2025 年 第 25 卷 第 20 期
2025, 25(20): 08583-12

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2406365
引用格式:王洁宁, 闫思卿, 孙禾. 基于 Stacking 集成学习的空管危险源数据分类[J]. 科学技术与工程, 2025, 25(20): 8583-8594.
Wang Jiening, Yan Siqing, Sun He. Air traffic management hazard data classification based on Stacking ensemble learning[J]. Science Technology and Engineering, 2025, 25(20): 8583-8594.

基于 Stacking 集成学习的空管危险源数据分类

王洁宁[1,2], 闫思卿[1,2*], 孙禾[1,2]
(1. 中国民航大学空中交通管理学院, 天津 300300; 2. 天津市空管运行规划与安全技术重点实验室, 天津 300300)

摘 要 在现代空管系统中,高效准确地识别和分类危险源文本数据对于保障飞行安全至关重要,空管危险源数据指的是那些可能影响航空安全的潜在因素、条件或事件的信息集合,然而现有的文本分类方法难以应对数据类别多样性和类别不平衡问题。当下迫切需要开发适用于空管系统的高效分类方法,以提高飞行安全水平。针对单一学习器用于空管危险源文本分类存在的类别分布较多,难以捕捉类别数据不平衡时的文本特征导致预测精度下降的问题,提出基于 Stacking 训练思想的、两次加权的改进集成模型。首先,参考双防机制对危险源和安全隐患完成类别划分;再采用词频-逆文档频率(term frequency-inverse document frequency,TF-IDF)算法提取预处理后的危险源文本特征完成向量化,并利用合成少数类过采样技术(synthetic minority over-sampling technique,SMOTE)和自适应合成过采样算法(adaptive synthetic sampling approach,ADASYN)分别随机生成向量化后的少数类文本,使文本数据集的类别分布趋于平衡;再从基学习器每折交叉验证的 F_1 分数加权和基学习器之间敏感性评估机制动态加权两方面改进 Stacking 集成模型,提高类别不平衡危险源文本的分类性能。在所构建的数据集上的实验结果表明:相较于 SMOTE + 改进集成模型,ADASYN + 改进集成模型的精确率、召回率和 F_1 分数分别提升 0.9、1.1 和 1.0 个百分点,较好地抑制处理多数类别过拟合的问题,实验结果验证了所提算法的有效性。

关键词 双防机制; 空管危险源; 文本分类; 自适应合成过采样算法(ADASYN); Stacking 集成模型
中图法分类号 TP391; **文献标志码** A

Air Traffic Management Hazard Data Classification Based on Stacking Ensemble Learning

WANG Jie-ning[1,2], YAN Si-qing[1,2*], SUN He[1,2]
(1. College of Air Traffic Management, Civil Aviation University of China, Tianjin 300300, China;
2. Tianjin Key Laboratory of Air Traffic Management Operation Planning and Safety Technology, Tianjin 300300, China)

[**Abstract**] Modern air traffic management systems necessitate efficient and accurate identification and classification of hazard-related text data to ensure flight safety. Air traffic control hazard data encompasses information on potential factors, conditions, or events that may adversely impact aviation safety. Existing text classification methods face challenges due to the diversity of data categories and imbalances within classes. An enhanced ensemble model based on the Stacking framework, incorporating a dual-weighting mechanism was proposed for improved performance. A dual-protection strategy was implemented to categorize hazards and safety risks systematically. The methodology employed the term frequency-inverse document frequency (TF-IDF) algorithm to extract and vectorize features from preprocessed hazard texts. To address class imbalance, the synthetic minority over-sampling technique (SMOTE) and adaptive synthetic sampling approach (ADASYN) algorithms were utilized to generate synthetic samples for minority classes. The Stacking ensemble model was refined by dynamically weighting the F_1 scores derived from cross-validation of base learners and integrating a sensitivity assessment mechanism across the ensemble. Experimental results on the constructed dataset demonstrate that the ADASYN-enhanced ensemble model achieves notable improvements in precision, recall, and F_1 scores by 0.9%, 1.1%, and 1.0%, respectively, effectively mitigating overfitting in majority classes. The proposed algorithm significantly enhances the classification performance of imbalanced hazard text categories, contributing to the advancement of safety risk management in air traffic control.

[**Keywords**] dual-protection mechanism; air traffic hazards; text classification; adaptive synthetic sampling approach (ADASYN); Stacking ensemble model

收稿日期: 2024-08-24; **修订日期**: 2025-04-24
基金项目: 国家重点研发计划(U2133207)
第一作者: 王洁宁(1966—),男,汉族,甘肃兰州人,博士,教授。研究方向:空管运行安全及空管系统仿真。E-mail:wang_jiening@aliyun.com。
*通信作者: 闫思卿(1999—),男,蒙古族,内蒙古呼和浩特人,硕士研究生。研究方向:自然语言处理与应用。E-mail:18747995182@163.com。

在现代民航领域,随着航空网络的扩大和航班数量的增加,危险源种类和数量的增多导致空管安全管理变得更加复杂。危险源分类不准确、信息共享不充分等问题层出不穷。精准的分类模型能够更好地预测空管风险事件[1],减少潜在危险源对空管系统的影响,保证系统能够更加高效地调度资源,提升空管运营效率。

随着当今人工智能技术的不断完善和自然语言处理技术的成熟[2-4],危险源数据自动分类技术得到了快速发展[5-6]。在民航空管领域,危险源数据(Hazard Data)指的是那些可能影响航空安全的潜在因素、条件或事件的信息集合。这些数据是识别和管理空中交通系统中风险的关键要素。Ma等[7]使用基于 Accimap 对危险源风险进行了分析,将自然语言处理(natural language processing,NLP)与矿山事故因果关系建模相结合,为危险源分析提供技术解决方案。相关技术的发展启发了民航领域的研究人员[8-9]。目前,针对空管危险源数据采用自然语言处理的研究[10]相对较少。郭九霞[11]采用基于 TFIDF-TextRank 算法对空管系统危险源小样本、多标签的数据样本进行文本分类。研究表明,自然语言处理技术在危险源分类中显示出其独特的优势。但目前研究中,算法大多基于单一学习器展开,对于建立空管危险源数据集成模型研究仍然不足。在各类民航数据中,空管危险源种类繁多,且不同种类之间数据量的差距较大,使得数据平衡化处理变得尤为重要。

鉴于此,在文本特征向量化和数据平衡化的现有工作基础上,提出空管危险源文本分类的 Stacking 集成模型[12]。采用能够有效识别并突出关键词的词频-逆文档频率(term frequency-inverse document frequency,TF-IDF)算法[13]提取预处理后的危险源描述文本的特征词并向量化,再运用自适应合成过采样算法(adaptive synthetic oversampling method,ADASYN)[14]自适应地生成新的合成样本,并结合改进 Stacking 集成模型构建空管危险源描述文本分类模型。该研究借助优化文本特征表征以及数据平衡策略,提升了危险源分类的准确性以及鲁棒性,为空管安全风险的实时预警、精准管控提供可靠的技术路径,还为高维度、不均衡文本数据的处理范式提供了跨领域迁移的参考价值。

1 相关工作

图 1 展示了一个系统化的数据处理和模型训练流程。首先,从某空管单位每季度监测的数据集中获取原始危险源数据,并进行数据预处理和文本向量化,包括缺失值处理和特征提取。实际生产过程中,危险源数据由空管单位多个部门上报,各部门主要关注自身风险,缺乏全局视野。尽管安管部门进行统一处理,但仍存在危险源部分重叠、分类标准不统一等问题。此外,危险源关注末端的生产管理,其分类主要基于部门的生产实际,缺少对危险来源的深层次分析。这导致现有分类在指导各部门工作时存在一定局限性。为了克服这些难题,采用 NLP 方法分析危险源数据显得尤为必要。NLP 可以帮助挖掘危险源背后的深层次原因,分类研究为全面的风险评估提供帮助。然而,实际生产中的危险源种类复杂、来源广泛,但由于数据表述简洁总量较少,撰写缺乏统一标准,属于低风格化的小样本问题。在使用 NLP 方法进行数据处理时,如何在不改变危险源原意的前提下进行数据集扩充是一个关键问题。

