改进的贝叶斯垃圾邮件过滤算法①

赵敬慧, 魏振钢

(中国海洋大学 信息科学与工程学院, 青岛 266100)

摘 要: 随着网络的不断发展, 电子邮件已成为人们生活中较为普及的通信手段, 相应地垃圾邮件也成为了困扰 E-mail 用户的主要问题, 因此研究如何更好的抑制垃圾邮件的滥发变得愈发紧迫. 在基于朴素贝叶斯算法的基础上提出了带有损失因子 k 的最小风险贝叶斯算法, 该算法通过调整 k 值, 来改善正常邮件的误判问题, 最大程度上减少用户的损失. 最后实验结果表明, 最小风险贝叶斯算法可以使垃圾邮件有着更好的过滤效果.

关键词: 垃圾邮件; 贝叶斯算法; 损失因子; 最小风险

Improved Bayes Algorithm for Filtering Spam E-Mail

ZHAO Jing-Hui, WEI Zhen-Gang

(College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China)

Abstract: With the continuous development of Internet, email has become a more popular forms of communication in people's life, and as well spam has become a major problem in E-mail users. Thus suppressing spam has become a rather urgent task. A minimum risk bayes algorithm based on the traditional bayesian is proposed on the basis of loss factor k, By adjusting the k value, the algorithm can improve the influence of spam false negative result, and reduce the loss of users at its best. At last, The results of the experiment indicates that the minimum risk bayes algorithm can make spam has a better filtering effect.

Key words: spam; bayes algorithm; loss factor; minimum risk

21 世纪的今天, 电子邮件成为了人们日常交流的主要工具. 这给我们的及时沟通提供了很大方便, 与此同时, 垃圾邮件也在飞速的增长. 大量的垃圾邮件给人们的生活带来了不同程度的困扰, 例如: 占用网络传输带宽、影响正常的网络通信、浪费人们的精力与时间等等. 所谓的垃圾邮件一般具有批量发送的特征. 其内容包括赚钱信息、成人广告、商业或个人网站广告、电子杂志、连环信等. 据中国互联网协会反垃圾邮件中心发布的 《2013 年第四季度中国反垃圾邮件状况调查报告》显示, 中国垃圾邮件的总量近几年持续攀升, 中国的电子邮件用户平均每周接收约 18.6 封垃圾邮件, 垃圾邮件的百分比占邮件总量的 47.3%, 每年给中国电子邮件用户造成的经济损失达几百亿元人民币. 因此, 利用有效的技术手段来阻挡垃圾邮件具有重要的实际意义.

为了更好的过滤垃圾邮件, 我们应该分析垃圾邮件快速增长的原因. 第一, 发送邮件的低成本(对于任何人来说, 只要想发送邮件, 他可以在零时间发送无数邮件); 第二, 邮件发送者的获利性(在发送数以万计的邮件中, 只要有寥寥无几的阅读者, 那么发送者就有机会获得收益). 基于这两点, 我们在一定程度上增加邮件发送者的发送成本, 可能会相应地减少垃圾邮件的传输. 然而, 如果想大力度的阻挡垃圾邮件, 就必须使用专业的技术手段.

世界各地成立了许多组织开展反垃圾邮件的工作. 目前几个著名的组织有 MAPS, ORBS, SPamCorp 等, 他们从技术角度着手解决垃圾邮件, 他们都各自维护了一个发送或转发垃圾邮件的数据库, 帮助用户过滤

① 基金项目:青岛市科技计划基础研究项目(11-2-4-1-(6)-JCH)

收稿时间:2016-01-20;收到修改稿时间:2016-03-17 [doi:10.15888/j.cnki.csa.005380]

© 中国科学院软件研究所 http://www.c-s-a.org.cn

垃圾邮件. 目前针对垃圾邮件的技术主要有三类: 基于 IP 的识别、基于行为的识别和基于内容的识别. 其中基于内容的识别是研究的主流, 是垃圾邮件过滤技术的研究趋势. 邮件过滤技术实质上把邮件分为垃圾邮件(spam)和正常邮件(ham), 因此就需要利用贝叶斯技术来预测收到的邮件是否为垃圾邮件.

