基于AdaBoost-Bayes算法的中文文本分类系统

徐 凯[1],陈平华[1],刘双印[2]
(1 广东工业大学 计算机学院,广东 广州 510003; 2 广东海洋大学 信息学院,广东 湛江 524088)

摘 要: 针对中文文本分类准确率低,分类算法低效不稳定问题,提出基于自适应提升朴素贝叶斯算法.该算法采用Naive Bayes和AdaBoost,并且通过优化组合结构,融合两种算法的优点.首先,使用SMEL序列组合成词算法对中文语料进行分词,提取文本特征词汇.然后,使用增强的贝叶斯分类器,通过较小的样本训练,提取出文本特征,生成训练分类矩阵.结合自适应提升算法对简单分类器进行加权,保证分类有平稳准确的效果.通过实验证明,该算法与其他算法相比,错误率更低,可以使分类准确率达到98%以上,而且F1值也优于其他分类算法.

关键词: 中文分词;文本分类;AdaBoost;Bayes
中图分类号: TP309.7 **文献标识码**: A **文章编号**: 1000-7180(2016)06-0063-05

A Chinese Text Classification System Based on Ada Boost-Bayes Algorithm

XU Kai[1], CHEN Ping-hua[1], LIU Shuang-yin[2]
(1 Faculty of Computer, Guangdong University of Technology, Guangzhou 510006, China;
2 College of Information, Guangdong Ocean University, Zhanjiang 524088, China)

Abstract: In view of the low accuracy of Chinese text classification algorithm, the classification algorithm is inefficient and the problem of low efficiency and low efficiency is proposed. Based on the adaptive algorithm, the proposed algorithm is proposed to improve the accuracy. The algorithm uses Bayes Naive and AdaBoost, and the advantages of the two algorithms are fused by the optimization of the structure. First, using the SMEL sequence of the word segmentation algorithm to segment the Chinese corpus and extract the feature words. Then, the enhanced Bias classifier is used to extract the feature of the text and generate the training classification matrix through the small sample training. Combined with the adaptive lifting algorithm, the simple classifier is weighted to ensure that the classification is stable and accurate. Experiments show that the error rate is lower than other algorithms, and the classification accuracy of the algorithm is more than 98%, and the F1 value is better than other classification algorithms.

Key words: Chinese word segmentation; text classification; AdaBoost; Bayes

1 引言

随着信息技术快速发展,海量的各种类型的数据需要管理,包括文本、声音、图像等.文本数据与声音和图像数据相比,占用网络资源少,更容易上传和下载,这使得网络资源中的大部分是文本形式[1].文本分类系统在给定的分类模型下,根据文本的内容通过机器学习方法挖掘文本信息,使文本分门别类,方便管理知识,文本分类技术成为信息处理领域最重要的研究方向之一[1].

文本分类算法有朴素贝叶斯(NB),最近邻(KNN),支持向量机(SVM)等[2].文献[3]指出,NB

收稿日期: 2015-07-14; **修回日期**: 2015-08-20
基金项目: 国家自然科学基金(61572144);广东省科技计划项目(2015A030401101);广东省教育部产学研项目(2013B090500127);广东省重大科技项目"科技业务综合管理系统流程再造及功能完善(2012B080500008)

算法运算速度最快，但分类精度较差. KNN 通过计算向量相似度，查找距离最近文档，缺点是计算复杂度过大，对维数敏感，分类性能不稳定. SVM 使用核函数把文本向高维空间变换从而得到全局最优解，且准确率最高，缺点则是核函数的选择至今还没有很好的理论依据，只有凭经验来选择核. 现有的中文文本分类算法还存在分类准确率不高，分类效率低的问题.

针对以上中文文本分类问题，本文提出了一种基于自适应提升朴素贝叶斯算法.

2 算法流程图

本系统主要分为几个部分：语料处理系统，朴素贝叶斯分类器，自适应提升算法容器. 开始时语料处理系统先加载数据文件，对文本进行分词处理. 接着计算文本热点词汇，统计出现频率最大的前若干个词汇，处理成词汇向量组. 使用朴素贝叶斯分类器对词汇向量进行训练学习得到训练参数，使用参数对测试集数据进行分类. 然后用自适应提升算法对较低错误率的分类器进行组合，最后得到最终分类结果. 基于 adaBoost-Bayes 的中文文本分类算法流程如图 1 所示.

