**基于朴素贝叶斯算法的垃圾邮件分类方法研究**

**摘要：**随着因特网的发展和普及，电子邮件也得到了广泛的使用。电子邮件在给人们带来了方便的同时也产生了一个新的问题，即大量垃圾邮件的出现。垃圾邮件作为电子邮件的副产物日益影响着人们的日常生活，因此研究垃圾邮件的分类具有重要意义。

基于此，本文实验采用5折交叉验证法，首先收集的一些邮件为样本，其次，提取文本特征，然后，使用朴素贝叶斯分类算法，通过训练得到垃圾邮件以及非垃圾邮件分类器，从而对测试集的邮件进行分类。实验证明：该方法具有良好的效果。

**关键词：**邮件分类，朴素贝叶斯，垃圾邮件，特征提取

**Abstract:** With the development and popularization of internet, E-mail has been widely applied. It also brings convenience to the people but have a new problem--there is a lot of spam. Spam is increasingly affecting people’s daily life, so research spam classification is very great significance.

Based on this, the experiment uses 5 fold cross validation method. Firstly collected some of the mail as a sample, secondly, extracted the text feature, then, used the naive Bayesian classification algorithm, through training to get spam and non-spam classifier, Test set of messages for classification. Experiments show that this method is a good for identifying spam email.

**Keywords:** email classification, naïve Bayesian, spam, feature select

**1绪论**

**1.1研究背景**

Internet的迅速发展，人与人的交往更加方便，电子邮件以其快捷低廉的特性逐渐成为人们信息交互的重要工具。人们用它来交流思想，传输文件，发表意见等，逐渐成为日常生活中不可缺少的通信工具。但是电子邮件在其给人们带来极大便利的同时也带来了一些负面影响，那就是我们每天收到的邮件有很大一部分是不请自来的。它们有些是商业广告，有些是政治宣传，有些是色情广告，还有一些甚至是病毒，这就是我们俗称的垃圾邮件。

根据美国nucleus research 公司公布的数据，全球每天大约有140亿封垃圾邮件在网上传播，相当于地球上每个人每天都要收到两封以上的垃圾邮件。

垃圾邮件给网民造成的经济损失是相当惊人的：据统计仅下载它们所花费的上网费和电话费用等费用，每年就会花掉全球网民94美元。作为垃圾邮件的发送方，价格是及其低廉的，通常是通过各种方式群发。而对于电子邮件服务提供商和用户来说，垃圾邮件却给他们带来了很大的危害和损失，而且如色情，电脑病毒等造成的损失更是造成难以评估。

**1.2贝叶斯研究简介**

贝叶斯的论文“关于几率性问题求解的理论”奠定了贝叶斯学派的基础。而后著名数学家laplace用贝叶斯理论导出的”相继律”使得贝叶斯理论受到人们的关注。但是由于当时贝叶斯方法在理论和实际的应用中还存在很多不完善的地方，因而在十九世纪并未被人接受。八十年代以后，人工智能的发展尤其是机器学习，数据挖掘等兴起，为贝叶斯理论的发展和应用提供了更为广阔的空间。

尽管对于贝叶斯学派哲学上的观点还是存在很多的异议，然而它的思想和方法在社会生活和生产实践中得到越来越广泛的应用确实不争的事实。尤其是近年来，贝叶斯方法以其独特的不确定知识表达形式，丰富的概率表达能力，综合先验知识的增量学习特性等成为当前数据挖掘众多方法中最为引人注目的焦点之一。

**1.3贝叶斯垃圾邮件过滤发展史**

1996年，Rvennie基于贝叶斯算法建立了一个用于邮件过滤的机器学习应用系统ifile[1]，利用贝叶斯算法对邮件进行分类。在建立ifile系统的过程中，Rennie注意到每个用户有不同的邮件集，并且组织邮件 的方式也各不相同，因此允许用户手工调整被误判的邮件。

1998年，Sahami利用贝叶斯算法对邮件进行过滤时，注意到垃圾邮件具有不同于合法邮件的特有属性：例如，在快速致富类的垃圾邮件中，除了邮件的文本中会含有许多free money之类的文本信息之外，还会有大量类似于 “!!!”的强调符号以及“$”这种表征线的符号。所以Sahamiliy利用朴素贝叶斯算法滤邮件时，手工加入了这些特定任务的域信息短语以及具有垃圾邮件特征的字符到过滤器中，提高了过滤垃圾邮件的精确度；另外，他还利用一个表征损失率的门槛值来降低合法邮件的误判率。

2001年，Matthew等人开发了一个垃圾邮件过滤器MEF。MEF可以在UNIX中过滤掉附件中有可执行程序的病毒邮件。该邮件过滤器首先对可执行程序的二进制码进行解码，然后把它与现有病毒的二进制码进行比较，利用朴素贝叶斯算法计算出它属于垃圾邮件的概率值，并据此作出决策。

