基于 Adaboost 改进的贝叶斯垃圾邮件检测算法

钱思杰

（东南大学信息科学与工程学院，江苏省 南京市 211100）

2018 年 3 月 1 日

摘　要　随着电子邮件的应用与普及，垃圾邮件的泛滥也越来越多地受到人们的关注。本文对基于贝叶斯的垃圾邮件过滤器的原理及其关键技术进行了详细的描述。针对朴素贝叶斯模型对分类信息过度简化和准确率低等缺点，通过引入 Adaboost 算法和对贝叶斯过滤器进行了改进，并进行了实验。实验结果表明，改进后的贝叶斯过滤器具有了更好过滤效果。

关键词　贝叶斯分类，Adaboost，垃圾邮件，朴素贝叶斯

An Improved Bayesian Spam Detection Algorithm Based on Adaboost

Qian Sijie

(School of information science and engineering,South East University; Nanjing; 211100)

Abstract　With the popularization and application of e-mail, the spread of spam is also attracting more and more attention. This paper describes the principle of Bayesian spam filter and its key technologies in detail. To overcome the shortcoming of over-simplified classification information and low accuracy of naive Bayesian models, the Adaboost algorithm is introduced and the Bayesian filter is improved and experiments are carried out. Experimental results show that the improved Bayesian filter has a better filtering effect.

Key Words　Bayesian classification, Adaboost, spam, Naive Bayes

1　引言

随着 Internet 的发展，电子邮件作为一种快捷、经济的通信方式得到了普及。邮件是当前 Internet 中最多的两项应用之一 [1]。当前网络上有超过 50% 的邮件是垃圾邮件，不仅占用了网络的带宽及邮件服务器的存储空间，同时也浪费了用户的时间和精力。人们提出了垃圾邮件问题的多种解决方法。目前主流的有这么几种：

(1) 针对邮件头信息进行过滤的方法如黑白名单法反垃圾邮件技术主机名反向验证技术;

(2) 协议改进的方法，比如①IRTF 提出三个在不放弃 SMTP 等协议的情况下对邮件地址进行校验的方案。②雅虎的 DomainKeys 方案。为邮件服务器编写出特定的检测软件，检测发送方的域合法性，并对邮件标上加密的验证标签，带有标签的邮件才为正常邮件。

(3) 针对邮件内容进行过滤的方法如各类贝叶斯关键元加权统计算法和它的变种;

第一类方法，相对简单并且已经被广泛使用，但是存在局限性，并不能阻止所有的垃圾邮件，还需要其他方法作为补充。

第二类方法，应用部署受到局限，很难在所有使用电子邮件的用户中推广。

第三类方法，即基于贝叶斯的过滤方法，具有智能学习功能，能够针对内容进行不断的学习和过滤，具有长久的适用性。国内外都已经有人使用该方法进行实验，取得了一定的成果 [2]。

2 相关技术介绍

2.1 贝叶斯定理

贝叶斯定理是由托马斯贝叶斯（1702-1761）提出的计算概率的一种方法。它是通过对某一事件过去发生概率情况的考察，大体可以判断出当前这一事件发生的概率。它的形式化表述为：设实验 E 的样本空间为 S，A 为 E 的事件，B_1，B_2，$\cdots$，B_n 为 S 的一个划分，且 $P(A) > 0$,$P(B_i) > 0(i = 1, 2, \cdots, n)$，则

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{N} P(A|B_j)P(B_j)} \tag{1}$$

2.2 贝叶斯过滤器

贝叶斯过滤器将垃圾邮件过滤问题看成是垃圾邮件文本的二值分类问题。对于每一个邮件样本，使用向量空间模型对其进行形式化描述，$d = (t_1, w_1; t_1, w_1; \cdots, t_i, W_i; \cdots; t_n, w_n)$,$t_i$ 是邮件选取特征项，w_i 是 t_i 的权值。垃圾邮件判定中的邮件类别：$c \in C = \{Spam, Ham\}$ 邮件分类器的任务就是计算邮件待分类邮件是垃圾邮件的概率，如果他超过某一阈值则认为该邮件是垃圾邮件。假设一封新邮件 $d_x = (x_1, x_2, \cdots, x_n)$ 根据贝叶斯理论，未知邮件属于垃圾邮件分类的概率计算公式为