由于朴素贝叶斯算法[1,2]是一种简单而又有效的分类方法, 故而在垃圾邮件过滤中得到了广泛的应用. 为了降低正常邮件被判断为垃圾邮件的损失, 通过对损失的控制来达到最好的分类效果, 本文引入了最小风险贝叶斯算法.

1　贝叶斯定理

贝叶斯定理是由英国数学家叶斯(1702—1761)提出的计算概率的一种方法. 1763 年, 在《论关机遇问题的求解》中发表了贝叶斯统计理论, 即通过对某一事件过去发生的概率情况的考察, 大体可以推断出当前这一事件发生的概率. 贝叶斯定理可以用一个数学公式表达, 即贝叶斯公式(Bayes Formula). 它的表述形式为: 设实验 E 的样本空间为 S, A 为 E 的事件, B1, B2, …, Bn 为 S 的一个划分, 且 P(A)>0,P (Bi)>0 (i=1,2,…,n), 则

$$P(Bi \mid A)=\frac{P(Bi)P(A \mid Bi)}{P(A)} \tag{1}$$

基于垃圾邮件作为研究模型的贝叶斯分类器是通过对邮件训练集的分类、加工来区分获得训练集中的垃圾邮件的特征模式,基于此模型的贝叶斯分类器用来检测、发现有用的信息来过滤掉垃圾邮件.

1.1 朴素贝叶斯算法

贝叶斯分类器[3-5]是一类常用的分类器, 最基本的形式是朴素贝叶斯分类器. 其原理是通过计算属于各个类别的概率, 将文本归为概率最大的一类.

假设文本集 D={d1,d2,…,dn}, 特征值集 W={W1,W2,…,Wm}, 另有变量 C={C1,C2,…,Ck}, 可以表示为 di={val(W1),val(W2),…,val(Wm)}; 如果 val(Wi)=1,则说明 Wi 存在于 di 之中, 且样本 di 属于类别 Cj 的条件要满足: P(C=Cj|d=di)=max{P(C=C1|d=di),P(C=C2|d=di),…,P(C=Ck|d=di)}即将 di 分类到概率值最大的相应类别中区. 计算 P(Cj|di)时, 利用贝叶斯公式:

$$P(Cj \mid di)=\frac{P(Cj)P(di \mid Cj)}{P(di)} \quad (j=1,2,\ldots,|C|) \tag{2}$$

式中, 根据全概率公式, 有

$$P(di)=\sum_{j=1}^{k}P(Cj)P(di \mid Cj) \tag{3}$$

其中 P(Cj)为 Cj 类的先验概率, P(di|Cj)是指类 Cj 中 di 发生的类条件概率, 即为似然函数. 对于同一篇文本, P(d=di)不变, 假设各个特征变量之间相互独立, 则有:

$$P(Cj \mid di)=\frac{P(Cj)\prod_{i=1}^{n}P(Wi \mid Cj)}{P(di)} \tag{4}$$

1.2 最小风险贝叶斯算法

垃圾邮件过滤实际上是一个二分类问题, 即对于每一个邮件样本, 都对其进行形式化描述 C={Spam,Ham},邮件分类器的任务就是计算待分类邮件是垃圾邮件的概率, 如果超过了正常邮件的概率或者某一阈值则认为该邮件为垃圾邮件. 根据贝叶斯公式, 即

$$P(C=Spam \mid d=di)=\frac{P(d=di \mid C=Spam)P(C=Spam)}{P(d=di)} \tag{5}$$

$$P(C=Ham \mid d=di)=\frac{P(d=di \mid C=Ham)P(C=Ham)}{P(d=di)} \tag{6}$$

P(C=Spam)、P(C=Ham)分别表示选取的邮件样本中垃圾邮件、正常邮件出现的概率; P(d=di|C=Spam)是指垃圾邮件中 di 中所有特征项同时出现的概率, P(d=di|C=Ham)是指正常邮件中 di 中所有特征项同时出现的概率. 当 P(d=di|C=Spam)大于 P(d=di|C=Ham)或者大于某一阈值时则认为该邮件为垃圾邮件.

