Statistics and Application 统计学与应用, 2021, 10(3), 365-372
Published Online June 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.103036

Hans汉斯

基于偏最小二乘回归的邮件分类问题研究

李会会

上海对外经贸大学，上海
Email: huihuilihoney@163.com

收稿日期：2021年5月6日；录用日期：2021年5月20日；发布日期：2021年6月1日

摘　要

本文基于最小二乘的主成分回归(PCR)方法对邮件进行分类，进一步使用偏最小二乘回归(PLS)对垃圾邮件识别分类。将PLS与PCR得到的分类准确度进行比较分析，考察PLS分类准确度百分比随分类截点变化的趋势，并得出两种方法下不同k值(主成分个数)对应的ROC曲线图，分析PLS与PCR方法识别和分类垃圾邮件的准确度和稳定性。

关键词

偏最小二乘回归，LGK双对角化，迭代算法，分类算法

Research on Mail Classification Problem Based on Partial Least Squares Regression

Huihui Li

Shanghai University of International Business and Economics, SUIBE, Shanghai
Email: huihuilihoney@163.com

Received: May 6th, 2021; accepted: May 20th, 2021; published: Jun. 1st, 2021

Abstract

This paper classifies emails based on the principal component regression (PCR) method of least squares, and further uses partial least squares regression (PLS) to identify and classify spam emails. The classification accuracy obtained by PLS and PCR was compared and analyzed. Then the trend of the percentage of classification accuracy of PLS with the classification cut-off point is examined, and the ROC curve corresponding to different k values (number of principal components) under the two methods is obtained. Finally, this article analyzes the accuracy and stability of PLS

文章引用: 李会会. 基于偏最小二乘回归的邮件分类问题研究[J]. 统计学与应用, 2021, 10(3): 365-372.
DOI: 10.12677/sa.2021.103036

and PCR methods to identify and classify spam.

Keywords

Partial Least Squares Regression, LGK Double Diagonalization, Iterative Algorithm, Classification Algorithm

1. 引言

随着网络越来越发达，各种信息更是无处不在，手机会收到广告信息，浏览器和播放器及各种网站上也会经常弹出垃圾信息。我们开始考虑研发了各种垃圾信息的识别系统去识别和阻拦一些对没用的信息。互联网的快速发展及电子邮件的应用也更加的方便快捷，垃圾邮件却不断地充斥我们的视野使得有意义的信息被覆盖，所以对垃圾信息的识别处理更受到重视。垃圾邮件的识别可以通过某些信息如：某种关键词出现次数、邮件的来源地址等。本文选取的是美国的一个邮件数据库，该数据中垃圾邮件的识别变量分成 57 个特征信息，通过区分垃圾邮件和非垃圾邮件 57 个变量不同取值情况来识别是否为垃圾邮件。对垃圾邮件的识别算法有很多，我们常见的就是建立回归模型来预测和识别垃圾邮件。

本文选择了偏最小二乘法建立回归模型，并与最小二乘主成分回归方法进行对比分析，李雪和孙建平(2017) [1]对偏最小二乘回归进行了改进，并提出了基于正交投影修正的偏最小二乘回归算法，以此建立了烟气含氧量预测模型，并通过数据验证发现改进后的算法建模有较高预测精度且采样过程更快。本文采用 LGK 双对角化实质上也是运用了类似正交投影的思想，通过 LGK 双对角化将自变量与因变量的矩阵转化为正交矩阵的乘积的形式，然后进行主成分分析。L. Elden (2014) [2]考虑多元线性回归，分析了计算投影到下维子空间的偏差最小二乘法，给出了 PLS 标准化理论分析。赵晓丹和徐燕(2014) [3]详细介绍了垃圾邮件的概念及分类识别原理，并进行了垃圾邮件处理技术的对比分析研究，包括预处理、特征选择和分类三大步骤。其中特征选择技术包括文档频率、信息增益、优势率等方法。毛雪莲(2019) [4]介绍了偏最小二乘法原理及算法的实现，偏最小二乘在数据的预处理中需要对数据进行标准化。陈龙等(2018) [5]、李雨亭(2018) [6]、黄鹤等(2018) [7]基于机器学习方法对垃圾邮件进行了检测、识别。段同庆等(2019) [8]、丁学利和任鹏(2020) [9]使用偏最小二乘法在其他领域进行运用；Keshav (2021) [10]对偏最小二乘法进行了详细介绍；在垃圾邮件的分类研究中还存在其他方法下的算法研究[11] [12] [13]。所以本文研究了基于偏最小二乘回归模型的垃圾邮件分类问题。

