贝叶斯文本分类器的研究与改进

史瑞芳
SHI Rui-fang

山西经济管理干部学院,太原 030024
Shanxi Economic Management Institute,Taiyuan 030024,China
E-mail:srfang0620@163.com

SHI Rui-fang.Research and improvement on Naive Bayes text classifier.Computer Engineering and Applications,2009,45(12):147-148.

Abstract: Naïve Bayes method is a simple and effective established probability categorization method at present.However,the problems on scattered data in methodology and Laplace smoothness method have some disadvantages.Therefore,author proposes to use uni-gram smoothness method to improve the condition and the effect on categorization by Bayes method.
Key words: Bayes text categorization;scattered data;smoothness

摘 要:朴素贝叶斯文本分类是目前公认的一种简单有效的概率分类方法,但该方法的数据稀疏问题以及所采用的 Laplace 平滑方法还不是最优,存在一定的缺陷。因此,用一元统计语言模型的平滑方法来改进数据稀疏状况,提高了分类效果。
关键词:贝叶斯文本分类;数据稀疏;平滑
DOI:10.3778/j.issn.1002-8331.2009.12.048 **文章编号:**1002-8331(2009)12-0147-02 **文献标识码:**A **中图分类号:**TP311

1 引言

随着 Internet 的飞速发展，越来越多的文本信息表现为电子文档的形式。面对如此庞大而且急剧膨胀的信息海洋,如何有效地组织和管理这些信息,并快速、准确、全面地从中找到用户所需要的信息是当前信息科学和技术领域面临的一大挑战。文档分类作为处理和组织大量文本数据的关键技术,可以在较大程度上解决信息杂乱的问题,方便用户准确地定位所需的信息和分流信息。因此,自动文本分类已作为一项具有较大实用价值的关键技术得到了广泛的关注,取得了很大的进展。

目前较为著名的文本分类方法有 Bayes、LLSF、SVM、KNN、决策树等。贝叶斯(Bayes)分类方法是一种最常用的有指导的方法。它以贝叶斯定理为理论基础,是一种在已知先验概率与条件概率的情况下的模式识别方法。目前,有不少的文本分类系统采用了 Bayes 算法,在邮件分类、电子会议、信息过滤等方面得到了较为广泛的应用。

2 贝叶斯分类器介绍

目前，贝叶斯分类器分两种：一种是朴素贝叶斯分类器(Naïve Bayesian Classifier)，它在很多领域都表现出优秀的性能。朴素贝叶斯分类器的“朴素”指的是它的条件独立性假设。它假设一个属性对给定类的影响独立于其他属性,即特征独立性假设。当假设成立时,与其他分类算法相比,朴素贝叶斯分类器是最精确的，但是文本属性之间的依赖关系是可能存在的。大量研究表明此时可以通过各种方法来提高朴素贝叶斯分类器的性能。另一种是贝叶斯网络分类器。可以考虑属性之间的依赖程度,其计算复杂度比朴素贝叶斯高得多,更能反映真实文本的情况。贝叶斯网络分类器实现十分复杂,目前还停留在理论的研究阶段。本文研究朴素贝叶斯文本分类器。

朴素贝叶斯分类算法有两种模型:多变量贝努里事件模型和多项式事件模型。这两种模型体现的是 $P(d/C_i)$估计的方法不同。本文采用多项式模型。

在多项式模型中,一篇文档被看作是一系列有序排列的词的集合。假定文章的长度对于给定类的影响是独立的,并且假定文档中的任何一个词与它在文中的位置以及上下文关系也是独立的。文档属于 C_j 类时特征词 w_i 出现一次的概率为 $P(w_i/C_j)$,文档中出现 x_i 次特征词 w_i 的概率为 $P(w_i/C_j)^{x_i}$,出现依这种次序排列的词的集合的概率为:$\prod_i P(w_i/C_j)^{x_i}$。

按照上面的估计,很多不同序列的特征词都会对应着同一篇文档,为了解决这个问题,这里采用多项式排列:

$$\binom{n}{n_1,n_2\cdots}=\binom{n}{n_1}\times\binom{n-n_1}{n_2}\times\cdots=\frac{n!}{n_1!\ (n-n_1)!}\times\frac{(n-n_1)!}{n_2!\ (n-n_1-n_2)!}\times\cdots=\frac{n!}{n_1!\ n_2!\ \cdots}$$

