二值文本分类中基于 Bayes 推理的特征选择方法

闫 鹏[1,2] 郑雪峰[1] 李明祥[1] 陈松华[2]

(北京科技大学信息工程学院 北京 100083)[1] (国家信息中心 北京 100045)[2]

摘 要 针对二值文本分类算法中的特征选择问题,本文提出了基于贝叶斯推理的评估函数算法来替代常用的、以 IG 或 MI 为评估函数的算法;同时,提出了以评估函数值的累计贡献率表示置信度,并以此确定特征选择维度的可量化的方法。对比实验显示,本文提出的新方法具有简便易行、高效实用的优点,此算法不仅对文本分类问题,对其它各类二值分类问题中的特征选择方法研究也都具有很好的参考、借鉴价值。

关键词 特征选择,数据挖掘,贝叶斯推理,文本分类

Feature Selection Method Based on Bayes Reasoning in Two-class Text Classification

YAN Peng[1,2] ZHENG Xue-feng[1] LI Ming-xiang[1] CHEN Song-hua[2]

(College of Information Engineering, Beijing University of Science and Technology, Beijing100083, China)[1]

(The State Information Center, Beijing 100045, China)[2]

Abstract Feature Selection is important for the text classification. The paper issued a new algorithm based on Bayes Reasoning to process the Feature Selection on alternative text classification. The experiments showed it had much better effect than the widely-used Mutual Information (MI) algorithm. And the paper also submitted a quantitative algorithm to decide the dimension of Feature Selection.

Keywords Feature selection, Data mining, Bayes reasoning, Text classification

1 引言

在进行文本分类算法研究时,要首先建立文本的"向量空间模型(Vector Space Model, VSM)",将原始文本转化为空间向量的形式供分类算法分析使用。但是,如果不加任何处理,在原始文本中直接提取特征词组成特征空间,必然会使得特征空间的维度过大,数据分析难以进行,产生所谓的"维灾难"。

因此,在建立 VSM 时,需要首先对原始文本形成的原始特征集进行技术处理,从中选择适当的特征词构成特征子集(或属性子集),这一处理工作通常称为"预处理"工作,它的任务主要有两点,一是确定"选择哪些特征词",二是确定"选择多少个特征词"。

原始文本的预处理工作是数据挖掘和机器学习等领域研究的难点之一。当前,预处理的处理方法主要有"维归约"和"特征选择"两种方法。

维归约的常用方法是使用线性代数技术,将数据由高维空间投影到低维空间,其中最常用到的是"主成分分析(Principal Components Analysis, PCA)法",但 PCA 方法的时间复杂度相当高,不适合在高速、大数据量的情况下使用。

"特征选择"是降低维度的另一个常用方法,它一般是通过数学方法,找出最具分类信息的特征词构成特征子集或特征空间,这种方法比较精确,人为因素的干扰较少,不足之处在于可能会丢失一些信息[1]。

评估函数法是常用到的特征选择方法之一,它是指对原始特征集中的每个特征词进行独立的评估并给定一个评估分值,然后排序,选取预定数目的最佳特征项构成特征子集。常用的评估函数有信息增益(Information Gain, IG)、互信息量(Mutual Information, MI),等等[2]。

1.1 信息增益法

信息增益法是一个基于熵的评估方法,其中,信息增益是指某特征项在文本中出现前后的信息熵之差。信息增益评估函数被定义为:

$$IG(w)=-\sum_{i=1}^{n}p(c_i)\log_2 p(c_i)+p(w)\sum_{i=1}^{n}p(c_i,w)\log_2 p(c_i,w)+p(\overline{w})\sum_{i=1}^{n}p(c_i,\overline{w})\log_2 p(c_i,\overline{w}) \quad (1.1)$$

其中:c_i 表示文本向量的类集,$i=1,\cdots,n$;w 表示特征词,$p(w)$为特征词 w 出现的概率,$p(\overline{w})$为特征词 w 不出现的概率,$p(c_i)$为文本向量属于 i 类的概率,$p(c_i,w)$表示特征词 w 在文本中出现且文本属于 i 类的概率,$p(c_i,\overline{w})$表示特征词 w 在文本中不出现且文本属于 i 类的概率。

