情报科学
第34卷 第7期 2016年7月

基于朴素贝叶斯的文本分类研究综述

贺 鸣,孙建军,成 颖

(南京大学 信息管理学院,江苏 南京 210023)

摘 要:文本自动分类是自然语言处理领域的重要分支之一,已经形成了大量的模型以及算法,其中基于朴素贝叶斯的相关研究是该领域持续的热点。本文对基于朴素贝叶斯的文本自动分类研究进行了系统的综述。探讨了多项式模型和多元伯努利模型等经典的朴素贝叶斯分类方法。重点分析了经典的特征选择方法以及包括ALOFT等在内的多种改进的特征选择方法。论文还对从加权、避免平滑等视角的NB改进算法进行了梳理。最后,提出了进一步改进NB的主要思路。

关键词:自动分类;朴素贝叶斯;特征选择;特征过滤

中图分类号:G254.9 **文献标识码**:A **文章编号**:1007-7634(2016)07-147-08

Text Classification Based on Naive Bayes: A Review

HE Ming, SUN Jian-jun, CHENG Ying

(School of Information Management, Nanjing University, Nanjing 210023, *China)*

Abstract: Automatic text classification is an important branch of natural language processing, and has already been formed amounts of models and algorithms, included Naive Bayes which is one of sustained research focus in this field. This article summarizes researches on automatic text classification based on Naive Bayes systematically,and discusses classic Naive Bayes methods, including multinomial model and multivariate Bernoulli model. This analyses on classical feather selection methods and some improved methods including ALOFT. And improved NB algorithms are sorted from avoiding smoothing and weighted aspects. Finally, this work presents main idea for NB further improved.

Keywords: automatic classification; Naive bayes; feature selection; feature filtering

1 引 言

信息技术的迅猛发展导致电子文档呈指数级增长,在信息海洋中快速、准确、全面地找到所需信息变得越来越困难。如何有效地组织和管理信息,如何快速区分有用和无用信息,如何满足用户的个性化需求,都面临着挑战。文本自动分类是处理和组织海量信息的关键技术,可以在很大程度上解决信息的无序问题。

文本分类是指,给定文档集$D=\{d_1,d_2,.....,d_n\}$,和一个类别集(标签集)$C=\{c_1,c_2,.....,c_n\}$,利用某种学习方法或算法得到分类函数f,将文档集D中的每一篇文档d_i映射到类别集C中的一个或者多个类别。文本自动分类始于20世纪50年代末,主要有布尔模型、概率统计模型以及向量空间模型。基于三个模型提出了诸多分类算法,其中朴素贝叶斯(Naive Bayes,NB)分类算法在所有分类算法中具有简单且性能优异的特点。NB从算法的提出到目前的成熟应用已经产生了丰硕的成果,有必要对其进行系统的梳理。

2 文献选取

2.1 文献选取原则

(1)本文选取NB文本自动分类研究文献,书评等非研究型文献排除在外。

(2)本文选取的研究对象为文本,非文本型研究文献不纳入本文。

(3)本文主要聚焦于NB文本自动分类特征选择以及算

收稿日期:2015-03-21

基金项目:国家社会科学基金重大招标项目(12&ZD221);国家科技支撑计划子课题(2011BAH30B01-04)

作者简介:贺 鸣(1991-),女,江苏人,硕士生,主要从事信息检索研究.

法研究文献,其他有关自动分类评价方法等文献不是研究重点。

(4)考虑到相关研究的权威性,英文文献主要选择SSCI中SCI数据库中的相关研究,中文文献主要选择图书情报类、计算机类以及中文信息处理类的核心期刊。

2.2 文献选取过程

英文文献主要选择于WoS,检索式为((TS=("naive bayes" or "bayes classif*")) and TS=("text classif*" or "text categor*")) AND 语种: (English),时间段选所有年份,选择进行检索的数据库有:SCI-Expanded、SSCI和A&HCI,结果共查找到文献179篇,通过浏览摘要确定26篇文献。

中文文献在CNKI和万方中进行文献检索,主题设定为"bayes"或"贝叶斯",并且主题中包含"文本分类",选择《中文信息学报》、《计算机学报》、《软件学报》、《计算机研究与发展》、《图书情报工作》、《情报学报》中的文献,通过摘要浏览选择6篇。

根据文章的引文又阅览其他相关的文献,又获得23篇研究文献,中英文合计55篇文献。

3 经典朴素贝叶斯文本分类方法

1960年Maron和Kuhns首先提出[1]了朴素贝叶斯分类方法,是一种基于概率模型的分类方法。Lewis[2]阐释了NB在信息检索和文本分类领域的应用。朴素贝叶斯的"朴素"得名于条件独立性和位置独立性两个基本假设。其中,条件独立假设是假设属性值之间相互独立,即词项之间不存在依赖关系;位置独立假设是指词项在文档中出现的位置对概率的计算没有影响。显然,这两个假设在实际文档中不成立,原因是文档中词项之间存在条件依赖关系,且词项在文档中出现的位置对分类的贡献也不同。不过,即使这两个假设导致朴素贝叶斯概率估计的效果较差,但在分类决策中的效果却非常好,因此得到了广泛使用。

建立朴素贝叶斯分类器有两种方法,其一是多项式模型(multinomial Naive Bayes),另一是多元伯努利模型(multivariate Bernoulli model),也称为二值独立模型。

3.1 多项式模型

多项式NB是一种生成式模型,文档d属于类别c的概率可以通过公式(1)[3]获得:

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c) \quad (1)$$

其中,P(c)是文档d属于类别c的先验概率,$<t_1,t_2,t_3,...,t_{nd}>$是文档d中的词项,n_d是d中所有词项的数量。在已知文档d属于每个类别c_i的先验概率后,需要找到文档d最有可能的类别,对于朴素贝叶斯而言,也就是具有最大后验概率(maximum a posteriori, MAP)估计值的类别,即公式(2)[3]。

$$c_{map} = \arg\max_{c \in C} \hat{P}(c|d) = \arg\max_{c \in C} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c) \quad (2)$$

公式(2)计算条件概率的乘积,可能会导致浮点数下界溢出,所以引入对数,见公式(3)[3]。

$$c_{map} = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} [\log \hat{P}(c) + \sum_{1 \le k \le n_d} \log \hat{P}(t_k|c)] \quad (3)$$

3.2 多元伯努利模型

多元伯努利模型中,对于词项表中的每个词项都对应一个二值变量,分别表示该词项在文档中存在与否,存在用1表示,不存在用0表示。$P(d|c)=P(<e_1,...,e_i,...,e_M>|c)$,$<e_1,...,e_i,...,e_M>$是一个M维的布尔向量,表示每个词项在文档d中是否存在,其C_{map}可以通过公式(4)[3]计算得到。

$$c_{map} = \arg\max_{c \in C} \hat{P}(c) \prod_{t_i \in V} \hat{P}(e_i|c) \quad (4)$$

3.3 小结

前面两种方法的主要区别在于:①文档表示是否考虑词项在文档中的出现频率。多项式NB不仅考虑词项是否出现,且需考虑词项出现的频率;而多元伯努利模型仅考虑词项是否出现。②文档中未出现的词项对于分类的作用不同。在多项式NB中,未出现词项不参与分类,而在多元伯努利模型中,对未出现词项也要进行建模,即未出现词项也要作为一个因子参加P(c|d)的计算以决定文档类别[3]。

相较于多元伯努利模型,多项式NB分类方法具有如下优点:①当有多个同等重要的特征联合起来对分类决策起作用时,多项式NB能够表现出更好的分类效果;②多项式NB对噪音特征和概念漂移具有更强的鲁棒性;③速度快。不论训练还是分类过程其时间复杂度都是线性的[3]。

抛开两种方法的具体区别,Rish等发现尽管朴素贝叶斯分类器的独立性假设欠合理,概率估计也不那么精确,但是其分类决策依然在研究与应用中取得了很好的效果。通过在分类错误数据集上的分布熵影响试验证实:在低熵情形下,朴素贝叶斯分类器具有很好的分类性能。特别是在以下两种情形表现更优,一是具有完全独立的特征,二是具有函数相关的特征。该研究还证实,朴素贝叶斯分类器的准确性与特征向量独立性的程度间没有直接关联;在朴素贝叶斯分类器中损失部分特征所包含的类别信息却可以获得更优的分类准确率[4]。

4 特征选择

文本分类研究中存在的高维特征以及噪声特征问题增加了研究与应用的难度。解决问题自然的途径就是特征降维(dimensionality reduction)。目前,特征降维主要有特征选择(feature selection)和特征抽取(feature exaction)[5-6]两种方法。两种方法的基本思想都是将可以提高分类正确率和减少计算量的原始特征挑选出来。

特征选择方法是指从众多的原始特征集中通过某种全局的规则或排序方法来选取最能反映类别统计特性的相关特征,以构成原始特征集的子集。特征选择有过滤/筛选(fil-

ter)和复选(wrapper)两种基本方式。复选[7]的基本思想是通过规则或是随机选择生成多个不同的特征子集,并用分类器来测试每一个子集。特征子集的个数可以由用户事先定义,或通过规则自动生成,或通过其他参数生成[8]。特征过滤方法,由于无需分类器测试每一个子集,从而比复选的速度快。特征过滤仅仅依赖于训练样本的统计特性,与分类算法无关;而复选方法则不仅要依赖训练样本的统计特性,还依赖于分类算法。虽然复选方法可以选出更合适的特征,但是计算复杂度非常高,所以在文本分类领域通常采用过滤的方法进行[9],后者也是本文综述的重点。

特征抽取方法,即依据某一原则构造输入特征空间到新的特征空间的变换,从而将分散在众多原始特征中的分类信息或鉴别信息集中于少量新的特征[9],具体方法有潜在语义标引(Latent Semantic Indexing,LSI),主成分分析(Principal Component Analysis,PCA)以及语义映射(Semantic Mapping,SM)等。囿于篇幅,特征抽取方法另文阐述。

4.1 经典特征过滤方法

特征过滤依据变量排序算法(Variable Ranking,VR)选取特征向量;VR算法则通过特征评价函数(Feature Evaluation Function,FEF)完成特征的全局排序;最后选出n个得分最高的特征用于文本分类[8]。经典的特征评价函数有互信息[10](mutual information,MI)、CHI统计量[11]、信息增益(IG)[12]、基尼指数(GINI)、文档频率(DF)、几率比(odds ratio,OR)等,具体阐释见参考文献[3]等教科书。

