華中科技大學

课 程 报 告

题目： 垃圾邮件的自动判别方法

院（系）： 计算机科学与技术学院

课程名称： 人工智能导论（专选课）

学　　号：

姓　　名：

任课老师：

报告日期：

目　录

1.绪论

1.1 课题背景及意义

现如今，随着网络科技的不断发展，信息传播速率逐年攀升，电子邮件由于其方便、快捷、低成本的特点逐渐脱颖而出，成为现代社会主要的通讯方式之一和互联网上最重要、最普及的应用之一，大大方便了人们生活、工作和学习。

但近年来，为了自身利益与某些特殊目的，某些个人或团体会向大量邮箱用户发送垃圾邮件，形成了影响力广，破坏性大的垃圾邮件问题。垃圾邮件不仅占用了大量的邮箱空间，极大地浪费了网络资源，还降低了网络使用效率，甚至还侵犯了用户的权利。

但由于垃圾邮件的难辨别性与数量巨大的性质，致使人工辨别、删除操作的难度较大，耗时耗力。因而，基于机器学习的自动判别垃圾邮件方法具有极大的作用与意义。

1.2 研究现状[1]

现阶段，国内外涌现各大邮箱平台，但总体来说，不外乎以下几种技术：1、关键词识别；2、IP 黑白名单；3、蜜罐技术；4、贝叶斯算法；5、评分算法；6、DNS 反向查找；7、 意图分析技术。具体介绍如下：

1、关键词识别它首先将垃圾邮件中一些特征性的字眼收集起来（比如打折、免费、促销等），形成一个大的数据库，当一封邮件发出来的时候就会自动匹配邮件头、邮件标题、邮件内容中与这些库里的关键词特征，如果有相类似的字眼，就会判定为垃圾邮件。

2、IP 黑白名单和第一种一样，它首先会把经常发送垃圾邮件的 IP 收集起来，形成一个黑 IP 库，俗称黑名单，只要是这个黑名单中发出来的邮件，自然会被判定为垃圾邮件，相反，如果 IP 被划定到白名单中，那你的邮件就会畅通无阻，但前提是制作规范的邮件。

3、蜜罐技术这是目前 QQ 邮箱等主流邮局采用最多的方法，它会在网络上分布很多的邮箱地址让你采集到，也会在 QQ 群等人流密集的地方设置一些邮箱，当你把这些邮箱采集到自己的数据库，并且发送了邮件之后，邮局马上会发现你在发未经许可的垃圾邮件，并且将这些垃圾邮件放到数据库中，之后发送的邮件都会在这个库里进行匹配，从而有效判定于垃圾邮件以及垃圾网址。

4、贝叶斯算法贝叶斯算法是目前世界上用的比较多的一种算法，它首先会收集大量的垃圾邮件和非垃圾邮件，建立垃圾邮件库和非垃圾邮件库，然后提取其中的特征字符串，对大量的网络发出的邮件进行匹配和甄别，垃圾邮件的识别率非常高。

5、评分算法这种方法是建立在关键字技术基础之上的，单一的关键字会出现大量的误判情况，为了解决这个问题，出现了多关键字检测的方式—评分。为每个可能在垃圾邮件中出现的关键字赋予分数，分数的多少要根据关键字在垃圾邮件中出现的可能性和严重性来决定。对一封邮件进行扫描，其中有一个关键字

[1] 相关技术引用“垃圾邮件的判定标准与识别方法”。

就加一定的分数，最后将所有的得分同设置好的阀值进行比较，从而有效判定出垃圾邮件。市场上大部分 ESP 都运用了此项方法。

6、DNS 反向查找在发邮件的时候，随意编造一个域名是非常容易的，如果采用阻断非法域名的方式来防止垃圾邮件的话。那么，用户可以说是被动到极点了，而且根本没有办法防止，因为那些域名都是根本不存在的。DNS 反向查找技术就是在收到邮件时对发件人的地址的真实性进行核查，防止 DNS 欺骗。这也是为什么正规的 ESP 都要求域名进行 spf 和 dkim 设置的原因。

7、 意图分析技术，垃圾邮件技术如今变得愈加复杂，许多垃圾邮件变得与正常的邮件几乎一样，在这些邮件中含有 URL 链接，这个链接往往指向一些不健康的网站，或某个商品促销的网站。ESP 为此创建了意图分析技术，构建了垃圾邮件 URLS 地址数据库。它检查邮件中的 URL 链接，确定邮件是否为垃圾邮件。