1.2 互信息量法

互信息量法也是特征选择时常用的评估方法,假设有特征项 w 和类 c,那么,w 和 c 的平均互信息量定义为:

$$MI(w;c)=\sum_{i=1}^{n}p(w,c_i)\log_2\frac{p(w,c_i)}{p(w)p(c_i)}+\sum_{i=1}^{n}p(\overline{w},c_i)\log_2\frac{p(\overline{w},c_i)}{p(\overline{w})p(c_i)} \quad (1.2)$$

其中,$p(c_i)$为文本属于 c_i 类的概率,$i=1,\cdots,n$,$p(w)$为文本包含特征词 w 的概率,$p(w,c_i)$为文本包含特征词 w 且属于

闫 鹏 博士研究生,高级工程师,主要研究领域为计算机应用、网络安全;郑雪峰 博导,教授,主要研究领域为网络与信息安全;李明祥 博士研究生,主要研究领域为网络安全;陈松华 高级工程师,研究领域为运筹学。

c_i 类的概率，$p(\overline{w})$为文本不包含特征词 w 的概率，$p(\overline{w},c_i)$为文本不包含特征词 w 且属于 c_i 类的概率[3,4]。

2 二值文本分类问题的特征选择方法讨论

2.1 二值文本分类问题的特点

所谓二值，是指只有两个可能的结果。二值分类问题较多值分类问题要简单得多，因为只有两个可选的类别，且二者必居其一，所以我们一般只需牢记我们最为关心的类别的最主要特征即可，例如，生活中我们判断某人是不是张三，我们只需要记住张三的典型特征，当遇到某人时，我们只需判断此人的特征与我们记忆里的张三的特征是否相符，如相符，便可认为此人可能是张三，如不相符，便认为此人不是张三。对此，我们并不需要再去记住所有非张三的人的特征。

又如，互联网上常见的恶意攻击、电脑里的病毒程序，等等，安全防护软件或杀毒软件只需存储这些恶意攻击或者病毒程序的典型特征，就可以采取相应的措施，而不需要再记下所有正常访问或所有正常程序的特征。

同样，对于文本的二值分类问题，我们也不必把两类文本的特征都记下来，相反，只需提取出两个类别中我们最为关心的类别的最主要特征，以此构成特征子集，在此子集中判断某文本是否与此特征相符，如果符合程度在我们预定的阈值之上，即可认为此新文本即属于此类别，否则，则判断此文本非此类别，即属于另一个类别。

因此，我们认为，二值文本分类问题，可以换一个角度，把它看成是一个“判断某文本是否属于某一类别的问题”。二值文本分类问题的特征选择，同样可以简化为：第一，主要研究我们最为关心的类别；第二，选择这一类别的最主要特征，以这些特征来构成特征子集，这种思想方法，即是矛盾论中“抓主要矛盾和抓矛盾的主要方面”的思想的体现，是一种高效易行的处理方法。

2.2 IG 或 MI 为二值文本分类特征选择评估函数的不足

IG 或 MI 的概念，其理论依据都是建立在信息论中的“信息量”的概念基础之上，在处理多值文本分类的特征选择问题时，以它们作为评估函数效果是理想的，但对于二值分类问题，则未能利用二值分类的特殊性，使问题简化。下面对此进行简单分析：

根据乘法公式：

$$P(A,B)=P(A|B)P(B) \tag{2.1}$$

且针对二值分类问题，$n=2$，所以

公式(1.1)还可以表示为：

$$\begin{aligned} IG(w)=&-\sum_{i=1}^{2}p(c_i)\log_2 p(c_i)+p(w)\sum_{i=1}^{2}p(w|c_i)p(c_i)\log_2(p(w|c_i)p(c_i))\\ &+p(\overline{w})\sum_{i=1}^{2}p(\overline{w}|c_i)p(c_i)\log_2(p(\overline{w}|c_i)p(c_i)) \end{aligned} \tag{2.2}$$

公式(1.2)还可以表示为：

$$\begin{aligned} MI(w;c)=&\sum_{i=1}^{2}p(w|c_i)p(c_i)\log_2\frac{p(w|c_i)}{p(w)}+\sum_{i=1}^{2}p(\overline{w}|c_i)\\ &p(c_i)\log_2\frac{p(\overline{w}|c_i)}{p(\overline{w})} \end{aligned} \tag{2.3}$$