针对这些特征过滤方法,Yang、Pedersen[13]和单松巍等(2003)[14]都指出IG和CHI计算消耗比较大,DF计算消耗很小,而且效果与IG和CHI相同,DF具有算法简单质量高的优点可以用来代替CHI和IG。Yang和Pedersen[13]通过实验发现DF、IG和CHI在评价一个词项上具有很强的相关性;MI比较偏向少数词项和对概率估计错误很敏感,所以MI的效果相对来说要比DF、CHI和IG要差一些。李纲等(2008)[15]的研究显示,在局部文本特征选取的应用中,当特征数小于300时,CHI统计算法的性能远优于MI算法,当特征数大于500时,MI性能提高很快,和CHI的性能相近。

4.2 改进特征选择方法

(1) ALOFT。Pinheiro等(2012)[5]提出了一种称为ALOFT(At Least One FeaTure)的启发式特征选择方法。该方法的基本思想是:用空间向量模型表示文档,仅考虑特征出现与否,不考虑出现频次;用FEF对特征进行全局权重计算;依次选取每个文档中FEF值最大的一个特征加入到新的特征集中,如果某文档FEF值最大的特征已经在特征集中,则略过。经过上述步骤即可得到参与分类的特征集。实验使用评价指标macro-F1和micro-F1,通过t检验的结果显示:97%的ALOFT分类效果与单独FEF持平或比FEF好(即3个数据集x 2种分类器x 2种评判指标x 5种FEF方法=60种组合,ALOFT不输于其中58种组合);62%(37 of 60)的ALOFT要远好于FEF。ALOFT的优点是能找到涵盖训练集所有文档的一个特征子集,且特征数目最优。

(2)基于二项假设检验的特征选取方法。Jieming Yang等(2011)[16]提出了一种基于二项假设检验的特征选取方法Bi-Test,将其运用于垃圾邮件的筛选。算法的基本思想是:说明如果一个特征在垃圾邮件中出现的次数大于在非垃圾邮件中出现的次数,则该特征更多具有垃圾邮件类别的信息;反之则具有更多非垃圾类别的信息;如果在两个类别中出现次数相同,则该特征不具有判别作用。实验环节提出了两个假设,假设一,每封邮件都是独立的,与其他邮件不相关;假设二,每个特征属于垃圾邮件或非垃圾邮件的概率是一个常数。所以一个特征属于垃圾邮件或非垃圾邮件这个事件是一个二项分布实验(即伯努利试验),计算见公式(5)。

$$P(t=m)=\binom{n}{m}\pi^{m}(1-\pi)^{n-m} \tag{5}$$

其中,P(t=m)表示特征t属于垃圾邮件的概率,其中m表示包含特征t的垃圾邮件的个数,n表示所有邮件数,π表示分类的概率。由于P(t=m)、m和n从训练集中可以得知,所以根据公式(5)可以得到π,然后根据π的大小来选择特征,π越大或越小,特征包含的类别信息就越多,π越接近0.5,则特征包含的类别信息越少,则舍弃。

(3)CMFS。除了Bi-Test之外,Jieming Yang等(2012)[17]还提出了CMFS(comprehensively measure feature selection)特征选择算法,旨在测量词项在类别内和类别间的显著性(公式(6))。

$$CMFS(t_k,c_i)=\frac{tf(t_k,c_i)+1}{tf(t_k)+|C|}\cdot\frac{tf(t_k,c_i)+1}{tf(t_k,c_i)+|V|} \tag{6}$$

其中,$tf(t_k,c_i)$表示词项t_k在类别c_i中出现的频率,$tf(t_k)$表示词项t_k在整个训练集中出现的频率,$tf(t,c_i)$表示所有词项在类别c_i中出现的频率和,|C|表示类别个数,|V|表示特征向量空间的所有词项数。为了考虑词项在全局的表现,可以通过公式(7)进行特征选取。

$$CMFS_{max}(t_k)=\max_{i=1}^{|C|}\{CMFS(t_k,c_i)\} \tag{7}$$

作者使用三个语料库(20-Newgroups、Reuters-21578和WebKB),将CMFS与其他6种常用特征选择算法(IG、CHI、DF、OCFS、DIA、GINI)进行比较。实验结果表明:使用NB分类器的CMFS特征选择效果优于IG、CHI、DF、OCFS以及DIA;在大部分情况下,CMFS优于GINI,仅在极个别情况下,略低于GINI。使用支持向量机(SVM)分类器时,CMFS明显优于DIA、IG、DF以及OCFS。

(4)基于最大边缘相关的特征选择方法。刘赫等(2012)[18]根据文本分类具有高维特征空间和高度特征冗余的特点,采用CHI统计量处理高维特征空间,利用信息新颖度处理高度特征冗余。基于此,提出了一种基于最大边缘相关(maximal marginal relevance,MMR)的特征选择方法。所谓的边缘相关即相关性和新颖性的一个线性组合。在文档检索的过程中边缘相关应取最大值因此该方法被称为最大边缘相关[19]。其中,文档d_i最大边缘相关的定义见公式(8)。

$$MMR(d_i)=\lambda Sim_1(d_i,Q)-(1-\lambda)\max_{d_j\in S} Sim_2(d_i,d_j) \quad (8)$$

其中，$Sim_1(d_i,Q)$计算文档d_i和查询Q之间的相似度，用于相关性的度量；$Sim_2(d_i,d_j)$计算文档d_i和d_j之间的相似度，用于新颖性的度量。当$\lambda\in(0,1)$时，MMR就是优化两者的线性组合。实验结果表明MMR_FS方法比CHI和IG特征选择方法更有效且能够提高NB、Rocchio和KNN等分类器的分类性能。

