mid C_j) \tag{11}$$

因为最后是要通过比较得出最大的概率值,所以都去掉先验概率的计算不会影响最后的分类结果。同时,去掉先验概率的计算可以大幅减少计算机的 I/O 操作,从而增加计算的速度。

(2)在实验中,定义了一个 double 类型的变量来存放计算出来的后验概率,每个特征词计算出来的后验概率往往都是一个非常小的数,有时候待分类的文本比较长,其特征词就比较多,根据朴素贝叶斯算法计算出来的概率值往往非常小,在实验中发现很多时候计算出来的后验概率都为 0.0,当预先定义好的类别数目比较少时,就很有可能影响最后的分类精度。为克服这种误差传播,为每个特征

词求解出来的后验概率引入一个放大系数 K，这不会影响实验结果，因为最后是要通过比较得出最大的概率值，适当放大一定倍数更有利于减少误差传播，提高分类精度。K 值的确定会在下面的实验中给出。分类函数为

$$\arg\max_{C_j \in C} \prod_{i=1}^{n} K * P(x_i \mid C_j) \tag{12}$$

4 实验结果与分析

构建朴素贝叶斯分类器需要大量的训练文本集和测试集，本文实验数据来自搜狗实验室提供的 reduced 版本互联网语料库，该版本包含财经、IT、健康、体育、旅游、教育、招聘、文化、军事等 9 类文档，每一类文档包含 1 990 篇文档，实验中随机从每一类文档中抽取出一定数量的文本作为训练文本，再从剩下的每一类文本中抽取一些文本作为测试文本，最后使用准确率作为评价指标，其中

$$\text{准确率} = \frac{\text{分类的正确文本数}}{\text{该类的文本总数}} \tag{13}$$

实验时所用机器型号为惠普 Compaq 6515b，机器主要配置为 AMD Turion 64 处理器，1 GB 内存，2.2 GHz 主频，使用 JAVA 编程语言，Myeclipse8.5 开发环境。实验分别对改进前后的算法进行验证，实验数据如表 1 所示。

表 1 分类结果对比

Table 1 Classification results comparison

文本类别	训练集数量	测试集数量	改进前准确率	改进后准确率	变化趋势
财经	700	400	0.900 0	0.912 5	↑
IT	650	370	0.897 3	0.897 3	
健康	610	330	0.915 2	0.924 2	↑
体育	580	310	0.912 9	0.935 5	↑
旅游	540	290	0.910 3	0.917 2	↑
教育	510	270	0.870 4	0.877 8	↑
招聘	530	300	0.863 3	0.886 7	↑
文化	550	310	0.877 4	0.890 3	↑
军事	570	320	0.937 5	0.937 5	

通过表 1 实验可以看到朴素贝叶斯分类器的准确率还是非常高的，去掉先验概率的计算对分类结果的影响并不是很大，除了军事类和 IT 类在改进前后的准确率相同之外，其他几类文本的准确率，改进后的分类算法都比改进前的分类算法有小幅度的提高。

此外，实验还从每一类文本中抽取出一篇文本作为测试文本，对其改进前后的计算时间进行了统计，结果如表 2 所示。计算时间因测试文本的字数而异。

表 2 计算时间对比

Table 2 Computer time comparison s

训练样本	改进前	改进后	加快时间
财经	216	207	9
IT	148	143	5
健康	270	263	7
体育	202	196	6
旅游	138	132	6
教育	125	118	7
招聘	155	151	4
文化	182	177	5
军事	224	216	8

从表 2 的实验结果可以看出，改进后的算法在计算速度上有了明显的提升，9 类测试文本在算法改进后所用的时间都比改进前要少。综合表 1，2 的实验结果，可以得出结论：改进后的算法效率更高。

在改进的算法中，为减小误差传播，为后验概率的计算引入了一个放大系数 K，因为最后利用分类函数计算出来的后验概率值非常小，结果有时会显示为 0.0。起初将 K 的值设置为 10，但在实验中发现，有些时候放大倍数过大时也会使后验概率的计算结果为 0.0。随机从 9 类文本中，每类随机抽取了 100 篇测试文本，其中包含一些篇幅比较长的文本，当计算一篇文本分别属于 9 种类别的后验概率时，9 个后验概率中只要出现一次 0.0，就将这篇文本标记出来，最后统计了每一类的 100 篇文本在扩大倍数分别为 3，4，5，6，7，8，9，10 时被标记出来的文本数，放大系数为 1～2 倍时对分类性能的影响太小，所以没有列出对应的统计结果。实验结果如表 3 所示。