从公式(2.2)、(2.3)可知，IG 或 MI 的数值大小，主要依赖于 $p(c_1)$，$p(c_2)$，$p(w)$，$p(\overline{w})$四个先验概率值和 $p(w|c_1)$，$p(w|c_2)$，$p(\overline{w}|c_1)$，$p(\overline{w}|c_2)$四个后验概率值。在比较不同特征词“w”的 IG 或 MI 值时，前四个先验概率一般采用估计的办法，都分别假定为 0.5，所以，它们的作用可以忽略，真正使计算结果产生差异的在于 $p(w|c_1)$，$p(w|c_2)$，$p(\overline{w}|c_1)$，$p(\overline{w}|c_2)$这四个后验概率值。

公式(2.2)、(2.3)对这四个后验概率值，采取了同等的处理方法，并没有考虑到它们的不同作用和意义。所以，根据文本二值分类问题中，“主要选择最为关心的类别的主要特征”的思想，以 IG 或 MI 为评估函数的处理方法忽略了二值分类问题的特殊性，使可以简化的问题复杂化，同时也使计算量增大，系统的效率降低。

不仅如此，系统花费如此大的计算代价而得到的 IG 或 MI 值只能用于特征选择阶段，在后续的分类阶段不再具有使用价值，这也进一步说明了以 IG 或 MI 为特征选择评估函数方法并不经济。

2.3 基于贝叶斯的二值文本分类问题的特征选择方法

2.3.1 特征评估函数

根据 2.1 小节的分析，文本的二值分类问题，可以看成是一个“判断某文本是否属于某一类别的问题”。所以在处理此类问题时，可以选择我们最为关心的类别中的最有代表性的特征词，构成特征子集。

以下假设 $c1$，$c2$ 为两个可能的类别，其中，$c1$ 为我们最为关心的类别，w 表示特征词，p 表示概率。

那么，我们是不是只选择较大的 $p(w|c1)$值对应的特征词 w 就可以了呢？其实不然，下面用简单的示意图来作说明。

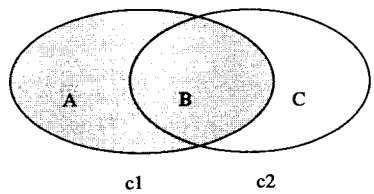

图 1 特征词与类别集合关系示意图

在图 1 中，$c1$，$c2$ 分别表示两个可能的类别，$A\cup B$，$B\cup C$ 分别表示类别 $c1$，$c2$ 中具有代表性的特征词集合，都具有很高的 $p(w|c_i)$，且 A，B，C 三者互不相交。

如图 1 所示，尽管 A，B，C 中的特征词 w 都具有很高的 $p(w|c_i)$ $(i=1,2)$值，但是，特征词 B 为 $c1$ 和 $c2$ 的交集，B 中的特征词在 $c1$ 和 $c2$ 中都经常出现，这样的特征词并不能体现出 $c1$ 和 $c2$ 的差异性，极端的例子如英文的冠词“the”，中文的“的、地、得”等等，因此它们对于分类问题的贡献是很小的，所以应该尽量选择位于 A 集合中的特征词，即应选择那些在 $c1$ 类中经常出现且在 $c2$ 类中较少出现的特征词，或者说，应该选择特征词 w，使 $p(w|c1)$与 $p(w|c2)$ 之差或之商尽量大。

本文提出针对二值文本分类问题的“基于贝叶斯推理的特征选择评估函数 $p(c1|w)$”即充分体现了这一思想。下面推导评估函数 $p(c1|w)$的计算表达式：

贝叶斯定理 设 $B_1,B_2,\cdots,B_n$ 为样本空间 S 的一个划分，且 $P(B_i)>0(i=1,2,\cdots,n)$，对于任一事件 A，$P(A)>0$，则有贝叶斯公式：

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n}P(A|B_j)P(B_j)} \tag{2.4}$$