$$P(S|d = d_x) = \frac{P(d = d_x|S)P(S)}{P(d = d_x)} \tag{2}$$

其中，$P(d = d_x|S)$ 指垃圾邮件中 d_x 中所有特征项 $(x_1, x_2, \cdots, x_n)$ 同时出现的概率；$P(S)$ 指选取样本中垃圾邮件的概率。当 $P(d = d_x|S)$ 大于 $P(d = d_x|H)$ 或大于某个阈值时则认为该邮件为垃圾邮件 [4]。

2.3 朴素贝叶斯模型

朴素贝叶斯模型建立在一般贝叶斯过滤器的基础上，通过假定各个因素之间不存在任何关联，即完全独立而得到一种简化后的贝叶斯模型 [3]。朴素贝叶斯模型的独立性大大降低了计算的复杂度，且具有较高的精确度。对于邮件中出现的所有词汇，考虑每个词汇出现事件的独立性，计算 $P(S|x_i)$ 的联合概率 $P(S|d_x)$ $d_x = x_1\ x_2\ \cdots\ x_n$：

$$p = \frac{p_1 p_2 \cdots p_n}{p_1 p_2 \cdots p_n + (1 - p_1)(1 - p_2) \cdots (1 - p_n)} \tag{3}$$

其中 p 即 $P(S|d_x)$，出现词汇 $d_x = (x_1, x_2, \cdots, x_n)$ 的邮件是垃圾邮件的条件概率；P_i 即 $P(S|x_i)$，出现词汇 x_i 的邮件是垃圾邮件的条件概率。

3 改进的 Adaboost 朴素贝叶斯算法

3.1 邮件预处理

我们可以将一封邮件看成是一个向量，该向量的每一维都是一个文本，因此，对邮件的分类实质上就是对文本的分类。在文本分类领域中，文本表示主要采用向量空间模型，它的基本思想是用向量表示文本，文本中的每个词作为向量的一维，将文本映射到向量空间中的一个向量。

本文在邮件表示的时候先对导入的邮件内容进行筛选预处理，首先是去除空字符串，并统一小写，然后去掉非字母非数字，只留下单词。最后是对训练样本进行标记，垃圾邮件标记为 1，正常邮件标记为 0，将标记和原来的文本内容结合就成了训练样本。

3.2 邮件特征提取

邮件特征提取是从邮件的特征集合中选择那些最能反映和代表该邮件主题内容的特征项。由于邮件中的词汇量很大，在用表示邮件之后，生成的向量空间是一个高维的特征向空间，有时甚至可达上万维，在这样一个高维向量空间上进行训练和分类是一件非常困难的事因此，需要对特征向量进行降维处理，这样不仅能提高分类效率，还能有效地避免“过学习”现象。

在对邮件进行预处理之后，统计所有训练样本中出现过的词汇，并且生成相应的词汇库；接着标记词汇库中的每隔词汇出现的次数；最后标记每封训练样本邮件中的单词在词汇库中出现的次数。

3.3 Adaboost 算法

Adaboost 是一种迭代算法，它利用样本权值来确定训练集的样本分布。一开始，所有样本都赋予相同的权值 $1/N$，从而使得它们被抽取的可能性一致 [5]。

Adaboost 算法通过每一轮更改训练样本的权值，能够自适应地改变训练样本的分布，并且根据每一轮的样本权值训练出一个基分类器，最后根据基分类器的投票决定待测样本的分类，这样有效地避免了单分类器对训练样本的敏感性。

在本文中，在计算联合后验概率的时候，引入了一个调整因子，作用是用来表示词汇表中某一个单词的“垃圾权值”，即 $d = (x_1, w_1; x_2, w_2; \cdots, x_i, w_i; \cdots; x_n, w_n)$，$x_i$ 是邮件选取特征项，w_i 是 x_i 的权值。其中 w_i 通过 Adaboost 算法迭代获取最佳值。原理如下：

