基于贝叶斯的垃圾邮件过滤算法的研究

李 雯,刘培玉
LI Wen,LIU Pei-yu

山东师范大学 信息科学与工程学院,济南 250014
School of Information Science and Engineering,Shandong Normal University,Ji'nan 250014,China
E-mail:lw_656@hotmail.com

LI Wen,LIU Pei-yu.Research on anti-spam e-mail filtering algorithm based on Bayesian.Computer Engineering and Applications,2007,43(23):174-176.

Abstract: Nowadays,the study on content-based spam filtering is one of the hot topics in the Internet security research area all over the world.And Bayesian classification method has expressed high accuracy.An improved algorithm based on Naïve Bayesian and Boosting method is proposed in this paper.The experiment results show that the algorithm has better performance.
Key words: spam;e-mail filtering;Bayesian algorithm

摘 要:基于内容的垃圾邮件过滤问题是 Internet 安全技术研究的一个重点问题,而基于贝叶斯的分类方法在垃圾邮件处理上表现出了很高的准确度,因此受到了广泛的关注。在朴素贝叶斯算法的基础上,提出了一种基于最小风险贝叶斯方法同 Boosting 算法相结合的邮件过滤改进算法,提高了分类的精确度。实验证明,算法在邮件过滤中有更好的表现。
关键词:垃圾邮件;邮件过滤;贝叶斯算法
文章编号:1002-8331(2007)23-0174-03 **文献标识码**:A **中图分类号**:TP393

1 引言

随着 Internet 的发展和应用,越来越多的商务、日常活动通过 Internet 才能进行,网络跟人们的生活越来越紧密。然而网络是双面的,人们在享受网络所带来的便利的同时,不可避免地接触到大量的不良信息,如色情、暴力、反动、邪教、赌博、病毒等,部分成人尤其是自制力不强的青少年学生沉迷在网上游戏、娱乐、色情世界里而不能自拔。因而如何管理网络用户尤其是青少年学生对 Internet 访问,而又不影响用户对网络的正常访问,越来越引起人们的关注。

为了过滤网络信息,使网络用户尤其是青少年学生远离非友善信息的侵扰,使得网络环境更加纯净、美好,网络信息过滤技术已经成为当前研究的热点[1-4]。而电子邮件(e-mail)以其方便、快捷、低成本的独特魅力成为人们日常生活中不可缺少的通信手段之一。但电子邮件给人们带来极大便利的同时,也日益显示出其负面影响,那就是我们每天收到的邮件中有很大一部分是那种"不请自来"的,它们或者是推销广告,或者是一些有害的不良信息,甚至还有病毒。这些垃圾邮件占用网络带宽,浪费网络资源,浪费用户宝贵时间及上网费用,其中有些垃圾邮件传播各样有害信息甚至病毒,直接对网络安全形成威胁。

2 朴素贝叶斯模型中的两种概率估计方法

在贝叶斯假设的基础上,文本可以看作是若干词汇的集合,可以认为文本是这些词汇按照一定的方式"产生"的。根据产生方式,朴素贝叶斯分类算法有两种概率估计方法:多变量贝努里事件模型[5](MBM,Multi-variate Bernoulli event Model)和多项式事件模型(MM,Multinomial event Model)。这两种模型在公式中的体现是 $P(d_x|c_j)$估计的方法不同。

2.1 多变量贝努里事件模型

在这种模型中,文本向量是布尔权重,也就是说,如果特征词在文本中出现,则权重为 1,否则,权重为 0,不考虑特征词的出现顺序,忽略特征词在文本中的出现次数。设特征数量为 n,将文本看作一个事件,这个事件是通过 n 重贝努里实验产生的,即某个特征出现或者不出现。这也是贝叶斯分类算法的一个传统模型。

设 B_{xt} 表示特征 w_t 在文本 d_x 中的出现情况,B_{xt}=0/1 表示特征出现/不出现,则有:

$$P(d_x|c_j)=\prod_{j=1}^{n}(B_{xt}P(w_t|c_j)+(1-B_{xt})(1-P(w_t|c_j))) \quad (1)$$

$P(w_t|c_j)$表示在属于类 c_j 的情况下 w_t 出现的概率。从上面公式可以看出,在多变量贝努里事件模型中,文本是所有特征的类条件概率之积,若特征在文本中出现,乘的项是 $P(w_t|c_j)$,若不出现,乘的项是$(1-B_{xt})(1-P(w_t|c_j))$。

$P(w_t|c_j)$的估计也采用文档频次:

基金项目:山东省自然科学基金(the Natural Science Foundation of Shandong Province of China under Grant No.Y2006G20);山东省科技公关计划(the Key Technologies R&D Program of Shandong Province of China under Grant No.053120126)。
作者简介:李雯(1982-),女,硕士研究生,主要研究领域为网络信息安全,Web 数据挖掘;刘培玉(1960-),男,教授,博士生导师,主要研究领域为网络安全,数据库系统。

$$P(w_t|c_j)=\frac{c_j\text{类中特征}\ w_t\text{在其中出现的文本数量}}{c_j\text{类的文本数量}} \tag{2}$$

对上面公式进行简单的平滑处理：

$$P(w_t|c_j)=\frac{1+c_j\text{类中特征}\ w_t\text{在其中出现的文本数量}}{2+c_j\text{类的文本数量}} \tag{3}$$

由此可以看出，多变量贝努里事件模型的特点在于：

(1)计算 $P(w|c_j)$和 $P(d_x|c_j)$的时候都不考虑特征在文本中的出现次数；

(2)对那些没有在文本中出现的特征，计算时乘以$(1-B_{xt})(1-P(w_t|c_j))$项。

应用于垃圾邮件过滤，只考虑两个类别：垃圾邮件和非垃圾邮件，设 $c=1$ 表示垃圾邮件类，$c=0$ 表示非垃圾邮件类，目标是计算：

$$P(c=1|d_x)=\frac{P(c=1)P(d_x|c=1)}{P(d_x)} \tag{4}$$

根据式(1)，有：

$$P(d_x|c=1)=\prod_{t=1}^{n}(B_{xt}P(w_t|c=1)+(1-B_{xt})(1-P(w_t|c=1))) \tag{5}$$

$$P(d_x|c=0)=\prod_{t=1}^{n}(B_{xt}P(w_t|c=0)+(1-B_{xt})(1-P(w_t|c=0))) \tag{6}$$

$$P(d_x)=\sum_{j=1}^{|C|}P(c_j)P(d_x|c_j)= P(c=1)P(d_x|c=1)+(1-P(c=1))(1-P(d_x|c=0)) \tag{7}$$

根据式(2)和式(3)，可以从训练集中估计 $P(c=1)$、$P(w_t|c=1)$和 $P(w_t|c=0)$；分类时，可以根据式(4)、式(5)、式(6)和式(7)计算待分类邮件属于垃圾邮件类别的概率。

2.2　多项式事件模型

这种模型中，与前面的多变量贝努里事件模型不同，要考虑特征词的出现次数。在这里，将每个特征词的出现看作“事件”，文本是这些事件的集合，假设这些事件之间是相互独立的。设文本 d_x 的长度(即文本中包含的词数)为$|d_x|$，N_{xt} 表示特征 t 在文本 d_x 中的出现次数，n 为特征数量，则 $P(d_x|c_j)$符合多项式分布(multinomial distribution)：

$$P(d_x|c_j)=P(d_x)*|d_x|!\ *\prod_{t=1}^{n}\frac{P(w_t|c_j)^{N_{xt}}}{N_{xt}!} \tag{8}$$

在训练集上估计时，采用词频：

$$P(w_t|c_j)=\frac{c_j\text{类的所有文本中特征词}\ w_t\text{的出现次数}}{c_j\text{类的所有文本中特征词总数}} \tag{9}$$

上面公式进行平滑处理变为：

$$P(w_t|c_j)=\frac{1+c_j\text{类的所有文本中特征词}\ w_t\text{的出现次数}}{n+c_j\text{类的所有文本中出现的特征词总数}} \tag{10}$$

应用于垃圾邮件过滤时，设 $c=1$ 表示垃圾邮件类，$c=0$ 表示非垃圾邮件类，根据式(4)和式(7)，有：

$$P(c=1|d_x)= \frac{P(c=1)P(d_x|c=1)}{P(c=1)P(d_x|c=1)+(1-P(c=1))(1-P(d_x|c=0))}= \frac{1}{1+\frac{1-P(c=1)}{P(c=1)}*\frac{P(d_x|c=0)}{P(d_x|c=1)}} \tag{11}$$

对同一封邮件 d_x，由式(8)得：

$$\frac{P(d_x|c=0)}{P(d_x|c=1)}=\prod_{t=1}^{n}\frac{P(w_t|c=0)N_{xt}}{P(w_t|c=1)} \tag{12}$$

通过实验比较两种模型发现贝努里事件模型在计算上较简便，而精确率也略高于多项式事件模型。因此本文中使用了贝努里分布模型。

3　基于最小风险的贝叶斯分类模型

贝叶斯分类模型中有两种决策规则[6,7]分别是：基于最小错误率贝叶斯决策和基于最小风险贝叶斯决策。其中基于最小错误率的贝叶斯决策是指使错误率为最小的分类规则，但是实际上有时需要考虑一个比错误率更为广泛的概念——风险，而风险又是和损失紧密相连的。对邮件的分类不仅要考虑到尽可能做出正确判断，而且还要考虑到做出错误判断时会带来什么后果。如果把正确邮件判为垃圾邮件放到垃圾箱中可能会使用户的重要信件丢失带来一定的损失；而如果本来就是垃圾邮件却判为正常，就会给用户带来许多麻烦。显然这两种不同的错误判断所造成损失的严重程度是有显著差别的，前者的损失比后者更严重。最小风险贝叶斯决策正是考虑各种错误造成损失不同而提出的一种决策规则。

