拉普拉斯平滑对朴素贝叶斯邮件分类器的影响机制研究

张叔超[1*] 李廷元[1]
ZHANG Shuchao LI Tingyuan

摘 要 针对朴素贝叶斯分类器在邮件分类任务中的特征表示和平滑处理问题，文章提出了一种改进的预处理流程和参数优化方案。传统的朴素贝叶斯分类器在处理文本数据时通常采用单一的分割方式进行特征提取，这种方式可能导致部分重要特征的丢失或噪声的干扰，从而影响分类效果。为有效提升特征提取的质量，设计了一种“逗号 - 空格”双重分割策略，这一策略能够更加细致地处理邮件内容中的结构性信息，从而提升特征的全面性和准确性。同时系统验证了拉普拉斯平滑对模型性能的影响机制，拉普拉斯平滑作为常用的平滑技术，能够有效避免在朴素贝叶斯模型中因某些特征未出现而导致的零概率问题。为验证改进方案的有效性，选取 5 574 封真实邮件数据集进行实验，实验结果表明，改进后的朴素贝叶斯分类器在准确率方面取得了显著提升，准确率达到了 97.40%，相比传统实现提升了 6.82%。

关键词 垃圾邮件；正常邮件；朴素贝叶斯算法；拉普拉斯平滑算法

doi：10.3969/j.issn.1672-9528.2025.07.009

0 引言

随着互联网技术的快速发展，电子邮件成为人们日常生活和工作中不可或缺的沟通工具之一。凭借即时传递信息的高效性和低成本优势，打破地域限制，让全球用户得以随时随地互通有无。但与此同时，电子邮件的普及也给不法分子和不良商家提供了可乘之机。利用这一平台大肆开展广告轰炸、网络诈骗等非法活动，给广大用户带来诸多困扰，严重影响了用户的使用体验和信息安全。

据《中国网民权益保护调查报告 2016》[1] 的数据显示，超过半数的网民认为电子邮件存在个人信息泄露的隐患，尤其是在垃圾邮件泛滥的现状下。2016 年上半年，网民们每周平均收到 18.9 封匿名的垃圾邮件，这些邮件通常是广告推销、钓鱼诈骗、恶意软件传播等形式的邮件，给用户带来了极大的困扰和风险。垃圾邮件的频繁出现，不仅浪费了用户宝贵的时间，还严重影响了其正常的工作与生活。甚至垃圾邮件所含的链接可能隐藏着病毒或恶意程序，用

1. 中国民用航空飞行学院 四川广汉 618307

[8] CAO H, WANG Y Y, CHEN J, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation[C]//Computer Vision-ECCV 2022 Workshops, Berlin: Springer, 2022: 205-218.

[9] MIAO Y H, ZHANG B Y, LIN J, et al. A review on the application of blind deconvolution in machinery fault diagnosis[J]. Mechanical systems and signal processing, 2022, 163: 108202.

[10] HAN Y T, ZHANG J H, JIANG Z H, et al. Is the area under curve appropriate for evaluating the fit of psychometric models?[J]. Educational and psychological measurement, 2023, 83(3): 586-608.

[11] 袁非牛，章琳，史劲亭，等 . 自编码神经网络理论及应用综述 [J]. 计算机学报，2019, 42(1): 203-230.

[12] DENG W J, SHI Q, LI J. Attention-gate-based encoder–decoder network for automatical building extraction[J]. IEEE journal of selected topics in applied earth observations and remote sensing, 2021,14:2611-2620.