在电子邮件的实际分类中, 有时对邮件的分类要考虑到做出错误判断时会带来的后果, 如果将垃圾邮件判为正常邮件会浪费用户宝贵的时间和精力, 然而, 如果把正常邮件判为垃圾邮件放到垃圾箱中可能会耽误用户的重要事情, 比如会议. 很明显, 错误的阻断一个正常邮件要比漏掉一个乃至几个垃圾邮件的代价大得多, 这也就是很多用户不愿轻易使用垃圾邮件过滤设备的原因. 因此, 我们需要一种可以使得损失尽量最小化的过滤算法, 即引入损耗因子 k 的改进算法

表 1　贝叶斯最小风险决策表

实际邮件类型	系统判为邮件类型	损失因子
垃圾邮件	垃圾邮件	0
垃圾邮件	正常邮件	1
正常邮件	垃圾邮件	k
正常邮件	正常邮件	0

© 中国科学院软件研究所 http://www.c-s-a.org.cn

根据表 1 可以认为把正常邮件错判成垃圾邮件的损失是把垃圾邮件判为正常邮件损失的 k 倍(理论上认为 k≧1), 只有当 P(C=Spam|di)/P(C=Ham|di)>k 时, 才判定邮件 di 为垃圾邮件. 又有 P(C=Spam|di)=1-P(C=Ham|di), 故有

$$P(C=Spam\,|\,di)>T \qquad (T=\frac{k}{1+k}) \tag{7}$$

当 P(C=Spam|di)>T 时, 可保证决策为垃圾邮件的风险比决策为正常邮件的风险小, 这种情况下分类器判定为垃圾邮件. 为了进一步减少垃圾邮件的错判情况, 在新邮件到来时, 我们进行更细致的分级判断

① 当 P(C=Spam|di)>T 时, 判定为垃圾邮件

② 当 P(C=Ham|di)> P(C=Spam|di)时,判定为正常邮件

③ 当 T>P(C=Spam|di)> P(C=Ham|di)时,判定为可疑邮件, 待用户人工进行再判断, 直到再次满足分类要求.

这种方法既保证了垃圾邮件的过滤效果, 又可以减少误判所带来的损失, 非常适用于对正确率要求比较严格的用户.

2 实验及评价

2.1 评价标准

表 2 邮件过滤系统判定情况分布

	实际为垃圾邮件	实际为正常邮件
判定为垃圾邮件	TP	FP
判定为正常邮件	FN	TN

a)查全率: R=[TP/(TP+FN)]×100% (8)

体现了模型识别垃圾邮件的能力, 即查全率越大, 漏网的垃圾邮件数量越少

b)查准率: P=[TP/(TP+FP)] ×100% (9)

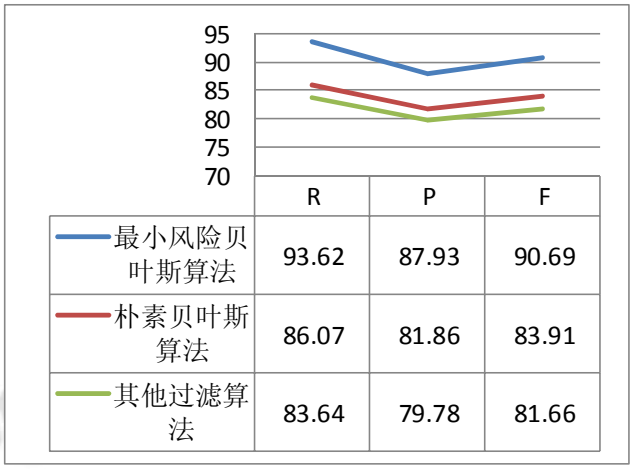
体现了模型检对垃圾邮件的能力, 即查准率越大, 正常邮件被误判为垃圾邮件的数量越少

c)调和率: $F=\frac{2\times R\times P}{R+P}\times 100\%$ (10)

由于在某些模型中查全率 R 和查准率 P 之间会相互影响(即一个大, 而另一个小), 因此实验将把 F 作为又一个重要性能评价指标. F 是 R 和 P 的调和平均, 是它们的综合体现[6-9].

