DOI:10.15948/j.cnki.37-1500/s.2011.05.079
山东省农业管理干部学院学报 2011 年 第 28 卷 第 5 期

# 一种对贝叶斯算法的改进算法分析

李 欣

(山东省农业管理干部学院,山东 济南 250100)

摘 要:贝叶斯理论在信息过滤中得到了很好的应用,他与 KNN、支持向量机等相比,加油高效、节约空间以及有利于实现个性化过滤的优势。但是也存在忽略了特征词之间的联系、忽略误判所带来的风险以及不支持增量学习机制等缺陷。本文着重分析贝叶斯算法在信息过滤方面的不足,然后针对这些不足,结合应用要求,提出改进的贝叶斯算法。

关键词:信息过滤;贝叶斯;特征词

中图分类号:F325 文献标识码:A 文章编号:1008-7540(2011)-05-0156-03

## 一、贝叶斯理论概述

### 1 简单贝叶斯

简单贝叶斯以其创始人 Thomas Bayes 的名字命名,它是一种基于概率分析的可能性推理理论。托马斯·贝叶斯 1761 年在《论有关机遇问题的求解》一文中,首次提出了贝叶斯统计理论,就是我们所说的简单贝叶斯,之所以说它是简单贝叶斯,就因为他仅仅根据已经发生的事件就能预测相关未来事件发生的可能性[1-3]。

贝叶斯理论假设是一种基于简单贝叶斯的假设机制,其定义如下:

如果正在发生的事件的结果不确定,那么我们如果要想评估未来要发生的事件的可能性,就只能用已经发生的相关事件发生的概率进行预测和评判,这也就是说如果我们已经知道某一事件在过去的试验中者发生的概率,我们就可以运用贝叶斯理论以计算出该事件在未来可能出现的概率。

下面给出贝叶斯理论的数学表达方式,也就是我们通常所说的贝叶斯公式。

定义 1 在随机试验中,为样本空间,如果我们对于中的任一个事件都能赋予一个实数,并且这个实数满足下面三个条件:

①0<=P(A)=1;

②P(S)=1;

③对于满足定义的任意两个两两互不相容的事件 A1、A2,存在一个概率 P,满足如下条件:

$$P(\sum_{i=a}^{\infty} A1)=\sum_{i=a}^{\infty} P(A1)$$

那么,我们把 P(A)为事件 A 发生的概率。

定义 2 条件概率

在随机试验 E 中,A,B 为任意两个事件,且 P(A)>0,称 P(A|B)是 A 发生下 B 的概率。计算公式如下:

$$P(B|A)=\frac{P(AB)}{(A)} \quad (1)$$

定理 1 乘法公式:设,则 P(AB)=P(B|A)P(A)

定理 2: 全概率公式

$B_1,B_2\cdots$为 S 的一个划分,且 P(Bi)>0(i=1,2,…,n),则对任一事件 A∪S,有:P(A)=P(A|B1)P(B1)+P(A|B2)P(B2)+……+P(A|Bn)P(Bn)(2)

定理 3: 贝叶斯公式

$B_1,B_2\cdots$为 S 的一个划分,且 P(Bi)>0(i=1,2,…,n),则对任一事件 A∪S,若 P>0,则有:

$$P(Bi|A)=\frac{P(Bi|A)P(B)}{\sum_{i=a}^{n} P(Bi|A)P(B)} \quad i=1,2,\cdots\cdots,n \quad (3)$$

### 2.贝叶斯方法的特点

简单贝叶斯在英文信息过滤中得到了很好的应用,他与 KNN、支持向量机等相比,存在以下几个独到之处:

(1)高效

相对于 SVM、Boosting 等方法需要多次扫描训练样本,贝叶斯训练阶段只需对所有训练样本扫描一遍并提取特征词,然后统计每个特征词在该样本中出现的次数即可,因此,从训练和分类所用的时间效用考虑,贝叶斯具有快捷、高效的特点。

(2)节约空间

贝叶斯在信息过滤的具体应用中,需要计算、统计和存储的只是特征词及其频率,并不存储实际文本信息,这就为计算机节约了大量的内存空间。

(3)有利于实现个性化过滤

应用贝叶斯分类方法实现的网络信息过滤器适合每个

作者简介:李欣,男,山东省农业管理干部学院计算机科学与技术系办公室主任。

用户单独在客户端训练和使用，因此，同其他方法相比，对于网络信息过滤器的个性化具有更加优厚的条件。

但是，在实际应用过程中，贝叶斯算法也存在很多局限性。

(1)贝叶斯算法在英文过滤中取得了较好的效果，但是在中文中的应用还需要进一步研究。

究其原因，这在于中文文本信息处理的独特性，英文中单词具有独立的意义，而中文中不同字和词的组合表达不同的意义，因此这就需要进行切词处理，而目前尚没有一种理想的切词方法能够取得满意的效果，这就需要进一步的研究工作。

