基于贝叶斯算法的垃圾邮件过滤技术研究

顾 玮
（徐州高等师范学校 江苏 徐州 221116）

摘 要 分析了垃圾邮件内容过滤技术，认识到垃圾邮件过滤技术与普通的文本分类和挖掘问题存在着很多不同。从邮件结构不同于普通文本出发，对基于贝叶斯的过滤方法进行了一系列改进，提出一种阈值调整算法，设计了集成加权模型，以充分利用邮件的结构信息。基于集成加权模型对邮件头和邮件正文分别建立模型，最后通过加权方法集成二者结果，对垃圾邮件进行过滤。通过在改进和扩展而设计的贝叶斯过滤器在最新的标准数据集上的测试结果表明，与经典的贝叶斯过滤器 Bogo 相比，过滤效果有较大的提高。

关键词 集成加权贝叶斯 最小风险贝叶斯 主动学习贝叶斯 特征选择 阈值调整
中图分类号 TP391 **文献标识码** A **文章编号** 6148

Research on Spam Filtering Technology Based on Bayesian Algorithm

Gu Wei
（Xuzhou Higher Normal School Xuzhou 221116）

Abstract This paper analyzes the content filtering technology of spam and realizes that there are many differences between spam filtering technology and common text classification and mining problems. In this paper, the filtering method based on bayes is improved, and the integrated weighted model is designed to make full use of the structure information of the mail. Based on the integrated weighted model, a model was established for the mail head and the body of the mail, and finally, the results were integrated with the weighted method to filter the spam. By designed to improve and expand the bayesian filter in the latest standard data sets on the test results show that compared with the classical bayesian filter Bogo, filtering effect has great improvement.

Keywords Integrated weighted bayes Minimal risk bayes Active learning of bayes Feature selection Threshold adjustment

一、引言

目前我国垃圾邮件泛滥，情况十分严重。全球前十大垃圾邮件大国之中，中国仅次于美国高居垃圾邮件大国第二。中国网民收到的垃圾邮件数量占全球的十分之一，并且这个数字每五个月会翻一番。随着垃圾邮件的泛滥，垃圾邮件过滤技术也在不断的发展，产生了许多对付垃圾邮件的方法。一些成熟的方法在客户端、服务器端等被广泛的应用，新的方法也在不断的产生。本文对贝叶斯模型在垃圾邮件过滤中的应用进行深入研究，并针对一些关键问题提出了基于贝叶斯的过滤方法一种阈值调整算法，设计了集成加权模型，对邮件头和邮件正文分别建立模型，最后通过加权方法集成二者结果，对垃圾邮件进行过滤[1]。通过在改进和扩展而设计的贝叶斯过滤器在最新的标准数据集上的测试结果表明，与经典的贝叶斯过滤器 Bogo 相比，过滤效果有较大的提高。

二、贝叶斯定理

贝叶斯定理的描述如下：

对于一个统计实验 ε，样本空间 s 是所有可能结果的集合，并且$\{B_1, B_2, \cdots, B_r\}$是 s 的一个划分。令$\{p(A);\ A\subseteq S\}$表示定义在 s 中所有事件上的一个概率分布，则对于 s 中的任意事件 A 和 B，有 $p(A)>0$，$p(B\mid A)=p(A\cap B)/p(A)$表示条件概率，即在已知 A 发生的情况下 B 发生的概率[2]。贝叶斯定理可以表示为：

$$p(B_i\mid A)=p(A\mid B_i)p(B_i)/p(A)(i=1,2,\cdots,r) \quad \text{公式(1)}$$

其中 $p(A)>0$，由全概率公式可得

$$p(A)=\sum_{j=1}^{r}p(A\mid B_j)p(B_j) \quad \text{公式(2)}$$

在公式(1)中，$p(Bi/A)$为后验概率，$p(A\mid B_i)$为似然概率，$p(B_i)$为先验概率。

三、改进贝叶斯过滤器的设计

1、特征提取模块

（1）文本块的划分。这种算法从文档的每一个位置开始，都取出一度为 k 的子字符串。这样在一个长度为 n 的文档中，一共包含 n-k+1 个这样的字符串。

（2）散列函数采用了 Karp-Rabin 算法

Karp-Rabin 算法是一种著名的模式匹配算法[3]。它通过把字符串的匹转换成相应的整数的匹配，而大大提高了匹配的效率。这种算法在计算"相应的整数"时，对长度为 k 的子字符串t_i,.....,t_{i+k-1}用了以下的散列函数