(5)改进的基尼指数(improved GINI)。Aggarwal等(1999)[20]开展了用基尼指数的杂度原理进行特征选择的研究，Shankar等(2000)[21]人提出了应用基尼指数进行文本特征加权的思路。Shang(2007)[22]提出了一种改进的基尼指数算法，即采用基尼指数的纯度原理，将之运用于文本的特征选择。纯度值越大，说明该特征对于分类的作用越大。改进的基尼指数的效果可以与CHI统计量和IG相媲美[22]；实验结果表明，Ogura[23]等学者改进的基尼指数(公式(9))实质上要比IG和CHI统计量的效果要好。

$$Gini(t_k)=\sum_{i=1}^{n}P(t_k|c_i)^2P(c_i|t_k)^2 \quad (9)$$

(6)基于不确定性度量的特征选择算法。Mengle等(2009)[24]提出了一种基于不确定性度量(ambiguity measure，AM)的特征选择算法。该算法旨在寻找所谓的确定性特征，即该特征具有很强的置信度使一篇文档只属于某一个类别，计算方法见公式(10)和(11)。

$$AM(t_k,c_i)=\frac{tf(t_k,c_i)}{tf(t_k)} \quad (10)$$

$$AM(t_k)=\max(AM(t_k,c_i)) \quad (11)$$

当词项AM值接近于1的时候，表示该词项属于某一类别的确定性越大；AM值接近0时，表示该词项属于某一类别的不确定性就越大，也可能该词项属于多个类别，有必要把这些词项过滤掉。论文使用5种语料集在NB分类器上完成了测试，将AM与改进的GINI等其他8种特征选择算法利用micro-F1值进行了比较。结果表明：在Reuters-21578、20NG和WebKB中，9种算法表现都不错，但AM最好；在OHSUMED和TREC 05 Genomics datasets中，9种算法都表现一般，但最优的依然是AM，其次的是改进的GINI。

(7)Best Terms(BT)。Fragoudis等(2005)[25]提出了一种称为Best Terms(BT)的特征选择算法。BT首先定义了积极特征和消极特征。特征t当且仅当满足公式(12)时，特征t被称为类别c的积极特征；t当且仅当满足(13)时，被称为类别c的消极特征。

$$P(c|t)>1/2\cdot p+1/2\cdot P(c) \quad (12)$$

参数p为了抵消P(c|w)>P(c)时出现。

$$P(\bar{c}|t)>1/2\cdot p+1/2\cdot P(\bar{c})$$
$$\Leftrightarrow[1-P(c|t)]>1/2\cdot p+1/2\cdot[1-P(c)] \quad (13)$$
$$\Leftrightarrow P(c|t)<1/2\cdot(1-p)+1/2P(c)$$

BT算法的核心有两步，第一步从每个类别文档中选出得分最高的积极特征；第二步从每个不是该类别，但是至少包含一个第一步选出来的积极特征的文档中，选出得分最高的消极特征。通过Reuters和Newsgroup上的测试，使用NB分类器，用microavg recall、microavg precision、microavg F和macroavg F1进行评价，发现BT的效果最优。

BT与其他特征选择算法的一个不同点是：不用设定特征的数量，只要确定参数p，BT就会自动选择合适的特征数目；由于算法无需考虑训练集文档的数量，且词项空间大小和类别数无关，因此BT算法具有线性时间复杂度。

(8)低损降维。宋枫溪等(2005)[9]基于贝叶斯分类器提出了一种低损降维(Low Loss Dimensionality Reduction，LLDR)的特征选择方法。设Y是原始特征集合，N为训练样本个数，r是需要选择的特征个数，基于低损降维的特征选择过程如下：

①对于每一模式类C，统计出训练集中该模式类的正例个数N_w；

②对于特征集合Y中的每一个特征t，统计出其在正例中出现的频数ptf(t)；

③计算：lldr(t,C)=max{ptf(t)/N_w, (N-ptf(t)/(N-N_w)}；

④取使得lldr(t,C)达到最大的前r个特征。

实验结果表明：LLDR的识别效果与MI和CHI相当，优于DF，同时LLDR的时间复杂度低于MI和CHI。

(9) Cluster Representation Quality (CRQ)。Schneider(2005)[26]提出了一种基于分布式聚类的思想的CRQ特征选择方法(公式(14))。

$$CRQ(t_k)=\frac{1}{n}\sum_{j=1}^{|C|}\sum_{d_i\in c_j}p_i(t_k)\left[\log p_j(t_k)-\log p(t_k)\right] \quad (14)$$

CRQ取值可以为负数、零或正数，由于CRQ有一个为零的理论阈值，所以可以以0为自然阈值进行特征筛选，即在推导的过程中发现当CRQ为负数时，该词项会被舍弃。

(10)SFS MI。Battiti(1994)[27]和Kwak等(1999)[28]分别提出了应用于非文本型信息的改进MI算法，Novovicova等(2004)[29]基于二者的工作，提出了用于特征选择的前向选择法(Sequential Forward Selection，SFS MI)。前向选择算法首先通过最大的MI选择一个最优的词项，然后每次增加一个词项，直到所选择的词项数量达到期望的k值。通过Reuters(Apte split)的测试，比较了两种SFS MI和IG的表现，实验证明两种SFS MI的F1值好于IG。SFS方法的缺点是并不能取得最优的词项子集，优点是可以表现词与词之间的依赖关系，因此SFS MI不仅要计算词项与类别的MI值$I(C,t_i)$，也要考虑词项与词项之间的MI值$I(t_i,t_j)$。