2.2 实验结果

本文实验数据集来源于 PUI 公共词料库和 UCI 机器学习数据库中的垃圾邮件数据库, 选择 1500 条邮件测试集进行实验, 分别采用朴素贝叶斯算法以及本文提出的最小风险贝叶斯算法测试.

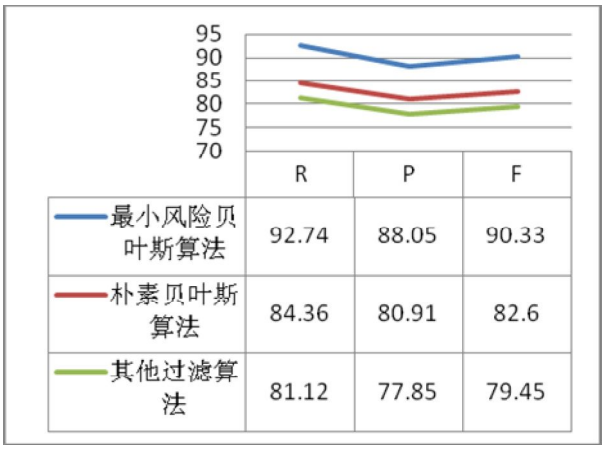


	R	P	F
最小风险贝叶斯算法	93.62	87.93	90.69
朴素贝叶斯算法	86.07	81.86	83.91
其他过滤算法	83.64	79.78	81.66

图 1 基于多种算法对数据库中垃圾邮件分类性能对比

图 1 可以看出朴素贝叶斯算法的查全率、查准率、调和率分别为 86.07%、81.86%、83.91%, 最小风险贝叶斯算法的查全率、查准率、调和率上升为 93.62%、87.93%、90.69%. 因此本文提出的最小风险贝叶斯算法在不考虑可疑邮件的情况下在三个指标上均优于朴素贝叶斯算法, 对比结果表明, 最小风险贝叶斯算法是一种有效的、分类精度高、误分率较好的垃圾邮件分类算法, 可以更好地满足垃圾邮件分类要求.

为了进一步测试本文提出的最小风险贝叶斯算法的有效性, 再次取邮件于 PUI 公共词料库和 UCI 机器学习数据库, 其中包含正常邮件 860 封和垃圾邮件 640 封. 将邮件分为 5 份, 每次取一定的份数作为测试集, 进行实验.



	R	P	F
最小风险贝叶斯算法	92.74	88.05	90.33
朴素贝叶斯算法	84.36	80.91	82.6
其他过滤算法	81.12	77.85	79.45

图 2 基于两种算法对数据库中邮件分类性能对比

© 中国科学院软件研究所 http://www.c-s-a.org.cn

从图 2 的对比结果看出朴素贝叶斯算法的查全率、查准率、调和率分别为 84.36%、80.91%、82.60%, 最小风险贝叶斯算法的查全率、查准率、调和率仍然全部上升, 分别为 92.74%、88.05%、90.33%. 再一次证明了本文提出的最小风险贝叶斯算法的有效性和高效性.

根据前面的讨论, T=k/(k+1), 可以按照用户的要求, 通过调整损失因子 k 的大小来控制阈值 T, 从而最终获得相对满意的结果. 实际应用中要想确定合适的 T 值需要有一定的经验和通过大量的实验, 往往还要根据所研究的具体问题, 分析误判决策所造成的严重程度等等. 采用最小风险贝叶算法[10-13]进行邮件过滤, 根据提供的阈值 T 不同产生不同的结果, 经过测试得出测试数据(表 3).

表 3　不同阈值 T 对性能的影响

k	T	R (%)	P (%)	F (%)
1.0	0.50	74.50	82.31	78.21
1.20	0.55	81.37	91.56	86.16
1.50	0.60	91.95	98.84	95.27
2.00	0.67	93.54	92.89	93.21
2.50	0.71	95.62	90.07	92.76

