Hans Journal of Data Mining 数据挖掘, 2025, 15(2), 151-158
Published Online April 2025 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2025.152013

Hans汉斯

基于SMOTE算法的垃圾邮件检测研究

张博航[1]，闫　嘉[2]

[1]西南大学附属中学，重庆
[2]西南大学人工智能学院，重庆

收稿日期：2025年3月7日；录用日期：2025年3月21日；发布日期：2025年4月3日

摘　要

垃圾邮件检测一直是大数据和人工智能领域的研究热点。本文对Kaggle平台上的垃圾邮件数据集，进行了从数据预处理、文本特征构建，到垃圾邮件检测模型构建的完整数据处理过程。由于在垃圾邮件数据集中正常邮件和垃圾邮件占比极度不均衡，故采用SMOTE算法对垃圾邮件进行数据扩充，之后采用逻辑回归、支持向量机、决策树和随机森林四种学习算法构建垃圾邮件检测模型。本文对比了SMOTE前后四种检测模型的性能，尤其比较了准确率、精确度、召回率和F1-Score几个指标，以及混淆矩阵。实验结果可见，SMOTE算法有效提高了垃圾邮件检出的准确度，基于SMOTE算法的垃圾邮件检测模型具有较好性能。

关键词

SMOTE，精确度，召回率，F1-Score，混淆矩阵

The Research of Spam Detection Based on SMOTE Algrithom

Bohang Zhang[1], Jia Yan[2]

[1]High School Affiliated to Southwest University, Chongqing
[2]College of Artificial Intelligence, Southwest University, Chongqing

Received: Mar. 7th, 2025; accepted: Mar. 21st, 2025; published: Apr. 3rd, 2025

Abstract

The detection of spam has always been a research hotspot in big data and artificial intelligence. This paper presents a complete data analysis process for the spam data set on the Kaggle, including data preprocessing, the construction of text feature, building the detection model of a spam. Due to the imbalance between ham and spam, the SMOTE algorithm is used to expand the spam data, then four

文章引用: 张博航, 闫嘉. 基于 SMOTE 算法的垃圾邮件检测研究[J]. 数据挖掘, 2025, 15(2): 151-158.
DOI: 10.12677/hjdm.2025.152013

learning algorithms such as logistic regression, SVM, decision tree and random forest are used to build the detection model of spam. The performance of four detection models is compared before and after SMOTE, especially the classification accuracy, precision, recall, F1-Score and confusion matrix. The experimental results show that SMOTE algorithm can effectively improve the accuracy of spam detection, and the spam detection model based on SMOTE algorithm has good performance.

Keywords

SMOTE, Precision, Recall, F1-Score, Confusion Matrix

1. 引言

随着互联网时代的来临，电子邮件已经成为人们日常工作学习中不可或缺的一部分。但是伴随着电子邮件的普及，大量的垃圾邮件也给人们带来了很多困扰。垃圾邮件(spam)是指未经收件人许可批量发送的无用信息，其内容主要包括商业广告、钓鱼邮件、不良信息等，有的垃圾邮件甚至携带病毒。垃圾邮件在传播过程中会占用网络带宽资源，降低了正常邮件的传输效率；占据用户邮箱的容量，浪费空间资源；降低用户的工作学习效率，分散用户的时间和精力；同时还可能传播病毒，威胁用户个人以及整个网络的安全。因此研究垃圾邮件的检测算法，对于遏制垃圾邮件的传播、构建垃圾邮件的过滤系统具有关键作用。