$$H_i = t_i * p^{k-1} + t_{i+1} * p^{k-2} ... t_{i+k-2} * p + t_{i+k-1}$$

（P 为常数） 公式（3）

如果对每一个子字符串的散列值都从头开始计算，代价是很高的，karp-Rabin 算法采用了以下方法来进简化：

$$H_{i+1} = p * H_i + t_{i+k} - t_i * p^k \quad \text{公式（4）}$$

由于 p^k 是一个常数，所以采用了以上的简化方法以后，每个指纹元素 Hi 的计算代价就只有两次乘法和加减法各一次，是非常低的。

2、特征选择模块

对一封邮件，我们分别提取邮件头，邮件正文特征，由于邮件头含有的关键特征少，邮件头在特征选择上提取 8 个特征，邮件正文在特征选择上选取 10 个特征，然后对邮件头和邮件正文分别得到类别分数$score_1$，和$score_2$然后对$score_1$，$score_2$分别赋予不同的权重 a 和 b，最后得出类别权重分数 score，计算 p 的公式如下所示：

$$\text{score}=(a * score_1 + b * score_2)/(a+b) \quad \text{公式（5）}$$

对于权重系数 a 和 b，我们考虑邮件头和邮件正文单独过滤时的历史准确率。设 N 表示邮件总数；N_{top}表示邮件头过滤器历史上判断正确的邮件数，也就是邮件头过滤器中判断正确的 Spam 与判断正确的 Ham 数量之和[4]。N_{body}表示邮件正文过滤器历史上判断正确的邮件数，也就是邮件正文过滤器中判断正确的 Spam 与判断正确的 Ham 数量之和。R_{top}表示邮件头过滤器的历史准确率；R_{body}表示邮件正文过滤器的历史准确率。则有：

$$R_{body} = \frac{N_{body}}{N},$$

$$a = \frac{R_{top}}{R_{top} + R_{body}}, \quad b = \frac{R_{body}}{R_{top} + R_{body}} \quad \text{公式（6）}$$

我们分别取 trec07p 邮件集前 10000,20000,40000,60000 封邮件，得到 a,b 的取值权值 a 的范围为 0.51-0.53，权值 b 的取值范围为 0.46-0.49。

3、联合概率计算

取邮件头类条件值 最大的 8 个特征，进行联合概率计算；取邮件正文类条件值最大的 10 个特征进行联合概率计算。联合概率计算方法如下：

$$f(w) = \frac{a * x + (n * p(w))}{a+n} \quad \text{公式（7）}$$

$$P = 1 - \sqrt[n]{(1-f(w)_1) * (1-f(w)_2) * ... * (1-f(w)_n)} \quad \text{公式（8）}$$

$$Q = 1 - \sqrt[n]{f(w)_1 * f(w)_2 * ... * f(w)_n} \quad \text{公式（9）}$$

$$S = \frac{1 + \frac{P-Q}{P+Q}}{2} \quad \text{公式（10）}$$

其中公式（7）计算了特征的平滑概率，公式（8）至公式（10）联合概率的计算公式，其中公式（10）将概率映射到 0-1 的区间。

4、集成概率计算

得到邮件头联合概率$score_1$和邮件正文联合概率$score_2$，计算集成加权概率$score$：

$$score = 0.52 score_1 + 0.48 score_2$$

5、阈值判断

贝叶斯算法在构造模型的时候，最初学习的样本比较少，所以存储关键词的 hash 表中关键词的数量也比较少，导致最初的判断不是很确定。随着学习的深入，hash 表中关键词数量增大，导致贝叶斯模型趋于稳定[5]。那么基于不断学习的过程，我们提出一个阈值调整自适应算法。

（1）若 sor(T)<stv(T)*20%，则 TH(T+1)=TH(T)*1.2

如果垃圾邮件判断准确率过低，则迅速提高阈值。

（2）若 hor(T)<htv(T)*20% ，则 TH(T+1)=TH(T)*0.8

如果正常邮件判断准确率过低，则迅速减少阈值。

（3）若 class(d^n)=spam，judge(d^n)=ham

则 TH(T+1)=TH(T)*0.98，若一封邮件的类别是垃圾邮件，而过滤器判断为正常邮件，则减少阈值。

（4）若 class(d^n)=ham，judge(d^n)=spam，则 TH(T+1)=TH

(T)*1.02 若一封邮件的类别是正常邮件,而过滤器判断为垃圾邮件,则增加阈值其中 sor(T)为在 T 时段正确检出的垃圾邮件的数量;stv(T)为在 T 时段检出的垃圾邮件的数量;hor(T)为在 T 时段正确检出的正常邮件的数量;htv(T)为在 T 时段检出的正常邮件的数量。class(d^n)为邮件 d_n 的实际类别,judge(d^n)为过滤器判断的邮件类别。