2. 研究问题

偏最小二乘(PLS)回归是一种面向高维数据的回归分类方法，在化学计量等领域应用尤其广泛。当自变量维数大于样本量且变量间相关性强时，PLS 方法常优于主成分回归(PCR)。

设 y: $n\times 1$ 为因变量，X: $n\times p$ 为自变量对应的设计矩阵。回归最小二乘问题是：

$$\min_{\beta\in R^p}\left\|X\beta-y\right\|_2 \tag{1}$$

当 p 较大时，常采取的处理是将 β 的解限定在某一个低维子空间中。即对于 $k\ll p$，给定一组近似正交基 $\Gamma k=(\gamma_1,\cdots,\gamma_k)$，代入 $\beta=\Gamma k\alpha$，最小二乘问题变为：

$$\min_{\alpha\in R^p}\left\|X\Gamma_K\alpha-y\right\|_2 \tag{2}$$

因此，此时的核心问题是如何确定Γ_k，效果良好的Γ_k应使$\left\|X\Gamma_k\alpha-y\right\|_2$的下限较小。

3. 研究方法

3.1. PCR

主成分分析在多元统计中是一种降维方法，通过消去自变量之间的多重共线性，同时又不损失太多的信息来达到降维目的。主成分分析是要找到自变量的线性组合$Z=X\Gamma$，要求Z方差和最大且不相关，而这个线性组合满足列正交的系数矩阵Γ被称为因子载荷矩阵，这些独立的变量被称为主成分，矩阵Z被称为得分矩阵。

PCR主要使用了SVD分解，对于$X=U\begin{pmatrix}D\\0\end{pmatrix}V^T$，令$\Gamma_k=(V_1,\cdots,V_k)$，即是我们要求的$\Gamma_k$，且$\Gamma_k$的选取仅与自变量数据有关。将$\Gamma$带入回归模型得到$y=X\Gamma_\beta+\varepsilon$，对系数$\beta$的估计采用最小二乘方法，得：$\hat{\beta}=D_k^{-1}U_k^Ty$，$\hat{y}=X\Gamma\hat{\beta}$。

3.2. PLS

偏最小二乘法运用LGK双对角化将自变量与因变量分解为两个正交阵和双对角阵的乘积，相较于PCR，PLS不仅考虑了自变量X的多重共线性，还考虑了自变量与因变量之间的多重共线性，对X与Y合成的矩阵进行降维，所以，可以达到比PCR更好的降维效果。

3.2.1. LGK双对角化的概念

偏最小二乘回归使用LGK双对角化将原始变量X,y合并为一个新的变量$Z=(y:X)$，对Z进行LGK双对角化，表示为：

$$Z=P\begin{pmatrix}B\\0\end{pmatrix}W^T \tag{3}$$

其中，$P=P^{(1)}P^{(2)}\cdots P^{(p)}$和$W=W^{(1)}W^{(2)}\cdots W^{(p)}$，分别为$n\times n$与$(p+1)\times(p+1)$的$p$个householder矩阵相乘构建的正交阵，矩阵$B:(p+1)\times(p+1)$为：

$$B=\begin{pmatrix}b_1 & a_1 & & & & \\ & b_2 & a_2 & & & \\ & & \ddots & \ddots & & \\ & & & b_p & a_p & \\ & & & & a_{p+1} & \end{pmatrix} \tag{4}$$

令Γ_k为矩阵$(w_2,w_3,\cdots,w_k)$删去第一行。

3.2.2. LGK双对角化的性质

记$p_i:n\times1$为矩阵P的第i列，$w_i:(p+1)\times1$为矩阵W的第i列，LGK双对角化性质如下：

- 令公式(3)左右同时左乘P^T，可以得到：

$$\left(p_1:p_2:\cdots:p_{p+1}\right)^TZ=BW^T=\begin{pmatrix}b_1 & a_1 & & & & \\ & b_2 & a_2 & & & \\ & & \ddots & \ddots & & \\ & & & b_p & a_p & \\ & & & & a_{p+1} & \end{pmatrix}\begin{pmatrix}w_1^T\\w_2^T\\\vdots\\w_{p+1}^T\end{pmatrix} \tag{5}$$