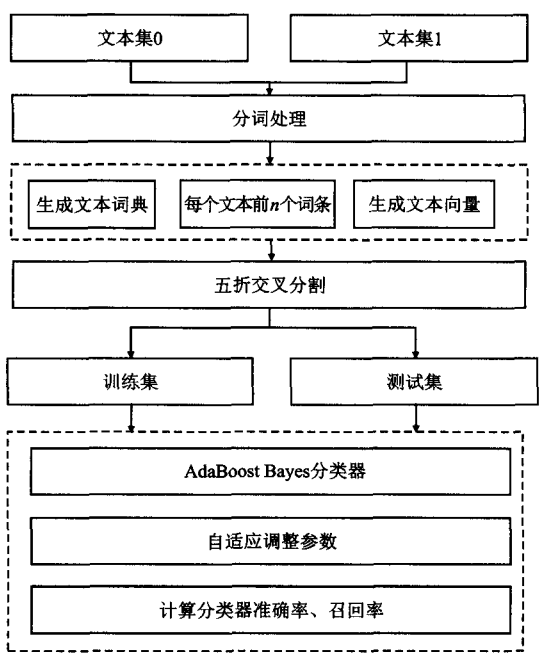

图 1 adaBoost-Bayes 中文文本分类算法流程

关于算法的说明如下所示.

分词算法：使用 HMM 分词算法，结合已经训练好的中文词典库对文本进行分词，添加停用词汇过滤不能反映文章主题特征的干扰词汇.

改进的朴素贝叶斯分类器：采用 n 折交叉对训练数据集进行分割，对分割后的语料结合朴素贝叶斯算法生成训练概率参数. 通过阈值过滤掉效果不好的分类器. 记录优化后的分类器生成分类器字典.

使用自适应提升算法：遍历分类器字典，采用 AdaBoost 算法组合分类器，用符号函数对结果进行计算，以及提升后的分类结果.

3 算法介绍

3.1 中文分词

基于人民日报语料生成中文词典库[4]，采用 Trie 树结构实现高效的词图扫描，生成句子中可能成词构成的有向无环图 DAG.

采用动态规划查找待分词材料中已经切分好的词语，计算词典中出现频率最小词语的频率作为该词频率. 最后得到最大概率路径，得到最大概率的切分组合.

对于词典中没有的词，使用 SMEL 序列组合成词的算法. 按照 SMEL 四个状态（即 S 开头，M 中间，E 结尾，L 独立）来标记. 得到一个概率最大的 SMEL 序列.

设置停用词汇，对文本中干扰词进行删除，不计入查找词汇. 最后统计文本词汇数量，对词汇频数进行排序，得到排序后的分词列表. 中文分词算法如下所示.

```
Input:    Chinese text,partitionNum
Output:   Partition word
1:        load(text,partitionNum)
2:        if(word not in stop_words)
3:            for(word in wordSet)
4:                word Number+1
5:        return list of Word(wordSet,partitionNum)
```

实际测试分词算法得到很好的效果，随机选择搜狗语料库财经类文章 090. txt，用文本编辑器打开后发现里面很多换行类字符，对语料产生干扰. 停用该类字符，分词后统计出该篇文章出现频率最多的前五个词汇：大盘，个股，牛市，涨幅，短线. 由结果可见分词器能很好地对中文文本进行分词和词频统计，规避干扰字符和无意义词汇.

3.2 贝叶斯分类器

朴素贝叶斯基于贝叶斯定理与特征条件独立假设，对于给定的训练数据集，首先基于特征条件独立假设学习输入输出的联合概率分布. 然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率

最大的输出 y[5].