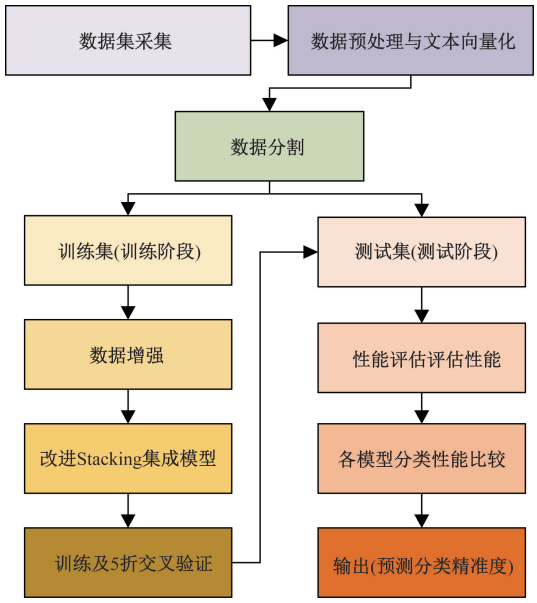


图 1 技术路线图

Fig. 1 Technical roadmap

在训练阶段,传统的 Stacking 集成模型(Stacking model, SM)[15]在小规模数据集上的训练速度较快且性能优异,但未考虑数据集不平衡问题,导致模型过拟合多数类数据,预测结果偏向多数类。此外,不同基学习器的性能差异显著,由于次级学习器无差别处理分类结果,可能导致整体模型性能下降。因此,需要优化基学习器分类结果的权重分配,以改善集成模型在不平衡数据集上的分类性能和鲁棒性。并采用过采样算法进行数据增强,增加训练数据数量及其多样性,以提高模型的泛化能力。并应用改进的 Stacking 集成模型,结合多种基

学习器的输出,进行超参数调优,提升模型性能。在训练过程中,使用5 折交叉验证方法,确保模型的稳定性和抗过拟合能力。在测试阶段,利用独立的测试集进行模型评估,通过一系列性能指标全面评估模型的分类能力,最终输出预测结果和分类分析报告。整个流程系统地涵盖了数据采集、预处理、模型训练、验证与评估的各个环节,确保模型的高准确度和鲁棒性。

2　基础知识

2.1　TF-IDF 算法

TF-IDF 是一种用于信息检索与文本挖掘的常用加权技术,用以评估一个词语对于一个文档的重要程度[16]。词语的重要性随着它在文件中出现的次数成正比增加,但随着其在语料库中出现的频率成反比下降。该算法通过计算文档中词语的权重,从而区分文档之间的类别。

TF(t,d)衡量了一个词语在文档中出现的频率,在文档中出现的次数 IDF(t,D)越多,它对文档的重要性就越大。然而仅使用词频可能会偏向于那些词数较多的文档,因此通常会进行归一化处理。衡量词语在整体文档中的重要程度,它的主要思想是:如果包含词条的文档越少,IDF(t,D)越大,则说明词语 t 具有很好的类别区分能力。

TF-IDF 可表示为

$$N_{i,j} = \mathrm{TF}(t,d)\,\mathrm{IDF}(t,D) = \frac{f_{t,d}}{\sum_{t\in d} f_{t,d}} \lg \frac{|D|}{1+\{d\in D; t\in d\}} \tag{1}$$

$$x_{\mathrm{new}} = x + \gamma(x_{nn} - x) \tag{2}$$

式中:$f_{t,d}$为词语 t 在文档 d 中出现的次数;$\sum_{t\in d} f_{t,d}$ 为文档 d 中所有词语出现次数的总和;$|D|$为语料库中文档的总数;$\{d\in D; t\in d\}$ 为包含词语 t 的文档数目,防止分母为0,通常会在分母上加 1;$N_{i,j}$ 为 t_i 词语在文档 d 中的权重,其中 $0<j<D$;x 为少数类样本;x_{nn}为 x 的近邻点;γ 为0 ~ 1的随机数;x_{new}为新生成的样本。

2.2　过采样算法

2.2.1　SMOTE 算法

SMOTE 算法通过在少数类的样本间插入人工合成的新样本来增加少数类的样本数量,改善数据的平衡性,从而提升分类模型的性能。其基本思想是对少数类样本进行分析并合成新样本添加到数据集中,算法包括以下步骤:对于数据集中的每一个少数类样本 x,计算其与少数类中其他所有样本之间的欧式距离,为每个少数类样本 x 选择 K 个最近邻样本(K 为邻近样本的数量),从 K 个最近邻样本中随机选择一个样本 x_{nn},在每个特征维度上,通过式(3)合成一个新的样本 x_{new},将新生成的样本加入数据集中,更换少数类样本点并重复以上步骤。

2.2.2　ADASYN 算法

ADASYN 算法是一种通过自适应地生成合成样本来平衡数据集的过采样技术,其核心思想是根据少数样本被多数样本覆盖的程度来动态调整合成样本的数量,算法流程包括一下步骤:根据少数和多数类样本的数量,计算数据集的不平衡比例,根据少数样本周围多数样本的密度,计算一个加权值,反映该少数类样本被多数类样本“覆盖”程度。加权值高的少数类样本将生成更多的合成样本,同时也会在每个少数类样本于其近邻少数类样本之间插值,合成新样本。公式化表达为,对于一个少数样本 x_i,选择其 K 个最近的少数类邻居,然后根据计算得到的加权值 G_i来生成新的合成样本,最终新样本 x_{new}通过以下公式生成:

$$x_{\mathrm{new}} = x_i + \gamma(x_{zi} - x_i) \tag{3}$$

式(3)中:x_{zi}为 x_i的一个少数类邻居。

3　算法设计

集成学习通过组合多个模型显著提升预测精度,其中异质集成算法 Stacking 特别突出。它整合了多种类型的模型,利用各自的优势增强泛化能力,并提高整体预测精度。在 Stacking 模型中,第一层由多个基学习器构成,这些学习器在预处理的数据集上进行训练;而第二层,即次级学习器,基于第一层学习器的输出进行训练,最终构建出一个性能强大的集成学习器。

3.1　基分类器

为了确保 Stacking 集成模型预测的准确性和较好的泛化能力,需要考虑基分类器种类的多样性和预测危险源描述文本的精准性,因此选择适合数据集规模较小的、分类性能较好的随机森林(random forest,RF)、支持向量机(support vector machines,SVM)[17]、极限梯度提升(extreme gradient boosting,XGBoost)、轻量级梯度提升机(lightweight gradient boosting machine,LightGBM)和 Stacking 集成模型融合的策略[18]。

3.2　改进的 Stacking 文本分类集成模型

提出一种基于交叉验证 F_1 分数和敏感性评估机制的动态加权策略改进的 Stacking 集成模型(improved stacking model,ISM)。对于每个基学习器进行交叉验证,通过对基学习器每折验证集上的 F_1 分

数进行精度加权,可以有效提高模型的泛化能力,更好捕捉数据的多样性。针对基学习器之间分类性能参差不齐的问题,设计一种动态权重调整基学习器的敏感性评估机制,该机制旨在衡量基学习器对特定输入数据的敏感度,允许集成模型更精细地调整各初级学习器的贡献,以最大限度提高对当前数据的预测准确性。最终,考虑到 Stacking 集成模型的第二层使用的数据集是由第一层各基学习器的训练集输出值组合而成,这会导致新训练集丢失部分原始数据集的信息。仅用元学习器来训练基学习器的输出数据,无法充分体现原始数据间的关联性。为了提升 Stacking 集成模型的性能,需要引入最优特征子集之间的关联性,通过特征间的相互作用来增强模型表现。