通过表 3 的实验结果可以看出，放大倍数集中

表 3 放大倍数实验结果

Table 3 Experiment results of amplification factor

类别	放大倍数							
	3	4	5	6	7	8	9	10
财经	9	8	5	5	5	6	7	7
IT	8	7	6	5	6	7	7	7
健康	10	7	4	5	5	7	8	9
体育	8	7	6	6	7	8	8	8
旅游	9	8	5	6	6	7	7	8
教育	9	7	7	6	7	8	9	9
招聘	8	6	4	5	6	6	7	8
文化	7	6	5	5	6	7	8	9
军事	6	4	4	5	5	6	6	7

于 4,5,6,7 时被标记出来的文本数相对较少,放大的效果比较好。其中,当放大系数为 5 时,取得的实验结果相对更好。

5　结束语

通过朴素贝叶斯算法构造文本分类器,是一种简单有效的方法,分类的准确率非常高,速度也相对较快。但由于朴素贝叶斯算法是一种基于机器学习的算法,它的准确率在很大程度上依赖于训练集,因此,如何确定训练文本集的数量将是今后的一个研究方向。在本次放大系数的实验中,只是初步选取了一些整数倍,初步确定了放大倍数的范围,具体放大到什么程度,也有待日后做进一步的研究。此外,当文本中包含一些较复杂的句式时,往往会影响分类的精度,这也将是今后一个主要的研究方向。

参考文献:

[1] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1): 1-9.

[2] Finn A, Kushmeick N, Smyth B. Genre classification and domain transfer for information filtering[C]// Proceedings of the 24th BCS-IRSG European Colloquium on Information Retrieval Research: Advances in Information Retrieval. UK: Springer, 2002: 353-362.

[3] Yu H, Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences [C]// Proceedings of the 2003 Conference on EMNLP. USA: ACL, 2003: 129-136.

[4] Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts [C]// Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Morristown, NJ, USA: ACL, 2004: 271-278.

[5] 刘丛山,李祥宝,杨煜普.一种基于近邻元分析的文本分类算法[J].计算机工程,2012,38(15):139-141.
Liu Congshan, li Xiangbao, Yang Yupu. Text classification algorithm based on neighborhood component analysis[J]. Computer Engineering, 2012, 38(15): 139-141.

[6] 张春英,王晶.一种新型加权朴素贝叶斯分类算法[J].微计算机信息:管控一体化,2010,26(10):222-224.
Zhang Chunying, Wang Jing. A new weighted naive Bayesian classification algorithm[J]. Microcomputer Information, 2010, 26(10):222-224.

[7] 钟将,孙启干,李静.基于归一化向量的文本分类算法[J].计算机工程,2011,37(8):47-49.
Zhong Jiang, Sun Qigan, Li Jing. Text classification algorithm based on normalized vector[J]. Computer Engineering,2011,37(8):47-49.

[8] 吕国云,赵荣椿,张艳宁,等.基于三音素动态贝叶斯网络模型的大词汇量连续语音识别[J].数据采集与处理,2009,24(1):1-6.
Lü Guoyun,Zhao Rongchun,Zhang Yanning. Continuous speech recognition for large vocabulary based on triphone DBN model[J]. Journal of Data Acquisition and Processing,2009,24(1):1-6.

[9] 盛骤.概率论与数理统计[M].北京:高等教育出版社,2010:26-33.

[10] Mitchell T M. 机器学习[M].北京:机械工业出版社,2003:126-128.
Mitchell T M. Machine learning[M]. Beijing: Chine Machine Press,2003:126-128.

[11] 中国科学院计算技术研究所. ICTCLAS 特色[EB/OL]. http://ictclas.org/index.html,2008/2013.
Institute of Computing Technology. ICTCLAS[EB/OL]. http://ictclas.org/index.html,2008/2013.

[12] 赵世奇,张宇,刘挺,等.基于类别特征域的文本分类特征选择方法[J].中文信息学报,2005,19(6):21-27.
Zhao Shiqi,Zhang Yu,Liu Ting. A feature selection method based on class feature domains for text categorization[J]. Journal of Chinese Information Processing,2005,19(6):21-27.

[13] 史岳鹏,朱颢东.基于类别相关性和优化的 ID3 特征选择[J].数据采集与处理,2011,26(2):231-234.
Shi Yuepeng, Zhu Haodong. Feature selection based on category correlation and improved ID3[J]. Journal of Data Acquisition and Processing,2011,26(2):231-234.

[14] 李晓欧,乐建威.基于小波预处理和贝叶斯分类器的 P300 识别算法[J].数据采集与处理,2011,26(4):420-423.
Li Xiaoou,Le Jianwei. P300 detection algorithm based on wavelet preprocessing and bayesian classification [J]. Journal of Data Acquisition and Processing, 2011,26(4):420-423.

作者简介:邸鹏(1989-),男,硕士研究生,研究方向:自然语言处理,E-mail:353967364@qq.com;段利国(1970-),男,博士,副教授,研究方向:中文信息处理。