基于以上技术，我将就某些技术做出进一步详尽解释与应用。

2.方法原理及其步骤

2.1 方法一：朴素贝叶斯算法

2.1.1 方法原理

为何可以用朴素贝叶斯方法来判定垃圾邮件：

基于一个事件，可能其为 A，也可能是 B，在后续过程中，也可能发生第三件事件，用 Rc 表示未发生，R 表示发生，则可以得到以下图例：

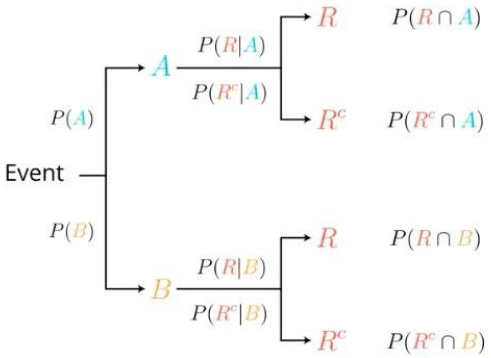

图 2-1　朴素贝叶斯原理

然后，我们可以通过计算先验概率，以及 A，B 发生条件下，R 发生或未发生的条件概率，进而通过贝叶斯公式进行计算，从而得出在 R 发生或未发生的条件下，A 或 B 发生的条件概率，即逆向思维。

故，朴素贝叶斯原理也可用于判断垃圾邮件，A、B 分别为垃圾邮件与非垃圾邮件，而此时的核心问题就在于，如何判定 R 事件。通过学习，我发现，R 事件即词汇表中各次出现的概率。因而，通过这一思想，可以得出某些词句出现时，推断该邮件是垃圾邮件的概率，进而进行判断。详细原理如下：

先规范用语，约定：[2]

符号	含义
S	垃圾邮件
H	非垃圾邮件
w(或 w_i)	单词
P(S)	邮件是垃圾邮件的概率（先验概率）
P(H)	邮件不是垃圾邮件的概率（先验概率）
$P(S\|w_1w_2...w_n)$	邮件包含 w_1，w_2，...，w_n 时，是垃圾邮件的概率
$P(H\|w_1w_2...w_n)$	邮件包含 w_1，w_2，...，w_n 时，不是垃圾邮件的概率
$P(w_1w_2...w_n\|S)$	一封垃圾邮件包含 w_1，w_2，...，w_n 的概率
$P(w_1w_2...w_n\|H)$	一封非垃圾邮件包含 w_1，w_2，...，w_n 的概率

表 2-1 符号的约定

[2] 详细原理引用“机器学习概论(1)：识别垃圾邮件”。

朴素贝叶斯基本公式：

$$P(y_i|x_1,x_2,...,x_n)=\frac{P(x_1,x_2,x_n|y_i)P(y_i)}{\sum_{j=1}^{m}P(x_1,x_2,x_n|y_j)P(y_j)}$$

在此问题中 x_i 是单词，y_i 代表垃圾邮件或非垃圾邮件。即上述 y_1 表示 A、B 事件，x_i 表示 R 事件。

原始公式可改写成，

$$P(S|w_1,w_2,...,w_n)=\frac{P(w_1,w_2,...,w_n|S)P(S)}{P(w_1,w_2,...,w_n|S)P(S)+P(w_1,w_2,...,w_n|H)P(H)}$$

通过这个公式，我们可以得到一封含有 w1,w2,…,wn 这些单词的邮件是垃圾邮件的概率是多少。即得知 P(R|A)。

为方便计算，假设 wi 之间相互独立，上式化简为：

$$P(S|w_1,w_2,...,w_n)=\frac{P(S)\prod_{x=1}^{n}P(w_x|S)}{P(S)\prod_{x=1}^{n}P(w_x|S)+P(H)\prod_{x=1}^{n}P(w_x|H)}$$

为了防止 P(wi|S)=0 而使得 P(S|w1,w2,…,wn)=0，故将使用线性平滑，即

$$P(w_x|S)=\frac{P(w_x,S)+\alpha}{P(S)+M\alpha}$$

其中，M 为单词数。

2.1.2 方法步骤[3]