【作者简介】

宋媛媛（1996—），女，山西晋城人，硕士研究生，研究方向：计算机技术、移动互联应用及开发技术。

李宏滨（1968—），男，山西晋中人，硕士，副教授，研究方向：智能图像处理。

汪慧娟（1999—），女，河南信阳人，硕士研究生，研究方向：计算机技术、移动互联应用及开发技术。

（收稿日期：2025-02-20 修回日期：2025-06-30）

户点击后， 致电脑感染病毒，泄露个人隐私信息或造成设备损坏。

此外，随着大数据时代的到来，广告商越来越善于运用大数据技术分析用户的兴趣和需求，通过精准的广告推送方式让垃圾邮件更难以识别[2]。这些广告邮件经过精心包装，看 似与正常邮件没有太大区别，容易误导用户点击。而用户一旦点击邮件中的链接或附件，可能会导致设备感染病毒，甚至使个人财产受到威胁。与此同时，垃圾邮件消耗了大量的网络流量，导致用户的网络使用体验下降，特别是在流量有限的情况下，影响尤为严重。

更为严重的是，垃圾邮件的泛滥还可能导致邮件服务器的超载[3]。大量垃圾邮件涌入服务器，不仅增加了服务器的负担，更导致邮件服务的宕机或无法正常使用，给用户和企业带来经济损失。对于企业而言，垃圾邮件不仅会浪费公司员工的时间，还可能影响公司与客户、合作伙伴之间的商业关系，甚至影响企业的声誉和信誉。假如企业的邮件系统被大量垃圾邮件占据，正常的业务沟通可能被延误或错过，从而导致合同延期、合作破裂等一系列问题，造成的经济损失不容小觑。

因此，垃圾邮件的处理和识别显得尤为重要。为了避免垃圾邮件对个人和企业带来损害，用户需要采取有效的防范措施。随着垃圾邮件问题的日益突出，各国政府相继出台了一系列法律法规，旨在保护个人信息不被滥用据分析。据统计，全球有 152 余个国家和地区已经颁布了与个人信息保护、隐私保护相关的法律法规，各类行业标准也相继出台[4]。总而言之，尽管电子邮件为人们的工作和生活带来了极大的便利，但垃圾邮件所带来的问题也不容忽视。在大数据时代，垃圾邮件的隐蔽性和危害性逐渐增加，因此，及时识别和处理垃圾邮件，不仅是保护个人信息和网络安全的必要措施，也是维护互联网健康发展的重要环节。

1 相关理论

1.1 朴素贝叶斯原理

贝叶斯算法是一种基于概率分析判断事件发生可能性的方法，其核心在于选择概率最高的结果进行分类。该算法以 18 世纪哲学家托马斯 · 贝叶斯的名字命名，通过运用贝叶斯定理，将先验信息与观测数据相结合，进而推算事件发生的概率。其应用场景十分广泛，不仅适用于分类问题，在概率推理、预测分析、异常检测等领域发挥重要作用。

条件概率[5]公式：在事件 A 发生的前提下，事件 B 发生的概率。用 $P(A)$ 表示事件 A 的发生概率；用 $P(B)$ 表示事件 B 的发生概率；条件概率 $P(B|A)$ 表示在 A 发生的情况下 B 发生的概率；$P(AB)$ 则表示 A 和 B 同时发生的概率：

$$P(B \mid A) = \frac{P(AB)}{P(B)} \tag{1}$$

全概率公式[6]：用于计算事件 B 的概率，其中样本空间 Ω 被划分为事件 $A_1, A_2, \cdots, A_n$（每个事件的概率 $P(A_i) > 0$）。用公式表示为：

$$\begin{aligned} P(B) &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \ldots + P(B \mid A_n)P(A_n) \\ &= \sum_{i=1}^{n} P(A_i)P(B \mid A_i) \end{aligned} \tag{2}$$

贝叶斯公式：基于条件概率和全概率公式。假设样本空间 Ω 被划分为事件 $A_1, A_2, \cdots, A_n$（每个事件的概率 $P(A_i) > 0$），且 B 是样本空间 Ω 中的一个事件，则贝叶斯公式为：

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}, i, j = 1, 2..., n \tag{3}$$

在机器学习中，通过训练数据 D 来确定最佳的假设 h。假设 h 的先验概率是 $P(h)$，而 $P(D)$ 表示在没有特定假设下计算的数据 D 的概率。条件概率 $P(D|h)$ 表示在假设 h 成立时计算的数据 D 的概率。后验概率 $P(h|D)$，则是在给定数据 D 后假设 h 成立的概率。贝叶斯法则提供了计算后验概率的方法：

