

基于改进的贝叶斯模型的中文网页分类器

秦兵 郑实福 刘挺 张刚 李生
(哈尔滨工业大学信息检索研究室, 哈尔滨 150001)
{qinb,zsf,tliu,bert,sli}@ir.hit.edu.cn

摘要

本文提出了一种改进的基于统计的中文网页的分类算法,通过对传统的基于计算相似度文本分类方法和基于贝叶斯模型文本分类算法的研究,我们对贝叶斯模型分类算法进行了改进,提出了利用类别密度函数似然比来增加特征词的可分性信息的算法。通过对计算相似度方法,贝叶斯方法及改进的贝叶斯方法对比实验表明,改进算法可以最大化类与类的间隔,因而具有较高的分类精确率和召回率。

关键字: 似然比, 贝叶斯方法, 文本概率模型, 网页分类

An Improved Bayes Classifier for Chinese Web Pages

Qin Bing Zheng ShiFu Liu Ting Zhang Gang Li Sheng
(Information Retrieval Laboratory, Harbin Institute Technology, Harbin, 150001)
{qinb,zsf,tliu,bert,sli}@ir.hit.edu.cn

Abstract

This paper presents a modified statistic Chinese web page classification algorithm. Similarity based method and Bayes model based method are the popular approaches for text classification. In this paper, we modified Bayes model method, then uses the probability likelihood ratio of each class to increase the separability of feature words vectors. In the following method: Similarity based method, Bayes model based method, modified Bayes model based method. The experiment shows that among these methods, the modified algorithm not only presents the maximum distance between classes, but also improves precision and recall.

Keywords: likelihood ratio, Bayes model, text probability model, web page classification

1. 引言

随着在线信息的迅速增长,对信息获取的速度和数量要求越来越高,使得文本分类成为日益重要的研究领域。文本分类技术可以用在许多方面,信息过滤,建立个人信息档案,在网上快速定位各类需要的信息等。

文本分类是将未知文本类别的文本自动划分到其所属的类别中。传统上,是由人来完

成的，如果是用手工来将不断增加和变化的文本加入到主体分类中，将是非常费时费力的，特别是 internet 上在线信息的增加，使自动文本分类作为信息处理领域的一个重要研究课题，成为了当前的一个研究热点。

在目前的自动文本分类方法主要有两种：一种是类似于专家系统的基于规则方法，通常需要手工建立规则库，并且修改起来非常困难。另一种是归纳学习方法，在已知类别的训练集基础上自动地建立分类器，采用归纳学习方法建立的分类器具有许多优势，分类器易于建立和更新。目前的文本分类研究中应用了许多统计分类和机器学习技术，如相关反馈，贝叶斯概率模型，神经网络，决策树，最近邻法，支持向量机等。本文在传统的计算相似度和贝叶斯模型进行分类的过程中，提出了一种改进的贝叶斯模型，即将特征词类别间似然比与贝叶斯方法相结合，提高了特征词区分性信息，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了提高。

本文其余部分组织如下：第二节文本分类中的贝叶斯方法及改进算法；第三节系统实现步骤；第四节实验结果；第五节结论。

2. 文本分类中的贝叶斯方法及改进算法

应用最广泛的文本分类的学习算法是在信息检索中发展起来的计算相似度的方法，在文本分类中是通过训练集学习，构成类别向量 C_j ，然后以向量的内积计算待分类向量 d' 与类别向量距离，对每一类都由一个特征向量 C_j ，由带类别标记训练集的得到的。用这些向量对待分文本 d' 分类，待分文本也表示成向量 d' ，然后分别计算与各类特征向量的余弦， d' 将属于余弦值最大的类别。见公式

$$H(d') = \arg \max_{C_j \in C} \cos(\overrightarrow{C_j}, \overrightarrow{d'}) \quad (1)$$