四、实验及性能分析

bogofilter 是一款开源国外垃圾邮件过滤器。Bogo 过滤器采用的也是朴素贝叶斯算法,每一个 token 为用空格隔开的短语,用 wordlis.db 记录 token 的出现次数,采用联合概率计算邮件为垃圾邮件的分数。为了评价本课题所研究的改进贝叶斯的过滤精度,我们和同样使用朴素 baeys 算法的bogo 过滤器进行比较,并且在离线和在线两种过滤模式下进行。

1、离线过滤模式下的性能比较

离线模式就是采取先训练、再过滤的模式。我们把 sewm2008 公开邮件集分成两部分,前 30000 作为训练集,后四万封作为测试集,用两个过滤器进行过滤得到的结果可以看出,本文研究的朴素贝叶斯过滤器在合法邮件误判率,非法邮件误判率以及 roc 参数上都要比 bogo 小,这说明在过滤精度上,本文所研究的过滤器要优于 bogo。但是由于 bogo 采取高效的数据存储技术,使得在邮件过滤效率上要优于朴素 baeys 过滤器。本文所研究的过滤器每封邮件的平均处理时间是 0.28s,虽然长于 bogo 的处理时间,但是在现实邮件过滤中,还可以接受。

2、在线过滤模式下的性能比较

在线过滤模式就是采用一边训练,一边测试,测试得到的邮件再加入到训练集中。为了比较本课题研究的朴素贝叶斯过滤器与 bogo 在在线过滤模式下的过滤精度,采用 sewm2016 公开邮件集作为测试邮件集,并采取立即反馈的测试方法,在线反馈模式中,本课题研究的过滤器在过滤精度上更加优于 bogo 过滤器,但是在过滤效率上,还是不如 bogo 过滤器。

五、结论

分析了垃圾邮件内容过滤技术,认识到垃圾邮件过滤技术与普通的文本分类和挖掘问题存在着很多不同。实现了改进的朴素贝叶斯过滤器,并且与同样使用贝叶斯算法的 bogofilter 过滤器进行实验对比。通过实验对比分析,可以知道,改进的朴素贝叶斯算法在在线立即反馈过滤和离线过滤精度上面,能够取得更好的效果,但是在过滤效率上,不如 bogofilter 过滤器。

参考文献

[1]宫秀军,刘少辉,史忠植. 一种增量贝叶斯分类方法[J]. 计算机学报,2012,56(6):3~5.

[2]Herber A. Simon, Glenn lea. Problem Solving And Rule Reduction[J] Knowledg And Cognition, 2014, 78(8):145–156.

[3]周水庚,关估红,俞红奇,胡运发. 基于 N–Gram 信息的中文文档分类研究 [J]. 中文信息学报. 2011, 34(1):34~39.

[4]宫秀军,孙建平,史忠植. 主动贝叶斯网络过滤器[J]. 计算机研究与发展,2012,39(5):6~9.

[5]周威成,马素霞,齐林海. 一种基于机器学习的垃圾邮件智能过滤方法[J]. 现代电力,2013,20(1):65~67.

作者简介

顾 玮,女,1981 年生,汉族,徐州高等师范学校教师,讲师,硕士研究生,研究数据库,数据挖掘,机器学习,系统开发等相关知识。

(上接第 48 页)

参考文献

[1] 汤洪宇,周国庆,唐 臣,黄西宁.一体化教学在《物流管理信息系统》课程中的应用 [J]. 物流科技,2015(4):50~52.

[2] 王 晶,孙 芳.利用微信公众平台实现辅助教学[J]. 农业网络信息,2014(10):124~126.

[3] 王雪芳.以学生为中心导向的《管理信息系统》教学探讨[J].新一代,2012(10):166~167.

[4] 李 艳. 基于创新型人才培养的管理信息系统教学改革探讨[J].湖北经济学院学报(人文社会科学版),2012(09):178~179.

作者简介

张利娜,1979 年出生,女,河南人,硕士研究生,副教授,研究方向为电子商务、信息技术教育,东南大学成贤学院。

郑桂玲,1982 年出生,女,湖南人,硕士研究生,讲师,研究方向为电子商务、数据挖掘,东南大学成贤学院。