(11)小结。ALOFT等改进的特征过滤算法大多使用了概率统计方法，其基本思想都是最大化具分类价值的特征值。各项研究的实验均表明，在诸如macro-F1、micro-F1以及准确率等指标上，这些改进算法都比其他经典特征取得了更优的分类效果。不过，通过对这些研究的引文分析发现引用较少，该现象说明大部分改进算法并没有获得广泛的认可。导致该现象的一个主要原因是，这些改进算法基于相异的文本集提出，而文本集各有其自己独特的文本特征，围绕私有特征的相关改进算法的普适性相对较差，因此力争提出

一个通用型的特征过滤算法将成为本文进一步研究的方向。

5 改进的朴素贝叶斯算法

为了更好地贴近NB的假设，许多研究者对文档表示进行了改进[26]，具体包括：抽取更多复杂的特征，例如，根据语法或同级的词组[30]；基于词典资源以强化语义关系[31]，不过遗憾的是这些方法都没有取得很好的效果。另一种改进思路是通过词项聚类[32]或转化特征空间完成抽取特征[33]，这些方法一定程度上提高了分类的准确率。

选择合适的概率统计模型是数据与分类器匹配更重要的方法，一些学者试图通过选择训练集估计的模型参数以提高分类器的性能[34]，下面是该方向的主要改进。

(1)加权NB算法。Li Y等(2012)[35]提出了一种加权NB算法，即提出了一种新的词项-类别依赖关系的度量值$R_{t,c}$(公式(15))，即用$R_{t,c}$表示词项-类别是积极依赖还是消极依赖。积极依赖表示更多包含词项t的文档属于c类；消极依赖，则反之。

$$R_{t,c}=\frac{O(t,c)}{E(t,c)} \tag{15}$$

O(t,c)表示包含词项t，并且属于c类的文档数，期望频率E(w,c)通过公式(16)计算。

$$E(i,j)=\frac{\sum_{a\in\{t,\neg t\}}O(a,j)\sum_{b\in\{c,\neg c\}}O(i,b)}{n} \tag{16}$$

公式(16)可用概率来表示，即为

$$R_{t,c}=\frac{p(t,c)p(\neg t,\neg c)-p(t,\neg c)p(\neg t,c)}{p(t)p(c)}+1 \tag{17}$$

由公式(17)可知，如果词项t和类别c依赖越弱，则$R_{t,c}$的值越接近于1；词项t和类别c积极依赖的程度越大，则$R_{t,c}$的值越大于1；词项t和类别c消极依赖的程度越大，则$R_{t,c}$的值越小于1。将$R_{t,c}$引入到多项式NB中可以获得加权词项概率(公式(18))。

$$\tilde{P}(t_k|c_j)=\begin{cases}\hat{P}(t_k|c_j)R_{t_k,c_j}.....if...R_{t_k,c_j}>1\\ \hat{P}(t_k|c_j).....if...R_{t_k,c_j}\le 1\end{cases} \tag{18}$$

论文从WebKB、Reuters-21578和Newsgroups中提取出五个数据集进行测试。实验结论表明：在五个数据集上，无论是基于F-measure还是准确率的测评，$R_{t,c}$加权的NB分类效果都优于经典的NB以及卡方加权的NB。

(2)TWCNB。Jason等(2003)[36]提出了TWCNB(Transformed Weight-Normalized Complement Naive Bayes)算法。与经典的多项式NB相比，TWCNB将类别c之外的类别用于估计类别c的参数。Kibriya等(2004)[37]在研究中也提到了TWCNB。公式(19)定义了词项权重(word weight)，式中N表示词汇表的大小，F_{kj}表示表示词项k在所有训练集中属于j类的文档。

$$t_{kc}=\log\left(\frac{1+\sum_{j=1}^{|C|}F_{kj}}{N+\sum_{j=1}^{|C|}\sum_{x=1}^{N}F_{xj}}\right),\ j\neq c\wedge j\in C \tag{19}$$

文档的分类见公式(5)-(6)，f_{ki}代表词项t_k在测试文档d_i中的频率。

$$class(d_i)=\arg\max_c[\log(\Pr(c))-\sum_t(f_{ki}t_{kc})] \tag{20}$$

由于log(Pr(c))对于整个式子来说影响很小，所以可以将公式(20)简化为公式(21)。

$$class(d_i)=\arg\max_c\sum_t(f_{ki}t_{kc}) \tag{21}$$

研究中还提出了一种不用将词项权重归一化的方法TCNB，性能与TWCNB的性能相当。通过4种数据集(20news18828、WebKB、IndustrySector和Reuters-21578)对MNB、TWCNB、TCNB和SMO的比较，发现MNB得到的结果最差。

(3)NB_TF和NB_TS。在NB中，平滑方法是最大似然估计不能解决训练集特征缺失问题的常用方法。不过，平滑方法缺乏坚实的理论基础，一些平滑策略效率欠佳且难以被理解，比如，GT平滑因为其复杂的平滑过程难以理解且效率较低[38]，再比如Laplace additive smoothing不考虑未登录词(out of vocabulary，OOV)特征[39]，而实验证明考虑OOV能取得更好的效果。基于此，Zhu等(2005)[40]认为如果在不牺牲分类准确率的前提下去除平滑方法，那么分类器会更有效率。据此，Zhu提出了两种新的策略：NB_TF和NB_TS以避开平滑方法。