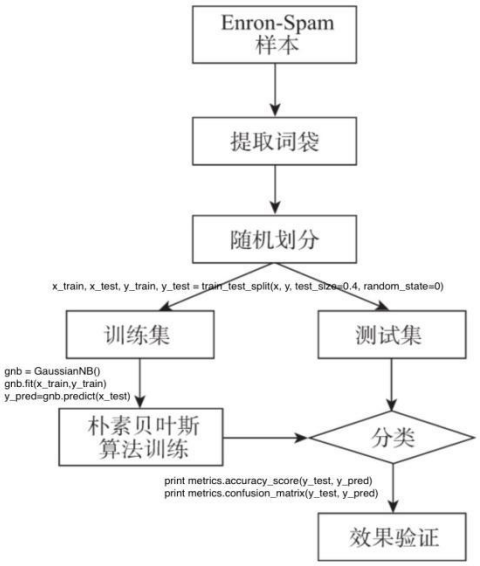

图 2-2 过程简要描述

[3] 参考“机器学习概论(1)：识别垃圾邮件”。

详细步骤：

1.处理相关数据：数据集提供了 60000 多条已经分词后的邮件，邮件中包含了邮件头（收发件人等信息）和内容。通过上述简述过程，则需要把这些邮件分为训练集和测试集两个部分，分类要求随机，但比例可以进行规定，再次规定 p%的数据作为训练集，其他为测试集。并通过数据集求出 P(S)和 P(H)，以便后续计算。

2.处理训练集：根据分词结果，统计每个词分别在垃圾邮件和非垃圾邮件中出现的次数，然后计算得到 $P(w_i|S)$和 $P(w_i|H)$。

3.处理测试集：首先提取这个邮件中的每一个词 w_i，带入式子中，即可求得 $P(S|w_1,w_2,...,w_n)$和 $P(H|w_1,w_2,...,w_n)$。如果 $P(S|w_1,w_2,...,w_n) > P(H|w_1,w_2,...,w_n)$则认为是垃圾邮件。

但有些特殊情况需要进一步优化：

- 加入邮件标题中的分词内容：考虑到邮件的标题中也会有有用的信息（比如发信方，时间，标题），因此可以将 Header 中的分词内容加入计算。
- 识别教育邮箱：将.edu 作为一个词加入到概率的计算中。
- 去除标点：句号和逗号等标点符号是出现频率很高但是对垃圾邮件识别没有什么用处的符号，可以在测试集和训练集的分词结果去除。
- 仅使用高频词：由于用于测试的邮件可能很长，有很多词，可能给计算造成负担，尤其时会导致精度不够。因此，可以在测试集上计算概率时，仅使用频率最高的一部分词（比如频率最高的 15 个）。

后续有对优化的进一步检验，发现某些特定的优化效果不佳，在此不做详细解释，但值得思考的是，各参数的改变对准确率的影响。

对参数的比较：

1. Smoothing 参数 α：采用无优化的方法，测试集 p=50 。

α	**准确度(%)**
0.005	98.57
0.01	98.44
0.05	98.06
0.1	97.61

表 2-2 α 对准确度的影响

$$P(w_x|S) = \frac{P(w_x, S) + \alpha}{P(S) + M\alpha}$$

分析：α 取值太大会导致正确率下降，原因是 α 过大时，计算 P(w|S)和 P(w|H)时分母远大于分子，导致计算对分子大小即该词在垃圾邮件和非垃圾邮件中的出现次数不敏感，因此准确率下降。

2.测试集比例大小：采用无忧化的方法。

测试集比例	准确度(%)
0.05	97.38
0.25	98.26
0.50	98.44
0.95	98.58

表 2-3　测试集比例大小对准确度的影响

分析：测试集越大，准确率越高。这是因为测试集越大，概率的估计越接近真实值，因此得到的估计值也更准确。另外，在测试集较小时准确率提高比较快，测试集较大时准确率提高比较慢。

2.2 方法二：SVM 算法

2.2.1 方法原理[4]