文本分类中常用的统计方法是利用文本的概率模型，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，纯粹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，词于词之间是独立的。具体算法如下：设集合 X 代表文本的内容，其中元素 x 表示训练文本 d 包含 x 的概率即为 $p(x|d)$ 通过全概率理论和贝叶斯理论，我们可以得到：

$$\begin{aligned} \Pr(C_j | d) &= \sum_{x \in X} \Pr(C_j | x, d) \Pr(x | d) \\ &= \sum_{x \in X} \frac{\Pr(d | C_j, x)}{\Pr(d | x)} \Pr(C_j | x) \Pr(x | d) \end{aligned} \quad (2)$$

为了便于处理假设 X 提供的关于 d 的信息足够多，包含 C_j 类别信息，即 C_j 的信息是冗余的，可以认为 $\Pr(d | C_j, x) = \Pr(d | x)$

则 (2) 简化为：

$$\Pr(C_j | d) \approx \sum_{x \in X} \Pr(C_j | x) \Pr(x | d) \quad (3)$$

对于 X 集的选择，元素 x 是通过 n -gram 模型计算的得到的，即 x 的概率是与特征集 F 前 n

个词的相关的，从中我们可以看到， n 的大小决定了分类算法的质量和精确度，在这里我们采用最简单的形式 $n=1$ 作为开始，可以想象，若 $n=|d|$ ， $\Pr(C_j|d)=\Pr(C_j|x)$ ，精确度将很高，但计算起来将非常复杂，我们采用纯粹贝叶斯假设，即词与词之间是独立的，即 $n=1$ 作为研究的起点，则(3) 就被简化为

$$\Pr(C_j | d) \approx \sum_{\omega \in F} \Pr(C_j | \omega) \Pr(\omega | d) \quad (4)$$

其中： $\Pr(w|d)$ 用训练文本 d 中 w 出现的次数来表示， $\Pr(\omega | d) = \frac{TF(\omega, d)}{\sum_{\omega' \in F} TF(\omega', d)} = \frac{TF(\omega, d)}{|d|}$

$\Pr(C_j|w)$ 表示当文本中有词 w 出现时，文属于第 j 类的概率。通过训练集由贝叶斯公式得到后验概率：

$$\Pr(C_j | \omega) = \frac{\Pr(\omega | C_j) \Pr(C_j)}{\sum_{C' \in C} \Pr(\omega | C') \Pr(C')} \quad (5)$$

其中：由训练语料的文本数我们可以得到 C_j 先验概率和后验概率 $\Pr(w|C_j)$ 的值：

$\Pr(C_j) = \frac{|C_j|}{|D|}$ ，其中 $|C_j|$ 为训练集中某一类的文档数， $|D|$ 训练集的总文本数

$\Pr(\omega | C_j) = \frac{DF}{|C_j|}$ ， DF 为 C_j 类中出现 w 的次数， $|C_j|$ 为训练集中该类的文本数,我们可以得

到文本类别贝叶斯判别式为：

$$H_{PrTFIDF}(d') = \arg \max_{C_j \in C} \sum_{\omega \in F} \frac{\Pr(\omega | C_j) \Pr(C_j)}{\sum_{C' \in C} \Pr(\omega | C') \Pr(C')} pr(\omega | d') \quad (6)$$

在对上述方法的实现中，文本特征均以 DF 来表示，通过我们实验表明， DF 是特征表示法中的一个简单、费用较低的工具，其效果与其它方法相当。

获取 DF 的过程，首先除去停用词，然后统计词频，这里采用的词频即某一个词 w 出现在该类文本中个数，给定阈值后，词频大于阈值的词作为该类文本的特征，形成特征向量 F 。在对篮球和足球分类中我们发现，对于不易分别的文本，会在两类特征向量中由很多相似的特征，且在两类文本中出现的概率均衡，区分性信息不好，如“得分”，在篮球和足球中都会经常出现，但仅通过这个词，我们不能文本属于哪一类，而两类中出现词频差别较大的词，即使词频较低，也具有很强的区分信息，因此我们提出了一种改进的分类方法，在贝叶斯分类的基础上，对区分性好的词增加权重，对分类性差的词降低权值，具体方法是：用出现的词频作第一次的排序，再用似然比对其分类性能作第二次处理，即对每一项加权 $\Pr(w|c_j)/\Pr(w|c_i)$ ，对两类问题，我们可以用两类密度函数似然比作因子，提高平均可分信息，推广到多类问题，我们可以用某一类的密度函数与各类密度函数的均值进行比较，对于在各类中表现均衡的词，即出现概率相近的词，得到权值的很小，对在各类中出现概率差别较大的词，其权值较大，从而提高了可分性信息。改进后的文本分类判别公式如下：