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \tag{4}$$

1.2 朴素贝叶斯算法

在机器学习中，朴素贝叶斯算法[5]是一种基于贝叶斯准则的分类方法。核心思想是通过计算各个特征在不同类别下的条件概率，然后利用贝叶斯定理结合先验概率来推断未知样本的类别。

朴素贝叶斯算法的“朴素”之处在于，假设所有特征之间是相互独立的，这一假设在实际问题中可能不完全成立，但简化了计算过程，使得该算法在很多实际应用中表现良好，特别是在数据特征较多的情况下。

由于其简单、高效且易于实现，朴素贝叶斯算法在许多任务中得到了广泛应用。例如，在垃圾邮件过滤系统中，朴素贝叶斯可以根据邮件中出现的特定词汇（如“免费”“优惠”等）来判断邮件是否为垃圾邮件。其会计算各个单词在垃圾邮件和非垃圾邮件中的出现概率，并结合邮件内容的实际特征来预测邮件的类别。

此外，朴素贝叶斯算法还广泛应用于文本分类、情感分析、推荐系统、疾病预测等任务中。尽管朴素贝叶斯假设特征之间的独立性较为理想化，但其高效性和良好的泛化能力使得它在大量实际问题中依然是一个非常有效的工具。

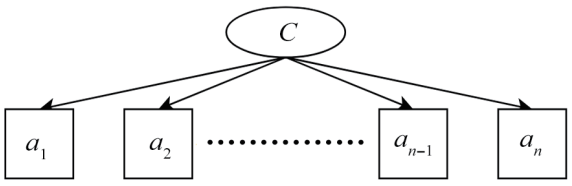

图 1 算法思想

可以看到，在朴素贝叶斯算法中（如图 1 所示），训练的实例被表示为特征向量 A 和决策类别变量 C。此算法假设特征在给定决策变量时彼此都独立。所以，对于其中每个特征向量 A 中的特征 a_k，如果这些特征是独立的，则条件概率 $P(A|C_i)$ 可以被分解为：

$$P(A|C_i)=\prod_{k=1}^{n}P(a_k|C_i),i,k=1,2,...,n \quad (5)$$

后验概率 $P(C_i|A)$ 是特征向量 A 属于决策类别变量 C_i 的概率。根据贝叶斯准则，这个概率用公式计算为：

$$P(C_i|A)=\frac{P(C_i)}{P(A)}\prod_{k=1}^{n}P(a_k|C_i),i,k=1,2,...,n \quad (6)$$

式中：类别 C_i 的先验概率 $P(C_i)$ 可以通过训练集中属于 C_i 的样本数量与训练集总量的比值来计算。

1.3 贝叶斯分类器

贝叶斯分类器 [7] 基于贝叶斯决策理论，是一种经典的概率统计方法，广泛应用于文本分类、垃圾邮件过滤、情感分析等领域。其核心思想是通过贝叶斯定理计算测试样本在各个类别下的后验概率，从而确定样本所属的类别。目前，常见的贝叶斯分类器有 4 种：Naive Bayes、TAN Bayes、BAN Bayes 和 GBN Bayes，通过在不同程度上引入特征间的依赖关系，进一步提高分类器的准确性。TAN Bayes 通过构建一个树形结构来建模特征间的依赖关系；BAN Bayes 则通过增强贝叶斯网络的结构来改进分类精度；GBN Bayes 则利用更复杂的图模型来描述特征间的依赖，提升了分类性能。

贝叶斯分类器的分类过程通常分为两个主要阶段：

（1）训练阶段。在这个阶段，分类器使用一组已知类别的样本集来学习特征与类别之间的关系。通过计算训练数据集中的每个类别的先验概率和各特征条件概率，分类器构建出一套可以用于分类的模型。

（2）测试阶段，在该阶段，利用已经训练好的贝叶斯分类器计算文本之中词汇的条件概率 [8]，并根据上文的贝叶斯公式计算后验概率，从而判别未知文本的类别。训练时间复杂度受样本量和特征依赖程度的影响。为了提高效率，常用的朴素贝叶斯分类器 [9] 对特征依赖进行简化优化，使分类过程更加高效。