2. 相关研究

垃圾邮件检测问题本质上是一个二值分类问题，邮件将被分为正常邮件和垃圾邮件。垃圾邮件检测主要采用基于邮件内容和行为两种检测技术，其中基于邮件内容的检测技术大量使用了机器学习和人工智能算法，是当前研究的热点。朴素贝叶斯算法由于其在文本分类领域的大量应用，从而使得该算法在垃圾邮件检测中也得到了大量使用。文[1] [2]都采用调整特征值的权重，通过不同的加权方式结合贝叶斯算法实现垃圾邮件检测，结果显示分类精度和准确率都有了显著提高。文[3]探讨了多种机器学习算法在垃圾邮件检测中的应用，包括朴素贝叶斯、支持向量机、多层感知机、卷积神经网络和循环神经网络，经过比较发现多层感知机和卷积神经网络效果最好。文[4]采用 KNN-SVM 算法实现垃圾邮件过滤，其中采用 KNN 对垃圾邮件训练样本进行选择，将训练样本缩减到 k 个，采用 SVM 对 k 个样本进行检测，结果表明 KNN-SVM 提高了垃圾邮件过滤的准确率，大幅度降低了虚警率，过滤速度较快，可以满足邮件处理的在线需求。随着深度学习的发展，大量深度学习算法也被应用于该领域。文[5]采用改进的卷积神经网络进行垃圾邮件识别，提出了一种 LSTM-Attention-CNN 混合模型，该模型首先利用 LSTM 提取文本上下文信息，用一个时序的输出向量作为特征向量，再进入 Attention 层计算每个单词的权重，之后在进入卷积层特征提取，池化层压缩降维，softmax 分类器计算输出类别。文[6]提出了一种新的结合卷积神经网络(CNN)、双向门控循环单元(Bi GRU)和注意力机制(Attention)的 CNN-Bi GRU-Attention 模型用于文本型邮件的分类，通过双向 GRU 与 CNN 全面提取邮件文本特征，包括局部特征和上下文关系特征，通过 Attention 提取对邮件分类结果影响较大的文本词条，有效提高了分类准确率。

垃圾邮件过滤同时也是一个经典的非平衡数据分类问题。在垃圾邮件处理的数据集中，正常邮件占

比远远大于垃圾邮件，这使得普通的分类器在识别少数类样本时性能不佳，因此提高少数类的识别精度成为垃圾邮件检测中关键问题。解决此问题的思路有两种，一是扩充少数类样本的数量，二是采用代价敏感的学习算法。文[7]采用 SMOTE 算法扩充少数类样本数量以消除正常邮件和垃圾邮件之间的数据不平衡性，并应用随机森林集成学习算法进行垃圾邮件识别。实验结果表明，本方法在召回率、精确度和 F1-Score 等多个分类指标上性能表现良好。文[8]针对沙尘暴气象数据，通过改进的过采样 SMOTE 算法提出基于边界因子的过采样改进，通过与其他几种过采样方法结合随机森林分类器进行对比研究，显示基于边界因子的过采样改进方法的有效性和鲁棒性。

综上所述，本文拟采用以 SMOTE 算法为基础的过采样技术扩充垃圾邮件样本集，实现正常邮件和垃圾邮件之间数量均衡，然后基于多种机器学习算法，如逻辑回归、支持向量机、决策树和随机森林等构建垃圾邮件检测模型，通过实验对比 SMOTE 前后模型的性能差异，提出一种基于 SMOTE 算法的垃圾邮件检测技术。

3. SMOTE 算法

SMOTE 算法是非平衡数据分类问题中最常采用的过采样方法[9]。这种算法的基本思想是，通过合成新的少数类样本来增加少数类样本的数量，从而平衡数据集。与简单的随机过采样不同，SMOTE 不是简单地复制少数类样本，而是通过插值的方式生成新的样本。具体来说，对于每个少数类样本，SMOTE 算法会从其 K 近邻中随机选择一个样本，然后在两者之间的连线上随机选取一点作为新合成的样本。这种合成新样本的方法有助于扩展少数类样本的特征空间，使模型能够更好地探索和学习少数类的特征，从而提高模型的分类性能。

SMOTE 算法的主要步骤如下：

1) 随机选择一个少数类样本：从少数类样本集中随机选择一个样本作为起点。

2) 确定 K 近邻：计算该样本的 K 个最近邻样本，这些近邻通常也是少数类样本。

3) 随机选择一个近邻：从上一步得到的 K 个近邻中随机选择一个样本。

4) 生成新样本：在起点样本和选定的近邻样本之间的连线上，根据一个 0 到 1 之间的随机数，生成一个新的合成样本。

重复步骤：重复上述步骤，直到生成足够数量的合成样本，以达到数据集平衡的目的。

4. 不均衡问题的评价标准

在不平衡数据分类时，通常称数目多的样本为负样本，数目少的样本为正样本。在非平衡数据学习中，主要目标是提高正样本的分类，同时对负样本保持合理的性能。为此，需要采用如下评价指标：