所以，有 $p_i^T Z = a_i w_{i+1}^T + b_i w_i$，即 $w_{i+1} = \frac{1}{a_i}\left(Z^T p_i - b_i w_i\right)$，$i = 1,2,\cdots,p$

- 令公式(3)左右同时右乘 W，可以得到：

$$\left(p_1 : p_2 : \cdots : p_{p+1}\right)^T Z = BW^T = \begin{pmatrix} b_1 & a_1 & & & \\ & b_2 & a_2 & & \\ & & \ddots & \ddots & \\ & & & b_p & a_p \\ & & & & a_{p+1} \end{pmatrix} \begin{pmatrix} w_1^T \\ w_2^T \\ \vdots \\ w_{p+1}^T \end{pmatrix} \tag{6}$$

所以，有 $Zw_i = a_{i-1} p_{i-1} + b_i p_i$，即

$$p_i = \frac{1}{b_i}\left(Zw_i - a_{i-1} p_{i-1}\right),\ i = 2,\ldots,p,(p+1) \tag{7}$$

- LGK 的P矩阵第一列有 $p = \frac{y}{b_1}$，且 W 矩阵具有如下形式：

$$W = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{W} \end{pmatrix},\ \text{则}\ b_1 = \|y\|_2,\ \ p_1 = \frac{y}{\|y\|_2},\ \ w_1 = (1,0,0,\cdots,0)$$

LGK 这三条性质为 P、B、W 矩阵的迭代算法提供了迭代初始值和迭代关系式。

4. 算法实现

4.1. P, W 矩阵的迭代算法

由 LGK 的性质关系知，P、B、W 三者之间的关系，其中，$P : n \times (p+1)$，$B : (p+1) \times (p+1)$，$W : (p+1) \times (p+1)$ 为正交阵。首先，定义两个空的矩阵 P、W，由第三条性质可知迭代初始值为：

$$w_1 = (100\cdots 0)^T,\ b_1 = \|y\|_2,\ p_1 = \frac{y}{\|y\|_2} = \frac{b}{y}$$

$$a_1 = \left\|z^T p_1 - b_1 w_1\right\|_2$$

将初始值填充空矩阵第一列，然后进行迭代运算得到第一次迭代结果为：

$$w_2 = \frac{1}{a_1}\left(z^T p_1 - b_1 w_1\right),\ b_2 = \left\|zw_2 - a_1 p_1\right\|_2$$

$$p_2 = \frac{1}{b_2}\left(zw_2 - a_1 p_1\right),\ a_2 = \left\|z^T p_2 - b_2 w_2\right\|_2$$

同理，将第一次迭代结果填充空矩阵的第二列，继续第二次迭代结果为：

$$w_3 = \frac{1}{a_2}\left(z^T p_2 - b_2 w_2\right),\ b_3 = \left\|zw_3 - a_2 p_2\right\|_2$$

$$p_3 = \frac{1}{b_3}\left(zw_3 - a_2 p_2\right),\ a_3 = \left\|z^T p_3 - b_3 w_3\right\|_2$$

…

最后，将以上每次迭代结果都相应的填充到 P、W 矩阵中，即可得到正交阵 P、W。但是在计算 a_i，b_i 存在一定的误差使得我们得到的矩阵是近似列正交。则 Γ_k 就是 $\tilde{W}$ 矩阵中选取前 k 列。

4.2. 邮件分类算法

本文选取 spam.data 数据集，其中包含 4601 封邮件(其中 1813 封为垃圾邮件)，58 个变量(其中第 58 个变量显示该邮件是否是垃圾邮件：是−1，否−0)首先，将该数据集分为训练集和测试集：

训练集：906 封垃圾邮件，1394 封非垃圾邮件

测试集：907 封垃圾邮件，1394 封非垃圾邮件

并对选取好的训练集和测试随机排序，在该问题中，$n=2300$，$p=57$，$X:2300\times58$，$y:2300\times1$。

Step 1：将训练集数据代入迭代算法中，通过迭代最终得到Γ_k，此时回归模型为：

$$y=X\Gamma_k\alpha+\varepsilon$$

Step 2：将$X\Gamma_k$看作新的变量，然后利用最小二乘法估计计算新的线性模型系数$\hat{\alpha}$，计算结果为：

$$\hat{\alpha}=\left(\left(X\Gamma_k\right)TX\Gamma_k\right)^{-1}\left(X\Gamma_k\right)^T y$$，预测值为：$\hat{y}=X\Gamma_k\hat{\alpha}$

