

文章编号: 1000-5013(2011)04-0401-04

# 一种改进的朴素贝叶斯文本分类方法

陈叶旺, 余金山

(华侨大学 计算机科学与技术学院, 福建 泉州 362021)

**摘要:** 针对网络中所存在的大量以网页等非结构化形式存在的文本资源, 提出一种改进的朴素贝叶斯分类方法. 首先, 通过卡方检验方法求文档特征并对文档降维, 提高特征词区分性信息; 然后, 以文本特征来代替原始词条进行朴素贝叶斯对类. 实验表明, 该方法不仅理论上易于建立和更新, 而且分类的精确率也得到提高.

**关键词:** 文本分类; 朴素贝叶斯方法; 文档特征; 卡方检验

**中图分类号:** TP 311.13 **文献标志码:** A

文本挖掘中最基本的两项工作就是分类和聚类, 几乎在所有文本挖掘的应用领域都离不开文本的分类和聚类<sup>[1]</sup>. 文本分类是文本挖掘的一个重要内容, 是指按照预先定义的主题类别, 为文档集中的每个文档确定一个类别. 通过自动文本系统把文档进行归类, 可以帮助人们更好地寻找需要的信息和知识. 随着文本信息的快速增长, 特别是 Internet 上在线文本信息的激增, 文本自动分类已经成为处理和组织大量文档数据的关键技术. 与此同时, 人们对于内容搜索的准确率、查全率等方面的要求会越来越高, 因而对文本分类技术需求大为增加, 如何构造一个有效的文本分类系统仍然是文本挖掘的一个主要研究方向. 近年来, 国内外研究人员对文本分类问题进行深入研究, 他们采用很多不同方法来构造分类器<sup>[2-6]</sup>. 在文本分类系统中, 文本被表示成一个文本特征向量, 文本特征用词来表示. 即文本表示采用 BOW 模型. 目前, 大多数文本分类系统都是使用这种文本特征表示方法等. 本文主要是以改进的朴素贝叶斯方法来实现资源分类.

## 1 基于文档词汇的朴素贝叶斯粗粒度分类

文本分类中常用的统计方法是利用文本的概率模型, 其基本思想是利用词和文本的联合概率估计文本所属类别的概率. 朴素贝叶斯假设文本是基于词的 Unigram 模型, 即文本中词的出现依赖于文本类别, 但不依赖于其他词及文本的长度. 也就是说, 词与词之间相互独立的. 因而在对文本进行分类前需要对文本进行分词.

分词工具主要基于中文基本词库和专业词库, 其词库可动态变换和加载. 如对于一段与农业有关的文本“黄瓜的叶子发霉有小黑点”, 经过处理和分词后, 可以得到的词汇集合为{黄瓜, 叶子, 发霉, 有, 小, 黑点}. 中文词库采用联合国粮食及农业组织(FAO)的中文农业叙词表和中文基本词库, 词汇数量分别为 37 060, 119 850 个, 前者优先于后者. 经过分词处理后, 按全概率理论和贝叶斯定理有

$$P(c \mid d) = \frac{P(c) \times P(d \mid c)}{P(d)}.$$

(1)

式(1)中:  $c$  为类别;  $d$  为一个文档, 分解为一个词汇向量  $d = (\omega_1, \omega_2, \dots, \omega_k)$ ;  $P(d)$  可以认为是一个常数, 在分类过程中不起作用;  $P(c)$  为文档属于这个类别的先验概率,  $P(c) = |c| / |D|$ ,  $|c|$  为类别为  $c$  的训练文档的数量,  $|D|$  为训练集文档总数;  $P(d \mid c) = \prod_{i=1}^{i=k} P(\omega_i \mid c)$ ,  $P(\omega_i \mid c)$  为词汇  $\omega_i$  在训练文档中属