1）特征向量构成之后先用朴素贝叶斯算法对特征向量初次训练，得到初始的特征向量训练集合及其类别，当处理完最后一封时，计算先验概率 $P(S)$ 及给定条件下的条件概率 $P(d_i|S)$ 和 $\mathrm{P}(d_i|H)$；

2）设定 Adaboost 的循环次数 count；

3）随机选择 n 个样本；

4）w_i 初始化为何词汇表大小相等的全一向量；

5）循环迭代 count 次，每次迭代的时候计算各个样本在 w_i 下的样本分类，如果分类出错，计算出错的程度；

6）比较 count 个出错率，找到最小的错误率和此时的词汇表、w_i、$P(S)$、$P(d_i|S)$ 和 $\mathrm{P}(d_i|H)$，即保存训练好的最佳模型的信息。

4 实验及评价

本文所有实验都是在笔记本（Intel CORE i7，2.90GHz，8.0GBRAM），使用 Python 实现。本文的实验样本来自网上开源的 5574 封邮件，其中使用 4574 作为训练集，其余的 1000 封作为测试集。

测试效果：5574 个样本，获取 Adaboost 算法训练的最佳模型信息（包括词词汇表、DS、$P(S)$、$P(d_i|S)$ 和 $\mathrm{P}(d_i|H)$），对 1000 个测试样本，分类的平均错误率约为：0.5%。

为了进行比较，本次实验还赠加了一组不用 Adaboost 算法的常规朴素贝叶斯分类模型，即可以看做每个单词的“垃圾权值”都相等，同样选用 5574 个样本，采用交叉验证，随机选取 4574 个作为训练样本，产生词汇列表（语料库），对 1000 个测试样本，分类的平均错误率约为：2.5%。

从结果上来看，加入了 Adaboost 的朴素贝叶斯算法相比较于一般的朴素贝叶斯算法在分类的准确率上有着明显的提升。

5 进一步展望

实验表明 Adaboost 算法的准确率较好。但实验过程中 Adaboost 算法的耗时较长，需今后进一步优化完善，同时 Adaboost 倾向于那些被错误分配的样本，因此，它很容易受到过拟合的影响。怎么样避免过分拟合是接下来待研究的问题。

一个改进的措施 [6] 是不再固定训练子集的大小，而是根据每个样本的权重和训练样本集容量的乘积上取整结果，决定每个样本被选入新训练子集的次数。一方面，使得训练子集都包含了所有的样本，没有信息遗失，提高了分类性能。另一方面，避免了训练子集中某一类别样本数目很多，其它类别样本数目很少甚至没有的情况，从而有效避免了过拟合和偏见问题。

另外一方面，由于朴素贝叶斯文本分类是建立在条件独立假设和位置独立假设的基础上，而现实的文本当中，这两个假设是不成立的。所以可以在朴素贝叶斯分类器的基础上增加特征间可能存在的依赖关系，以此削弱朴素贝叶斯的独立假设。

6 结论

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。本文通过引入 Adaboost 算法对朴素贝叶斯算法进行改进，实验证明使贝叶斯过滤器取得了更好的过滤效果。文中虽然对某些问题做了探讨，但仍有许多亟待解决的问题值得拓展研究。

参考文献

[1] 刘心宇. AdaBoost 和主动学习方法在邮件分类中的应用研究 [D]. 哈尔滨理工大学,2016.

[2] 张江霞. 基于 Adaboost 算法的控制图模式识别应用研究 [D]. 南京理工大学,2013.

[3] 严超, 王元庆, 李久雪, 张兆扬.AdaBoost 分类问题的理论推导 [J]. 东南大学学报 (自然科学版),2011,41(04):700-705.

[4] 李茹, 刘培玉, 朱振方. 基于 AdaBoost 的最小风险贝叶斯的垃圾邮件过滤算法 [J]. 济南大学学报 (自然科学版),2011,25(01):19-22.

[5] 陈松峰, 范明. 利用 PCA 和 AdaBoost 建立基于贝叶斯的组合分类器 [J]. 计算机科学,2010,37(08):236-239+256.

[6] 卢婷. 基于 AdaBoost 的分类器学习算法比较研究 [D]. 华东理工大学,2014.