(2)贝叶斯算法忽略了特征词之间的联系

该算法在具体应用中只考虑特征词出现的先验概率，并没有考虑特征词之间的关系，这就不能够完整的表达邮件本身的内容。针对这一问题，近年来出现了诸如贝叶斯网络、树增广简单贝叶斯等改进的理论和方法，但是，相关研究还有待进一步深入。

(3)忽略了误判所带来的风险

在信息过滤处理过程中，由于其要求实时性以及准确性，这就导致二者产生矛盾，因为我们将合法网页误判为垃圾网页所带来的损失要远远的大于垃圾网页判为合法网页，而传统贝叶斯就带来这一问题。

(4)不支持增量学习机制

应用于信息过滤的贝叶斯方法是根据训练样本的类先验概率和特征项的类条件概率来预测后验概率，从而得出新网页类别而进行过滤的。这就带来新主题网页的过滤误差。要解决这一问题，我们只能通过重新学习来实现，这就浪费了大量时间。

3.贝叶斯理论在信息过滤中的应用

最先使用贝叶斯算法进行信息过滤的是 Sahami，他利用贝叶斯算法对信息进行分类，并结合事先搜集和构造的一系列特征词来提高过滤精度。其存在很多问题，下面针对其缺陷进行分析并提出改进方案。

二、传统贝叶斯缺陷

1.应用分析

简单贝叶斯算法具有很多优越性，简单而且性能优越，同时也具有易于实现、运算速度快等优点，所以在信息过滤中具有极为广泛的应用。我们也可以看出，它存在一些缺陷。

表 1 所示数据为简单贝叶斯实验数据，表中数据为 CCERT 提供的通用过滤语料库，同时结合研究者在学习和工作中搜集的一部分文本，数量为 2000。并且随机抽取合法内容 100 个以及垃圾内容 300 个供 400 个文本作为测试数据。

表 1 简单贝叶斯过滤结果

| | 系统判定为合法的内容 | 系统判定为不合法的内容 | 总数 |
|---|---|---|---|
| 实际为合法的内容 | 89.25 | 10.75 | 100 |
| 实际为不合法的内容 | 12.25 | 287.25 | 300 |
| 总数 | 101.50 | 298.50 | 400 |

由上表可以看出，测试集中的测试语料被错分为 101.5 和 298.5，其中分别有 10.75 篇和 12.25 篇的误判，系统精确率为 94.25%，召回率为 95.92%。

实验数据说明了其缺陷：

(1)过多的简化失去了很多语义信息，使分类效果不理想。

(2)垃圾信息被误判为合法邮件浪费了用户的精力，由于过滤没有达到预期效果，导致这些垃圾信息给用户的工作、学习和生活带来问题。

(3)该算法假定特征相互独立，也就是说假设问题满足贝叶斯应用条件。但事实上，这种条件独立的假设不会在所有应用领域都成立[4,5]。

2.理论分析

在简单贝叶斯文本分类算法中：

$$P(c_j|d_x)=\frac{P(c_j)P(d_x|c_j)}{P(d_x)}\propto P(c_j)P(d_x|c_j)\propto(4)$$

$P(c_j)$是类的先验概率，$P(d_x|c_j)$是类条件概率，$P(d_x)$对同一篇文本，不变。

$$P(d_x|c_j)=P(t_1|c_j)*P(t_2|c_j)*\Lambda*P(t_n|c_j)=\prod_{i=1}^{n}P(t_i+cj)(5)$$

设 $dx$ 表示为特征集合$(t_1,t_2,t_n)$，$n$ 为特征个数，假设特征之间相互独立。$P(c_j)$和 $P(t_i+cj)$都可以利用训练集估计。

由此公式可以看出如果各个类别的训练模型中的特征词都是单类别词汇，将导致公式(5)对于某些类别的概率计算是无意义的，本文对公式(5)进行了改进，改进后的公式解决了上述问题。