3.2.1 基学习器的 F_1 分数精度加权计算

对于某一基学习器,在进行 K 折交叉验证时,会按照 K 折数生成 K 个不同的测试集。这些测试集用于评估基学习器的预测结果并进行加权计算,以评估初级学习器在 K 折交叉验证测试集上的预测精度,并采用 F_1 分数值作为衡量指标,具体流程如下。

假设含有 N 个 Z 维向量的样本数据共有 M 个分类,样本数据可表示为

$$X = \{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\},\quad x_n \in \mathbf{R}^Z, y_n = (1,2,\cdots,M) \tag{4}$$

训练过程中,初级学习器 u 的 K 折交叉验证会计算出样本的 x 类分布向量 $\boldsymbol{p}_{u,m}^{(k)}(x)$,计算过程为

$$\boldsymbol{p}_u^{(k)}(x) = [\boldsymbol{p}_{u,1}^{(k)}(x),\boldsymbol{p}_{u,2}^{(k)}(x),\cdots,\boldsymbol{p}_{u,m}^{(k)}(x)],\quad u = 1,2,3 \tag{5}$$

式(6)中:$\boldsymbol{p}_u^{(k)}(x)$ 为判断为第 m 类的概率;m 为类别;u 为初级学习器的编号;K 为交叉验证的折数。

通过记录初级学习器 u 在每一折 K 的测试集上计算得到的 F_1 分数值 R_u^k,然后对于每一折 K 的 F_1 分数值,计算其精度权重 $\rho_{u,k} = \dfrac{R_u^k}{\sum_{k=1}^{k} R_u^k}$。根据 K 折交叉验证中计算得到的精度权重 $\rho_{u,k}$,对初级学习器 u 在 Stacking 集成模型的第一层训练集 D_{train} 的预测输出类分布向量 $\boldsymbol{p}_u^{(k)}(x)$ 进行赋权,得到加权后类分布向量为

$$\boldsymbol{p}_u^{(k)}(x) = \rho_{u,k}[\boldsymbol{p}_{u,1}^{(k)}(x),\boldsymbol{p}_{u,2}^{(k)}(x),\cdots,\boldsymbol{p}_{u,m}^{(k)}(x)] \tag{6}$$

对于每个初级学习器 u,将其在所有 K 折交叉验证中的加权类分布向量合并,作为该学习器总体预测输出,合并每个基学习器的输出矩阵得到加权预测输出矩阵为

$$\boldsymbol{p} = \Big[\sum_{k=1}^{K}\rho_{1,k}\boldsymbol{p}_1^{(k)}(x),\sum_{k=1}^{K}\rho_{2,k}\boldsymbol{p}_2^{(k)}(x),\cdots,\sum_{k=1}^{K}\rho_{u,k}\boldsymbol{p}_u^{(k)}(x)\Big] \tag{7}$$

3.2.2 动态权重调整的敏感性评估机制

假设一个基学习器 u 对于输入数据 x 的预测输出为概率分布 $T(y_i|x)$,其中 y_i 为风险类别,通过该基学习器 u 计算其对每个样本的敏感性得分,数据的敏感性 $S_u(x)$ 可定义为

$$S_u(x) = -\sum_i T(y_i|x)\lg T(y_i|x),\quad u = 1,2,3 \tag{8}$$

通过训练剩余的基学习器,计算其对每个样本的敏感性得分,为了在不同基学习器之间比较敏感性得分,需要对得分进行归一化处理,确保它们在同一量级上,归一化的敏感性得分可表示为

$$S_u^*(x) = \frac{S_u(x) - \min[S_u(x)]}{\max[S_u(x)] - \min[S_u(x)]} \tag{9}$$

式(10)中:$S_u^*(x)$ 为归一化后的敏感性得分;$S_u(x)$ 为原始敏感性得分;$\max[S_u(x)]$ 和 $\max[S_u(x)]$ 分别为在所有历史数据上计算得到的最大和最小敏感性得分。

基于归一化的敏感性得分,为每个基学习器的类分布向量 $\boldsymbol{p}_u^{(k)}(x)$ 分配动态权重,这个权重可以根据当前待分类的文本数据动态调整,基学习器 u 的动态权重可表示为

$$W_u = \frac{S_u^*(x)}{\sum_u S_u^*(x)} \tag{10}$$

将每个基学习器的合并输出矩阵赋予动态权重值,并以此作为次级学习器的输入特征,总体加权预测输出矩阵为

$$\boldsymbol{p}_{\text{total}} = \Big[\sum_{k=1}^{K}W_1\rho_{1,k}\boldsymbol{p}_1^{(k)}(x),\sum_{k=1}^{K}W_2\rho_{2,k}\boldsymbol{p}_2^{(k)}(x),\cdots,\sum_{k=1}^{K}W_u\rho_{u,k}\boldsymbol{p}_u^{(k)}(x)\Big] \tag{11}$$

3.2.3 针对原始数据的扩充

空管单位提供的危险源数据直接关联运行风险,总量较少且表述简洁。由于数据涉及多个运行部门,数据重叠和分类标准不统一等问题使得对危险源的分析更加困难。为了避免危险源分类效果不清晰,甚至生成虚构的危险源,暂时不具备使用生成式模型(如 GAN)的条件,只能在不增加现有数据的前提下进行数据集扩充。

在训练元学习器时,需将原始数据集与基学习器的输出组合,并引入最优特征子集,合并后输入下一级的元学习器,具体流程如下。

(1)将全部特征数据按照 4∶1的比例划分为训练集 D_{train} 和测试集 D_{test}。

(2)分别对所有基学习器的训练集和测试集进行 K 折交叉验证,再次训练以不同训练集和测试集组成的数据,并得到的输出结果 D^{*}_{train} 和 D^{*}_{test}。

(3)将所有基学习器的 D^{*}_{train} 和 D^{*}_{test} 分别赋动态权重 W_u,合并得总体加权预测输出矩阵 $\boldsymbol{D}^{*}_{\text{train}u}$ 和 $\boldsymbol{D}^{*}_{\text{test}u}$,并与初始的 D_{train} 和 D_{test} 合并,形成完整的训练集 M_{train} 和测试集 M_{test}。将 M_{train} 作为第二层学习器的训练集训练元学习器。再用该元学习器对测试集 M_{test} 进行预测,预测结果即为最终的输出结果。样本扩充下的集成学习结构如图 2 所示。

训练模型过程中,采用精度加权和样本扩充的 Stacking 算法。进行预测时,首先,将样本数据输入第一层学习器,得到 5 个基学习器的预测结果。接着,将每个预测结果与相应的基学习器权重相乘,得到加权结果,并将这些加权结果与样本数据结合,形成新的样本数据。最后,这些新样本数据与