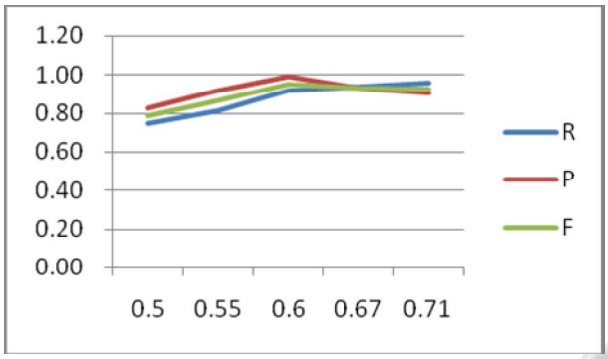


图 3　不同阈值 T 对性能的影响

我们从图 3 不难看出一定的规律: 随着阈值的增大, 查全率也在相应的增加, 而查准率却在增加之后逐渐降低, 调和率在对查全率和查准率进行调和之后也遵循着先增加再降低的规律. 这表明引入损失因子 k 之后, 漏判垃圾邮件的概率降低了, 但同时如果 k 值太大, 正常邮件被判定为垃圾邮件的概率会相应增加, 因此, 为了取得较好的性能指标, 要选取合适的损失因子.

在本文中经过多次选取 k 的大小, 最终确定几个有代表性的阈值, 从折线图中可以看出, 当阈值T=0.60 时, 各项性能指标相对较好, 所以选用该阈值对应的损失因子 k=1.50 时可以得到较满意的分类效果.

3　结语

基于贝叶斯的垃圾邮件过滤器是目前比较高效的垃圾邮件过滤技术之一, 它已经开始广泛的使用到垃圾邮件过滤领域[14,15]. 本文在对朴素贝叶斯过滤器分析的基础上, 针对朴素贝叶斯算法的缺陷并结合损失最小化的思想, 同时根据垃圾邮件的特性, 对朴素贝叶斯算法做了进一步改进, 提出了最小风险贝叶斯算法, 该算法能够通过调整损失因子 k 值, 使得正常邮件错判成垃圾邮件概率最小化, 从而最大程度减少用户的损失. 实验证明虽然该算法取得了更好的过滤效果, 但是还有很多问题亟待解决, 因此, 要想使得邮件过滤系统更加成熟化, 我们还需进行更深入的研究.

参考文献

1 李翔鹰,叶枫.一种基于多贝叶斯算法的垃圾邮件过滤方法.计算机工程与应用,2006,42(31):114–116.
2 王涛,裘国永,何聚厚.新的基于最小风险的贝叶斯邮件过滤模型.计算机应用研究,2008,25(4):1147–1149.
3 王美珍,李芝棠,吴汉涛.改进的贝叶斯垃圾邮件过滤算法.华中科技大学学报(自然科学版),2009,(8):27–30.
4 邓慧.基于关联规则的垃圾邮件分类模型.计算机应用与软件,2015,32(8):320–323.
5 Thiago SS,Walmir MC. A review of machine learning approaches to spam filtering. Expert Syst Appl. 2009, 36(7): 10206–22.
6 薛松,张钟澍,殷知磊.贝叶斯算法在反垃圾邮件应用中的改进方案.成都信息工程学院学报,2009,24(4):351–355.
7 罗倩,秦玉平,王春立.反垃圾邮件技术综述.渤海大学学报(自然科学版),2008,29(4):385−389.
8 王新艳.基于行为的垃圾邮件过滤技术研究.计算机光盘软件与应用,2015,18(3):176–177.
9 宋文,张明新,彭太乐.图像型垃圾邮件过滤技术研究综述.计算机系统应用,2011,20(10):255–258.
10 计宏.改进贝叶斯垃圾邮件过滤技术的研究.计算机测量与控制,2013,21(8);2181–2184.
11 吴志军.基于内容过滤的反垃圾邮件系统研究.无线互联科技,2015,10(14):121–122.
12 王忠建,张树舰,李颖.一种改进的基于贝叶斯的垃圾邮件过滤方法.黑龙江科技信息,2014,10(21);175–175.
13 王红玲.改进的贝叶斯算法在垃圾邮件过滤中的应用.信息通信,2013,(9):85–86.
14 Sun GL, Sun HY. Spam filtering: Online naive Bayes based on TONE.中兴通讯技术:英文版,2013,(2);51–54.
15 王斌,潘文峰.基于内容的垃圾邮件过滤技术综述.中文信息学报,2005,19(5):1–10.

© 中国科学院软件研究所 http://www.c-s-a.org.cn