NB_TF通过在分类前增加测试文档以调整分类器，适合于在线分类。使用训练集D和测试文档d_i建立模型，并将该文档集称为E，得到公式(22)和(23)，|V*|表示在测试文档中考虑到OOV特征的词项数量。

$$P(c_k)=\frac{\sum_{d_x\in E}b_{xk}}{|E|} \tag{22}$$

$$P(t_j|c_k)=\frac{\sum_{d_x\in E}d_{xj}b_{xk}}{\sum_{i=1}^{|V^*|}\sum_{d_x\in E}d_{xi}b_{xk}} \tag{23}$$

将公式(22)和(23)中的$P(c_k)$和$P(t_j|c_k)$带入经典NB就可以决定文档d_i的所属类别。

NB_TS通过训练阶段将整个测试集应用于分类器以提高性能，在批处理的分类中更有效率。NB_TS使用训练集D和测试集T建立模型，并称该文档集为E，计算公式为(22)和(23)。

基于mini_newsgroups corpus语料库，采用Micro-Values、Macro-Precision、Macro-Recall和Macro-F1值等四种方法评价Laplace平滑、SGT、NB_TF和NB_TS的分类表现。结果显示，NB_TS在四种评价指标上都取得了最好的结果；在语料库Reuters-21578的测试结果显示Micro-Values、Macro-Recall和Macro-F1值三个测量指标NB_TS最优。

(4)基于集成学习的改进方法。集成学习是当今机器学习的热点之一，它将多个同质或异质的基分类器通过某种规则进行集成，得到最终的集成分类器。集成学习中Boosting算法以及改进的AdaBoost算法被广泛使用，许多学者将其与NB算法相结合以提高分类器的性能。

Liu 等(2005)[41]提出了一种改进的 FloatBoost 算法以增强 NB 分类器,称之为 DifBoost 算法。这个算法结合分而治之的策略(Divide and Conquer Policy),在训练阶段将输入空间分成一些子空间,这些不同的子空间影响基分类器的生成,最后通过加权基分类器得到最终的分类器。实验证实该方法的有效性。Zhu J 等(2004)[42]提出了称为 NB_FLB 的改进 NB 算法。该算法针对无监督学习,提出一种特征学习自引导算法(feature learning bootstrapping,FLB),即通过少数种子词项的学习以获得各个类别的特征,然后将其运用于 NB 算法。

尽管 Boosting 算法被证明能有效提高基于机器学习分类器的准确率,但是由于 NB 比较稳定,基分类器的准确度差异不大,所以总体而言 Boosting 在 NB 中的应用成效并不特别突出[43-44]。Lewis 提出了一种基于不确定性选择抽样[45]的改进 Boosting 算法(AdaBUS)以增强 NB 分类器,在 Reuters-21578 上的测试结果显示,该算法可以有效提高分类器的准确率。

(5)其他改进方法。除了上述技术路线之外,还有学者从其他角度提出了 NB 算法的改进思路。Jiang(2012)[46]提出可以通过有区别的实例权重以改进 NB 算法的思路,称之为有区别的加权朴素贝叶斯(discriminatively weighted Naive Bayes)。在每一次迭代的过程中,根据估计的条件概率损失,将不同的训练实例有区别的赋予不同的权重。Ganiz 等(2011)[47]提出了高阶 NB(Higher Order Naive Bayes,HONB)的方法,与其他假设数据实例无依赖关系不同,该方法利用了特征在不同实例中的高阶关系。

Lee 等(2010)[48]提出通过自动计算文档依赖(automatically computed document dependent,ACDD)权重因子以及解决 NB 中当训练集类别倾斜所致分类准确率低的问题。计算 ACDD 的目的在于通过基于调整在每个类别被分类文档密度的概率值,以提高分类器的性能,使分类错误率最小化。Kang 等(2006)[49]提出了一个 RNBL-MN 算法,该算法目的在于为序列分类建立一个基于 NB 分类器的树。树中的每个 NB 分类器都是一个多项事件模型(multinomial event model)。

NB 方法中通过使用参数平滑技术以避免空估计值对计算的影响。Andrés-Ferrer 等(2004)[50]认为可以减少参数域以替代参数平滑,该思路引发了约束域最大似然估计(constrained domain maximum likelihood estimation)问题,针对该问题 Andrés-Ferrer 等提出了一个迭代算法以获取最优解。Nigam 等(2000)[51]设计了一种期望最大化(EM)和 NB 分类器集合的算法以处理混合有标签文档和未标签文档的文档集。

(6)小结。上述研究中学者们通过对朴素贝叶斯模型进行加权以及对平滑方法进行改进等以提高分类器的性能。近年来,随着集成学习研究与应用的不断深入,朴素贝叶斯方法被应用到集成学习中,比如,将期望最大化算法和朴素贝叶斯算法相结合对算法进行改进的学者群在不断增大。还有学者从文档表示层面对朴素贝叶斯算法进行改进,如 Kim 等[52]提出了基于泊松分布的朴素贝叶斯文本分类模型,该模型假设每篇文档都通过一个多元泊松模型生成,将每篇文档的词频归一化信息用于泊松参数的估计,并用多项式分类器将训练集视为一个巨大的文本以估计参数。除了上述途径之外,从其他更合理的视角对朴素贝叶斯分类器进行改进将成为本文进一步研究的方向。