Step 3：利用测试集去代入预测模型，将得到的预测值与真实值比较，计算相应的分类准确率(真阳率、真阴率)，图示不同 k 值对应的分类准确率以及 ROC 曲线图，分析该方法的整体识别分类效果。

5. 结果展示与分析

5.1. 不同 k 值下分类准确率分析

根据选取的 k 值的不同，即主成分个数的不同，所对应的分类准确率也不同。本文选择的指标为平均分类准确率、真阳率(垃圾邮件分类准确率)和假阳率(非垃圾邮件分类准确率)。不同的 k 值下的分类准确率为如表 1 所示：

Table 1. Classification accuracy under different k values
表 1. 不同 k 值下的分类准确率

K 值	分类准确率	真阳率	真阴率
5	0.9174 (0.720)	0.9008 (0.465)	0.9283 (0.887)
10	0.9174 (0.791)	0.8986 (0.613)	0.9297 (0.907)
15	0.9174 (0.812)	0.8986 (0.641)	0.9297 (0.923)
21	0.9179 (0.834)	0.8997 (0.677)	0.9297 (0.937)

上表中括号中是利用 PCR 方法取相同 k 值时对应的分类准确率，用 PLS 计算得到的各种分类准确率都很高，当 k = 5 之后 PLS 方法对应的分类准确度已经稳定。同等情况下，PCR 得到的分类准确度还未稳定，且值较低。所以下面给出了真阳率和真阴率随 k 值变化的趋势图如图 1 所示。

由图 1 可知，PLS 方法得到的真阴率和真阳率在 k = 5 之后就稳定了，这与表 1 数据刚好对应，由于在之后准确率与 k 值的关系不大，可以选择更小的 k 值来进行建模和预测。

5.2. 不同分类截点下分类准确率分析

不同分类截点会使分类的结果不同，相应影响各种分类准确率的差别，所以本文选择分类截点 cut 在(−1.6, 2.2)之间，画出了对应的平均分类准确率 cr、真阳率(垃圾邮件分类准确率) tpr 和真阴率(非垃圾邮件分类准确率) tnr 的变化趋势如图 2 所示：