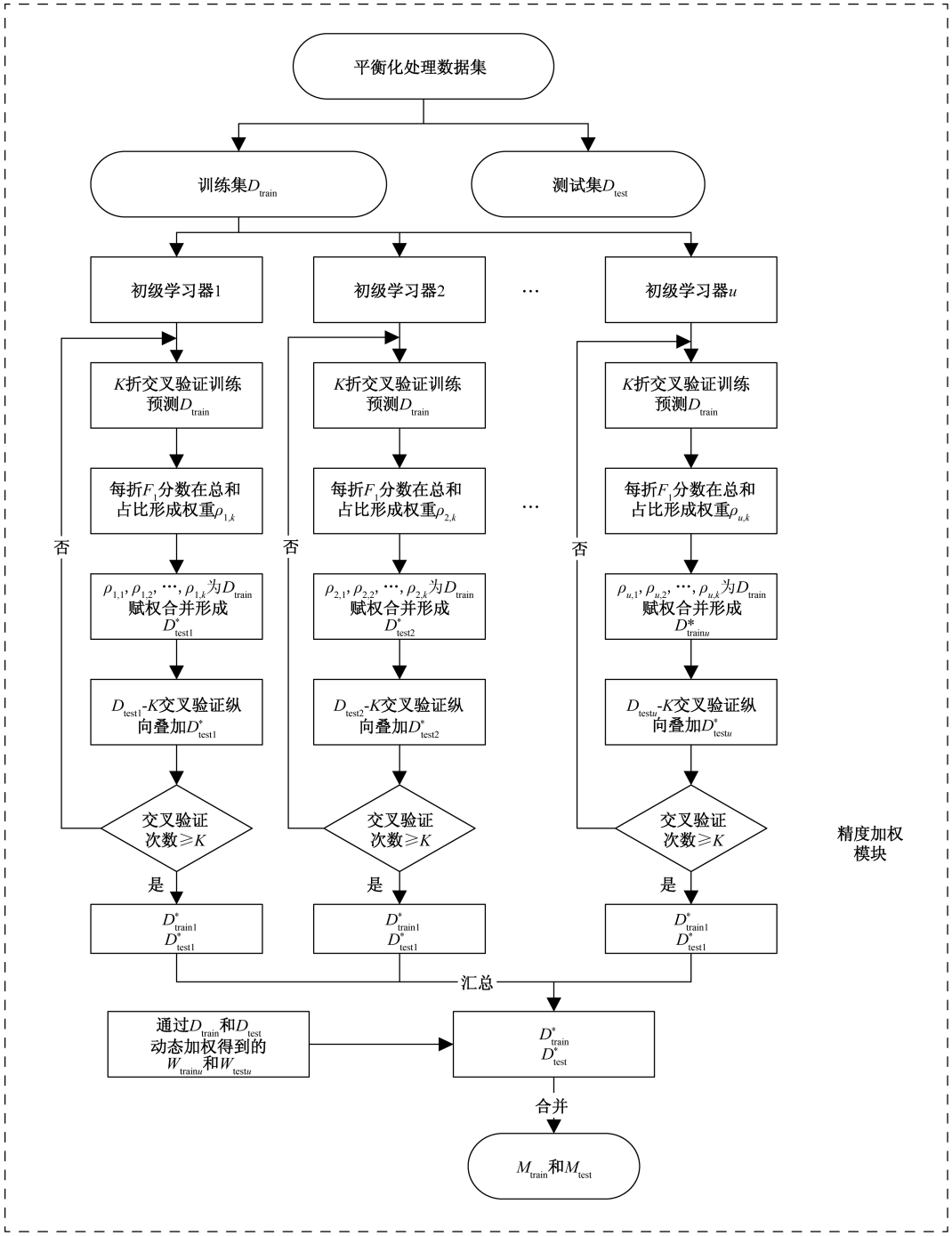


图 2　精度加权下的基学习器流程

Fig. 2　Workflow of base learner under precision weighting

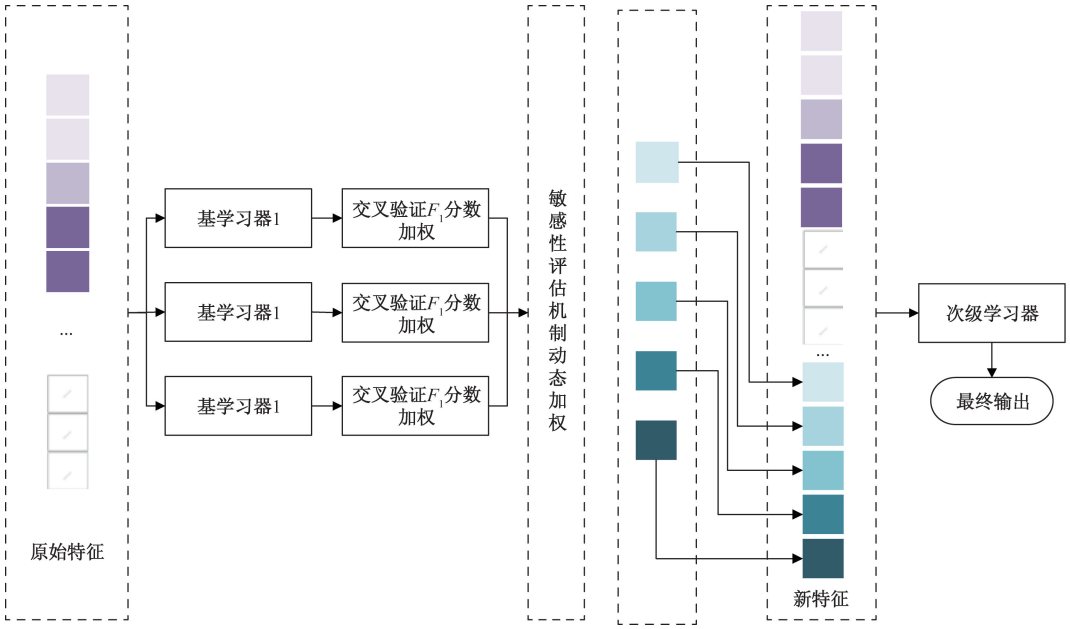


图 3 样本扩充下的集成学习结构图

Fig. 3 Structure diagram of ensemble learning with sample augmentation

原始数据集一起作为元学习器的输入,输出整个集成模型的预测结果,如图 3 所示。

4 实验流程与结果分析

4.1 实验环境与流程

实验在配备 Intel Core i7-9700K 处理器的计算机上进行的,所有代码均采用 Python 3.8 编程语言实现,运行操作系统为 Windows 10 Professional (64 位),确保软件环境的一致性和实验结果的可复现性。

空管危险源描述文本分类集成模型的实验流程可以大致分为 3 个部分,文本预处理、不平衡数据处理、模型组合与加权优化,如图 4 所示。文本预处理包括空管危险源特征词表建立、Jieba 分词及文档向量化,不平衡数据处理包括分析数据类别分布、过采样算法及可视化,模型组合与加权优化包括基学习器和元学习器组合的性能比较、模型训练和性能对比。

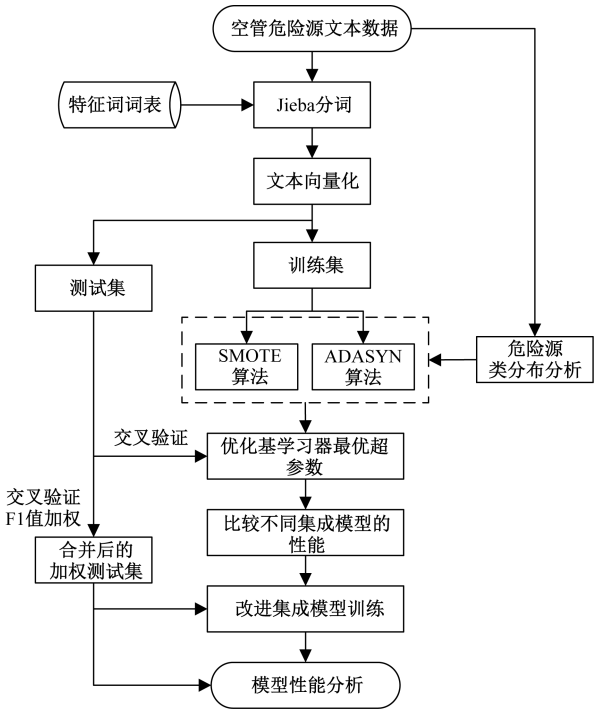


图 4 实验流程

Fig. 4 Experimental procedure

4.2 空管危险源数据预处理

4.2.1 数据来源

本次测试收集某地区空管分局危险源清单 4 个季度 1 107 条文本作为数据源。参照《空中交通管理安全管理体系(SMS)建设指导手册》[19]和《民航安全风险分级管控和隐患排查治理双重预防工作机制管理规定》[20]对危险源分类的指导意见,结合双重防御机制的管理要求,将危险源数据分为 17 个小类,如表 1 所示。