3.1　基于最小风险的贝叶斯规则

最小风险贝叶斯规则是：已知先验概率 $P(C_j)$及类条件概率密度 $P(d_x|C_j)$，损失函数为 $\lambda(\alpha_i,c_j)(i=1,2,\cdots,a)$，决策空间由 a 个决策 α_i 组成，λ 表示当真实状态为 C_j 而所采取的策略是 α_i 时所带来的损失，这样可以得到一般的决策表如表 1 所示：

表 1　一般决策表

决策	损失					
	c_1	c_2	…	c_j	…	c_M
α_1	$\lambda(\alpha_1,c_1)$	$\lambda(\alpha_1,c_2)$	…	$\lambda(\alpha_1,c_j)$	…	$\lambda(\alpha_1,c_M)$
α_2	$\lambda(\alpha_2,c_1)$	$\lambda(\alpha_2,c_2)$	…	$\lambda(\alpha_2,c_j)$	…	$\lambda(\alpha_2,c_M)$
…	…	…	…	…	…	…
α_i	$\lambda(\alpha_i,c_1)$	$\lambda(\alpha_i,c_2)$	…	$\lambda(\alpha_i,c_j)$	…	$\lambda(\alpha_i,c_M)$
…	…	…	…	…	…	…
α_a	$\lambda(\alpha_a,c_1)$	$\lambda(\alpha_a,c_2)$	…	$\lambda(\alpha_a,c_j)$	…	$\lambda(\alpha_a,c_M)$

根据贝叶斯公式，后验概率为：

$$P(c_i|d_x)=\frac{P(c_j)P(d_x|c_j)}{P(d_x)}\ \text{其中}\ P(d_x)=\sum_{j=1}^{|C|}P(c_j)P(d_x|c_j)$$

由于引入了“损失”的概念，在考虑错误判断所造成的损失时，就不能只根据后验概率的大小来做决策，而必须考虑所采取的判断是否使损失最小。对于给定的 d_x，如果采取决策 α_i，从决策表可见，对应于决策 α_i，可以在 M 个 $\lambda(\alpha_i,c_j)$，$j=1,2,\cdots,M$ 值中任取一个，其相应概率为 $P(c_j|d_x)$。因此在采取决策 α_i 情况下的条件期望损失 $R(\alpha_i|d_x)$为：

$$R(\alpha_i|d_x)=E[\lambda(\alpha_i,c_j)]=\sum_{j=1}^{M}\lambda(\alpha_i,c_j)P(c_j|d_x),i=1,2,\cdots,a \tag{13}$$

把采取决策 α_i 的条件期望损失 $R(\alpha_i|d_x)$称为条件风险。对得到的 a 个条件风险值 $R(\alpha_i|d_x)(i=1,2,\cdots,a)$进行比较，找出使条件风险最小的决策 α_k 即 α_i：

$$R(\alpha_k|d_x)=\min_{i=1,2}R(\alpha_i|d_x)$$

则 α_k 就是最小风险贝叶斯决策。

在考虑错判带来的损失时，希望损失最小。如果在采取每一个决策或行动时都使其条件风险最小，则对所有的 d_x 作出决策时，其期望风险也必然最小。这样的决策就是最小风险贝叶斯决策。

3.2 朴素贝叶斯同基于最小风险的贝叶斯分类结果比较

从中国反垃圾邮件联盟(http://www.anti-span.org.cn/)中提供的中文邮件语料库中随机选取了2 000封邮件作为训练样本,对400封作为测试集的邮件(其中垃圾邮件为300封,合法邮件为100封)进行分类。其中基于朴素贝叶斯邮件过滤算法的实验结果(经过多次实验得出的结果取其平均值)见表2。

表2 基于朴素贝叶斯邮件过滤算法的实验结果

实际类别	系统判断的类别		
	合法邮件	垃圾邮件	总数
合法邮件	89.25	10.75	100
垃圾邮件	12.25	287.75	300
总数	101.50	298.50	400

由表2知对400封邮件进行分类测试,其中有10.75篇合法邮件被判为垃圾邮件,而有12.25篇垃圾邮件判为合法邮件。系统精确率为94.25%,合法邮件的召回率为89.25%。

