DOI:10.11737/j.issn.1003-0239.2024.03.007

基于多模型组合的类别不平衡海洋数据质量控制方法

宋巍[1]，张贵庆[1]，谢京容[1]，董明媚[2]，岳心阳[2]，杨扬[2]

（1.上海海洋大学信息学院，上海 201306；2.国家海洋信息中心，天津 300171）

摘 要： 提出一种多模型组合的两层海洋数据质量控制框架，选择了多种常见分类算法作为基学习器对数据质量标签进行初级预测，再经过投票法或堆叠（Stacking）法确定海洋数据质量的标识符；针对类别不平衡问题，结合自适应下采样策略，降低数据的不平衡比率，并结合Focal Loss损失函数，提升模型对难分类样本的识别能力。以来源于国际综合海洋大气数据集的海表温度和气温数据为例进行质量控制验证，结果表明：投票法或堆叠法对极少类的错误样本分类的F1 score（精确率和召回率的加权调和平均值）在海表温度数据上可达到0.980 6和0.981 2，在气温数据上可达到0.998 5和0.998 3。

关键词： 质量控制；海洋气象数据；集成学习；类别不平衡

中图分类号：P731.11 **文献标识码：**A **文章编号：**1003-0239（2024）03-0061-10

0 引言

随着卫星遥感、海洋测绘等信息采集技术的完善及其在海洋领域的广泛应用，海洋数据的种类和数量呈现爆发式增长，海洋探索与研究进入了大数据时代。海洋大数据逐步成为实施海洋强国战略、开发海洋资源、拉动海洋经济、维护海洋权益的重要基础。国际综合海洋大气数据集（International Comprehensive Ocean Atmosphere Data Set，ICOADS）是一个综合性的、全球范围内的海洋和大气数据集。它收集了1662年10月至今的全球海洋气象资料[1]，但其中不乏相当一部分的欠准甚至错误数据，原因包括但不限于环境的异常变化、数据传输过程出错及浮标本身老化腐蚀等因素。提升数据质量的重要性应该得到重视，以促进海洋工程[2]和海洋科学的可持续发展。

海洋数据质量控制（Quality Control，QC）是指采用一定资料处理方法、模型和参数，判断资料质量的可靠性与准确性，并进行质量标识的处理过程[3]。目前，包括中国、美国在内的多个国家的政府组织或科研机构根据各地区实际情况，制定了相应的海洋数据质量控制标准或规范，针对不同数据类型设计不同的质量控制手册和检测方法。基于传统方法的海洋数据质量控制主要包括范围检验、一致性检验、漂浮检测、尖锋检测等测试方法[4-9]。

机器学习的本质是对数据进行分类、识别和预测[10]，而质量控制的本质是对海洋观测数据进行识别和分类。鉴于机器学习与质量控制两者间的共性，可以将机器学习的思想运用到海洋观测数据的质量控制当中。2011年，TIMMS等[11]提出使用模糊逻辑实现数据质量的连续评估，并在澳大利亚霍巴特的Derwent河口监测的温度和电导率传感器的实时平台上测试了该自动数据质量评估框架，结果与领域专家人工标识结果的误差线具有良好的一致性。ZHOU等[12]针对中国东海海底观测系统数据的质量问题，提出了一种结合统计分析和领域专家知识的数据质量检测和修复方法。该方法由3部分组成：通过一组测试先对数据进行预处理，然后利用自回归积分滑动平均模型（Autoregressive Integrated Moving Average Model，ARIMA）进行数据离群

收稿日期： 2023-05-05。
基金项目： 国家重点研发计划项目（2021YFC3101601）；上海市科委部分地方高校能力建设项目（20050501900）。
作者简介： 宋巍（1977-），女，教授，博士，主要从事海洋大数据分析研究。E-mail：wsong@shou.edu.cn

点检测并标注可疑数据点，最后通过数据插值方法填充缺失和可疑数据。2019年，随机森林方法被应用于Argo浮标的温、盐延迟模式质量控制[13]，对每个剖面数据提取下列特征：观测值与气象学平均值之间差值的特征、观测值与气象学平均值之间的气候学标准差数特征、气象学方格中观测数量的特征，并同时研究了网络尺度对检测精度的影响。2021年，刘玉龙等[14]构建了深度学习的多层感知机（Multilayer Perceptron，MLP）和深度神经网络（Deep Neural Network，DNN）分类模型用于快速识别西太平洋海域的海温数据质量，对错误和问题数据的召回率达到87%～92%。同年，MIERUCH等[15]将人工神经网络运用到传统的质量控制算法当中，在地中海大型数据集的测试中，能够正确检测出超过90%的错误温度观测数据。