假定共有 n 个词,则 $n=\sum_i x_i$,有公式:

$$P(x/C_j)=\binom{n}{x_1,x_2\cdots}\prod_i P(w_i/C_j)^{x_i}=n!\ \prod_i\frac{P(w_i/C_j)^{x_i}}{x_i!}$$

多项式模型的参数是特征词出现的概率 $P(w_i/C_j)$,其中,

作者简介:史瑞芳(1974-),女,讲师,主要研究方向:计算机应用。
收稿日期:2008-09-25 **修回日期:**2009-01-08

$\sum_i P(w_i/C_j)=1$。

$P(w_i/C_j)$的值可以从训练中估计：

$$P(w_i/C_j)=\frac{count(w_i,C_j)}{\sum_i count(w_i,C_j)}$$

其中 $count(w_i,C_j)$表示特征词 w_i 出现在 C_j 类文档中的次数，$\sum_i count(w_i,C_j)$表示 C_j 类文档中出现的所有特征词的总次数。为了避免 $P(w_i/C_j)$等于 0，对其的估计采用 Laplace 平滑技术：

$$P(w_i/C_j)=\frac{count(w_i,C_j)+\delta}{\sum_i count(w_i,C_j)+\delta|V|}$$

其中，$|V|$表示特征词表中的总单词数，δ 可以是任意的一个非零数，常取 $\delta=1$。这种平滑技术认为在没发生任何事件之前，每个事件都有一定的发生概率 $1/|V|$。对于在某个类中没有出现的特征词，为了避免零概率会将所有的词频普加一个常数(常取 1)，将所有没有出现的词的概率视为相等的，这样会导致未出现的事件概率过高。因此，提出用一元语言模型的平滑技术 Jelinek-Mercer 对其进行改进。

3　基于统计语言模型的平滑技术

3.1　统计语言模型

当用变量 W 代表一个文本中顺序排列的 n 个词，即 $W=w_1,w_2,\cdots,w_n$，则任意词序列 W 在文本中出现的概率为 $P(W)$。如果任意一个词 w_i 的出现概率只同它前面的 $n-1$ 个词有关时的语言模型叫做 N 元语言模型(N-gram)，即为下式：

$$P(W)=P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_n|w_1w_2\cdots w_{n-1})$$

当 $N=1$ 时为一元模型(uni-gram)，即认为每个词的出现概率是独立的，$P(w_1,w_2,\cdots,w_n)=P(w_1)P(w_2)\cdots P(w_n)$，这种假设与朴素贝叶斯分类方法相同，所以用 uni-gram 模型的平滑方法来改善贝叶斯文本分类器。

3.2　uni-gram 模型的 Jelinek-Mercer 平滑方法

Jelinek-Mercer 方法围绕极大似然估计模型的线性插入法，用系数 $\lambda\in[0,1]$来控制各个模型的影响。通过乘系数$(1-\lambda)$降低可出现词的概率。公式如下：

$$P_\lambda(w_i/C_j)=(1-\lambda)P_{ml}(w_i/C_j)+\lambda P(w_i/C)$$

则改进的贝叶斯分类公式如下：

$$C_j^*=\arg\max_{C_j\in C}\{P(C_j/d)\}=\arg\max_{C_j\in C}\{P(d/C_j)P(C_j)\}=$$

$$\arg\max_{C_j\in C}\{P(d/C_j)\}=\arg\max_{C_j\in C}\{n!\prod_i\frac{P(w_i/C_j)^{x_i}}{x_i!}\}=$$

$$\arg\max_{C_j\in C}\{\frac{n!}{\prod_i x_i!}\prod_i((1-\lambda)P_{ml}(w_i/C_j)+\lambda P(w_i/C))^{x_i}\}$$