4.1. 混淆矩阵

Table 1. Confusion matrix
表 1. 混淆矩阵

	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

表 1 中，TP 为正确分类的正类样本数；TN 为正确分类的负类样本数；FP 为误分为正类的负类样本个数；FN 为误分为负类的正类样本个数。

4.2. 分类准确率

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{1}$$

对于样本均衡的数据集，常用准确率(公式 1)来评价分类性能。但是对于一个样本不均衡的数据集，该评价指标则存在较大问题。试想，对于一个只有 3%正类样本的不均衡数据集，分类器若将所有样本分类为负类，仍然有 97%的准确率，但是没有一个少数类样本被正确分类。显然这种分类器的性能是比较糟糕的。在不均衡数据集的分类问题中经常采用 F1-Score 来评价分类器的性能，该指标重点考察了少数类的识别精度，同时兼顾分类器的整体识别率。

4.3. F1-Score

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) \tag{2}$$

公式(2)中 Precision 被称为精确率，Recall 被称为召回率。它们的定义如下：

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{3}$$

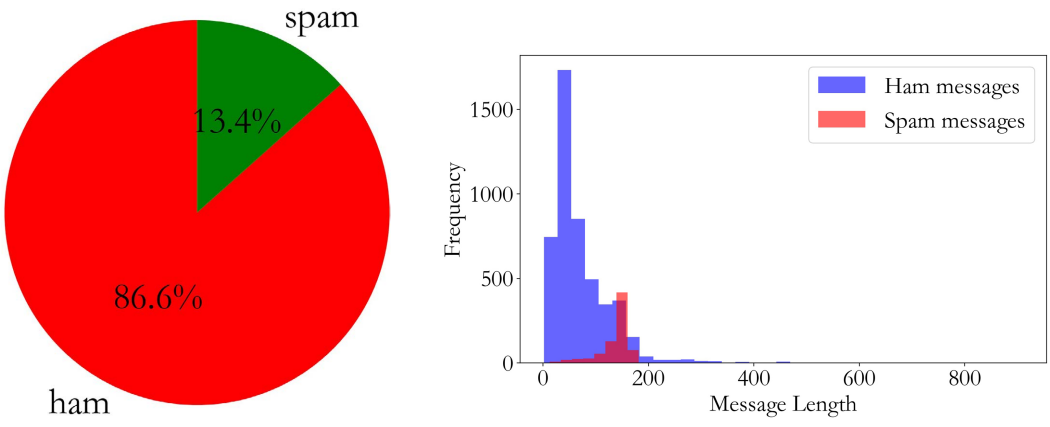
$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{4}$$

其中，Precision 表示判定为垃圾邮件的实例中真正为垃圾邮件的比例；Recall 表示垃圾邮件中实际检测出垃圾邮件的比例。

5. 实验及结果分析

5.1. 数据集介绍

实验的数据集来自于 Kaggle 竞赛社区的垃圾邮件数据集，包含 5574 条邮件文本。数据集的字段包括“label”和“message”，其中 message 是邮件的文本内容，label 则标注了该文本是垃圾邮件(spam)或者正常邮件(ham)。可以看到 ham 类一共有 4825 条数据，非重复数据有 4516 条；spam 类一共有 747 条数据，非重复数据一共有 653 条。图 1(a)展示了垃圾邮件和正常邮件的分布，可以看到垃圾邮件占比较小，待处理的文本数据类别不均衡。图 1(b)展示了垃圾邮件和正常邮件在文本长度上的直方图，可以看到垃圾邮件倾向于有更多的字符。



(a) (b)

Figure 1. The distribution of ham vs. spam. (a) The pie chart of ham vs spam; (b) The histogram of message length ham vs spam