4.2.2 文本特征向量化

空管危险源清单中的文本数据涉及空中交通管理相关的各种潜在危险源,其中包含大量民航领域专业词汇。收集该空管分局不正常事件清单、安全隐患清单和风险通告等相关信息,建立空管危险源特征词词表。使用 Jieba 分词工具对空管危险源描述文本进行分词,分词效果如表2所示。利用

表 1　危险源数据类别

Table 1　Categories of hazard data

风险类别	要素	类别	举例	数量
危险源	管理	空域划设不合理 a_1	区域南移交接点附近冲突加剧	75
		组织规划不完善 a_2	通用航空与民航运输保障存在冲突	72
		航班量大幅增长 a_3	区域航班大流量高密度运行	60
危险源	人员	业务能力不足 b_1	新放单管制员工作流程不熟练	91
		协调配合不足 b_2	主副班监控配合难度大	44
		人员操作失误 b_3	情报岗位人员工作状态不稳定	70
危险源	运行环境	军民航相撞 c_1	军航与民航航空器同航线飞行	60
		跑道侵入 c_2	维护人员误入机场保护区	41
		规定间隔 c_3	指挥错误造成尾流小于规定间隔	60
		外部干扰 c_4	无线电干扰	46
		鸟击意外 c_5	航空器进场航班接地时遭遇鸟击	24
		恶劣天气 c_6	雷雨台风等复杂天气运行风险	27
危险源	设备运行	设施遭破坏 d_1	机房内通信电缆被老鼠啃咬损坏	49
		软硬件设备异常 d_2	流量系统设备运行不稳定	165
		外场设备异常 d_3	机场多点定位系统基站设备故障	42
安全隐患		违规违章 e_1	施工吊装机未严格按照程序执行	98
		风险控制措施弱化 e_2	部分 UPS 系统蓄电池老化	83

表 2　分词效果展示

Table 2　Display of word segmentation results

原始文本数据	分词结果
军航活动期间航空器绕飞雷雨,临场决策难度较大,如果处置不当可能造成飞行冲突	军航活动、期间、航空器、绕飞、雷雨、临场、决策、难度、较大、如果、处置、不当、可能、造成、飞行冲突

Jieba 分词结果构建危险源语料库,作为 TF-IDF 算法的输入数据。TF-IDF 算法统计每个词条在语料库中的出现频率及包含该词条的文档数量,然后根据 TF-IDF 模型计算 F_T(文档词频)和 F_{ID}(逆文档频率)的乘积,生成单个文档特征向量 d_j,并将所有文档的特征向量拼接成文档特征矩阵 $\boldsymbol{Z}$。

4.3　不均衡数据处理

空管危险源清单的各种类分布极不平衡,如图 5所示,“软硬件设备异常”d_2小类比“恶劣天气”

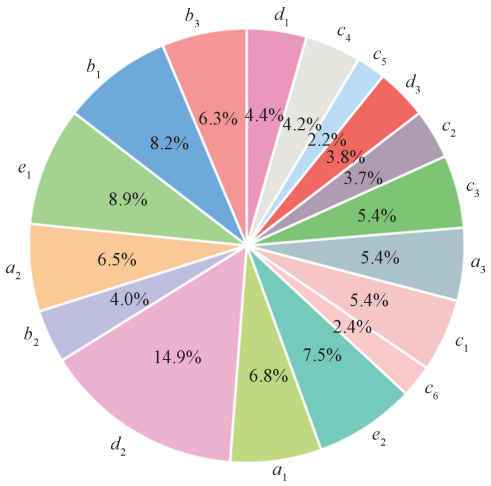


图 5　危险源类别分布

Fig. 5　Distribution of hazard categories

c_2小类的出现频率高出约 8 倍,同时还明显多于“鸟击意外”等其他类别。这种显著的类别不平衡现象会削弱分类算法性能。近年来,平衡化的过采样技术已被广泛应用于多个领域[21]以提高分类的准确性。为解决此问题,采用 SMOTE 和 ADASYN 算法来分别增强数据集中的少数类样本,目标是实现各类别样本量的平衡,趋近 1∶1。

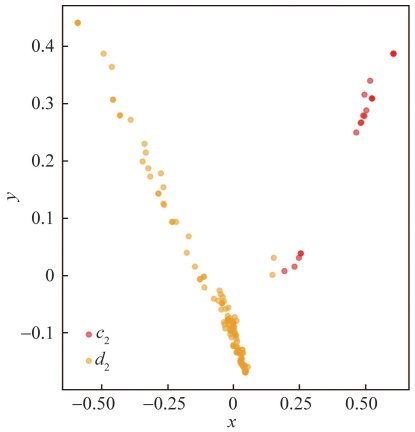
原始数据集按照 4∶1的比例被分割为训练集和测试集,以 d_2 与 c_2 训练集为例,通过 SMOTE 和 ADASYN 算法进行数据增强,平衡 d_2 与 c_2 类数据数量,并采用奇异值分解法(singular value decomposition,SVD)降维,使处理结果可视化,如图 6 所示。

将 d_2、c_2的数据降维映射到二维平面如图 6 所示,原始分布中 d_2 类的数据量远多于 c_2 类,SMOTE 算法处理后,边界区域内 c_2 类随机生成新的样本,数据量较原始数据明显增多;ADASYN 算法处理后,边界区域内 c_2 数据量增多的同时,分布更加均匀,帮助模型更好地学习少数类别的特征分布,增强泛化能力,红色区域外的 c_2 数据量与原始数据中的 c_2 数据量基本一致。

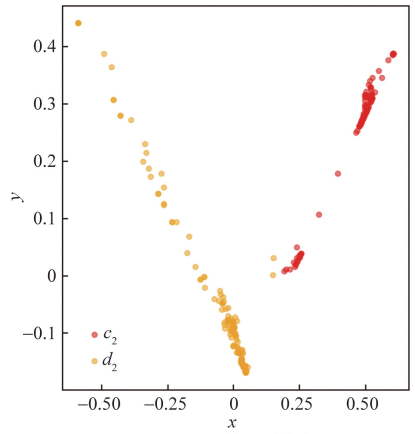
4.4　改进集成模型的训练

4.4.1　超参数选择

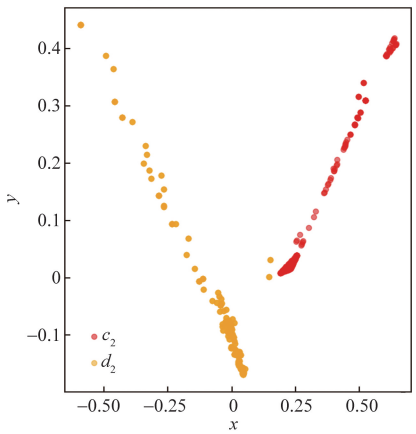
改进 Stacking 集成模型(improve stacking model,ISM)由 RF、SVM、XGBoost 和 LightGBM 4 种学习器组成,这 4 种单一学习器的分类性能会直接影响集成模型的整体性能。因此,通过贝叶斯优化算法优化单一学习器的超参数,从而提高集成模型的分类能力。SVM 的关键超参数为核函数(Kernel)的类型和惩罚系数 C 的数值范围;RF 的超参数为决策树最大深度(max_depth)和决策树数量(n _estimators);XGBoost 的超参数为决策树数量(n _esti-



(a) 原始分布



(b) SMOTE处理后分布



(c) ADASYN处理后分布

图 6 过采样算法可视化

Fig. 6 Visualization of oversampling algorithms

mators)、决策树最大深度(max_depth)、学习器贡献缩减系数(learning_rate)、训练每棵树的样本比例(subsample)和正则项(penalty)的选择;LightGBM的超参数为单一决策树上的叶子数(num_leaves)、决策树最大深度(max_depth)、学习器贡献缩减系数(learning_rate)和特征抽样的比例(feature_fraction)。