朴素贝叶斯法对条件概率分布做出了条件独立的假设. 条件独立的假设：

$$P(X=x \mid Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)} \mid Y=c_k)=\prod_{i=1}^{n}P(X^{(i)}=x^{(i)} \mid Y=c_k)$$

朴素贝叶斯算法的参数估计

$$P(Y=c_k \mid X=x)=\frac{P(X=x \mid Y=c_k)P(Y=c_k)}{\sum_k P(X=x \mid Y=c_k)P(Y=c_k)}$$

对训练数据集进行朴素贝叶斯参数计算，算法如下所示.

```
Input:    trainMatrix, trainClasses
Output:   p0Vect, p1Vect, pClass
1:        pClass = sum(trainClasses)/numTrainDocs
2:        for i in range(numTrainDocs):
3:            if trainClasses[i] == 1:
4:                p1Num += trainMatrix[i]
5:                p1Denom += sum(trainMatrix[i])
6:            else:
7:                p0Num += trainMatrix[i]
8:                p0Denom += sum(trainMatrix[i])
9:        p1Vect = p1Num/p1Denom
10:       p0Vect = p0Num/p0Denom
11:       return  p0Vect, p1Vect, pClass
```

对每个类别，$\sum_k P(X=x \mid Y=c_k)P(Y=c_k)$ 都相等，所以分类器最后形式可以表示为：

$$\text{BayesClass}=\text{argmax}\{P(Y=c_k)\prod_{i=1}^{n}P(X^{(i)}=x^{(i)} \mid Y=c_k)\}.$$

最终的朴素贝叶斯分类器算法如下所示.

```
Input:    trainMatix , trainClass , testMatrix
Output:   classList
1:        for i in len(testMatrix):
2:            wordVector = testMatrix[i]
3:            p1 = sum(wordVector * p1V) + pClass
4:            p0 = sum(wordVector * p0V)+1-pClass
5:            if p1 > p0:
6:                classList. append(1)
7:            else:
8:                classList. append(0)
9:        return classList
```

3.3　AdaBoost 算法

朴素贝叶斯算法具有局限性，由于假设分类特征类是条件独立的，虽然假设可以简化算法，同时也会降低分类的准确率.

提升(boosting)方法是一种常用的统计学习方法，应用广泛且有效. 在分类问题中，它通过改变训练样本的权重，学习多个分类器，并将这些分类器进行线性组合，提高分类的性能[6].

AdaBoost 提升算法基于这样一种思想：对于一个复杂任务来说，将多个专家的判断进行适当的综合所得出的判断，要比其中任何一个专家单独的判断好. 给定一个训练样本集，求弱分类器要比求强分类器容易. 提升方法就是从弱学习算法出发，反复学习，得到一系列弱分类器，然后组合这些弱分类器，构成一个强分类器.

这样，对提升方法来说，有两个问题需要解决：一是在每一轮如何改变训练数据的权值或概率分布，二是如何将弱分类器组合成一个强分类器. 第一、提高那些被前一轮弱分类器错误分类样本的权值，降低那些被正确分类样本的权值. 这样一来，那些没有得到正确分类的数据，由于其权值的加大而受到后一轮的弱分类器的更大关注. 第二、加大分类误差率小的弱分类器的权值，使其在表决中起较大的作用，减小分类误差率大的弱分类器的权值，使其在表决中起较小的作用.

AdaBoost 算法理论思想如下[7]. 假设给定一个二类分类的训练数据集：

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$

输出：分类器 .

① 初始化训练数据权值分布

$$D_1=(w_{11},\cdots,w_{1i},\cdots,w_{1n}),w_{1i}=\frac{1}{n},i=1,2,\cdots,n$$

② 计算 $G_m(x)$ 在训练数据集上的分类误差率

$$e_m=P(G_m(x_i)\neq y_i)=\sum_{i=1}^{n}w_{mi}I(G_m(x_i)\neq y_i)$$

③ 由前向分步算法可以推导出 $G_m(x)$ 系数训练数据集 T 上的指数损失最小的系数

$$\alpha_m=\frac{1}{2}\log\frac{1-e_m}{e_m}$$

④ 更新训练数据集的权值分布

$$D_{m+1}=(w_{m+1,1},\cdots,w_{m+1,i},\cdots,w_{m+1,n})$$

$$w_{m+1,i}=\frac{w_{mi}}{Z_m}\exp(-\alpha_m y_i G_m(x_i)),i=1,2,\cdots,n$$

这里的 Z_m 是规范化因子

$$Z_m = \sum_{i=1}^{n} w_{mi} \exp(-\alpha_m y_i G_m(x_i))$$

⑤ 最终分类器

$$G(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

以下算法是 AdaBoost-Bayes 算法实现：

```
Input:    trainMatrix, trainClasses, testMatrix
Output:   predictValueMat
1:        if weightedError < minError:
2:            minError = weightedError
3:            alpha =0.5 * log((1.0-minError)/minError)
4:            bestBayes['alpha'] = alpha
5:            weakClassArr.append(bestBayes)
6:        classEst = Algorithm3(bestBayes, testMatrix)
7:        aggClassEst+=weakClassArr[alpha] * classEst
8:        predictValueMat = sign(aggClassEst)
9:        return  predictValueMat
```