三、贝叶斯算法的改进

1.改进方法

利用前面章节的朴素贝叶斯概率公式可以分别计算出任一待分类邮件 $d_x$ 属于合法邮件和垃圾邮件的概率：$P(c_j|d_x)(c_j$ 为垃圾邮件类)和 $P(c_0|d_x)(c_0$ 为合法邮件类)。在传统的方法中，一般当 $P(c_j|d_x)>P(c_0|d_x)$时，就判定邮件 $d_x$ 为垃圾邮件，否则就判为合法邮件。但是，这种判断方式并不精确，会产生较高的误判率和漏判率。在上一节中我们分析了朴素贝叶斯算法存在的不足之处，它使分类过程丧失了很多有用的信息，从而很可能导致误判，造成严重的后果，因此直接应用它时分类偏差会比较大。

因此，为了能更加谨慎、准确地识别出垃圾邮件，减少由于把合法邮件判为垃圾邮件而造成的损失，设 $\frac{P(c_1|d_x)}{P(c_0|d_x)}>\theta$ 当时，即当一封邮件 $d_x$ 为垃圾邮件的概率是合法邮件概率的 $\theta$ 倍时，将其判定为垃圾邮件。当 $\theta$ 值越大，其为垃圾邮件的可能性就越大。

又由全概率公式得到 $\frac{P(c_1|d_x)}{P(c_0|d_x)}>\theta$ 可以表示为：

$$\frac{P(c_1|d_x)}{1-P(c_1|d_x)}>\theta$$

$$P(c_1|d_x)>\frac{\theta}{1-\theta}=k(6)$$

也就是当时，将判为垃圾邮件。

2.算法的效果分析

上述方法可以用分类器进行衡量，具体细节在本节加以表述：

(1)实验数据集

本文通过从复旦大学语料库中选取部分语料和从网上搜集部分文本，最终选取农业、艺术、计算机、环境、历史、政治这六个类别，文中对分类器分别进行封闭性能测试和开放性能测试。在进行开放性能测试时为了更客观的评价分类器性能的好坏，文中将选取的 715 篇测试文本混合放在一个文件夹下来进行分类，然后根据分类结果将各(下转第 170 页)

成长,与成年球员相比,年轻球员更需要比赛。

2.4 中国足球学习目的不明确，没有形成自己的风格特点。大家都知道青少年阶段正是打基础,接受技战术最重要的阶段，这也直接反映到今后国家整体的足球风格的形成。中国足球为了冲出亚洲，走向世界，长久以来都处于摸索之中。学过巴西、效仿德国、推崇过所谓的“疯狗精神”、还喊出学习巴萨、西班牙的口号。主教练是换了一个又一个,球队的风格也一直在变,结果是要技术没技术,要身体没身体。从本次国青国少的表现可以看出,他们在场上不是过于单打独斗而脱离全队的战术体系,就是水平发挥起伏不定,一场好一场坏。近看我们的邻国韩国和日本,他们可以说是现在亚洲足球的代表,但是二者也都皆有独到的不同的风格。韩国队攻守都极为出色,他们的打法已经和世界接轨,对抗世界二流球队一点都不难。如在南非世界杯与希腊队的比赛中,他们在技战术运用上表现出来的成熟,令对手为之叹服。韩国足球能够一直强于中国的原因也是他们因地制宜的结果。日本球员的身体条件与其他球队的队员相比没有任何优势可言,但现在他们练就的技术却在亚洲一枝独秀。由于历史上和巴西颇有渊源,日本足球也将巴西足球视为老师,不仅多次招入巴西后裔,对于细腻足球的钻研更是远超亚洲各队。

3、建议

3.1 我国青少年足球培养过程中,从竞赛、训练管理体制和运行机制上改变了注重比赛结果、比赛成绩和青少年早期成人化、专业化训练的急功近利思想;突出以训练为中心,以技术辅导为手段,以提高技术、技能、作风、职业道德为根本,以培养具备综合优秀素质为目的的核心思想。

3.2 在竞训体制的实际运作过程中强调符合竞技体育运动发展规律;符合教育规律;符合现代足球职业化发展规律;符合全国青少年足球运动员培养输送成材规律;符合全国足球运动管理体制和运行机制改革的发展方向。