贝叶斯优化后的各模型主要参数的调节范围和最优值如表 3 所示。

表 3 集成模型超参数设置

Table 3 Hyperparameter settings for the ensemble model

模型	参数	调节范围	最优值
RF	决策树数量	np. linspace(10, 500, 20)	253
	决策树最大深度	range(3,20,1)	16
SVM	核函数	('linear','pole','rbf')	rbf
	惩罚系数	range(0.1,2.0,0.3)	1.188 7
XGBoost	决策树数量	np. linspace(10, 300, 20)	258
	决策树最大深度	np. linspace(1, 10, 10)	7
	贡献缩减系数	np. linspace(e^{-3}, 0.3,10)	0.081 6
	每棵树样本比例	np. linspace(0.6, 1.0, 5)	0.767 8
	正则项	(L_1,L_2)	L_2
LightGBM	决策树的叶子数	range(10, 80, 5)	65
	特征抽样的比例	np. linspace(0.5, 0.9, 5)	0.860 7
	决策树最大深度	range(3, 10, 1)	111
	贡献缩减系数	np. linspace(e^{-3}, 0.3,10)	0.058 1

注:range(a,b,c) 表示计数从 a 开始,计数到 b 结束,c 表示每次跳跃的间距;np. linspace(a,b,c)表示产生一组由 a ~ b 的等差数列,这组数的个数即为 c;np. logspace(a,b,c)表示产生一组由 10 的 a 次方到 10 的 b 次方的等差数列,这组数的个数即为 c。

4.4.2 Stacking 集成模型的融合

本实验主要研究内容是对超参数优化后的 RF、SVM、XGBoost 和 LightGBM 4 个模型进行融合,并观察精度加权后的模型是否有更好的预测效果。首先采用交叉验证计算 4 个学习器单一训练效果,取评测指标的平均值作为分类性能的结果,结果如表 4 所示。使用不同的次级学习器构造出来的 Stacking 集成模型性能差距较大,因此将这 4 种模型每次选取其中一种作为 Stacking 集成框架第二层中的次级学习器,其余 3 种作为第一层的基学习器,计算不同次级分类器下融合模型的准确率、召回率和 F_1 分数,同样采用 5 折交叉验证、最优超参数,并取每次评测指标的平均值为分类结果,整理结果如表 5 所示。选取性能衡量更全面的 F_1 分数,将 8 个学习器训练所有危险源类别的 F_1 分数结果用热力图表示,如图 7 所示。

表 4 单一学习器分类模型对比

Table 4 Comparison of classification models of single learners

单一学习器	精确率	召回率	F_1 分数
RF	0.810 6	0.816 3	0.813 4
SVM	0.774 5	0.734 2	0.753 8
XGBoost	0.829 1	0.827 2	0.826 3
LightGBM	0.846 3	0.847 0	0.843 4

表 5 集成模型分类性能对比

Table 5 Comparison of classification performance of ensemble models

次级学习器	精确率	召回率	F_1 分数
RF	0.830 4	0.826 5	0.828 4
SVM	0.817 3	0.802 5	0.809 8
XGBoost	0.842 7	0.841 6	0.842 4
LightGBM	0.860 7	0.857 1	0.858 9

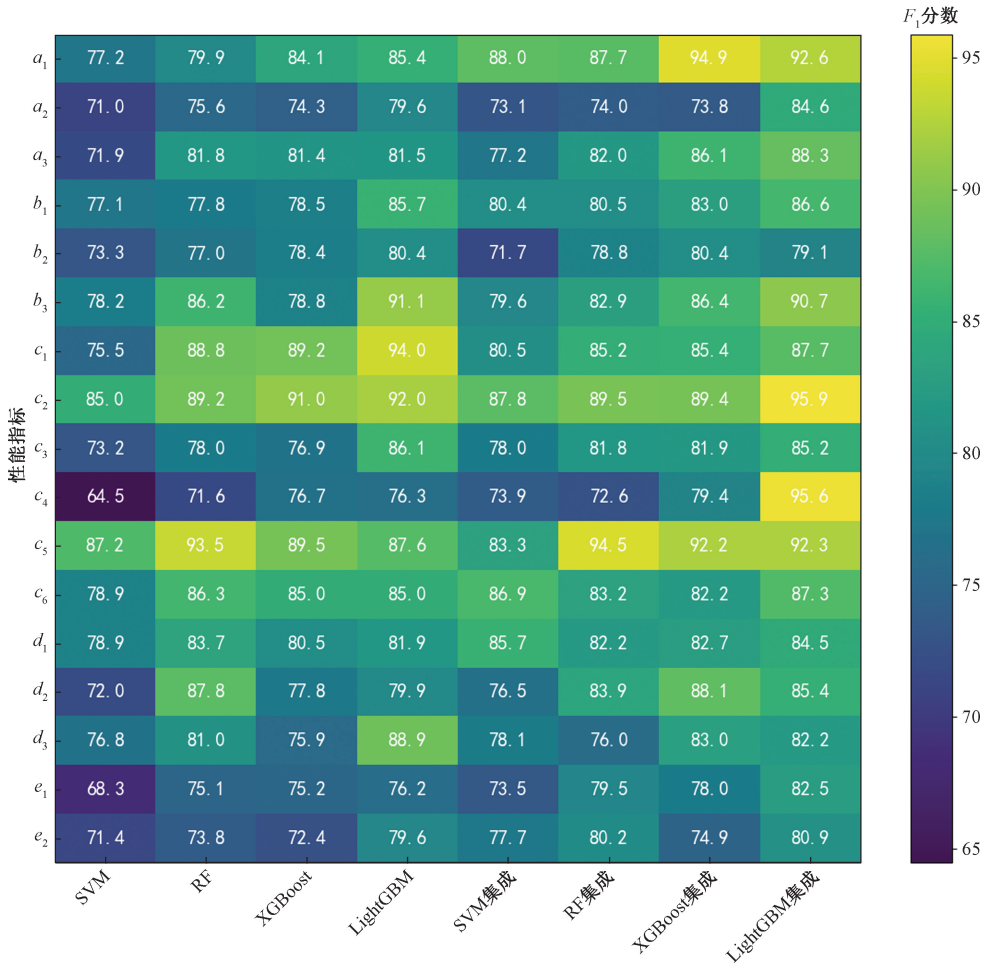


图 7　各学习器 F_1 分数结果对比热力图

Fig. 7　Heatmap comparison of F_1 scores across learners

通过表 5 和表 6 综合评估了单一学习器与集成模型的性能后,观察到 Stacking 集成模型相对于单一学习器在精确率、召回率和 F_1 分数上都展现了更优的表现。特别是以 LightGBM 作为次级学习器时,集成模型在性能指标上领先于其他单一学习器和次级学习器,这一结果强调了 LightGBM 的优异性能以及其在 Stacking 集成模型中作为次级学习器的有效性。

图 7 以热力图形式直观揭示不同学习器在 F_1 分数方面的表现差异。热力图中浅色调的格子代表更高的 F_1 分数,反映了模型在相应类别上的优异分类性能。对比不同学习器的热力图模式,发现集成模型的颜色普遍比单一学习器更浅,表明集成模型在大部分类别上都取得了更高的 F_1 分数。特别是以 LightGBM 作为次级学习器的集成模型,在整个热力图中呈现了最浅的颜色块,说明其在所有类别上的性能均有显著提升。

进一步地,通过审视图 8 中的混淆矩阵,以 LightGBM 作为次级学习器的 Stacking 模型展示出高度的准确性。尽管某些类别存在轻微的误分类,如类别 10 和类别 11,这种情况可以通过后续的特征工程和模型调优来缓解。总体来说,这个模型表现出其在多类别危险源分类任务上的泛化能力和鲁棒性。在此基础上,提出对该 Stacking 集成模型进行精度加权的策略,以进一步优化模型性能。这种方法将基于各类别的 F_1 分数,调整模型权重,旨在提升模型在各个单独类别上的识别能力,特别是对那些易于被混淆的类别。