文献利用朴素贝叶斯算法设计了垃圾邮件过滤系统SpamCop[2]。该系统能够识别大约92%的垃圾邮件，并且有1.16%的错纠率。SpamCop系统在关键词的选取原则上进行了改进，忽略掉空格，连续序列的字母和数字，以及除掉上述字符之外的不超过长度为3的连续序列字符。并且该系统对特征类概率的计算使用了m估计，采用了如下的公式：

 (1-1) (1-2)

 (1-3)

其中，表示特征Token属于垃圾类别的后验概率。代表该关键字在垃圾邮件中出现的次数，表示关键词在正常出现的次数。表示正常邮件的个数，代表垃圾邮件的个数。K代表待检邮件中不同关键字的个数，从而解决了零概率问题。

文献提供了一个有效的过滤垃圾邮件的贝叶斯方法[3]，该过滤器抓住了99.5%的垃圾邮件，且错纠率低于0.03% 。该过滤器对垃圾邮件和正常邮件集分别建立两个哈希表，用于统计相应得语料库的关键词及其出现的次数。计算每一个关键词的概率，使用以下的公式：

 (1-4)

其中b代表该关键词在垃圾邮件集中出现的次数，g代表该关键词在正常邮件集中出现的次数，nbad代表垃圾邮件的总数，ngood代表正常邮件的总数。在分母中的系数2是一个推荐的经验值，用于减少把正常邮件当作垃圾邮件的概率。在计算联合概率时，过滤器使用如下的公式：

 (1-5)

其中（i=1，2，…，n）表示已经计算出来的第i个关键词的概率。

当使用公式(1-5)计算一个邮件属于垃圾邮件的概率时，会经常遇到数据稀疏问题，即如果邮件中包含一个新出现的特征项，则不管这封邮件中包含的其他特征项的条件概率有多高，都会导致条件概率的零概率问题。这是个不容忽略的问题。其中文献[3]给出了一个零概率平滑公式，较好的解决了零概率问题。公式如下所示：

 (1-6)

其中a是一个可调节的常量，n为包含特征w的垃圾邮件和正常邮件的总合，x是初始概率，当n=0时，f(w)等于初始概率，随着n的增大，f（w）越来越接近p（w）。根据经验，一般设置初始概率为0.52，a值为0.0178。文献[4]对公式1-5也进行了改进。给出了计算关键词联合概率的新方法。

 (1-7)

 (1-8)

 (1-9)

S的值介于-1和1之间，高的值意味着垃圾邮件，低的意味着正常邮件。0意味着介于两者之间。

上述利用朴素贝叶斯构建的过滤器大多在贝叶斯的计算公式上进行改进，没有考虑到邮件过滤与普通的文本分类的区别，另一方面，这些传统的贝叶斯方法都是基于最小错误率的决策方法，没有考虑到合法邮件与垃圾邮件具有的不同特性，即合法邮件误判为垃圾邮件可能给用户带来更大的损失。另外传统的贝叶斯学习算法使用给定的训练样本学习分类参数，它所处理的训练样本必须是带有类别标注的，并且随机选择样本，被动接受这些样本的信息。

**2 相关技术简介**

**2.1 贝叶斯定理**

贝叶斯定理是决策逻辑学的一个分支，使用理论统计学研究概率推论，即根据已经发生的事件来预测将来可能发生的事件。贝叶斯理论假设，如果过去试验中事件的出现的概率己知，那么根据数学方法可以计算出未来试验中事件出现的概率。贝叶斯理论指出，如果事件的结果不确定，那么量化它的唯一方法就是事件的发生概率。

垃圾邮件过滤其实就是邮件分类问题，把邮件分为垃圾邮件和正常邮件。这就可以应用贝叶斯定理，通过对己经正确分类的邮件来预测新接收的邮件是否为垃圾邮件。

贝叶斯定理的描述如下：

对于一个统计实验，样本空间s是所有可能结果的集合，并且

是s的一个划分。令{p(A)； A 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，表示条件概率，即在己知A发生的情况

下B发生的概率。贝叶斯定理可以表示为：

 (i=1， 2，…，r) (2-1)

其中p(A) >0，由全概率公式可得

  (2-2)

在公式(2-1)中，p(Bi / A)为后验概率，为似然概率，为先验概率。

**2.2朴素贝叶斯的原理**

贝叶斯方法是垃圾邮件过滤中一个重要的方法，该方法的实质是把邮件确定为垃圾邮件或者正常邮件，这是一个分类问题。

设有m个样本空间，邮件d中有n个特征项，对于给定的类(k=1,2,…,m)，d属于类的概率为

 (2-3)

通过贝叶斯概率公式可得：

 (k=1,2,…，m) (2-4)