其中，所有文本的类的集合表示成 $C=(C_1,C_2,\cdots,C_j,\cdots)$，$P(C_i/d)$表示 C_i 类的后验概率，$P(d/C_i)$表示条件概率，$P(C_i)$表示先验概率，n 表示文章的总词数，x_i 表示词 w_i 在文本 d 中出现的次数，满足 $n=\sum_i x_i$。式中数学符号 $\arg\max_{C_j}$ 表示对不同的类 C_j 计算条件概率 $P(d/C_j)$的值，从而使 C_j^* 成为条件概率值最大的类。

上述公式描述的算法是用 Jelinek-Mercer 平滑方法来代替 Laplace 平滑，即用公式$(1-\lambda)P_{ml}(w_i/C_j)+\lambda P(w_i/C)$来计算 $P(w_i/C_j)$，其中有待定的参数 λ，用 λ 来调整在 C_j 类中没有出现的词 w_i 的概率分配，λ 可以在 0 到 1 之间任意取值。待定的参数 λ，究竟取什么样的值才能有较好的分类效果，这将在实验中进行分析。

4　实验及结果分析

4.1　测试方法

采用标准的数据集-新闻组(20-NewsGroup)在数据挖掘工具 WEKA 系统上进行测试。数据集分为典型类(Classic)、不同类(Different)、相同类(Same)以及相似类(Similar)几个类别的数据集合。

采用 k 分交叉评价(k-fold cross-validation)方法选择训练集和测试集：将初始样本集 T 分成 10 份 $T=\{T_1,T_2,\cdots,T_{10}\}$，选择其中 1 份作为测试集，剩余的 9 份作为训练集，这样每一份都依次轮流作为测试集，运行分类算法，最后将 10 次测试的平均值作为最终结果。采取准确率和召回率作为测试指标。

4.2　采用 Jelinek-Mercer 平滑时参数 λ 的选择

参数 λ 是用来调整在某个类中没有出现的词的概率的，取值范围是从 0 到 1。λ 取 0 和 1 是两种极限情况，取 0 时相当于不采用平滑技术，只是用极大似然估计来计算特征词 w_i 在类 C_j 中出现的概率；取 1 时相当于每个词都当作稀疏数据来处理。为了寻找合适的取值，取 0 到 1 之间的值，间隔 0.1，然后分析、比较不同的取值对分类性能的影响，选择分类性能最好的值作为 λ 的取值。

图 1、图 2 是 Jelinek-Mercer 平滑中参数 λ 的不同取值对准确率和召回率影响的表现：

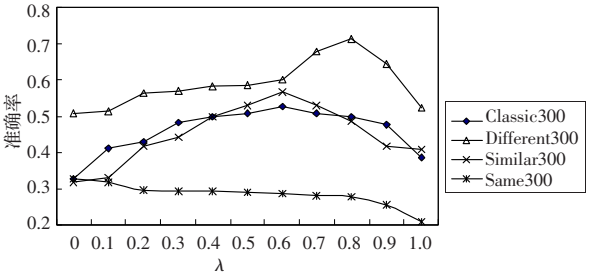

图 1　Jelinek-Mercer 平滑中 λ 与准确率的关系

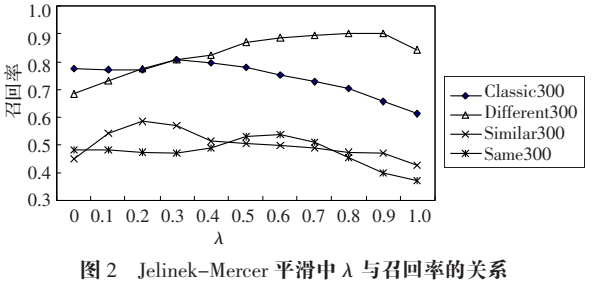

图 2　Jelinek-Mercer 平滑中 λ 与召回率的关系

由图 1、图 2 可知，对于不同的数据集，λ 会在不同的取值处达到准确率的最高值，综合所有的测试数据，当 $\lambda=0.6$ 或 $\lambda=0.8$ 时效果较好。而对于召回率，λ 的理想取值不是很集中，分散在 0.2 到 0.9 之间。综合考虑准确率和召回率的效果，取 $\lambda=0.8$。

4.3　改进后的贝叶斯分类器与原分类器的性能比较

经过了上面平滑技术的参数选择的分析，下面来对贝叶斯

(下转 197 页)