于类别  $c$  的概率. 即有

$$P(w_i | c) = \frac{\sum_{j=1}^{j=|c|} t(w_{i,j})}{\sum_{m=1}^{m=|c|} l(c_m)}. \tag{2}$$

式(2)中: $t(w_{i,j})$ 为词汇  $w_i$  在第  $j$  个训练文档中出现的次数; $\sum_{m=1}^{m=|c|} l(c_m)$ 为类别为  $c$  的训练文档的总长度; $|c|$ 为类别为  $c$  的训练文档的数量. 综合式(1),可以得到

$$P_w(c | d) = P(c) \times \prod_{i=1}^{i=k} P(w_i | c) = \frac{|c|}{|D|} \times \prod_{i=1}^{i=k} \frac{\sum_{j=1}^{j=|c|} t(w_{i,j})}{\sum_{m=1}^{m=|c|} l(c_m)}. \tag{3}$$

式(3)中: $P_w(c|d)$ 为基于词汇统计的朴素贝叶斯概率.

如上所述,基本贝叶斯分类法对文档中出现的所有词汇进行统计. 然而,当需进行分类文档的数量较大时,其词汇向量往往达到数十万,多数词汇的  $P(w|c)$  相当小几乎为 0,可以看成是一个巨型稀疏矩阵. 因而可以通过一些方法进行必要过滤,以大量减少不必要运算.

## 2 基于文档特征的朴素贝叶斯粗粒度分类

为进行有效过滤,需先对文档做特征选择,然后根据文档特征进行概率统计,以达到降维效果. 卡方 (Chi-Square) 检验的主要思想是: 词条与类别之间符合  $\chi^2$  分布, 词条的  $\chi^2$  统计量表示词条对某个类别的贡献大小. 统计量越高, 词条和类别之间的独立性越小、相关性越强, 即词条对此类别的贡献越大.

特征选择的方法是  $\chi^2$  统计值<sup>[7]</sup>. 即在所有训练文档中,对所有与类别  $c$  相关的词汇  $t$  按  $\chi^2$  值进行排序,有

$$\chi^2(t, c) = \frac{N \times (A \times D - C \times B)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}. \tag{4}$$

式(4)中: $t$  为词汇; $c$  为文档分类; $N$  为训练文档总数; $A$  为在所有属于  $c$  类的训练文档中  $t$  出现的次数; $B$  为在所有不属于  $c$  类的训练文档中  $t$  出现的次数; $C$  为所有属于  $c$  类但没有  $t$  出现的训练文档数; $D$  为所有即不属于  $c$  类也没有  $t$  出现的训练文档数.

对于所有的  $\chi^2$  值,选定一个阈值  $H$ ,以获得一个词汇集合  $F = \{t | \chi^2(t, c) > H\}$ ,并以集合  $F$  中的所有词汇来作为类别  $c$  的特征. 那么,将式(1)中的  $P(d|c)$  按文档特征来计算概率值,则为

$$P(d | c) = \prod_{i=1}^{i=k} P(f_i | c) = \prod_{i=1}^{i=k} \frac{\sum_{j=1}^{j=|c|} t(w_{i,j})}{\sum_{m=1}^{m=|c|} l(c_m)}. \tag{5}$$

式(5)中: $f_i$  为文档  $d$  中出现的属于类别  $c$  的第  $i$  个特征值, $k$  为特征总数; $\sum_{j=1}^{j=|c|} t(w_{i,j})$  为特征  $f_i$  在所有类别为  $c$  的训练文档中出现的次数; $P(f_i|c)$  为特征  $f_i$  属于类别  $c$  的概率. 综合式(1),则有

$$P_f(c | d) = P(c) \times \prod_{i=1}^{i=k} P(f_i | c) = \frac{|c|}{|D|} \times \prod_{i=1}^{i=k} \frac{\sum_{j=1}^{j=|c|} t(w_{i,j})}{\sum_{m=1}^{m=|c|} l(c_m)}. \tag{6}$$

式(6)中: $P_f(c|d)$ 为基于特征统计的朴素贝叶斯概率.