在大多数海洋环境数据中，通过QC检验的数据量通常至少比未通过QC检验的数据大两个数量级[16]，因此样本不均衡问题是海洋质控中需要考虑的因素。本文将海洋观测数据的质量控制问题定义为对类别极度不平衡数据的二分类问题，基于集成学习思想，提出一种多模型组合的海洋观测数据质量控制方法。选择多种常见分类算法应用于ICOADS的海表温度（Sea Surface Temperature，SST）和气温（Air Temperature，AT）数据集，结合自适应下采样模块，将分类结果进行对比分析，选取效果较好的决策树（Decision Tree，DT）、随机森林（Random Forest，RF）以及LightGBM（Light Gradient Boosting Machine）作为基模型构建二级组合模型。将组合模型在不同子数据集上进行分类和预测，并与相同条件下的基模型进行比较并得出结论。

1 多模型组合海洋数据质量控制方法

在实际应用中，很难构建出各个方面表现都好的单模型，本文提出一种基于多模型组合的质量控制方法。该方法采用两层框架的结构（见图1），第一层经过自适应下采样模块后，由多个异构的基学习器作为质量标签的初级预测模型，得到伪海洋数据质量预测结果；第二层由多组伪海洋数据质量预测结果构成向量，作为投票器（Voting）或元学习器的输入，通过判别后得到最终海洋数据质量标签预测结果，并给出相应的质量控制标识符。

1.1 自适应下采样的基模型训练

自适应下采样的方法是由LIU等[17]于2020年提出，本文将其应用于对海洋数量质量标识基分类模型的训练，以降低数据类别极度不平衡问题的影响。

设由已知海洋观测数据构成的训练数据 $\boldsymbol{X}=\{(\boldsymbol{x}_i,y_i)\,|\,i=1,2,\cdots,m\,|\}$，其中 $\boldsymbol{x}_i\in R^{1\times l}$ 是有l个特征值的观测数据样本向量，包含类别为1（即正确数据；大样本量）的子集为 $\boldsymbol{N}$，类别为0（即错误数据；小样本量）的子集为 $\boldsymbol{P}$。

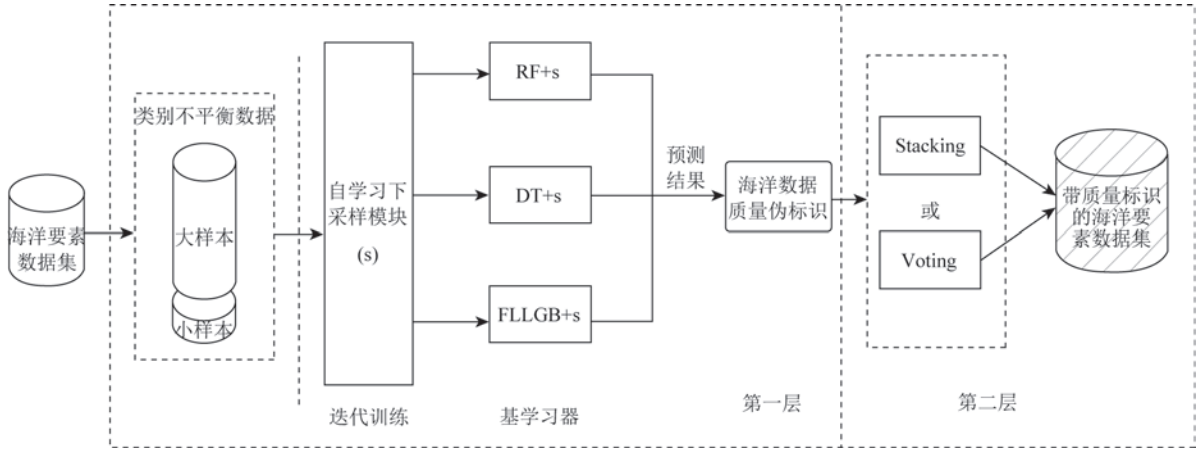

图1 多模型组合质量控制方法

Fig.1 Multi-model combination quality control method

定义分类硬度函数H为分类器的样本分类误差函数,本文用交叉熵计算,假设F是一个训练好的分类器,则样本(x,y)相对于F的分类硬度可表示为函数$H(x,y,F)$。

对于任一分类器f,初始分类器f_0采用随机采样方式得到的数据集进行训练,然后自适应数据下采样的方式对分类器进行迭代训练,迭代次数为n。第i次迭代的具体步骤为:

①将大样本的子集$\boldsymbol{N}$分为k个数据块,第l个子块表示为B_l,定义为:

$$B_l=\left\{(x,y)\left|\frac{l-1}{k}\leqslant H(x,y,F)<\frac{l}{k}\right.\right\} \tag{1}$$

②计算第l个子块的平均贡献硬度h_l:

$$h_l=\sum_{s\in B_l}H(x_s,y_s,F_i)/|B_l|,\ \forall l=1,2,\cdots,k \tag{2}$$

③更新自适应系数α:

$$\alpha=\tan\left(\frac{i\pi}{2n}\right) \tag{3}$$

④计算第l个子块的非归一化采样权重p_l:

$$p_l=\frac{1}{h_l+\alpha},\ \forall l=1,2,\cdots,k \tag{4}$$

⑤从第l个子块下采样的样本量为:

$$\frac{p_l}{\sum_m p_m}\cdot|\boldsymbol{P}| \tag{5}$$

⑥使用下采样得到的数据子集对f_i进行训练。更新分类器$F_i(x)$的计算公式为:

$$F_i(x)=\frac{1}{i}\sum_{j=0}^{i-1}f_j(x) \tag{6}$$

重复以上步骤n次,输出最终训练好的分类器F。

1.2　基学习器

1.2.1　决策树

决策树[18](DT)是一种基于树结构形式来进行决策的有监督学习算法。它将数据集分成多个子集,每个子集对应一个决策树节点,最终通过节点的判定确定数据集的类别。在分类时,算法根据DT的结构和节点属性,将新的数据分配到相应的子节点,最终得到新数据的类别。

1.2.2　随机森林

随机森林(RF)算法由多个DT组成,主要用于分类问题[19]。不同DT之间没有关联。每个DT都是通过随机选取一部分特征和训练样本来训练的,这样可以减少DT之间的相关性,提高分类的多样性。

设D_i为不同的决策树,在进行海洋数据的质量控制任务时,当有新的观测数据样本X进入时,森林内的每一棵D_i会分别进行判断和分类,得到自己的分类结果,通过投票的方式得到RF最终的分类结果。

1.2.3　LightBGM

LightGBM[20]是一个实现梯度提升回归树(Gradient Boosted Decision Trees,GBDT)算法的框架,其使用了基于直方图算法的加速训练过程,减少了内存消耗。本文使用了3个针对数据不平衡的损失函数与之相结合。

Focal Loss[21](FL)是一种针对类别不平衡问题的损失函数,其在交叉熵损失函数的基础上引入了类别权重因子α和系数γ,用来调整类别权重以此提升模型的分类准确率。计算公式为:

$$\mathrm{Loss}_{\mathrm{FL}}=\begin{cases}-\alpha(1-\hat{y})^{\gamma}\ln\hat{y}, & y=1\\ -(1-\alpha)\hat{y}^{\gamma}\ln(1-\hat{y}), & y=0\end{cases} \tag{7}$$

式中:$\hat{y}$表示预测结果;类别权重因子$\alpha\in(0,1)$用来调节不同类别的样本的权重大小,可以通过增加少数类样本权重来平衡正负样本的重要性;系数$\gamma(\gamma>0)$则用来减小易分样本的权重。对于正类样本$y=1$,预测结果越接近1表示样本越容易区分,同时调节因子$(1-\hat{y})^{\gamma}$的值越小,从而降低损失函数值,使算法更加关注难以区分的样本。LGB可以和FL损失函数结合,通过对系数α和γ的调节,使模型更关心负样本,进一步提高分类模型的鲁棒性。

1.3　组合方法

常用的模型组合方法包括:①投票法(Voting),即对多个基学习器的预测结果进行投票,按照少数服从多数的原则分为软投票(根据预测的概率结果加权投票)和硬投票(直接根据预测类别投票);②平均法,即对多个学习器的预测结果进行平均,包括算术平均、几何平均和加权平均等;③堆叠法[22](Stacking),是将多个基学习器的分类结果组合作为新的输入,放入一个元学习器进行学习,经过元学习器的分类结果被视为最终输出结果。由于平均法容易忽略不同模型之间的差异性,无法充分利