图 1. 垃圾邮件和正常邮件的分布情况。(a) 垃圾邮件和正常邮件的饼图；(b) 垃圾邮件和正常邮件在文本长度分布直方图

5.2. 实验过程介绍

实验过程如图 2 所示，包括了数据准备，文本预处理构建特征向量，划分训练集和测试集，对训练集采用 SMOTE 算法进行均衡处理，采用四种常用分类算法构建垃圾邮件检测模型以及对各模型的数据进行比较五个阶段组成。在数据准备阶段，对数据进行了简单的清洗，没有空缺值和异常值，有一定的重复数据，但是重复数据本身也是邮件数据的一部分，故而没有删除。对不同类别的数据进行了简单的可视化，如图 1(a)所示，可以发现垃圾邮件占比较小，数据集是典型的不均衡数据集。文本预处理阶段的主要任务是完成邮件特征矩阵的构建。这里主要使用了自然语言处理技术，包括了去除标点符号和停用词，采用 TF-IDF 算法构建文本特征向量。对文本特征向量划分训练集和测试集，75%的数据作为训练集，25%的数据作为测试集。之后，采用 SMOTE 算法对训练集中不同类别的样本进行均衡化，扩充垃圾邮件类样本数量。最后，分别采用四种分类算法，即逻辑回归、支持向量机、决策树和随机森林对 SMOTE 前后的数据进行分类对比分析。