4 实验设计与结果分析

4.1 实验环境和数据集

本实验用 PC 配置如下：处理器：Intel(R) Core (TM) CPU i5-4210H CPU @ 2.9 GHz；内存 (RAM)：8.0 GB；操作系统：Windows 8.1 企业版 64 位.

本实验采用搜狗语料库进行测试[8]，语料库地址 http://www.sogou.com/labs/dl/c.html. 搜狗语料库是中文文本分析常用的一个数据集. 该数据集是大量经过人工整理分类的新闻语料. 包括几十个类别的语料，约为十万篇文档的规模[9].

本文采用的算法可以应用在二分类的问题上，所以从搜狗语料库中选取了两个类别的文章进行实验，分别是 C000007(汽车类)和 C000008(财经类). 两类文章分别有 8 000 篇，总共 16 000 篇中文文本语料.

4.2 实验流程和结果分析

为了生成最佳的词汇向量，需要对数据进行实验测试. 随机选取 n 篇文本进行测试，采用朴素贝叶斯分类器进行训练分类，进行多次实验测试在不同分词个数情况下准确率，得到不同分词数分类准确率对比如图 2 所示.

从图 2 中可以看到，当文本单词个数为 5 时，文本识别的准确率就开始上升到 80%左右，而当文本单词个数到了 10 个之后，准确率就在 90%上下波动；并没有随着词数的增加一直保持增加. 所以选择合适的分词个数，可以在保证分类准确率的基础上，尽量减少向量维数，提高分类算法性能.

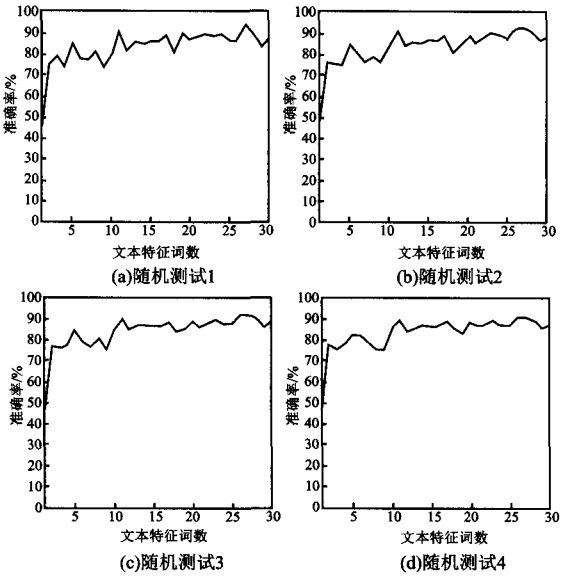

图 2 不同分词数分类准确率对比

使用朴素贝叶斯作为 AdaBoost-Bayes(ABNB) 算法的对比组进行实验. 对比四种不同算法错误率，随机选取 100 到 2 000 个语料集进行测试，采用五折交叉验证方式，进行多次实验求取准确率和召回率的均值，得到文本集数量为 1 000 时的算法准确率，召回率和 F1 值如表 1 所示.

表 1 1 000 个文本时算法对比

算法	准确率/%	召回率/%	F1 值
ABNB	0.978 7	0.901 2	0.938 4
NB	0.923 5	0.821 4	0.869 5
SVM	0.940 1	0.910 1	0.924 9
KNN	0.851 2	0.821 2	0.835 9

文本集数量从 100 到 2 000 时，F1 值对比图如图 3 所示. 实验表明，AdaBoost-Bayes 算法具有优良的分类效果，分类的性能很稳定，错误率基本上保持在 2%上下. F1 值也明显优于其他算法.