先对两种核函数进行简要介绍。

使用线性核函数的 SVM 算法：（线性情况）

首先，得到两种不同的散点图。

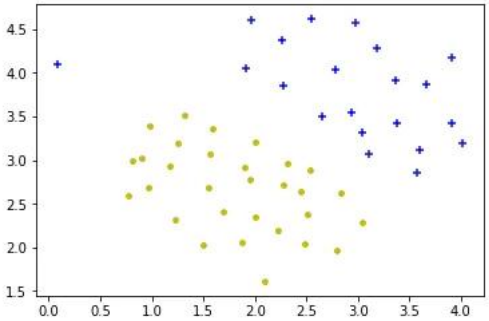

图 2-3 两种不同的散点图

通过使用 SVM 模块，利用 SVM 算法实现线性分类，不同的 C 值对决策边界有不同的影响。

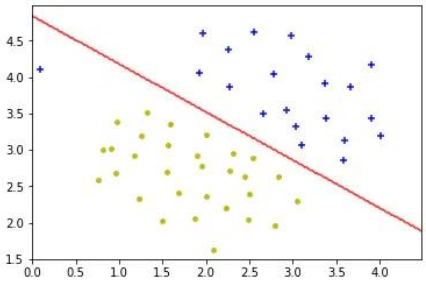

图 2-4 C=1 时,决策边界

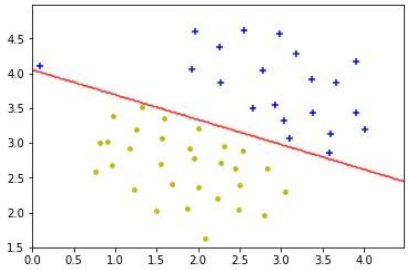

图 2-5 C=1000 时,决策边界

可以由以上图片看出，C 太大，可能会导致过拟合问题。

使用高斯核函数的 SVM 算法：（非线性情况）

对于非线性的分类任务，常用带有高斯核函数的 SVM 算法来实现，然后绘制非线性决策边界。

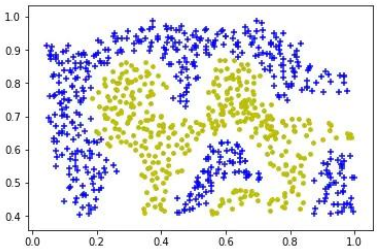

图 2-6 初始散点图

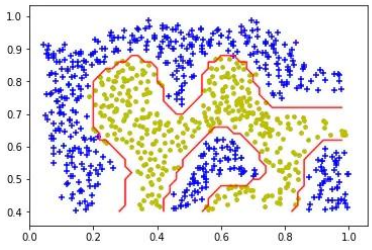

图 2-7 非线性决策边界

通过 SVM 算法，可以构造邮件过滤器，具体实现，如下所示。

[4] 原理引用自“用 SVM 算法构造垃圾邮件分类器”。

2.2.2 方法步骤

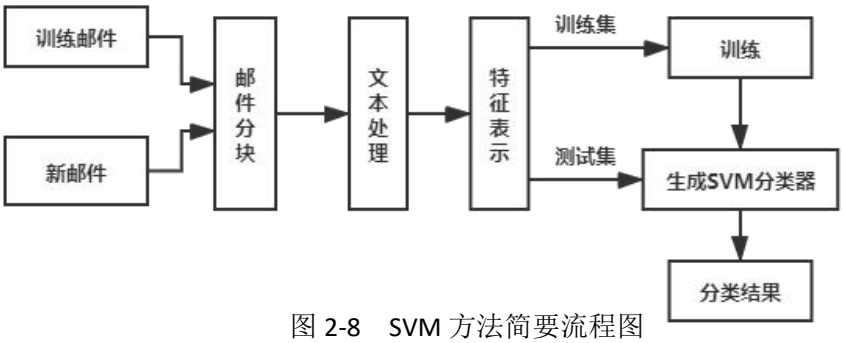

图 2-8　SVM 方法简要流程图

一封邮件格式如下所示，需要对文本邮件做必要处理。

```
> Anyone knows how much it costs to host a web portal ?
>
Well, it depends on how many visitors you're expecting.
This can be anywhere from less than 10 bucks a month to a couple of $100.
You should checkout http://www.rackspace.com/ or perhaps Amazon EC2
if youre running something big..

To unsubscribe yourself from this mailing list, send an email to:
groupname-unsubscribe@egroups.com
```

图 2-9　文本邮件范例

1.把整封邮件的大写字母全部转化为小写；
2.移除所有 HTML 标签（超文本标记语言）；
3.将所有的 URL 替换为‘httpaddr’；
4.将所有的地址替换为‘emailaddr’；
5.将所有数字替换为‘number’；
6.将所有美元符号($)替换为‘dollar’；
7.将所有单词还原为词根。

处理后的邮件格式如下：

```
Preprocessing sample email (emailSample1.txt) ...
==== Processed Email ====
anyon
know
how
much
it
cost
to
host
a
web
portal
well
it
depend
```

图 2-10　处理后的文本邮件

接下来，就需要我们筛选合适的单词来对邮件进行分类。

为了选择一个恰当方式对邮件进行分类，不能选择出现频率很小的词汇，因为那样会导致过拟合问题，如同上述 C 值。所以，我们需要选择一些出现频率较高的词汇。按照在垃圾邮件词汇库中至少出现了 100 次以上的单词的标准，将符合标准的词添加到词汇表中，最终，词汇表中有 1899 个单词。