用他们之间的互补性，因此本文暂不考虑此组合方法。

Voting将多个模型的预测结果综合考虑，减少由于个别模型预测错误而导致的整体预测结果偏差，增加了模型的多样性，提高了预测准确率。本文使用软投票（Soft Voting），其考虑了每个基学习器模型的预测概率或置信度，相比于多数投票法可以更加准确地组合预测结果。假定已有训练好的T个基模型，对于样本x_i，第t个模型预测概率为$p_{it}^{j} \in [0,1]$，j表示质量标识的类别，则软投票策略最终的预测结果为加权概率最大的类别。计算公式为：

$$\hat{y}_i = \arg\max_j \sum_{t=1}^{T} w_k p_{it}^{j} | [y_t = j] \tag{8}$$

式中：w_t表示第t个模型的权重，满足$w_t \geqslant 0$，$\sum_{t=1}^{T} w_t = 1$。本文中T个基模型的预测性能相差不大，因此w_t的权重设置相同。

Stacking倾向于减小偏置，能很好地缓解弱学习器的高偏置问题，因此在基于多模型组合的海洋数据质量控制方法中可以作为融合集成方法，元学习器选择结构简单的逻辑回归模型防止结果过拟合。设已有训练好的T个基学习器$\mathcal{L}_1, \mathcal{L}_2, \cdots, \mathcal{L}_T$，对于任一质量标识未知的数据样本$x$，Stacking方法预测其质量标识的输出结果为$L(\mathcal{L}_1(x), \mathcal{L}_2(x), \cdots, \mathcal{L}_T(x))$。

2 实验结果

2.1 数据集构建

海表温度和气温对海洋学研究具有重要影响，当其发生变化时，会影响海洋的物理和化学性质，从而影响海洋工程的稳定性和可靠性。通过研究气温和海表温度的变化，可以预测未来气候的趋势和变化。因此，本文对SST和AT数据进行质量控制，以验证方法的有效性。使用的数据来源于ICOADS官网（下载地址：https://icoads.noaa.gov/），时间跨度为2020年3月—2021年12月。数据的质量标签由专业分析员根据质量控制标准结合自动质控结果进行人工审查后给出。

训练集和测试集的具体构建过程如下：

首先，考虑时间及地理位置对海气数据的重要影响，将实验数据的特征分解为6个维度：月、日、小时、纬度、经度及海表温度（或气温）。

其次，针对数据集中的质量类别严重不平衡问题，通过欠采样方法构建训练集。对于数据类别不平衡问题，大量研究主要集中在数据预处理层面以及分类器层面。数据预处理层面解决该问题主要采用欠采样和过采样方法，欠采样的是通过减少不平衡数据中多数类的样本量来控制类间数据量的比例，过采样则与之相反，主要通过增加不平衡数据中少数类的样本量来达到平衡数据分布的目的。考虑到数据集整体基数较大且尽可能使用原始数据，而过采样需要添加一些人工数据可能会对结果产生干扰，因此本文选择使用随机欠采样的方法来降低训练集样本的平衡分布。同时，考虑到季节对海洋要素数据也会造成一定的影响，训练集从2020年3月、6月、9月以及12月的SST和AT中抽取，分别代表春、夏、秋、冬4个季节。为了消除随机因素，从以上4个月的原始数据集中进行3次随机抽取，且限制正确和错误样本量的比例为20∶1，整合为训练集STrain和ATrain，每个训练集的总样本量都为630 000条。

最后，为了保持真实海洋要素数据的不平衡性，测试集没有对样本的不均衡进行处理，分别在2021年全年数据内抽取50 000条数据，构成Test1—Test4。这些数据集的平衡比例都是不确定的，数据不平衡率（Imbalance Ratio，IR）的定义为：

$$\mathrm{IR} = \frac{\min(count(\boldsymbol{X}))}{\max(count(\boldsymbol{X}))} \tag{9}$$

式中：$count(\boldsymbol{X})$为总样本$\boldsymbol{X}$中不同类别对应的样本数量；IR为少数类样本量与多数类样本量的比值。测试样本集的不平衡率见表1。

2.2 评价指标

在类别不平衡的分类问题中，选择合适的评价指标对于分类效果的判定尤为重要。分类精度常用准确率（Accuracy）来评价，然而，在具体的海洋数据质量控制方面，数据类别严重不平衡，决策边界容易偏向多数类。此时模型的准确率并不适合不平衡分类问题。

本文使用精确率（Precision）、召回率（Recall）以及F1 score独立应用于正（正确）、负（错误）标签上。各评价指标的计算公式为：

表 1　测试数据集中不同类别质量标识数量

Tab.1　The number of different category quality identifiers in the Test datasets

类别	Test1		Test2		Test3		Test4	
	SST	AT	SST	AT	SST	AT	SST	AT
1-正确	48 745条	497 77条	49 663条	49 137条	49 591条	49 635条	49 382条	48 977条
0-错误	1 255条	523条	337条	863条	409条	365条	618条	1 023条
IR	0.025 1	0.010 5	0.006 7	0.017 5	0.008 2	0.007 3	0.012 5	0.020 8