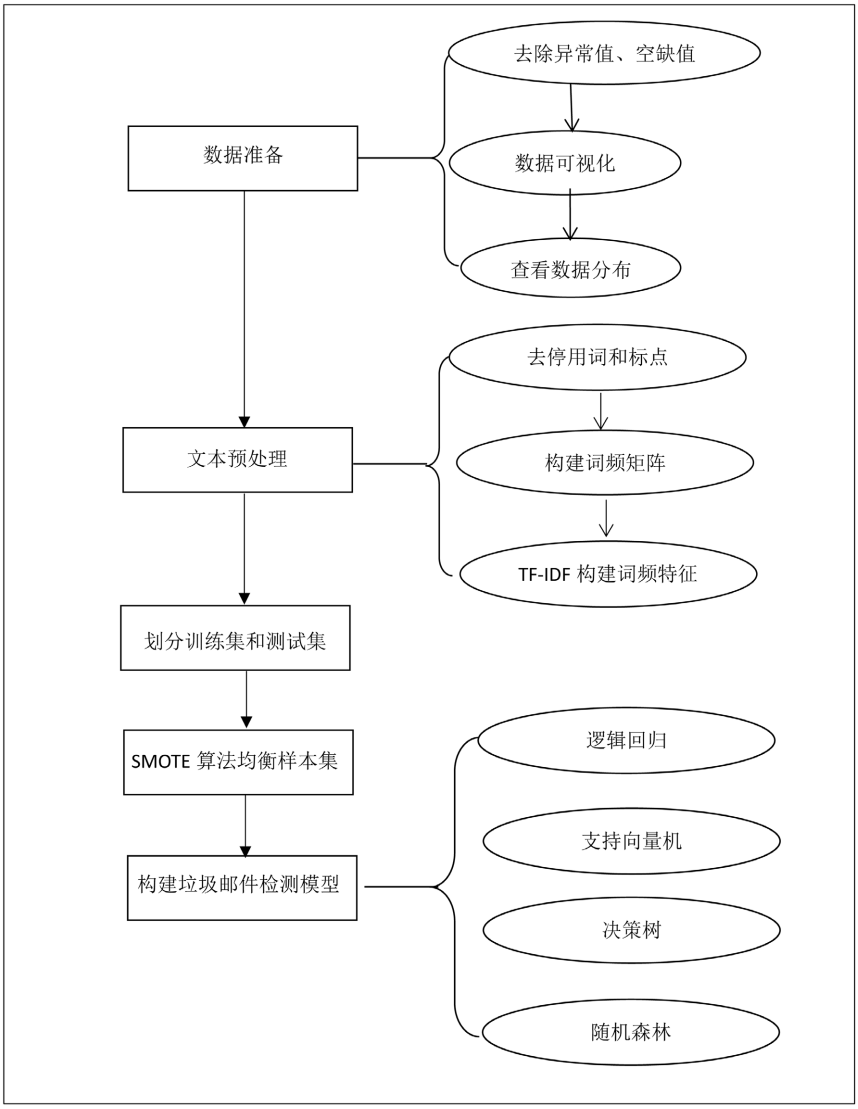


Figure 2. The process of experiment
图 2. 实验过程

5.3. 实验结果分析

对于 SMOTE 前后的训练数据，分别采用四种分类算法(逻辑回归、支持向量机、决策树和随机森林)建立垃圾邮件的检测模型。不同的分类模型分别采用分类准确率、召回率、精确率和 F1-Score 作为评价指标。四种检测模型的性能比较如表 2 所示。

Table 2. Comparison of four spam detection models
表 2. 四种垃圾邮件检测模型的性能比较

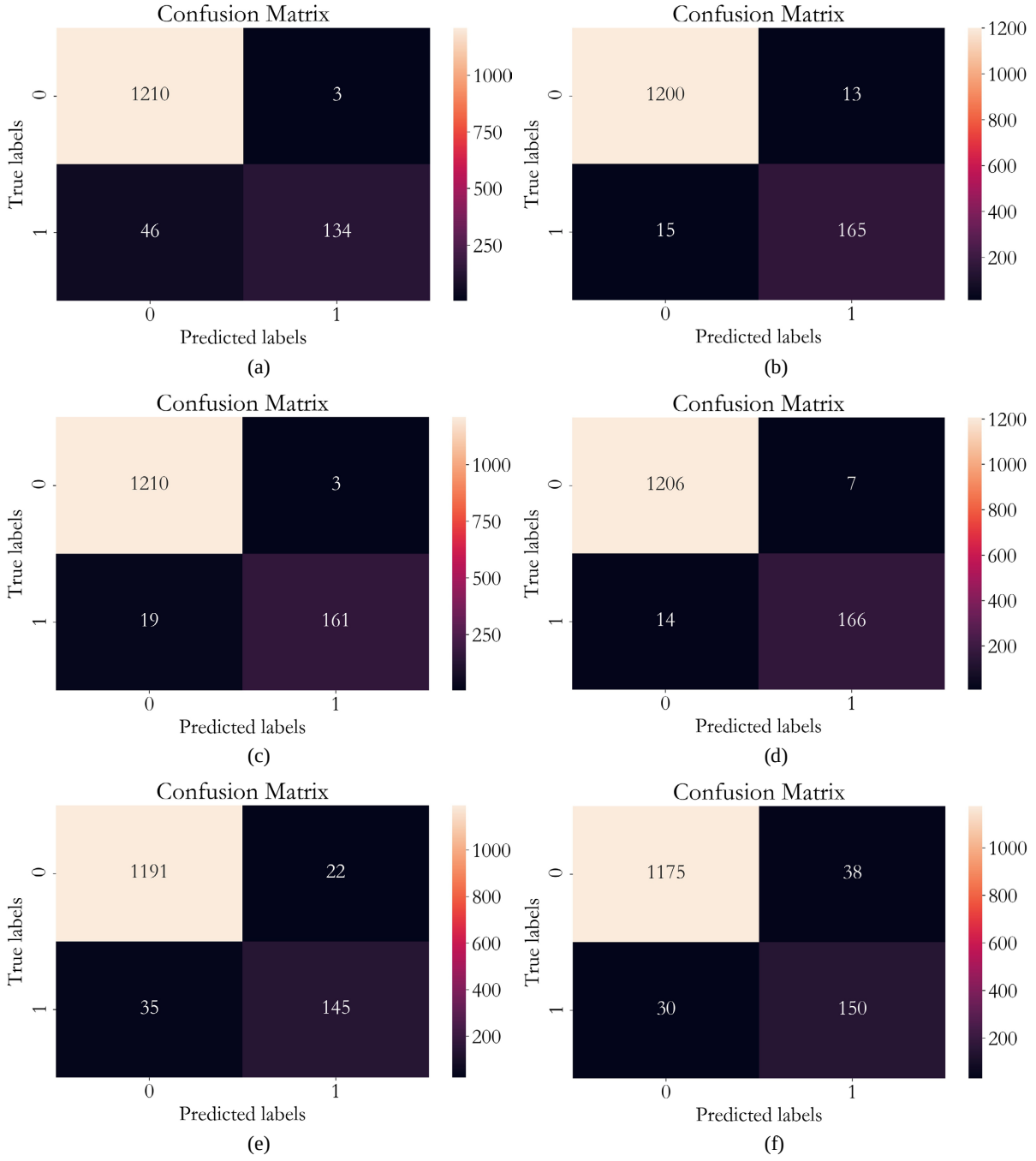
	未 SMOTE	SMOTE
逻辑回归	Accuracy = 0.9648 Recall = 0.7444 Precision = 0.9781 F1-Score = 0.9128	Accuracy = 0.9792 Recall = 0.9111 Precision = 0.9266 F1-Score = 0.9534
支持向量机	Accuracy = 0.9842 Recall = 0.8944 Precision = 0.9817 F1-Score = 0.9635	Accuracy = 0.9842 Recall = 0.9167 Precision = 0.9593 F1-Score = 0.9642
决策树	Accuracy = 0.8777 Recall = 0.7722 Precision = 0.8633 F1-Score = 0.8947	Accuracy = 0.8843 Recall = 0.8 Precision = 0.7912 F1-Score = 0.8825
随机森林	Accuracy = 0.9103 Recall = 0.8222 Precision = 0.9867 F1-Score = 0.9416	Accuracy = 0.9167 Recall = 0.8333 Precision = 1.0 F1-Score = 0.9484