1.4 拉普拉斯平滑算法

垃圾邮件过滤问题本质上是一个二分类问题，即将电子邮件分为垃圾邮件和非垃圾邮件两类。在使用朴素贝叶斯算法解决这个分类问题时，通常会先将邮件中的文本数据进行特征提取，使用 TF-IDF 的方法来提取文本的特征 [10]。当使用朴素贝叶斯算法去解决分类问题时，在构造出的朴素贝叶斯分类器上利用处理后的数据集进行训练时可以发现有可能出现某些特征的概率为零的情况，无论是在全文检索中某个字出现的概率，还是在垃圾邮件分类中，这种情况明显是不太合理的，无法因一个事件未观察到就认为该事件的概率是零，拉普拉斯平滑处理正是在处理这种情况时会用到的。

拉普拉斯平滑 [11] 指的是，假设 N 为表示训练邮件数据集总共有多少种类别，N_i 表示训练邮件数据集中第 i 总共有多少种取值。则训练过程中在算类别的概率时分子加 1，分母加 N，算条件概率时分子加 1，分母加 N_i 。

$$P(C)=\frac{|A_c|+1}{|A|+N}$$
$$P(X_i|C)=\frac{|A_c,x_i|+1}{|A|+N_i} \quad (7)$$

假设在垃圾邮件文本分类中，有 3 个类，C_1、C_2、C_3，在指定的训练样本中，某个单词 P_1，在各个类中观测计数分别为 0、980、20，P_1 的概率为 0、0.98、0.02，对 3 个量使用拉普拉斯平滑的计算方法为：

1/1 003=0.001，981/1 003=0.978，21/1 003=0.021 （8）

2 数据预处理

为了让计算机能够有效地读取和处理垃圾邮件文本，需要将原始文本数据转换成格式化的数据结构。这一转换过程可以通过多种方法完成，例如进行词向量化和特征提取 [12] 等操作。这一过程被称为文本预处理，通常包括多个步骤，如去除无关的空行、对文本进行分词处理、去除停用词、标点符号的处理等。通过这些预处理步骤，文本数据可以被整理成计算机能够理解和分析的格式，从而为后续的垃圾邮件分类和分析奠定基础。

传统方案：直接空格分割（易产生错误切分，如 "word1, word2" → ["word1, word2"]）。

本文方案：处理 "hello, world" → ["hello", "world"]，避免空字符串干扰。

2.1 去除文本空行

相关代码及解释

```
tokens = []
for comma_part in content.split(',')
    stripped = comma_part.strip()
```

把字符串 content 按逗号分隔成若干个子字符串 , 去除每个子字符串的前后空格，可以根据需求将每个去除空格后的子字符串添加到 tokens 列表中。

2.2 分词

将一个句子按照语法规则拆分为若干个词组的操作称为分词 [13]。分词有英文和中文两种，研究重点各有不同。中文分词以汉字为单位，目的是将句子切分成独立的词组；而英文分词以字母为单位，目的是将英文文本切分成独立的单词。本次垃圾邮件数据处理主要涉及英文分词。

相关代码及解释

```
if stripped:
tokens.extend(stripped.split(' '))
```

对 stripped 进一步处理，根据空格拆分 stripped 字符串，并将每个拆分出来的部分逐个添加到 tokens 列表中。

2.3 数据预处理结果对比

训练集预处理示例，如图 2 所示。

```
样本1（正常邮件）
原始内容： It vl bcum more difficult..
处理结果： ['It', 'vl', 'bcum', 'more', 'difficult..']
```

图 2 训练集邮件数据预处理

测试集预处理示例，如图 3 所示。

```
样本1（正常邮件）
原始内容： I've reached home finally...
处理结果： ["I've", 'reached', 'home', 'finally...']
```