有了词汇表，可以将预处理的电子邮件中的每个单词，通过查表的方式，组成数字文本。所谓单词索引列表就是每个单词所对应的数字索引所组成的列表，具体如右图所示：

图 2-11 数字索引表

邮件的文本经过处理后，可以得到以下处理文本：

```
Word Indices:
[  86  916  794 1077  883  370 1699  790 1822 1831  883  431 1171  794
 1002 1893 1364  592 1676  238  162   89  688  945 1663 1120 1062 1699
  375 1162  479 1893 1510  799 1182 1237  810 1895 1440 1547  181 1699
 1758 1896  688 1676  992  961 1477   71  530 1699  531]
```

图 2-12　由词汇表处理后的邮件范例

按照上述步骤，就需要提取邮件中的特征：

此过程中，需要从中得到特征向量 x，具体思路是：如果词汇表中第 i 个单词出现在邮件中，则用 $x_i=1$ 表示，否则用 $x_i=0$ 表示。最后，可以得到一个 x 的 n 维向量，如下所示：

$$x = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^n.$$

然后，进行训练 SVM 算法的步骤：

对邮件特征变量进行提取后，可以利用 p%比例的训练样本和 1-p%比例测试样本训练 SVM 算法，每个原始的邮件将会被转化为一个 $x \in R^{1900}$ 的向量（词汇表中有 1899 个词汇，$x_0=1$ 会被添加到向量中，最后，得到的向量包含 1900 个数字）。载入数据集之后，用变量 y=1 表示垃圾邮件，而 y=0 表示非垃圾邮件就可以训练 SVM 算法了。

最后，载入测试集数据，利用 SVM 算法构造的分类器，得到 98.9%的训练精度。

3.方法比较

SVM 方法在调试的时候需要尝试变换核与核参数。朴素贝叶斯方法在测试中的准确率到达了 98%，并且训练的复杂度低于 SVM。

因此就目前的认知来看，朴素贝叶斯方法原理更简单，实现与使用的效率比 SVM 更高。

SVM 方法的优缺点：

优点：

1、使用核函数可以向高维空间进行映射。

2、使用核函数可以解决非线性的分类。

3、分类思想很简单，就是将样本与决策面的间隔最大化。

4、分类效果较好。

缺点：

1、对大规模数据训练比较困难。

2、无法直接支持多分类。

朴素贝叶斯方法的优缺点：

优点：

1、朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。

2、对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练。

3、对缺失数据不太敏感，算法也比较简单，常用于文本分类。

缺点：

1、需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。

2、由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。

3、对输入数据的表达形式很敏感。

4.总结与展望

4.1 总结

本文通过对朴素贝叶斯算法和 SVM 算法的自然语言介绍，提出了两种自动判别垃圾邮件的方法，并对两种方法进行了简单的比较，并对二者的优缺点进行了简略的说明。其中，朴素贝叶斯算法在训练复杂度和效率方面要优于 SVM 算法，这也是为什么如今多数使用朴素贝叶斯算法来判别垃圾邮件的原因。

另外，对于朴素贝叶斯算法判别垃圾邮件的方法，首先需理清各概率之间的联系，并且基于一个特定的框架下，是否能进一步衍生出优化的方案，十分考验思维能力。特别的，对某些参数的研究实验，使得结论更为直观，更加便于理解。对于 SVM 算法判别垃圾邮件的方法，重点在于对核函数的运用，并在原理中介绍了两种核函数，分别用于线性与非线性的情况，对文本的预处理也十分的关键。

4.2 展望

垃圾邮件的泛滥是一个世界难题，虽然人们越来越重视研究自动判别垃圾邮件技术，也推出了一些新的方法与手段, 但是狡猾的垃圾邮件制造者为谋取私利，千方百计地修改垃圾邮件,使得垃圾邮件过滤系统无法发现或检测到此类邮件。因此，要把垃圾邮件阻挡在外，单靠垃圾邮件自动判别技术是无法解决的，还需要有关部门的重视和参与，通过宣传或者立法的形成，利用法律手段对垃圾邮件制造者进行制裁。只有大家都自觉行动起来，利用先进的技术手段武装网络系统，以完善的管理制度和法律法规为依托，双管齐下，才能从根本上消除垃圾邮件。

参考文献

[1] 匿名.疯的世界.“垃圾邮件的判定标准与识别方法”.CSDN 网站，https://blog.csdn.net/CSDN515/article/details/51426815

[2] 简书网站.“机器学习概论(1)：识别垃圾邮件”

[3] 匿名.“用 SVM 算法构造垃圾邮件分类器”.简书网站

[4] 百度百科.SVM 与朴素贝叶斯算法