$$H_{PrTFIDF}(d') = \arg \max_{C_j \in C} \sum_{\omega \in F} \frac{\Pr(\omega | C_j)}{\Pr(\omega | C')} \times \frac{\Pr(\omega | C_j) \Pr(C_j)}{\sum \Pr(\omega | C') \Pr(C')} pr(\omega | d') \quad (7)$$

3. 系统实现

整个系统的实现是在 Linux 下用 KDevelop1.2 开发的。系统的流程如下图所示：

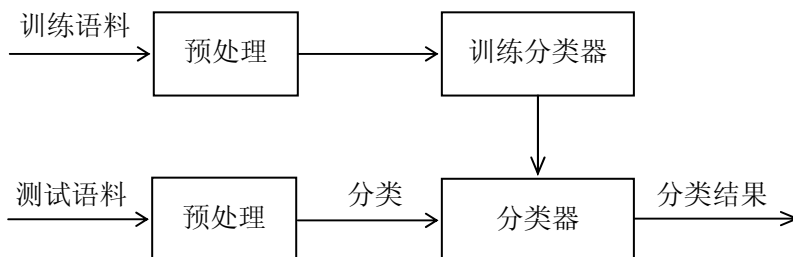


图 1 系统流程图

系统包括两方面：训练部分和测试部分。训练和测试都包含预处理过程。具体过程如下：

由于我们的语料都是从 Internet 上下载的网页，格式多种多样，为了确保实验的顺利进行，首先要对语料格式进行处理，去掉网页中多余的标记。使语料满足实验要求的格式。然后再对语料进行分词，分词是预处理的主要任务，分词的效果好坏会直接影响到后面的其他处理过程。我们的已有的分词模块能够解决大部分的组合歧义和交集型歧义，同时还有较强的未登陆词的识别功能，能够很好的对人名、地名进行识别，这就为进一步的处理提供了基础。

在训练过程中，首先拿出一部分已经分好类的语料作为训练语料。把不同类的语料存在不同的目录下。然后通过统计每个词在这些语料中出现的频率 $DF/DocNum$ 来提取特征向量 (F)。DF 为文档频率，DocNum 为训练文档的总数。对 $DF/DocNum$ 取一个阈值，将所有 $DF/DocNum$ 大于这个阈值词加入特征向量中。 $DF/DocNum$ 的值越小，得到的特征词就越多。然后计算每个特征词在各类文档中出现的概率 $\Pr(\omega | C_j)$ ，同时，为了计算 $\Pr(C_j)$ 还需要记录每类文档的文档数。训练的最后结果保存在两个文件中，这两个文件的结构如图 1 所示。

表 1 训练结果文件的结构

训练结果文件一：

文档类	特征向量	$\Pr(\omega C_j)$
第一类	$\omega 1$	$DF(\omega 1)/DocNum1$
	$\omega 2$	$DF(\omega 2)/DocNum1$
	。 。 。	。 。 。
第二类	$\omega 1$	$DF(\omega 1)/DocNum2$
	$\omega 2$	$DF(\omega 2)/DocNum2$
	。 。 。	。 。 。

训练结果文件二：

文档类	文档数
第一类	DocNum1
第二类	DocNum1
。	。
。	。
。	。

在测试过程中，我们要根据训练的结果构造出一个分类器。在构造分类器的时候首先要读入训练过程中得到的两个文件，存放到一个文档链表（DocClassList）中。这个链表的结构如图 2 所示。