4.5　空管危险源文本数据分类测试

实验采用改进 Stacking 集成模型,分别在原始训练集、SMOTE 算法和 ADASYN 算法处理后的训练集中进行模型训练,再使用测试集进行性能评测,并以传统 Stacking 集成模型作为对比模型,验证改进 Stacking 集成模型的分类性能。分类模型的评测指标采用精确率 P、召回率 R 和 F_1 分数 F_1,其表达式分别为

$$P = \frac{1}{n}\sum_i \frac{w_i}{z_i} \tag{12}$$

$$R = \frac{1}{n}\sum_i \frac{w_i}{t_i} \tag{13}$$

$$F_1 = \frac{2PR}{P+R} \tag{14}$$

式中:n 为数据类别总数;i 为数据类别;w_i 为模型正确判断 i 类数据的数量;z_i 为模型判断为 i 类数据的数量;t_i 为数据集中 i 类数据的总量。

为避免测试结果的偶然性,所有分类器模型的训练采用 5 折交叉验证,并取每次评测指标的平均值作为分类性能的结果。分类模型性能如表 6 所示。

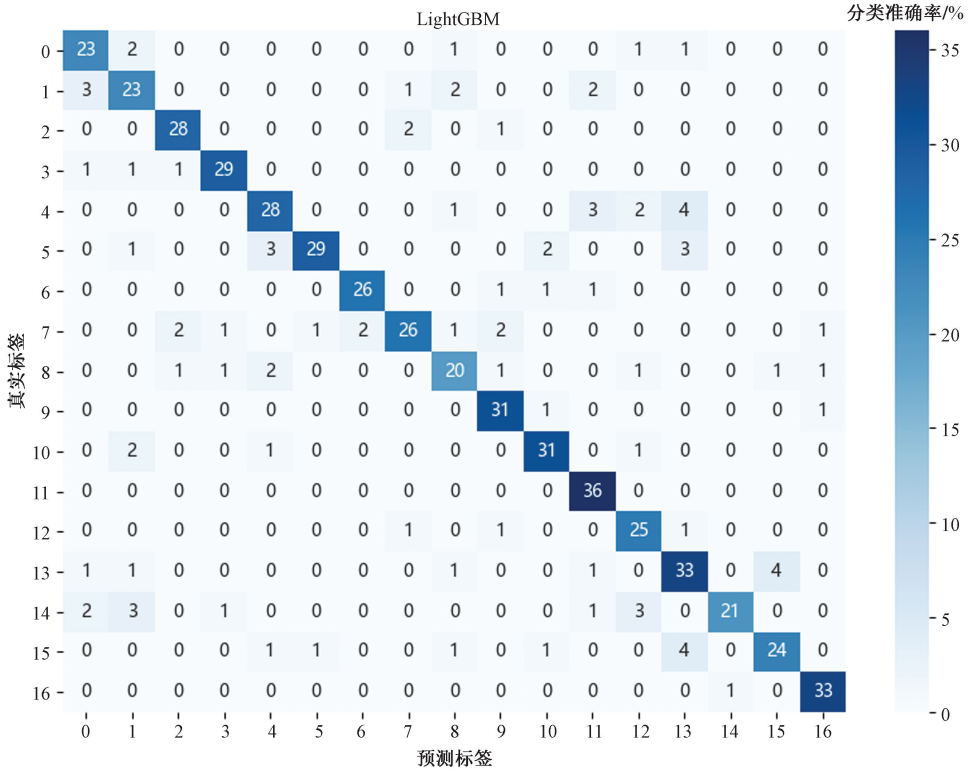


图 8 Stacking 模型的混淆矩阵

Fig. 8 Confusion matrix of the Stacking model

表 6 各模型分类性能

Table 6 Classification performance of each model

指标	精确率/%				召回率/%				F_1 分数/%			
类别模型	SM	ISM	SMOTE + ISM	ADASYN + ISM	SM	ISM	SMOTE + ISM	ADASY + ISM	SM	ISM	SMOTE + ISM	ADASYN + ISM
a_1	88.3	90.2	88.9	95.2	75.8	85.4	87.0	86.9	81.6	87.7	87.9	90.9
a_2	74.2	82.7	94.1	83.5	73.0	83.7	86.9	91.0	73.6	83.2	90.4	87.1
a_3	80.0	86.8	96.7	98.0	76.4	83.1	94.4	92.7	78.2	84.9	95.5	95.3
b_1	79.3	85.2	92.5	93.1	82.7	86.3	86.2	85.7	80.9	85.7	89.2	89.2
b_2	74.1	85.1	86.3	84.6	72.2	82.9	84.8	82.6	73.1	83.9	85.5	83.6
b_3	77.6	83.1	84.6	91.7	83.0	86.9	89.1	93.5	80.2	85.0	86.8	92.6
c_1	85.5	87.4	92.0	91.7	79.3	82.9	89.7	86.9	82.3	85.1	90.8	89.2
c_2	77.8	81.5	90.9	92.3	75.8	80.5	89.3	92.3	76.8	81.0	90.1	92.3
c_3	72.3	80.7	82.2	80.9	77.6	82.0	80.9	85.3	74.9	81.3	81.5	83.0
c_4	66.7	81.4	85.1	87.5	75.0	79.2	86.9	92.3	70.6	80.3	86.0	89.8
c_5	72.8	84.2	94.7	93.8	90.2	90.7	94.8	95.1	80.6	87.3	94.8	94.4
c_6	73.1	80.6	91.6	93.2	87.0	82.9	92.7	90.3	79.4	81.7	92.1	91.7
d_1	86.4	90.4	94.2	95.6	75.3	80.8	92.7	90.8	80.5	85.3	93.4	93.1
d_2	80.1	87.6	85.4	86.9	75.7	86.3	85.5	87.7	77.8	86.9	85.4	87.3
d_3	71.9	75.5	84.7	86.1	80.4	80.6	87.3	89.2	75.9	78.0	85.9	87.6
e_1	77.3	86.6	83.5	85.4	76.7	82.6	80.4	83.9	77.0	84.6	81.9	84.6
e_2	72.0	87.0	86.2	90.6	78.1	88.4	87.7	89.5	74.9	87.7	86.7	90.0
均值	77.0	84.5	89.1	90.0	78.5	83.8	88.1	89.2	77.5	84.1	88.5	89.5

由表 6 可知,数据集未进行过采样处理时,精度加权 Stacking 集成模型相较于传统 Stacking 集成模型,整体分类性能显著提升,平均精确率、召回率和 F_1 分数分别提升 7.5%、5.3% 和 6.6%,由此可见,在训练危险源描述文本时,精度加权集成模型更加有效。SMOTE + 精度加权集成模型相较于未经 SMOTE 处理的精度加权集成模型,精确率、召回率和 F_1 分数分别提升 4.6%、4.3%、4.4%,SMOTE 算法能够提升整体类别的分类性能,尤其对于 c_2、c_6 等少数类分类性能提升十分显著,其中 c_2(占比 3.7%)、c_6(占比 2.4%)的 3 项指标相较于未经 SMOTE 处理的精度加权模型数值分别提升 9.4%、8.8%、9.1% 和 11.0%、9.8%、10.4%,但存在某些多数类性能小范围下降的情况,其中 d_2(占比 14.9%)、e_2(占比 7.5%)的 3 项指标分别下降 2.2%、0.8%、1.5% 和 0.8%、0.7%、1.0%,存在过拟合的情况。ADASYN + 精度加权集成模型与 SMOTE + 精度加权集成模型相比,精确率、召回率和 F_1 分数分别提升 0.9%、1.1%、1.0%,并且多数类分类性能基本不会发生指标下降的情况,综上 ADASYN 算法与精度加权集成模型更适用于处理空管危险源数据。