采用基于最小风险贝叶斯算法进行邮件过滤,根据表3提供损失因子值的不同产生不同的结果。每次测试过程中改变损失因子,对邮件进行分类得到结果如表4所示。

表3 决策表

决策	损失	
	C_1	C_2
α_1	0.0	0.3
α_2	0.6	0.0

表4 根据不同损失因子值得出的结果

λ	实际类别	系统判断的类别		
		合法邮件	垃圾邮件	总数
0.3	合法邮件	90.25	9.25	100
	垃圾邮件	13.00	287.00	300
	总数	103.25	296.25	400
0.6	合法邮件	92.50	7.50	100
	垃圾邮件	13.75	286.25	300
	总数	105.25	293.75	400

由表4可知当λ的值为0.3时,有9.25篇合法邮件被判为垃圾邮件,而有13篇垃圾邮件判为合法邮件。系统精确率为94.31%,合法邮件召回率为90.25%;当λ的值为0.6时,有7.5篇合法邮件被判为垃圾邮件,而有13.75篇垃圾邮件判为合法邮件。精确率为94.68%,召回率为92.5%。

由此可见与基于朴素贝叶斯的邮件过滤算法相比基于最小风险的贝叶斯邮件过滤算法在精确率和合法邮件召回率的性能方面有一定提高,即当引入损失因子后精确率增加。合法邮件被系统误判为垃圾邮件的可能性减少。但同时,垃圾邮件的召回率有所下降,即垃圾邮件被系统误判为合法邮件的可能性增加。

4 基于最小风险贝叶斯模型的提升

由于在邮件的过滤过程中朴素贝叶斯算法没有考虑合法邮件被错判为垃圾邮件的情况。而采用最小风险贝叶斯分类使得系统的召回率有所下降,即垃圾邮件被系统误判为合法邮件的可能性增加。因此,在这里本文引入了提升方法(boosting)。即在朴素贝叶斯算法的基础上,提出了基于最小风险贝叶斯方法同Boosting算法相结合的邮件过滤算法。在尽量不使合法邮件错判为垃圾邮件的同时尽可能地提高分类的精确度。

提升方法[8](boosting)总的思想是学习一系列决策行动,在这个序列中每个决策对它前一个决策导致的错误判断例子给予更大的重视。尤其是在学习完决策行动α_k之后,增加了由α_k导致判断错误的训练例子的权重值,并且通过重新对训练例子计算权值,再学习下一个决策α_{k+1}。这个过程重复T次。最终的分类器从这一系列的决策中综合得出。

在这个过程中,每个训练例子被赋予一个相应的权值,如果一个训练例子被分类器错误分类,那么就相应增加该例子的权重,使得下一次学习中,分类器对该例代表的情况更加重视。

5 三种算法性能比较

垃圾邮件过滤中,除了公共的语料,还需要定义一些指标来评价垃圾邮件过滤器的效果。这些指标一般都是从文本分类和信息检索领域借鉴过来的。

设测试集合中共有N封邮件,为方便叙述,先定义几个变量,见表5,其中$N=A+B+C+D$。

表5 垃圾邮件系统判定情况分布 封

	正确答案是Spam	正确答案是非Spam
系统判定为Spam	A	B
系统判定为非Spam	C	D

定义如下几个评价指标:

(1)召回率(Recall):$Recall=\frac{A}{A+C}*100\%$,即垃圾邮件检出率。这个指标反映了过滤系统发现垃圾邮件的能力,召回率越高,"漏网"的垃圾邮件就越少。

(2)精确率(Accuracy):$Accuracy=\frac{A+D}{N}*100\%$,即对所有邮件的判对率。

仍采用表2中的400封邮件作为测试集,总结三种过滤算法的实验结果比较如表6所示:

表6 三种算法的比较

项目	算法				
	朴素贝叶斯	基于最小风险贝叶斯		提升的最小风险贝叶斯	
		λ=0.3	λ=0.6	λ=0.3	λ=0.6
精确率/%	94.25	94.31	94.68	94.96	95.43
召回率/%	89.25	90.25	92.50	92.75	93.50

从表6可以看出,基于最小风险的朴素贝叶斯提升算法在邮件过滤中,无论是在精确率还是在召回率上都优于其它两种算法。实验结果的数据对比充分说明了本文的方法能提高垃圾邮件过滤的整体性能。

6 结束语

本文基于最小风险的贝叶斯方法提出一种新的邮件过滤算法,以减少合法邮件的误判率。实验结果表明,该算法在邮件过滤性能方面具有较好的动态调整能力,取得了较为满意的结果,从而提高了邮件过滤的质量。(收稿日期:2007年5月)

参考文献:

[1] Robertson S E.The probability ranking principle in IR,readings in information retrieval[M].[S.l.]:Morgan Kaufmann,1997:281-286.

[2] Salton.Automatic text processing:the transformation,analysis and retrieval of information by computer[M].[S.l.]:Addison-Wesley Inc,1989.

(下转183页)