6 结 语

改进朴素贝叶斯分类器的方法主要有两种,一种是构建更有效的特征集,另外一种是削弱朴素贝叶斯的独立假设。

介于特征的好坏会直接影响分类效果,所以文本分类的进一步改进除了算法层面之外,对特征选择以及构建方面还可以有改进空间。除了对特征进行选择、降维外,也可以立足于影响文本分类最底层、最根本的因素:文本表示中的特征项。一般常用的特征项主要是单一的字词,而这些孤立的字词往往不能很好的表示文本的内容特征,其实可以使用一些结构稳定、语义完整的短语进行特征表示,对文本分类的效果会有所提高[53]。

除了本文上述的改进算法之外,由于朴素贝叶斯文本分类是建立在条件独立假设和位置独立假设的基础上,而现实的文本当中,这两个假设是不成立的。所以许多学者采取了削弱这两种假设以改进朴素贝叶斯分类器的思路,提高了分类效果。在朴素贝叶斯分类器的基础上增加特征间可能存在的依赖关系,比如采用 bigram 或 n-gram 模型以削弱朴素贝叶斯的条件独立假设,再比如 Peng 等[54]提出的结合 CAN 的 Bayes 模型,该模型将 n-gram 模型与朴素贝叶斯模型相结合,通过马尔科夫链表示属性间的依赖关系,以此削弱朴素贝叶斯的独立假设。此外,还出现了贝叶斯网络和树状贝叶斯方法。贝叶斯网络的最大特点是可以表示属性间的依赖关系,它由一个有向无环图和条件概率表组成,即它可以表示文本特征间的依赖关系,打破了条件独立假设,但是贝叶斯网络的学习却依旧很困难。文献【55】提出一种树状朴素贝叶斯(Tree-Augmented Naive Bayes, TAN)模型,该模型是对朴素贝叶斯的一种改进,它放松了朴素贝叶斯中的独立性假设条件,将贝叶斯网络表示依赖关系的能力与朴素贝叶斯的简易性相结合,以增强分类性能。

后继的研究还可以考虑文章结构以放宽位置独立假设,根据特征出现位置的不同给特征赋予不同的权重等方法,对分类方法进行改进,特别是对于医学等具有相对固定文本结构的文档应该会具有较好的效果。也可以考虑在当前个性化很强的时代,根据用户的信息行为改进文本分类,根据用户或其信息需求等的不同,加以调整使展现的结果更贴合用户需求。

参考文献

1 Maron M E, Kuhns J L. On relevance, probabilistic

indexing and information retrieval[J]. Journal of the ACM (JACM), 1960, 7(3): 216-244.

2 Lewis D D. Naive (Bayes) at forty: The independence assumption in information retrieval[C]//Machine learning: ECML-98. Springer Berlin Heidelberg, 1998: 4-15.

3 Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. Cambridge: Cambridge university press, 2008.

4 Rish I. An empirical study of the naive bayes classifier [J]. IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, 3(22): 41-46.

5 Pinheiro R H W, Cavalcanti G D C, Correa R F, et al. A global-ranking local feature selection method for text categorization[J]. Expert Systems with Applications, 2012, 39(17): 12851-12857.

6 奉国和, 郑 伟. 文本分类特征降维研究综述[J]. 图书情报工作, 2011, 55(9): 109-113.

7 Kohavi R, John G H. Wrappers for feature subset selection[J].Artificial intelligence,1997,97(1): 273-324.

8 Corrêa R F, Ludermir T B. Improving self-organization of document collections by semantic mapping[J]. Neurocomputing, 2006, 70(1): 62-69.

9 宋枫溪, 高秀梅, 刘树海, 等. 统计模式识别中的维数削减与低损降维[J]. 计算机学报, 2005, 28(11): 1915-1922.

10 Lewis D D. An evaluation of phrasal and clustered representations on a text categorization task[C]//Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1992: 37-50.

11 Schütze H, Hull D A, Pedersen J O. A comparison of classifiers and document representations for the routing problem [C]//Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1995: 229-237.

12 Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.

13 Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C] //ICML. 1997, 97: 412-420.

14 单松巍, 冯是聪, 李晓明. 几种典型特征选取方法在中文网页分类上的效果比较[J]. 计算机工程与应用, 2003, 39 (22): 146-148.

15 李 纲, 夏晨曦, 郑 重. 局部文本特征选取算法的比较和改进研究[J]. 情报学报, 2008, 27(4): 506-511.

16 Yang J, Liu Y, Liu Z, et al. A new feature selection algorithm based on binomial hypothesis testing for spam filtering [J]. Knowledge-Based Systems, 2011, 24(6): 904-914.

17 Yang J, Liu Y, Zhu X, et al. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization[J]. Information Processing & Management, 2012, 48(4): 741-754.

18 刘 赫, 张相洪, 刘大有, 等. 一种基于最大边缘相关的特征选择方法[J]. 计算机研究与发展, 2012, 49(2): 354-360.

19 Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries[C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998: 335-336.

20 Aggarwal C C, Gates S C, Yu P S. On the merits of building categorization systems by supervised clustering[C]//Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining,ACM, 1999: 352-356.

21 Shankar S, Karypis G. A feature weight adjustment algorithm for document categorization[C] //KDD-2000 Workshop on Text Mining, Boston, USA,2000.

22 Shang W, Huang H, Zhu H, et al. A novel feature selection algorithm for text categorization[J]. Expert Systems with Applications, 2007, 33(1): 1-5.

23 Ogura H, Amano H, Kondo M. Feature selection with a measure of deviations from Poisson in text categorization[J]. Expert Systems with Applications, 2009, 36(3): 6826-6832.