根据贝叶斯定理，评估函数 $p(c1|w)$可以表示为：

$$p(c1|w)=\frac{p(w|c1)p(c1)}{p(w|c1)p(c1)+p(w|c2)p(c2)} \quad (2.5)$$

在实际应用时，通常假设 $p(c1)$和 $p(c2)$相等，所以评估函数 $p(c1|w)$还可以再化简为：

$$p(c1|w)=\frac{p(w|c1)}{p(w|c1)+p(w|c2)} \quad (2.6)$$

其中，$c1$ 表示我们最为关心的文本类别，$c2$ 表示另一文本类别，w 表示某特征词，p 表示概率值，$p(w|c1)$，$p(w|c2)$分别表示特征词 w 在 $c1$ 或 $c2$ 类中出现的条件概率；那么，评估函数 $p(c1|w)$的物理意义可以理解为当特征词 w 在文本中出现时此文本为 $c1$ 类的概率；从公式(2.5)可以推导，$p(c1|w)$值与 $p(w|c1)$与 $p(w|c2)$之商或之差呈递增关系，限于篇幅，推导过程从略。

在确定以 $p(c1|w)$值为评估函数之后，分类算法应该选择尽量大的若干个 $p(c1|w)$值对应的特征词组成特征空间。

2.3.2 *确定特征空间维度的量化方法*

关于特征空间的维度问题，通常是通过试验的方法来确定，但不同的文献之间差异很大，实际应用过程中难以掌握。为此，本文提出一种可量化的“基于评估函数贡献率”计算方法如下：

令：w 表示某特征词，$f(w)$表示特征词 w 的评估函数值，$q(x)$表示评估函数 x 的贡献率，

设：按降序排列，共有 n 个评估函数值 $f(w_i)$，分别对应 n 个特征词 w_i，$i=1,\cdots,n$，且 $f(w_i)>f(w_{i+1})$，

则：评估函数 $f(w_i)$的贡献率 $q(w_i)$为：

$$q(w_i)=f(w_i)/\sum_{i=1}^{n}f(w_i)*100\% \quad (2.7)$$

在实际过程中，我们可以根据精确度需要，确定一个累计贡献率的临界值，例如 50%，80%或 90%，等等，然后在表 1 中选择与累计贡献率临界值相对应的前“k”个特征词构成特征空间，从而实现在一定的置信度下的特征选择处理工作。

表 1 评估函数贡献率及累计贡献率计算方法

序号 (k)	$f(w)$ (降序)	$q(w)=f(w)/\sum_{1}^{n}f(w_i)*100\%$	置信度(累计贡献率) $\sum_{1}^{k}q(w_i)$
1			
…			
n			

表 1 中，特征选择的“置信度”可以用“累计贡献率”来表示，通过计算累计贡献率的方法，我们可以对特征选择处理的可信度有一个量化的了解和控制。

2.3.3 *特征评估函数在朴素贝叶斯分类算法中的作用*

一般情况下，特征选择的评估函数只用于特征选择阶段，对后续分类工作没有使用价值，但是本文提出的基于贝叶斯的特征选择评估函数(见公式(2.5)或(2.6))则不然，尤其是对于朴素贝叶斯(NB)分类算法，可直接使用，所以，相比其它评估函数，这种基于贝叶斯的特征选择评估函数具有更高的使用价值。

以下推导直接使用 $p(c1|w)$评估函数的 NB 分类算法的表达式：

设经过特征选择处理后，在特征空间$\{W1,W2,\cdots,Wn\}$中，文本向量$\vec{x}=\langle w1,w2,\cdots,w_n\rangle$，$C=\{c1,c2\}$表示文本可能属于的两个类别(设 $c1$ 是我们关心的类别)，那么，根据贝叶斯定理，有：

$$p(c1|\vec{x})=\frac{p(\vec{x}|c1)p(c1)}{p(\vec{x}|c1)p(c1)+p(\vec{x}|c2)p(c2)} \quad (2.8)$$

$$p(c2|\vec{x})=\frac{p(\vec{x}|c2)p(c2)}{p(\vec{x}|c1)p(c1)+p(\vec{x}|c2)p(c2)} \quad (2.9)$$

根据朴素贝叶斯分类算法的基本原理[4,5]，NB 分类算法的常用表达式为：

$$p(c1|\vec{x})/p(c2|\vec{x})>\lambda \quad (\lambda>=1) \quad (2.10)$$

即当公式(2.10)成立时，此文本向量属于 $c1$ 类。其中，“λ”为用于判断的阈值。

又，根据公式(2.5)，可以推导出：

$$\frac{p(w|c1)}{p(w|c2)}=\frac{p(c2)}{p(c1)}\cdot\frac{p(c1|w)}{1-p(c1|w)} \quad (2.11)$$