3.3 在竞赛、训练的具体实施过程中,建立了分级管理、分级竞赛、分级指导的制度,加强了中国足协的宏观管理力度,充分调动了会员协会和俱乐部的积极性。

3.4 在青少年足球的年龄设置上与世界接轨;形成了 U-9、U-11、U-13、U-15、U-17、和 U-19 六个年龄段有机衔接的年龄结构体系。


参考文献

[1].黄光亮 从国青队和国少队的失利看中国足球的现状[J].体育科技 2011 年(第 32 卷)第 2 期

[2].李志刚 日韩足球震惊中国足球神经数据显亚洲足球进步[N].齐鲁晚报 2010·6·30

[3].国景涛 中德青少年足球人才培养模式的比较研究[C].山东师范大学 硕士论文 2011 年 6 月 1 日

[4].张磊 国青适龄注册球员仅 90 人冲击世青赛失败或注定[N].新京报 2010·10·18

[5].曹卫华,梁殿乙,韩舌 我国青少年足球竞训体系与管理体制[J].沈阳体育学院学报 第 23 卷第 5 期 2004 年 10 月

[6]韩　勇,王　蒲我国足球后备人才培养体系的研究》[J].天津体育学院学报 2001,16(1):35-36.

[7]郁　静 试论青少年足球竞赛市场的开发》[J].西安体育学院学报,2002,19(3):17-19.


编辑:张小玫

(上接第 157 页)个待测文本放入相应的类别文件夹下。文中选用中科院的 SharpICTCLAS 分词系统 1.0 进行文本切词,采用期望交叉熵来进行特征选择。

$$G(t)=-\sum_{i=1}^{m}P_r(C_i)logP_r(C_i)+P_r(t)\sum_{i=1}^{m}P_r(C_i+t)logP_r(C_i|t) \quad (7)$$

(2)评估指标

设测试集合中共有 N 篇测试文本,为方便叙述,定义如下几个变量,其中 N=A+B+C+D,其中 A 表示测试集中属于 $c_j$ 的文档被分类器正确判为 $c_j$ 的样本数,B 表示测试集中属于 $c_i$ 的文档被分类器误判为的样本数,C 表示测试集中属于 $c_j$ 的文档被分类器误判为 $c_i$ 的样本数,D 表示测试集中属于 $c_i$ 的文档被分类器正确判为 $c_i$ 的样本数。则类别 $c_j$ 的各种评估指标为:

查准率(Precision):$P=\frac{A}{A+B}\times 100\%$

查全率(Recall):$P=\frac{A}{A+C}\times 100\%$

$F_1$ 测试值:$F_1=\frac{P\times R\times 2}{P+R}$

(3)实验结果及分析

本文对上面所提到的六个类别,分别使用贝叶斯、中心向量以及改进贝叶斯进行分类。

表 2 分类效果实验准确率比较

| 类别 | 贝叶斯 | Rocchio | 改进方法 |
|---|---|---|---|
| 农业 | 85.5% | 79.17% | 89.76% |
| 艺术 | 74.34% | 63.84% | 81.88% |
| 计算机 | 96.39% | 86.67% | 88.98% |
| 环境 | 98.55% | 95% | 96.84% |
| 历史 | 43.61% | 41.72% | 76.4% |

通过表 2 可以看出,使用改进贝叶斯能从整体上提高各个类别的性能。

四、小结

简单贝叶斯算法因为其计算时简单、精确度高,具有坚实的理论基础而得到了广泛应用,它建立在“贝叶斯假设”的基础之上:假定所有的特征之间互相独立。但是实际上,在生活中这种独立性却很难存在。而且由于在信息过滤过程中简单贝叶斯算法没有考虑合法网页被错判为垃圾网页的情况。因此,本文引入一种改进方法,在尽量不使合法网页错判为垃圾网页的同时,以减少垃圾网页的误判率,并且尽可能地提高分类的精确度。为了证明本文算法过滤的优越性,通过实验给出本文算法和其他过滤算法之间的性能比较。


参考文献:

[1] 阮彤. 信息过滤模型与算法的研究[D]. 中科院软件研究所, 2001.

[2] 林鸿飞. 中文文本过滤的逻辑模型[D]. 东北大学, 2006.

[3] 牛洪波. 基于文本分类技术的信息过滤方法的研究[M]. 哈尔滨理工大学, 2009.

[4] 陈剑敏. 基于 Bayes 方法的文本分类器的研究与实现[D]. 重庆:重庆大学计算机学院, 2007.

[5] Neal,R.M, Hinton. A new view of the EM algorithm that justifies incremental, sparse, and other variants, In Learning in Graphical Models[M]. [S.l.]: Hewer Academic Publishers, 1998: 355-368.


编辑:冯惟榘