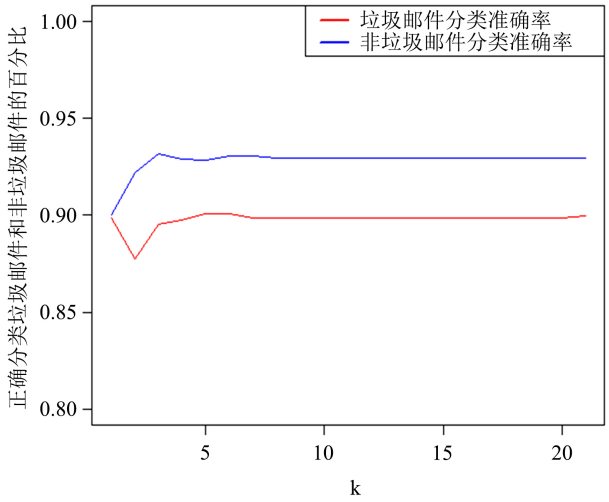

Figure 1. Accuracy rate change trend graph
图 1. 准确率变化趋势图

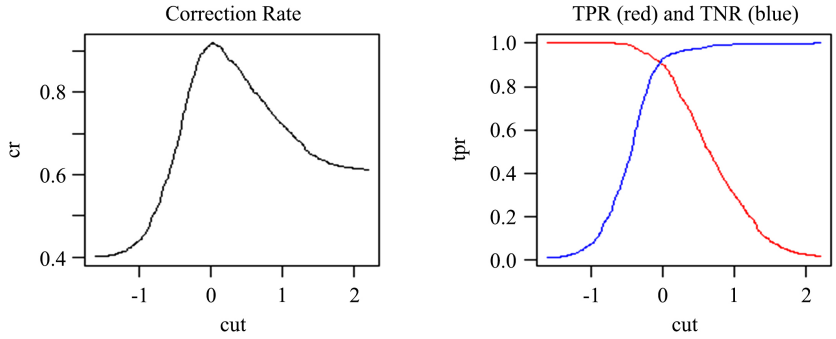

Figure 2. Classification accuracy percentage line chart
图 2. 分类准确率百分比折线图

在左图中，横坐标是分类截点的值，纵坐标表示了总的平均分类准确率，从左图中可以看出总的分类准确率随 cut 的增大而增大，当 cut 值取 0 附近时，分类准确率达到最高，之后又呈现一种下降趋势。右图中，横坐标也是分类截点，纵坐标为真阳率或真阴率值，图中红色线是真阳率，蓝色的是真阴率。cut 的取值越大，真阴率越大即非垃圾邮件分类正确率越高，cut 值越小，真阳率越大即垃圾邮件分类正确率越高。所以在具体操作中可以根据我们追求的目标而选择不同的分类截点。

5.3. 不同 k 值对应的 ROC 曲线图

受试者工作特征曲线(receiver operating characteristic curve，简称 ROC 曲线)，又称为感受性曲线。受试者曲线就是以假阳率为横轴，真阳率为纵轴所组成的坐标图，和受试者在特定刺激下由于采用不同的判断标准得出的结果画出的曲线。在该问题下，ROC 曲线具有如下作用：

- ROC 曲线能容易地查出任意界限值时对垃圾邮件的识别能力。
- 选择最佳的诊断界限值。ROC 曲线越靠近左上角，实验的准确性就越高，最靠近左上角的 ROC 曲线的点是错误最少的最好阈值，其假阳性和假阴性的总数最少。
- 两种及以上不同诊断试验对疾病识别能力的比较。越靠近左上角的曲线识别的准确度越高。

两种方法下取不同 k 值对应的 ROC 曲线如图 3 所示：

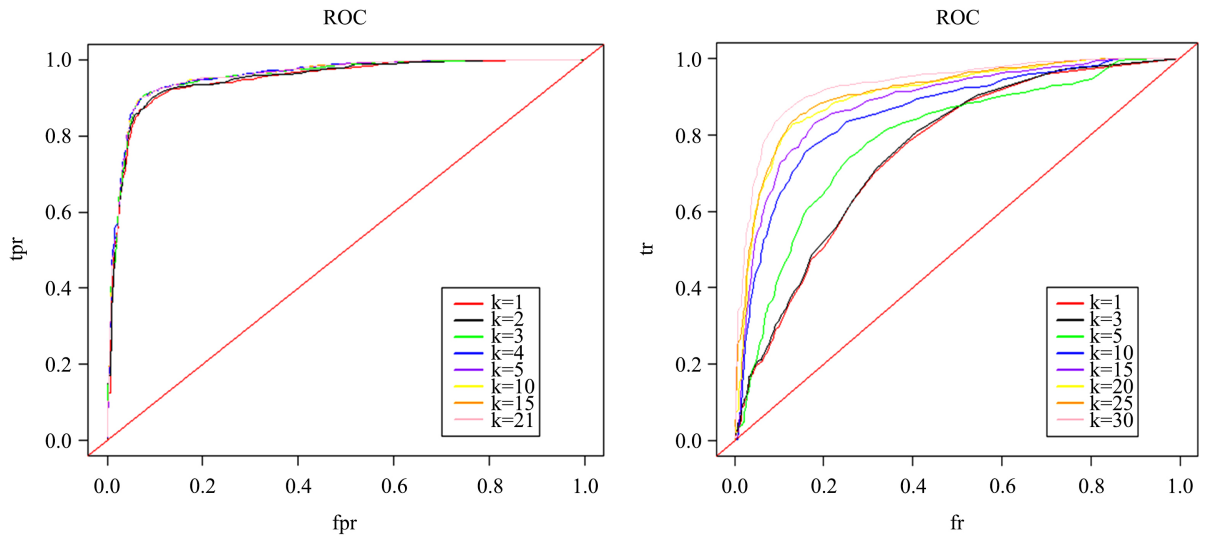

Figure 3. ROC graph
图 3. ROC 曲线图

上图分别是 PLS 和 PCR 两种方法对应的 ROC 曲线图，两图的横坐标都是假阳率，纵坐标都是真阳率。在左图中，可以很明显看到随着 k 的变化，整个 ROC 曲线图很快达到了稳定状态，当 k 取 5 时真阳率的变化趋势几乎不再变化。而在右图中，从 $k = 1$ 直到 $k = 30$ 真阳率的变化趋势一直在变化，甚至到 30 时还没有收敛，也即是 k 值的变化对真阳率的影响较大。除此之外，左边的 ROC 曲线整体比右边的靠近左上角，也即是 PLS 方法比 PCR 方法在对垃圾邮件的识别中更加有效，而且 PLS 在 k 值取很小时就已经达到了很高的识别准确度，所以在对垃圾邮件识别问题中 PLS 方法更加的有效。