其中：

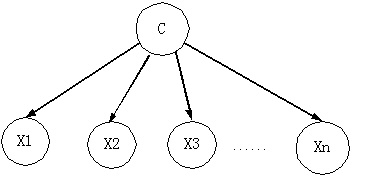
 (2-5)

公式(2-4)中的分母p(d)与类别无关，因而在公式(2-3)中比较最大值的时候可以忽略，所以只需计算概率和即可划分邮件d的类别。

公式(2-4)中，为先验概率，很容易计算，但的计算比较困难，特别是在特征项的数量较大，且特征项之间相依程度较高时，其计算将是极其费时间的。为了简化计算，引入了条件概率独立假设，即假定各特征项之间是相互独立的，这就是朴素贝叶斯过滤器，那么公式(2-5)就可以转换为：

 (2-6)

朴素贝叶斯过滤器的结构如图所示：



**图 2-1 朴素贝叶斯过滤器结构图**

朴素贝叶斯过滤器主要是利用先验概率求出后验概率，并且根据训练样本集构造过滤器，过滤器根据邮件的后验概率对文本进行分类。

**2.3 贝叶斯过滤器**

Bogofilter是目前比较流行的贝叶斯过滤器，它的主要原理是朴素贝叶斯理论。Bogofilter建立垃圾邮件和非垃圾邮件贝叶斯概率模型，在贝叶斯原理的实现上，加入了Paul Graham 关于垃圾邮件的过滤理论。该理论大体思想是，在已知的垃圾邮件中，一些单词出现的频率较高，而在合法邮件中，另一些单词出现的频率较高。运用一些众所周知的数学知识，对于每个单词，可以生成一个”垃圾邮件指示性概率”。根据消息中所办含一组词，可以用另一个简单的数学公式来确定文本消息的整体垃圾邮件概率。

Bogofilter将由空格隔开的单词作为特征，并且对特征进行更加严格的定义，譬如，去除单纯包含数字的特征，对于$20-25这种形式的价格范围，被标记为两个关键词，$20和$25等。Bogofilter还使用了平滑技术，来加强过滤器的过滤精度。

在过滤效率上，Bogofilter采取有效的数据表示，和高效的数据存储技术，获得比较高的过滤效率。Bogofilter使用高性能的Berkerly DB 数据库。Berkerly DB是历史悠久的嵌入式数据库系统，其小巧，可靠，性能高。Berkerly DB 比SQL SERVER 等数据库性能要高10-20倍。

**2.4贝叶斯过滤器的优缺点**

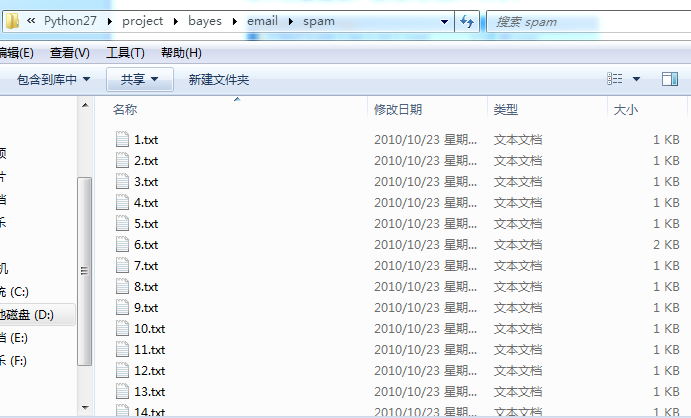
朴素贝叶斯算法，其条件概率独立假设，虽然忽略了特征的条件依赖性，但是，在许多实际应用领域中取得了很好的结果，而且其计算简单，降低了算法的复杂性。

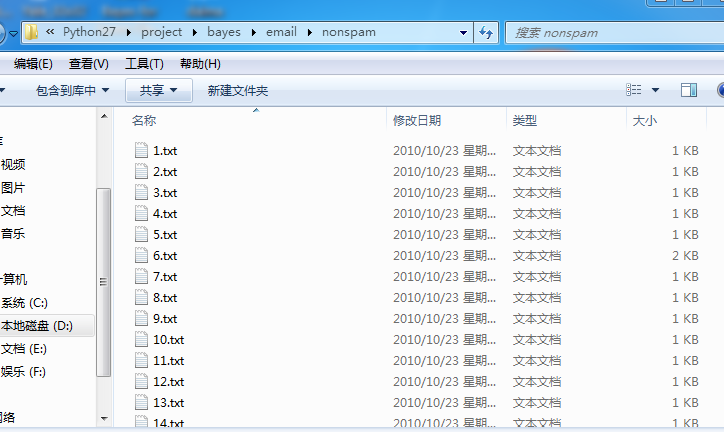
不同与对普通文本的分类，在邮件过滤方面，朴素贝叶斯过滤器也存在一些缺陷。因为邮件分类是一种二类分类问题，且邮件不同于普通文本，有其独特的结构。例如：邮件由邮件头和邮件正文组成，普通的文本分类朴素贝叶斯算法会将邮件头和邮件正文不加区别，而没有考虑到邮件头和邮件正文对邮件类别影响程度的不同。另外邮件分类是一种二类分类问题，两个分类之间是互相影响的，例如合法邮件的误判率低，而非法邮件的误判率高，说明判定一封邮件为非法邮件的阈值过高，为合法邮件的阈值过低。