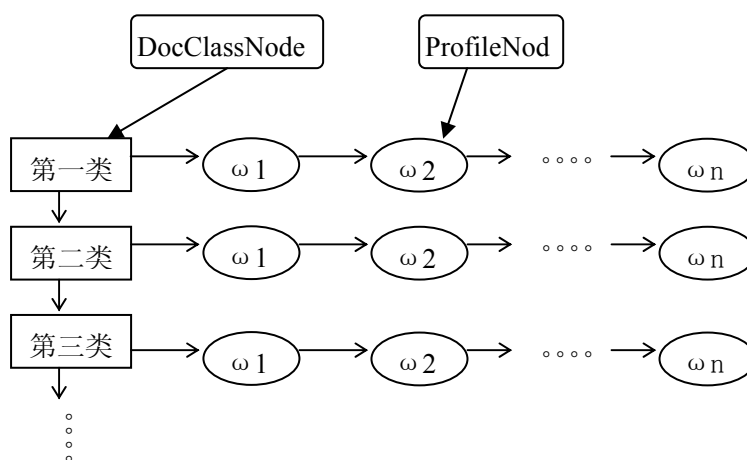


图 2 文档类链表的结构

在每个 ClassNode 中记录了文档的类型和这类文档的出现概率 $\Pr(C_j)$ 还有一个指向特征向量列表（ProfileList）的指针。在 ProfileList 的 ProfileNode 中记录了这个特征词在这类文档中出现的概率 $\Pr(\omega | C_j)$ 。当这个结构建立起来的时候，我们的分类器也就构造成功了。接下来就可以对文本进行分类。每输入一个测试文档 d' ，就要计算每个特征词在 d' 中出现的概率 $\Pr(\omega | d')$ 。然后我们根据判别式就可以计算出该文档属于各类的概率，并将测试文本判到概率最大的一类。

4. 实验结果

我们从新浪网(<http://www.sina.com.cn>)下载了大量的网页进行分类实验，由于新浪已经对这些网页进行了分类，这样我们就能方便地得到丰富的训练语料和测试语料。我们的实验主要在足球和篮球这两类问题之间进行。足球和篮球这两类问题是相对比较难分的问题，因为这两类问题中有很多相似特征。象足球和汽车这样的分类问题就简单多了。

实验一：测试训练语料的规模对分类效果的影响

我们用贝叶斯分类器对足球和篮球的分类问题进行实验。在这个实验中，分别取了不同数量的训练语料来进行测试。这些训练语料中足球类文档和篮球类文档的数量相等。实验结果如图 3 所示，从这个实验的结果我们可以看出随着训练语料的增多，分类效果就越来越好。但是到一定程度后（当训练文档多于 2000 篇时，也就是每类训练文档多于 1000 篇的时候），训练语料的规模对分类效果的影响就不大了。