从表 2 可见，逻辑回归算法中 SMOTE 后，模型的 Recall 得到了较大提升，从 0.7444 增加至 0.9111，F1-Score 也从 0.9128 提升至 0.9534。决策树算法 SMOTE 后 Recall 也有一定提升，从 0.7722 提升至 0.8，但是精确度由 0.8633 下降至 0.7912，F1-Score 从 0.8947 降低至 0.8825。另外两种分类器，SMOTE 前后的性能差异不大。

从整体的分类性能上看，支持向量机的分类性能最优，Accuracy 达到 0.9842；其次是逻辑回归，Accuracy 达到 0.9792；随机森林和决策树的性能较差，尤其是决策树模型，Accuracy 只能达到 0.8843。

线性分类器，即逻辑回归和支持向量的准确率要高于树模型分类器，如决策树和随机森林，原因可能是特征向量采用的是词袋模型，每个特征之间的关联度不大，适合使用线性分类器，故线性模型的分类性能优于树模型。

图 3 展示了四种分类器在 SMOTE 前后的混淆矩阵。从图 3 左边列上(a)、(c)、(e)、(g)对比右边列上的(b)、(d)、(f)、(h)是 SMOTE 前后四种分类算法的混淆矩阵(见表 1)，可以看到 SMOTE 后从实际的垃圾邮件中检测出垃圾邮件的比例是有提高的，充分说明 SMOTE 算法能够有效提升垃圾邮件检测能力。