6. 结论

偏最小二乘法的运用与最小二乘相比具有很大的优势。最小二乘是通过最小化误差的平方和来寻找数据的最佳函数匹配，利用最小二乘法可以简便地求得未知的数据，并使求得的数据与实际数据之间误差平方和最小。偏最小二乘法能够在自变量存在严重多重共线性的条件下进行回归建模；允许在样本点个数少于变量个数的条件下建模。本文中是基于偏最小二乘利用 LGK 双对角化设计了迭代算法和分类算法，对垃圾邮件分类问题进行了研究，并且分析了选取不同个数的主成分下分类准确度变化，还与 PCR 方法进行了比较，发现 PLS 的稳定性及识别垃圾邮件的准确性远远高于 PCR。在最小二乘法主成分回归的迭代中，主成分个数需要选择将近 30 才能达到较好的效果，但是对于偏最小二乘法仅需 5 个主成分就可以达到比最小二乘更好的效果。所以，相比之下本文所用的基于偏最小二乘法的邮件分类算法更加的有效、准确度更高、速度更快。

致　谢

感谢老师对本研究课题的指导，对于该研究的理论基础与编程问题比较困难，在这一过程中，老师与同学帮助了我很多，通过一步一步地推导与演练最终完成算法的实现。此外，也感谢论文评阅老师们的辛苦工作。

参考文献

[1] 李雪, 孙建平. 一种改进的偏最小二乘回归方法研究 [J]. 仪器仪表用户, 2017, 24(5): 16-19+28.

[2] Eldén, L. (2004) Partial Least-Squares vs. Lanczos Bidiagonalization-I: Analysis of a Projection Method for Multiple Regression. *Computational Statistics and Data Analysis*, **46**, 11-31. https://doi.org/10.1016/S0167-9473(03)00138-5

[3] 赵晓丹, 徐燕. 垃圾邮件分类技术对比研究[J]. 信息网络安全, 2014(2): 75-80.

[4] 毛雪莲. 多重共线性问题的偏最小二乘估计[J]. 科技视界, 2019(27): 152-153.

[5] 陈龙, 梁意文, 谭成予. 基于自适应性分类器的垃圾邮件检测[J]. 计算机工程, 2018, 44(5): 194-200.

[6] 李雨亭. 基于深度学习的垃圾邮件文本分类方法[D]: [硕士学位论文]. 太原: 中北大学, 2018.

[7] 黄鹤, 荆晓远, 董西伟, 吴飞. 基于 Skip-gram 的 CNNs 文本邮件分类模型[J]. 计算机技术与发展, 2019, 29(6): 143-147.

[8] 段同庆, 鲁瑞, 史新军, 刘红伟, 邓晓伟, 马骏. 偏最小二乘回归在探索 PCI 治疗冠心病患者预后影响因素中的应用[J]. 中国卫生统计, 2019, 36(6): 824-828.

[9] 丁学利, 任鹏. 基于偏最小二乘回归的空气质量数据校准研究[J]. 廊坊师范学院学报(自然科学版), 2020, 20(1): 9-14.

[10] Keshav, K. (2021) Partial Least Square (PLS) Analysis. *Resonance*, **26**, 429-442. https://doi.org/10.1007/s12045-021-1140-1

[11] 王琦, 吴钟扬, 黄陈蓉, 潘磊. 基于词嵌入与生成对抗网络的垃圾邮件分类算法[J]. 南京工程学院学报(自然科学版), 2018, 16(3): 20-27.

[12] 吴小晴, 万国金, 李程文, 林梦思, 曹书强. 一种改进 TF-IDF 的中文邮件识别算法研究[J]. 现代电子技术, 2020, 43(12): 83-86.

[13] 徐梦龙, 黄家旺. 朴素贝叶斯算法在垃圾邮件过滤方面的应用[J]. 网络安全技术与应用, 2018(7): 46-47.