由于采用概率分类,一个文档  $d$  可以同时属于两个以上分类,即取其按概率值排序的前  $N$  个类别作为文档  $d$  的分类.

## 3 评测实验与分析

在 Java 环境下,使用 Eclipse 作为开发平台,实验主要分为相对独立的两步.

(1) 基于词汇统计的朴素贝叶斯和卡方文本特征选择. 使用 Weka 开源软件包中提供的相应算

法,对朴素贝叶斯稍作修改,使其按输入特征值来做自动分类.从 3 类文档集中选出一部分做为训练集,并按训练集相应本体元知识做简单的人工分类,如表 1 所示.

为简化工作,按这些文档资源所在网站的分类作为人工分类结果,除去作为训练种子的文件,剩下的都用来作为测试数据集.对于所有经过粗粒度分类的文档,按式(5)取  $N=1$ ,即取最大概率值作为一个文档的自动分类结果.通过这个自动分类结果与原先文档所处的人工分类作比较,得到查准率( $R_p$ )和查全率( $R_r$ ).

(2) 对于粗粒度分类正确的结果,选出其中一部分,再用本体实例来进标注.

用贝叶斯分类器对表 1 中的 3 种文档集进行分类实验,分别取了不同数量的训练语料来进行测试,结果如图 1 所示.从图 1 的结果可以看出,随着训练语料的增多,分类效果就越来越好;但到一定程度后,训练语料的规模对分类效果的影响不大.

对于两种贝叶斯分类方法的实验测试,结果如表 2 所示.表 2 中: $n$  为平均每个文档特征数.经过  $\chi^2$  阈值  $H$  的调整,两种贝叶斯方法的查准率( $R_p$ )和查全率( $R_r$ )相差不多,与文[7]报告的结果接近,说明少数词汇对文本分类起到关键作用.

两个方法时间开销,如表 3 所示.从表 3 可知,训练时间改进方法时间开销有所增加.因为要多做卡方值计算,经过算法优化后,卡方时间开销随文档数量增加而平缓增长,如表第 2 列与第 4 列所示;然而,测试时间却达到一个数量级减少,如表第 3 列与第 5 列所示.这说明基于特征统计的贝叶斯方法实现了较好的时间性能.

表 1 训练文档集分类

Tab. 1 Training documents classification

语料库名称	训练文档数量	类数目	词汇总量
花卉知识	800	4	794 601
新浪国际足球新闻	800	4	1035 783
农作物病虫害知识	800	4	854 823

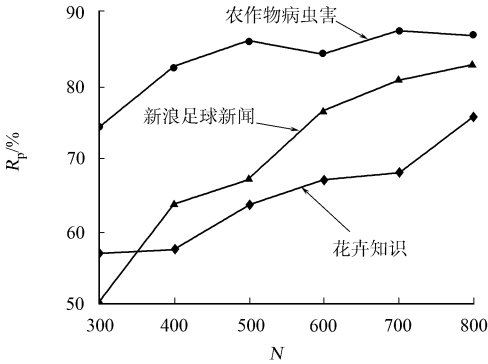


图 1 训练语料的规模对分类结果的影响

Fig. 1 Effect of training corpus size on the results of classification

表 2 两种文档分类方法的测试结果统计

Tab. 2 Statistics of the test results for both document classification

语料库名称	基于词汇贝叶斯		基于特征贝叶斯			
	$R_p/\%$	$R_r/\%$	$R_p/\%$	$R_r/\%$	$H$	$n$
花卉知识	78.3	78.9	85.7	77.2	2.34	16
新浪国际足球新闻	81.8	82.1	89.4	79.9	1.57	34
农作物病虫害知识	82.3	81.5	88.8	80.5	1.57	30