根据朴素贝叶斯分类算法的条件独立性假设和公式(2.8)、(2.9)、(2.11)，公式(2.10)可以表示为：

$$\frac{p(c1|\vec{x})}{p(c2|\vec{x})}=\frac{p(\vec{x}|c1)p(c1)}{p(\vec{x}|c2)p(c2)}=\frac{p(c1)}{p(c2)}\cdot\prod_{i=1}^{n}\frac{p(w_i|c1)}{p(w_i|c2)}$$

$$=\left(\frac{p(c2)}{p(c1)}\right)^{n-1}\cdot\left(\prod_{i=1}^{n}\frac{p(c1|w_i)}{1-p(c1|w_i)}\right)>\lambda \quad (2.12)$$

一般情况下，假设 $p(c1)=p(c2)=0.5$，所以，公式(2.12)还可以简化为：

$$\prod_{i=1}^{n}\frac{p(c1|w_i)}{1-p(c1|w_i)}>\lambda \quad (2.13)$$

从公式(2.12)、(2.13)、(2.13)可知，对于文本二值分类问题，在朴素贝叶斯分类器中，基于贝叶斯的 $p(c1|w_i)$指标不仅可以作为评估函数用于特征选择，还可以直接用于文本的二值分类的判别模型，是一个非常简单而高效的指标，具有非常理想的实用价值。

3 算法验证

3.1 算法验证说明

垃圾邮件判别是一个典型的文本二值分类问题，它包含垃圾邮件和正常邮件两个类别，其中，垃圾邮件是我们最为关心的类别。本文以垃圾邮件判别问题为算法验证的测试对象，分别以文中提出的“基于贝叶斯的评估函数”(用 Bayes 表示)(见 2.3.1 节)和“以平均互信息量 MI 为评估函数”(用 MI 表示)(见 1.2 节)两种方法进行特征选择，以 2.3.2 节设定置信度的方法确定特征空间的维度，最后通过朴素贝叶斯分类算法(公式(2.13))判断新邮件是否为垃圾邮件。

本测试将对比分析这两个评估函数在准确率和效率两个方面的实际效果，从而验证基于贝叶斯的特征评估函数法的实际价值。

3.2 测试环境

硬件配置：dell PowerEdge 2600 PC 服务器(标准配置)

操作系统及应用软件：RH9.0，Oracle9i，perl 5.8

3.3 测试数据(语料集)

下载 lingspam 语料库(http://www.iit.demokritos.gr/skel/i-config/downloads/ lingspam_public.tar.gz)，然后对压缩文件解包，本实验选取 lemm_stop/part2 文件夹下的共计 289 封邮件作为训练样本，选取 lemm_stop/part1 文件夹下的共计 289 封邮件作为测试样本。

3.4 测试结果

结果见表 2。

表 2　两种特征选择评估函数法测试结果的比较

阈值(λ)	置信度(%)	互信息量(MI)评估函数法 维数	运算时间(s)	准确率(%)	错误率(%)	垃圾邮件拦截率(%)	垃圾邮件精确率(%)	基于 Bayes 推理的评估函数法 维数	运算时间(s)	准确率(%)	错误率(%)	垃圾邮件拦截率(%)	垃圾邮件精确率(%)
	80	9896	12.078	85.47	14.53	12.50	100.00	8337	4.104	96.54	3.46	79.17	100.00
10	60	7376	11.944	83.74	16.26	2.08	100.00	5914	3.965	97.58	2.42	87.50	97.67
	40	3688	11.816	83.39	16.61	0.00	-	2280	3.979	96.54	3.46	89.58	89.58
	80	9896	11.963	85.47	14.53	12.50	100.00	8337	4.026	97.23	2.77	85.42	97.62
5	60	7376	11.884	84.08	15.92	4.17	100.00	5914	4.059	97.23	2.77	89.58	93.48
	40	3688	11.775	83.39	16.61	0.00	-	2280	3.980	94.81	5.19	93.75	78.95
	80	9896	11.982	87.54	12.46	25.00	100.00	8337	4.149	97.58	2.42	93.75	91.84
2	60	7376	11.930	86.16	13.84	16.67	100.00	5914	4.034	96.89	3.11	95.83	86.79
	40	3688	11.795	83.39	16.61	0.00	-	2280	3.931	90.66	9.34	95.83	64.79