24 Mengle S S R, Goharian N. Ambiguity measure feature - selection algorithm[J]. Journal of the American Society for Information Science and Technology, 2009, 60(5): 1037-1050.

25 Fragoudis D, Meretakis D, Likothanassis S. Best terms: an efficient feature-selection algorithm for text categorization [J]. Knowledge and Information Systems, 2005, 8(1): 16-33.

26 Schneider K M. Techniques for improving the performance of naive Bayes for text classification[M].Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, 2005: 682-693.

27 Battiti R. Using mutual information for selecting features in supervised neural net learning[J]. Neural Networks, IEEE Transactions on, 1994, 5(4): 537-550.

28 Kwak N, Choi C H. Improved mutual information feature selector for neural networks in supervised learning[C]// International Joint Conference on,Neural Networks, IEEE, 1999: 1313-1318.

29 Novovi ová J, Malík A, Pudil P. Feature selection using improved mutual information for text classification[C]//Structural, syntactic, and statistical pattern recognition. Berlin: Springer Berlin Heidelberg, 2004: 1010-1017.

30 Mladenic D, Grobelnik M. Word sequences as features in

text-learning[C]//In Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98),1998.

31 Hidalgo J M G, Rodriguez M B. Integrating a lexical database and a training collection for text categorization[J]. 1997,(9): 39-44.

32 Dhillon I S, Mallela S, Kumar R. A divisive information theoretic feature clustering algorithm for text classification [J]. The Journal of Machine Learning Research, 2003, (3): 1265-1287.

33 Torkkola K. Linear discriminant analysis in document classification[C]//IEEE ICDM Workshop on Text Mining, 2001: 800-806.

34 Kim S B, Rim H C, Yook D, et al. Effective methods for improving Naive Bayes text classifiers[C]//PRICAI 2002: Trends in Artificial Intelligence. Springer Berlin Heidelberg, 2002: 414-423.

35 Li Y, Luo C, Chung S M. Weighted Naive Bayes for Text Classification Using Positive Term-Class Dependency[J]. International Journal on Artificial Intelligence Tools, 2012, 21(1).

36 Rennie J D, Shih L, Teevan J, et al. Tackling the poor assumptions of naive bayes text classifiers[C]//ICML,2003, (3): 616-623.

37 Kibriya A M, Frank E, Pfahringer B, et al. Multinomial naive bayes for text categorization revisited[C]//AI 2004: Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2005: 488-499.

38 Good I J. The population frequencies of species and the estimation of population parameters[J]. Biometrika, 1953, 40 (3-4): 237-264.

39 Peng F, Schuurmans D. Combining naive Bayes and n-gram language models for text classification[M]. Berlin: Springer Berlin Heidelberg, 2003: 335-350.

40 Zhu W, Lin Y, Lin M, et al. Removing smoothing from naive bayes text classifier[M].Berlin:Springer Berlin Heidelberg, 2005: 713-718.

41 Liu X, Yin J, Dong J, et al. An improved floatboost algorithm for Naïve bayes text classification[M]. Berlin:Springer Berlin Heidelberg, 2005: 162-171.

42 Jingbo Z, Wenliang C, Tianshun Y. Using seed words to learn to categorize Chinese text[C]//Advances in Natural Language Processing. Springer Berlin Heidelberg, 2004: 464-473.

43 Kim H, Kim J, Ra Y. Boosting Naive Bayes text classification using uncertainty-based selective sampling[J]. Neurocomputing, 2005, (67):403-410.

44 Kim H, Kim J. Combining active learning and boosting for naïve bayes text classifiers[C] //Advances in Web-Age Information Management. Springer Berlin Heidelberg, 2004: 519-527.

45 Lewis D D, Catlett J. Heterogenous Uncertainty Sampling for Supervised Learning[C]//ICML. 1994: 148-156.

46 Jiang L, Wang D, Cai Z. Discriminatively weighted naive bayes and its application in text classification[J]. International Journal on Artificial Intelligence Tools, 2012, 21(21): 3898-3898.

47 Ganiz M C, George C, Pottenger W M. Higher order Naive Bayes: A novel non-IID approach to text classification [J]. Knowledge and Data Engineering, IEEE Transactions on, 2011, 23(7): 1022-1034.

48 Lee L H, Isa D. Automatically computed document dependent weighting factor facility for Naïve Bayes classification [J].Expert Systems with Applications, 2010, 37(12): 8471-8478.

49 Kang D K, Silvescu A, Honavar V. RNBL-MN: A recursive naive bayes learner for sequence classification[M].Berlin: Springer Berlin Heidelberg, 2006: 45-54.

50 Andrés-Ferrer J, Juan A. Constrained domain maximum likelihood estimation for naive Bayes text classification[J]. Pattern Analysis and Applications, 2010, 13(2): 189-196.

51 Nigam K, McCallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine learning, 2000, 39(2-3): 103-134.

52 Kim S B, Han K S, Rim H C, et al. Some effective techniques for naive bayes text classification[J]. IEEE transactions on knowledge and data engineering, 2006, 18(11): 1457-1466.

53 刘 华. 基于关键短语的文本分类研究[J]. 中文信息学报, 2007, 21(4): 34-41.

54 Peng F, Schuurmans D. Combining naive bayes and n-gram language models for text classification[J]. Advances in information retrieval lecture notes in computer science, 2003, (2633): 335-350.

55 Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine learning, 1997, 29(2-3): 131-163.

(责任编辑:徐 波)