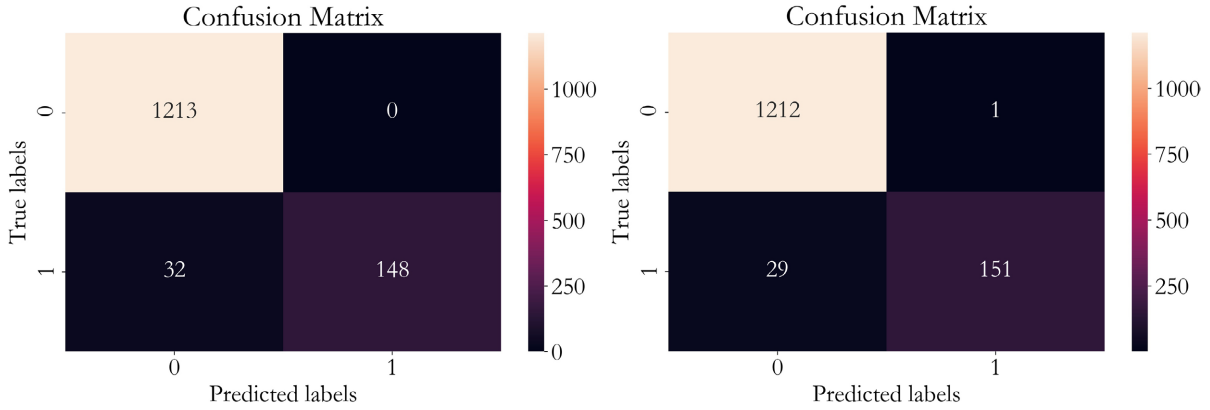


Figure 3. The confusion matrix of four models. (a) The confusion matrix of LR before SMOTE; (b) The confusion matrix of LR after SMOTE; (c) The confusion matrix of SVM before SMOTE; (d) The confusion matrix of SVM after SMOTE; (e) The confusion matrix of DT before SMOTE; (f) The confusion matrix of DT after SMOTE; (g) The confusion matrix of RF before SMOTE; (h) The confusion matrix of RF after SMOTE

图 3. 四个模型的混淆矩阵。(a) 未 SMOTE 逻辑回归混淆矩阵；(b) SMOTE 后逻辑回归混淆矩阵；(c) 未 SMOTE 支持向量机混淆矩阵；(d) SMOTE 后支持向量机混淆矩阵；(e) 未 SMOTE 决策树混淆矩阵；(f) SMOTE 后决策树混淆矩阵；(g) 未 SMOTE 随机森林混淆矩阵；(h) SMOTE 后随机森林混淆矩阵

6. 结束语

在垃圾邮件检测问题中，不均衡样本的分类识别一直是建模的重点。本文采用 SMOTE 算法扩充垃圾邮件样本数量以达到样本均衡的目的，之后分别采用四种分类算法——逻辑回归、支持向量机、决策树和随机森林来构建垃圾邮件检测模型。为了衡量垃圾邮件检测的效果，分别采用分类准确率、精确率、召回率和 F1-Score 作为模型性能的评价指标，结果显示 SMOTE 算法有效提升了垃圾邮件检测的精度，尤其在逻辑回归算法中 SMOTE 算法表现最优。在后续研究中，进一步考虑将样本分布等信息与 SMOTE 算法相融合，进一步提升 SMOTE 算法的性能。

参考文献

[1] 韩雪. 贝叶斯优化在垃圾邮件过滤中的应用研究[J]. 徐州工程学院学报(自然科学版), 2023, 38(2): 77-83.

[2] 王斯琴. 改进朴素贝叶斯算法在垃圾邮件过滤中的应用[D]: [硕士学位论文]. 重庆: 重庆师范大学, 2020.

[3] 冯军军, 李力. 机器学习在垃圾邮件过滤中的实现[J]. 电脑知识与技术, 2021, 17(8): 154-155.

[4] 林荫. 基于 KNN-SVM 的垃圾邮件过滤模型[J]. 现代电子技术, 2016, 39(23): 90-92, 97.

[5] 宋丹. 基于改进的卷积神经网络的垃圾邮件过滤方法[D]: [硕士学位论文]. 淮南: 安徽理工大学, 2021.

[6] 俞荧妹. 基于深度学习的垃圾邮件检测方法[D]: [硕士学位论文]. 上海: 东华大学, 2023.

[7] 丁伟民, 徐文钊. 一种基于 SMOTE 和随机森林的垃圾邮件检测算法[J]. 潍坊学院学报, 2020, 20(2): 14-15.

[8] 赵喆梅. 基于过采样的不平衡数据分类方法研究[D]: [硕士学位论文]. 兰州: 兰州交通大学, 2023.

[9] Chawla, N.V., Bowyer, K.W., Hall, L.O., *et al*. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953