**3 实验步骤简介与结果分析**

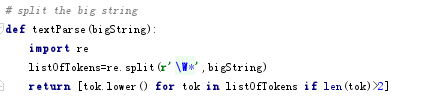
**3.1实验步骤**

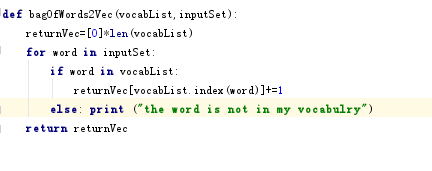
1）分别收集 150 封垃圾邮件和150封非垃圾邮件，建立垃圾邮件和非垃圾邮件的数据集。如下图所示：150封垃圾邮件及150封非垃圾邮件。

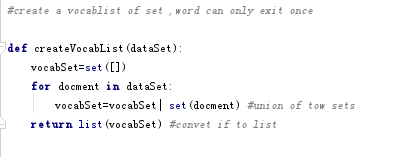




2）将垃圾邮件和非垃圾邮件组成一个文本，然后对其进行分词处理，去除掉里面重复单词，保证每个单词的唯一性，此处，在进行分词处理时只考虑长度大于 2 的单词，这些单词组成了词汇字典。 实现代码如下图所示：







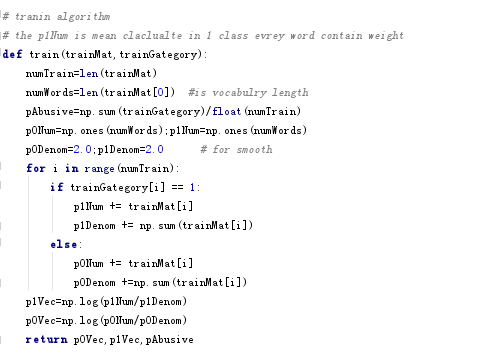
3）对每封邮件进行分词处理，从而建立特征向量矩阵，该矩阵包含 300 个行向量，每个行向量代表一封邮件，每个行向量的列数等于词汇字典的长度，而每个元素具体的取值则取决于该邮件中的单词在词汇列表中出现的次数，每出现一次该元素的值自动加1，若没出现，则该元素的值为0； 实现代码如下图所示：



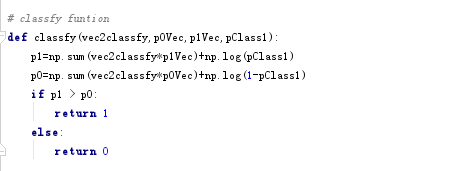
4）计算垃圾邮件的先验概率，即用所有的垃圾邮件除以所有邮件的个数，记为Pg，正常邮件的概率为1-Pg；

5）计算词汇列表中每一个单词在垃圾邮件中出现的概率，用行向量P1v表示，及在正常邮件中出现的概率，用行向量P0v表示；

步骤4）及步骤5）的实现代码如下图所示：

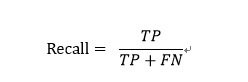


6）对某封邮件而言，根据 P0v， P1v 分别计算出邮件在垃圾邮件中出现的概率以及正常邮件中出现的概率，然后比较这两个概率，若 该邮件在垃圾邮件中出现的概率大于其在非垃圾邮件中的概率，则该封邮件属于垃圾邮件。实现代码如下图所示：



7）将整个数据集分为5份，每份60封邮件，取4份作为训练集，1份作为测试集，取3份作为训练集，1份作为测试集，………，求出上述测试集数据的TP , TN, FP, FN，然后根据

D:\amuse\qq\397012758\Image\C2C\7WTVSD)R35EMHTR]`[3V@64.png



得到：平均 Accuracy，Recall。实现代码如下图所示：

**参考文献：**

1. JasonD. M.Rennie: An Application of Machine Learning to E –Mail Filtering[J]. The Sixth ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining, 2000(8): 20-23
2. Patrick Pantel, Dekang Lin. A Spam Classification & Organization Program[EB/OL]. http://citeseer.ist.psu.edu/199924.html, 1998
3. Paul Graham. Better Bayesian filtering[EB/OL]. http://paulgraham.com/better.html, 2003
4. Gary Robinson. Gary Robinson rants on spam detection[EB/OL]. http://radio.webbogs.com/0101454/stories/2002/09/24/oldSpamDetection.html, 2002