说明：

1) 设 N_H：测试集中合法邮件总数；N_S：测试集中垃圾邮件总数；N_{H-H}：合法邮件被判定为合法邮件的总数；N_{S-S}：垃圾邮件被判定为垃圾邮件的总数；N_{H-S}：合法邮件被判定为垃圾邮件的总数；N_{S-H}：垃圾邮件被判定为合法邮件的总数。

则表中：

准确率：$Acc=(N_{H-H}+N_{S-S})/(N_H+N_S)$；

错误率：$Err=(N_{H-S}+N_{S-H})/(N_H+N_S)$；

垃圾邮件拦截率(Spam Recall)：$SR=N_{S-S}/(N_{S-S}+N_{S-H})$；

垃圾邮件精确率(Spam Precision)：$SP=N_{S-S}/(N_{S-S}+N_{H-S})$[6]

2)表中的运算时间只包括分类算法进行训练学习和对测试邮件分类的全部用时，而不包括对邮件样本进行抽词、形成原始特征集的预处理用时(626.622 秒)，因其与算法无关 。

3.5　结果分析

根据表 2 的数据，可以得到这两种方法的以下几个折线图。

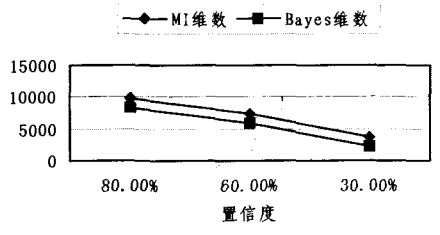

图 2　两种算法的特征选择维数比较

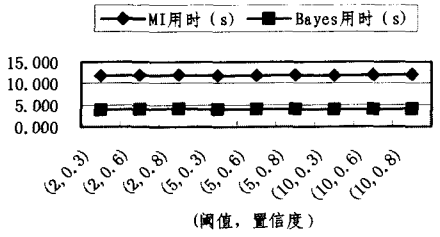

图 3　两种算法的运算用时比较

以上三个折线图所示，基于 Bayes 的特征选择算法在效率(运算用时)、性能(准确率、拦截率)和特征选择的维度等方面，都明显优于常用的基于 MI 的算法，实验充分显示了新算法的优势与价值。

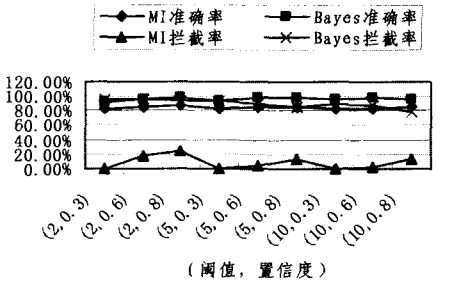

图 4　两种算法的性能比较

结束语　与基于 MI 的方法相比，本文提出的基于 Bayes 的算法在效率和性能等各方面都明显占优，从而证明了基于 Bayes 的特征选择方法在文本二值分类问题上具有相当高的价值。

下一步，我们准备分在三个方面再进行一些研究工作：第一，进一步对比、分析在相同的维度条件下本算法在准确率和效率方面有无优势；第二，研究本方法在判别网络入侵检测方面的实用价值；第三，关于多值分类问题的特征选择方法优化研究。

参 考 文 献

[1] Tan Pang-Ning, Stenbach M, Kumar V. Introduction to Data Mining (M) (数据挖掘导论). 范明，范宏建，译. 北京：人民邮电出版社，2006(5)：30-33

[2] 和亚丽，陈立潮. Web 文本挖掘中的特征选取方法研究(J). 计算机工程，2005，31(5)：181-182，190

[3] 孙丽华，谢仲华，陈荣伶. 信息论与纠错编码 (M). 北京：电子工业出版社，2005(3)：14-36

[4] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. Spyropoulos: An Evaluation of Naive Bayesian Anti-Spam Filtering (C)// Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona, Spain, 2000: 9-17

[5] Lai C C. An empirical study of three machine learning methods for spam filtering(J), Knowledge-Based System (J)(2006), doi: 10.1016/j.knosys.2006.05.016

[6] Zorkadis V, Karras D A, Panayotou M. Efficient information theoretic strategies for classifier combination, feature extraction and performance evaluation in improving false positives and false negatives for spam e-mail filtering (J). Neural Networks, 2005, 18: 799-807