5　结论

针对空管危险源不平衡数据处理的问题,综合考量了各基学习器的分类特性,提出并验证了一种精度加权的 Stacking 集成模型。得出如下结论。

(1)通过赋予基学习器预测结果以精度加权设计的权重,提出的精度加权的 Stacking 集成模型的效果显著优于单一基础学习器以及传统 Stacking 集成模型。精度加权 Stacking 中的权重分配考虑了每个基学习器的特性,使得次级学习器更准确地识别类别间的分布特征,从而在不平衡数据训练上取得更好的分类性能。实验结果表明,精度加权Stacking 集成模型相较于传统 Stacking 集成模型,平均精确率、召回率和 F_1 分数分别提升 7.5%、5.3% 和 6.6%,所提出的方法在提高分类准确度方面具有明显的效果。

(2)针对民航空管危险源数据规模小、数据不平衡的问题,提出一种引入最优特征子集的数据集扩充方法。该方法通过将原始数据集与基学习器的输出组合,作为元学习器的输入。在不增加现有数据的前提下,有效增强了数据集的多样性和数量。避免了使用生成式模型生成新的数据,从而防止了虚构危险源的出现,确保分类效果清晰准确。

(3)通过精确的文本分类模型,能够更有效地预测和识别空管系统中的潜在风险和危险源,从而为预防措施提供科学依据。准确的分类结果还可优化资源调度,使应急响应更加及时和有效,提升整体空管运营效率。进一步地,精准的文本分类能增强空中交通管制自动化系统的可靠性和安全性,减轻管制员的工作负荷,降低人为错误率。不仅在技术上取得了突破,更在实际应用中对提升空管安全和运行效率具有重要意义。

(4)尽管所使用的数据集在评估危险源文本分类性能方面具有一定的代表性,但它不足以涵盖民航领域文本分类的全貌。因此,未来的研究将考虑引入更多样化的数据集,如民航安全报告,以提高模型的泛化能力。这将使该模型框架有望在民航领域的文本分类中得到广泛应用,进一步提升安全管理和预警的整体效能。

参　考　文　献

[1] Guzanek P, Borucka A. An analysis of factors affecting the number of safety incidents in civil aviation[J]. Safety & Defense, 2021, 7(2): 105-118.

[2] Silvestri S, Islam S, Papastergiou S, et al. A machine learning approach for the NLP-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem[J]. Sensors, 2023, 23: 651.

[3] 张昱, 冯亚寒, 丁千惠. 融合 Word2Vec 词嵌入的多核卷积神经网络音乐歌词多情感分类方法[J]. 科学技术与工程, 2024, 24(20): 8598-8605.
Zhang Yu, Feng Yahan, Ding Qianhui. Multi-emotion classification method for music lyrics based on multi-kernel convolutional neural networks integrated with Word2Vec word embedding[J]. Science Technology and Engineering, 2024, 24(20): 8598-8605.

[4] Kounte M R, Tripathy P K, Bajpai H. Analysis of intelligent machines using deep learning and natural language processing[C]// 4th International Conference on Trends In Electronics and Informatics. Tirunelveli: IEEE, 2020: 956-960.

[5] 刘丹, 王晓兰, 邢胜. 面向不平衡数据分类的最近邻三角区域合成少数类过采样技术[J]. 科学技术与工程, 2018, 18(28): 215-219.
Liu Dan, Wang Xiaolan, Xing Sheng. Nearest neighbor triangular region synthetic minority oversampling technique for imbalanced data classification[J]. Science Technology and Engineering, 2018, 18(28): 215-219.

[6] Tanguy L, Tulechki N, Urieli A, et al. Natural language processing for aviation safety reports: from classification to interactive analysis[J]. Computers in Industry, 2016, 78: 80-95.

[7] Ma Z, Chen Z S. Mining construction accident reports *via* unsupervised NLP and accimap for systemic risk analysis[J]. Automation in Construction, 2024, 161: 105343.

[8] 刘旭, 张艳, 邓少阁, 等. 基于 *K*-means 算法的民航事故结构化分析[J]. 科学技术与工程, 2024, 24(30): 13210-13217.
Liu Xu, Zhang Yan, Deng Shaoge, et al. Structured analysis of civil aviation accidents based on the *K*-means algorithm[J]. Science Technology and Engineering, 2024, 24(30): 13210-13217.

[9] Robinson S D. Temporal topic modeling applied to aviation safety reports: a subject matter expert review[J]. Safety Science, 2019, 116: 275-286.

[10] 王洁宁, 张聪俊, 张钰涵. 民航不安全事件报告危险源识别模型[J]. 安全与环境学报, 2020, 20(1): 186-192.
Wang Jiening, Zhang Congjun, Zhang Yuhan. Hazard source identification model for civil aviation unsafe event reports[J]. Journal of Safety and Environment, 2020, 20(1): 186-192.

[11] 郭九霞. 基于自然语言处理的空管系统危险源文本分类方法研究[J]. 安全与环境学报, 2022, 22(2): 819-825.
Guo Jiuxia. Research on the text classification method for air traffic control system hazard sources based on natural language processing[J]. Journal of Safety and Environment, 2022, 22(2): 819-825.

[12] 巩家铭, 李康妹, 胡俊, 等. Stacking 集成学习应用于近视矫正中的角膜塑形镜临床验配[J]. 东华大学学报, 2024, 41(2): 184-194.
Gong Jiaming, Li Kangmei, Hu Jun, et al. Stacking ensemble learning applied to the clinical fitting of orthokeratology lenses for myopia correction[J]. Journal of Donghua University, 2024, 41(2): 184-194.

[13] Hou Z, Xiong M, Wang H, et al. Civil aviation safety risk intelligent early warning model based on text mining and multi-model fusion[J]. Journal of Aerospace Engineering: Part G, 2023, 237(10): 2402-2427.

[14] He H B. Adaptive synthetic sampling approach for imbalanced learning[C]//Proceedings of the 2008 IEEE International Joint Conference on Neural Networks. Hong Kong: IEEE, 2008: 1322-1328.

[15] Garg R, Oh E, Naidech A, et al. Automating ischemic stroke subtype classification using machine learning and natural language processing[J]. Journal of Stroke and Cerebrovascular Diseases, 2019, 28: 2045-2051.

[16] Yang E, Long Z. Research on the weighting method based on TF-IDF and apriori algorithm[C]//IEEE 6th International Conference on Information Systems and Computer Aided Education. Dalian: IEEE, 2023: 1003-1005.

[17] 潘娇, 李超, 彭文忆, 等. 基于随机森林和支持向量机的云南省土地利用分类[J]. 科学技术与工程, 2024, 24(17): 7043-7051.
Pan Jiao, Li Chao, Peng Wenyi, et al. Land use classification in Yunnan Province based on random forest and support vector machine[J]. Science Technology and Engineering, 2024, 24(17): 7043-7051.

[18] 许惠. 基于 NLP 方法实现文本分类识别[D]. 大连: 大连理工大学, 2022.
Xu Hui. Text classification and recognition based on NLP methods[D]. Dalian: Dalian University of Technology, 2022.

[19] 中国民用航空局. 空中交通管理安全管理体系(SMS)建设指导手册[M]. 3 版. 北京: 中国民用航空局, 2011.
Civil Aviation Administration of China. Air traffic management safety management system(SMS) construction guideline[M]. 3rd ed. Beijing: Civil Aviation Administration of China, 2011.

[20] 中国民用航空局. 民航安全风险分级管控和隐患排查治理双重预防工作机制管理规定[EB/OL]. (2022-08-31)[2024-08-01]. https://www.caac.gov.cn/PHONE/XXGK_17/XXGK/GFXWJ/202209/t20220914_215318.html.
Civil Aviation Administration of China. Regulations on the management of dual prevention mechanism for civil aviation safety risk grading control and hidden danger investigation and management[EB/OL]. (2022-08-31)[2024-08-01]. https://www.caac.gov.cn/PHONE/XXGK_17/XXGK/GFXWJ/202209/t20220914_215318.html.

[21] Xiang Y, Xie Y. Imbalanced data classification method based on ensemble learning[C]//International Conference in Communications, Signal Processing, and Systems. Singapore: Springer, 2018: 18-24.