表 3 两种文档分类方法的时间开销统计

Tab. 3 Time overhead statistics for both document classification

ms

语料库名称	基于词汇贝叶斯		基于特征贝叶斯	
	训练时间	测试时间	训练时间	测试时间
花卉知识	254 752	34 640	227 533	9 021
新浪国际足球新闻	309 801	43 172	297 632	11 212
农作物病虫害知识	247 542	71 212	281 704	10 413

4 讨论

以上结果表明,使用本文方法进行分类,具有较高的查准率和查全率.方法的效率主要受以下 3 个方面因素的影响.(1) 本体知识本身质量.包括知识表达方式、内容全面性;(2) 文档质量.包括文档内容文字表达、段落排版、有无错别字、文档格式等;(3) 文档解析器质量.若解析器不能正确解析文档内容,则语义标注无从谈起.

使用系统中的几种文档解析器,分别解析一定量的相应格式的文档,提取文档内容,再进行对比.对

比方式是人工把文档内容提出来,与解析器提出的内容进行字符串比较.结果表明,html 和 xml 解析器解析文档质量较好,其平均解析准确度分别为 87.5%,89.3%,基本上能抓取出文档主要内容.doc 解析器次之,其平均解析准确度为 79.4%.这是因为提取不出的 word 中存在一些特殊字符,图、表格式,或者可能应为经过加密而不能打开等原因.pdf 解析器解析效果较差,其平均解析准确度只有 48.6%.主要原因是一些 pdf 文档质量不是很好,其特殊的排版格式和编码方式也造成解析困难.但是,经过 latex 和 word 转化而成的 pdf 文档同样能有较好的解析结果,一般能达到 doc 解析器的水平.因此,本系统解析器品质有待提高.

文本分类是文本挖掘的一个重要内容,是指按照预先定义的主题类别,为文档集合中的每个文档确定一个类别.通过自动文本系统把文档进行归类,可以帮助人们更好地寻找需要的信息和知识.文中提出的基于贝叶斯分类的改进方法不仅理论上易于建立和更新,而且分类的精确率也得到了提高.

参考文献:

[1] 喻小光,陈维斌,陈荣鑫.一种数据规约的近似挖掘方法的实现[J].华侨大学学报:自然科学版,2008,28(3):370-374.

[2] SEBASTIANI F. Machine learning in automated text categorization[J]. ACM Computing Surveys,2002,34(1):1-47.

[3] HAO Li-li,HAO Li-zhu. Automatic identification of stop words in Chinese text classification[C]// Proceedings of the 2008 International Conference on Computer Science and Software Engineering. Washington D C;IEEE Computer Society,2008:718-722.

[4] LEWIS D D,RINGUETTE M. A comparison of two learning algorithms for text categorization[C]// Third Annual Symposium on Document Analysis and Information Retrieval. Las Vegas:[s. n.],1994:81-93.

[5] YANG Yi-ming,LIU Xin. A re-examination of text categorization methods[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM Press,1999:42-49.

[6] 黄萱菁,吴立德,石崎洋之,等.独立于语种的文本分类方法[J].中文信息学报,2000,14(6):1-7.

[7] YANG Yi-ming,PEDERSEN J O. A comparative study on feature selection in text categorization[C]// Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco:Morgan Kaufmann Publishers Inc,1997:412-420.

An Improved Text Classification Method Based on Bayes

CHEN Ye-wang, YU Jin-shan

(College of Computer Science and Technology, Huaqiao University, Quanzhou 362021, China)

**Abstract:** There are huge amount of unstructured text resources in internet, a refined Naïve Bayes based text categorization method is proposed in this paper for classifying these resources. Firstly, this method refines text by calculating the features of the text in order to improve the text's recognizability, and then Naïve Bayes is used to classify these resources based on these features instead of the original words. The experiments show that the new method is easy setting up and renew in theory, and the accurate rate of the classification is also improved.

**Keywords:** text categorization; Naïve Bayes; text feature; Chi-Square test

(责任编辑:钱筠 英文审校:吴逢铁)