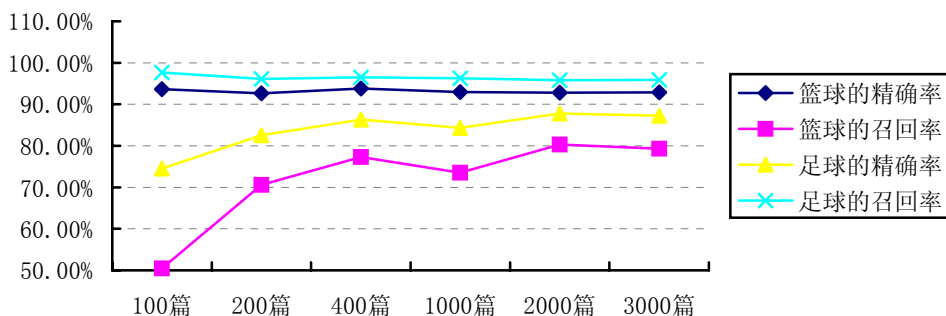


图3 训练语料的规模对分类结果的影响

实验二：DF/DocNum 的阈值对分类效果的影响

在提取特征词的时候，是由 DF/DocNum 的值来决定的。DF/DocNum 的阈值越小，提取出来的特征词就越多，那么分类效果是否也会越好呢？带着这个疑问，我们进行了如下实验：在其他条件一样的情况下，在训练的过程中我们分别取了不同的阈值，并且取得了不同的分类效果。实验结果如图 4 所示。从图中我们可以看出 DF/DocNum 的阈值越小，分类的效果就越好。但是随着 DF/DocNum 的阈值得降低，特征词的个数就急剧的上升，这样就会影响到分类的速度。所以 DF/DocNum 阈值也不能取太低，取 0.01 的时候速度已经非常慢，所以我们在速度和精度之间找一个平衡点。

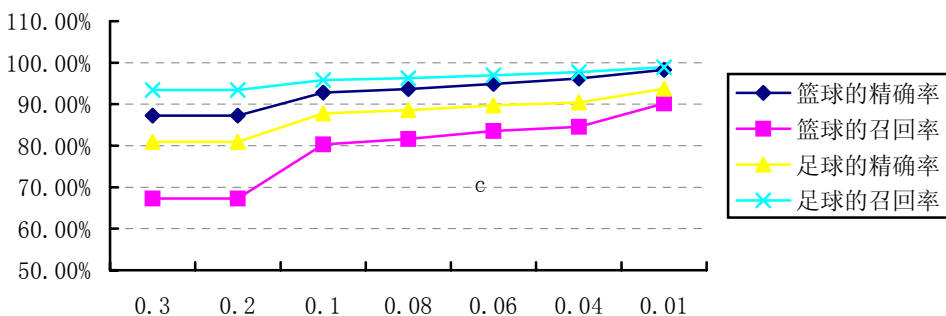


图4 DF/DocNum 阈值对分类效果的影响

实验三：三种分类方法的对比

在这个实验中我们对这三种分类方法进行了比较。训练语料中包括了 1000 篇足球文档和 1000 篇篮球文档。测试语料中有 2000 篇篮球文档和 2000 篇足球文档。DF/DocNum 的阈值为 0.1。实验结果如下表所示。从这个对比实验中我们可以看出，贝叶斯分类法优于相似度分类法，改进的贝叶斯分类法又优于贝叶斯分类法。

表 2 相似度分类法

	精确率	召回率
足球类	66.7%	71.0%
篮球类	79.1%	75.7%

表 3 贝叶斯分类法

	精确率	召回率
足球类	92.8%	80.3%
篮球类	87.8%	95.8%

表 4 改进的贝叶斯分类法:

	精确率	召回率
足球类	94.1%	93.1%
篮球类	95.3%	96.0%

5. 结论

通过上述三种分类方法的对比实验,我们看到,改进后的算法在分类精确率和召回率都优于其他两种分类器,尤其对于相似类别分类具有明显的优势。同时我们通过其他大量的实验发现,对那些很容易区分的问题(比如:足球和汽车),这几种分类器都能做到很高的精度,因而改进后贝叶斯分类法对分类效果的提高不是很明显。因为对那些容易区分的问题,各类中含有相同特征词的可能性较小,其特征词的区分性已经很好,因此加入似然比之后对分类效果的提高不是很明显。但是对那些比较难分的问题(比如:教育和科技,足球和篮球)效果是非常明显的,极大的提高了分类的准确率和召回率。

参考文献

1. Thorsten Joachims. "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization"
2. Yiming Yang and Jan O.Pedersen. "A Comparative Study on Feature Selection in Text Categorization"
3. David D. Lewis and Marc Ringuette. "A Comparison of Two Learning Algorithms for Text Categorization"
4. Yiming Yang. "An Evaluation of Statistical Approaches to Text Categorization". 1998.
5. 边肇祺, 张学工等。《模式识别》, 2000