5 结束语

中文的文章主题分析与分类新方法在当今海量数据的背景下显得尤为重要，本文提出 SMEL 序列组合成词的算法，在此基础上，采用朴素贝叶斯分类器结合自适应提升算法设计中文文本分类器[10]. 该算法具有显著优点：第一，分类准确率高，通过较小的样本训练，提取出特征经过加权得到更精确分类效果；第二，该算法在样本数量增加过程中，使用基

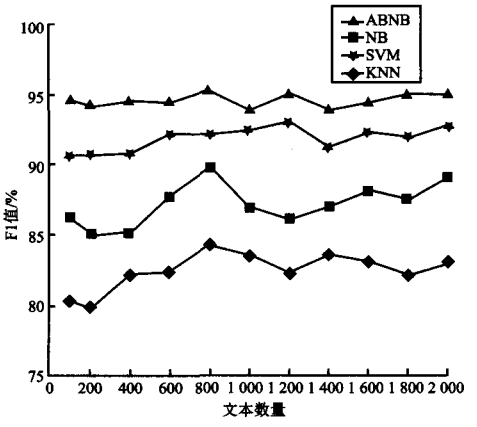

图 3　算法 F1 值对比

于朴素贝叶斯的高效分类算法结合自适应提升算法，保证了 F1 值的平稳. 实验结论表明分类器的设计合理，有优良的分类性能. 可以广泛地应用在文章主题分析、垃圾邮件过滤、政府文档归类等应用上. 本文有些地方还需要更做进一步的研究，比如机器学习技术在分词库建立的应用，多分类系统的设计，还有对海量数据进行处理技术. 这些将会在后续内容中进一步探讨.

参考文献：

[1] 靳小波. 文本分类综述[J]. 自动化博览，2006(增刊 1)：24-29.

[2] 苏金树，张博锋，徐昕，基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859.

[3] 芦苇，彭雅. 几种常用文本分类算法比较与分析[J]. 湖南大学学报：自然科学版，2007，34(6)：67-69.

[4] 彭瑜. 基于语法的分词系统的设计与实现[D]. 成都：电子科技大学，2013.

[5] Eyheramendy S, Lewis D, Madigan D. On the naive bayes model for text categorization[C]// The 9th International Workshop on Artificial Intelligence and Statistics. USA, Key West, 2003

[6] 于玲，吴铁军. 集成学习：Boosting 算法综述[J]. 模式识别与人工智能，2004，17(1)：52-59.

[7] 李航. 统计学习方法[M]. 北京：清华大学出版社，2012：75-90.

[8] 李湘东，曹环，黄莉. 文本分类中训练集相关数量指标的影响研究[J]. 计算机应用研究，2014，30(11)：3324-3327.

[9] 陈飞. 基于条件随机场方法的开放领域新词发现[J]. 软件学报，2013，24(5)：1051-1060.

[10] 杨长春. 基于文本相似度的微博网络水军发现算法[J]. 微电子学与计算机，2014，31(3)：82-85.

作者简介：

徐　凯　男，(1987-)，硕士研究生. 研究方向为推荐系统、数据挖掘. E-mail：504087493@qq. com.

陈平华　男，(1967-)，教授. 研究方向为云计算、Web 挖掘、推荐系统.

刘双印　男，(1977-5)，博士，教授. 研究方向为智能计算、智能信息处理、物联网.

(上接第 62 页)

[7] 毛力，樊养余，王慧琴，等. 基于 PSO-BF 优化算法的关系数据库水印算法[J]. 计算机应用研究，2014，31(5)：1484-1487.

[8] Shen H, Zhang M. Bacterial foraging optimization algorithm with quorum sensing mechanism[C]// Applied Mechanics and Materials. Singapore, 2014 (556): 3844-3848.

[9] Wan M, Li L X, Xiao J H, et al. Data clustering using bacterial foraging optimization[J]. Journal of Intelligent Information Systems, 2012, 38(2): 321-341.

[10] 刘丽轻，丁巧林，张铁峰，等. 数据预处理方法对模糊 C 均值聚类的影响[J]. 电力科学与工程，2011，26(8)：24-27.

作者简介：

闫　婷　女，(1990-)，硕士研究生. 研究方向为人工智能、医学信息处理.

谢红薇(通讯作者)　女，(1962-)，博士，教授. 研究方向为人工智能、医学信息学. E-mail：xiehongwei@tyut. edu. cn.