图 3 测试集邮件数据预处理

3 实验

3.1 实验设置

数据集：5 574 封英文文本邮件（训练集与测试集比例为 8:2）。

评估指标：准确率、提升幅度。

对比基线：相同数据下未使用平滑的朴素贝叶斯。

3.2 数据预处理效果验证

双重分割使有效特征增加 0.9%，提升了分类边界清晰度，如表 1 所示。

表 1 分割策略对比（测试集）

分割方式	特征维度
单阶段空格分割	17 437
双重分割（本文）	17 595

3.3 综合实验结果

综合实验结果如表 2 所示。

表 2 实验结果

模型版本	准确率 /%	提升幅度 /%
基线（无改进）	90.58	—
仅改进预处理	91.48	+0.9
完整改进模型	97.40	+6.82

4 结语

随着互联网和大数据的发展，电子邮件的使用者日益增加，垃圾邮件也不断增多。在极大程度上阻碍电子邮件的发展，还给人们带来了较大困扰。所以，本文主要研究通过设计“逗号 - 空格”双重分割策略和系统验证拉普拉斯平滑对模型性能的影响机制对原有模型进行改进，实验结果表明准确率提升不错。但仍有一些改进之处：

（1）可以通过建立决策树模型 [14]，计算训练误差和测试误差，画出对应的决策树，并将结果可视化。

（2）未来可扩展支持多语言混合邮件处理、研究动态平滑因子调整策略以及开发基于用户反馈的增量学习机制。

参考文献：

[1] 中国互联网协会 .2016 第四季度中国反垃圾邮件状况调查报告 [J]. 互联网天地 , 2016, 3(7):89-90.

[2] 李敬瑶 . 反垃圾邮件过滤技术方法的研究 [J]. 福建电脑 , 2016, 32(10):61-62.

[3] 尹勇 . 垃圾邮件的危害与防范 [J]. 科协论坛 (下半月), 2013(1): 103-104.

[4] 陈卫兵 , 范志文 , 俞檑芳 . 如何做好个人金融网络安全信息保护 [J]. 金融科技时代 ,2025,33(1):21-26.

[5] 王斌 . 基于朴素贝叶斯算法的垃圾邮件过滤系统的研究与实现 [J]. 电子设计工程 ,2018,26(17):171-174.

[6] 陆青梅 . 基于贝叶斯算法的垃圾邮件过滤研究 [D]. 太原 : 中北大学 ,2008.

[7] 张坤 , 陈曦 , 宋云 , 等 . 一种 TAN 分类器改进方法 [J]. 计算技术与自动化 , 2019, 38(1): 55-61.

[8] 林士敏 , 田凤占 . 用于数据采掘的贝叶斯分类器研究 [J]. 计算机科学 , 2000,27(10):73-76.

[9] 王鹿 . 基于贝叶斯分类的垃圾邮件过滤技术研究 [D]. 上海 : 上海工程技术大学 , 2020.

[10] 柯西军 , 文家俊 , 徐海文 , 等 . 基于 BERTopic-Word 2Vec 的民航旅客评论主题挖掘 [J]. 数学的实践与认识 , 2025, 55(3): 155-166.

[11] 李文婷 , 肖蓉 , 杨肖 . 通过拉普拉斯平滑梯度提高对抗样本的可迁移性 [J]. 计算机科学 , 2024,51(S1):938-943.

[12] 俞荧妹 . 基于深度学习的垃圾邮件检测方法 [D]. 上海 : 东华大学 , 2023.

[13] 韩伟 . 不规范英文文本分词系统的设计与实现 [D]. 大连 : 大连理工大学 , 2015.

[14] 刘芬 . 基于内容的图像垃圾邮件过滤技术研究 [D]. 合肥 : 中国科学技术大学 ,2010.

【作者简介】

张叔超（2000—），通信作者（email: 2786849576@qq.com），男，安徽铜陵人，硕士，研究方向：目标检测、图像识别。

李廷元（1967—），男，四川眉山人，硕士，教授，研究方向：民航计算机应用、AI 算法。

（收稿日期：2025-03-09 修回日期：2025-07-04）