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

式中(以正样本为例):TP表示被预测为正类且实际为正类的样本数量;FP表示被预测为正类但实际为负类的样本数量;FN表示被预测为负类但实际为正类的样本数量;TN表示被预测为负类且实际为负类的样本数量。而负样本的精确率是指分类器不将负样本标记为正样本的能力,召回率是指分类器查找所有负样本的能力,与Precision相比,应该更加关注Recall这个指标,它反映了错误样本被漏检的情况,值越大说明漏检率越低。F1 score则是精确率和召回率两个指标的综合体现,更能反应模型分类的整体情况。

2.3　实验结果分析与讨论

本文首先以SST数据为例,在其训练集和测试集上进行了基学习器的筛选、自适应下采样模块的评估以及加上损失函数的效果对比;然后,在SST和AT数据上进行质量控制方法的性能检验,并对结果进行分析和讨论。

2.3.1　基学习器筛选

为了使集成方法比任何构成它的单独的方法更准确,基学习器必须尽可能准确和多样。因此,本文对多种类型的分类器进行实验和选择。尽管对数据集做了处理,但仍然存在数据集的不平衡问题,即正类样本较多,考虑到实际情况,正样本的分类精度往往可以达到99%以上,数据量更少的负样本的预测结果是更需要关注的。

为了避免不同数据集和不同基学习器之间分类效果差异性的影响,首先使用单个模型在训练集上进行训练,在测试集Test1上测试并进行多次重复实验,比较多次实验结果,负样本5次的实验结果均值见表2。

表2　不同基分类器在Test1上的负样本分类结果

Tab.2　Classification results of different base classifiers on Test1 for negative samples

模型	Precision	Recall	F1 score
SVM	0.887 0	*0.465 9*	0.610 9
Bayes	*0.103 2*	0.611 3	*0.176 5*
DT	0.957 8	**0.946 3**	0.952 0
RF	**0.996 2**	0.926 7	0.960 2
MLP	0.752 3	0.777 6	0.767 4
KNN	0.681 9	0.884 3	0.770 0
LGB	0.981 7	0.942 3	**0.961 6**

注:粗体表示最高值;斜体表示最低值。

从表2中可以看出,RF、DT和LGB分别在3个评价指标上取得了最好的效果。支持向量机(Support Vector Machine,SVM)、K近邻(k-Nearest Neighbor,KNN)、MLP和朴素贝叶斯算法(Bayes)的分类效果较差,几项评价指标的结果都相对较低,并且SVM和MLP模型训练时间较长。

根据上述分析结果,在多模型组合的海洋数据质量控制方法中,第一级基分类器不再使用SVM、KNN、MLP和朴素贝叶斯这4种分类器。DT可以对大样本数据进行预测,且有不错的分类效果。RF和LGB具有严谨的数学推论,在很多领域都得到了使用。后续将使用RF、DT以及LGB作为第一层的学习器。

2.3.2　自适应下采样模块性能评估

自适应下采样模块已被证明在公开的欺诈检测数据集上取得了较好的效果,它适用于绝大多数的传统分类器。为了验证该模块在海洋要素数据

集上的可靠性，本文分别使用DT、RF、MLP、KNN和LGB这5种算法，并添加自适应下采样模块，在4个测试集上分别进行分类实验。由于本文更关注负样本，因此，评价指标选取负样本的召回率。实验结果见图2。

由图2可见，与原始的分类算法相比，添加了自适应子模块后，不同测试集上不同模型针对负样本的分类准确率都有提升，平均可以提高1个百分点左右，其中，MLP对该模块较为敏感，这可能是由于它分类精度较低，自适应下采样模块对其提升较大。由此可见，该模块适用于不平衡海洋要素数据集，为本文的后续框架提供了可行性保障。

2.3.3 不同设计的LGB模型的实验对比

本文还使用了LGB作为基础模型，并结合不同的损失函数，包括：FL、在梯度级别上考虑解决类别不平衡问题的EL（Equalization Loss）[23]、修正LGB模型中传统的交叉熵损失函数为不同形式的加权损失函数WL（Weighted Loss）以及自适应下采样模块（+s），比较不同模型在负类准确率方面的表现。研究评估了不同的损失函数在该模型的二分类任务中的性能表现，其中，WL函数对负样本的惩罚是正样本的5倍。在训练集上训练，并在Test1上测试，近5次实验结果的均值见图3。

由图3可知，仅采用LGB模型，其在负类准确

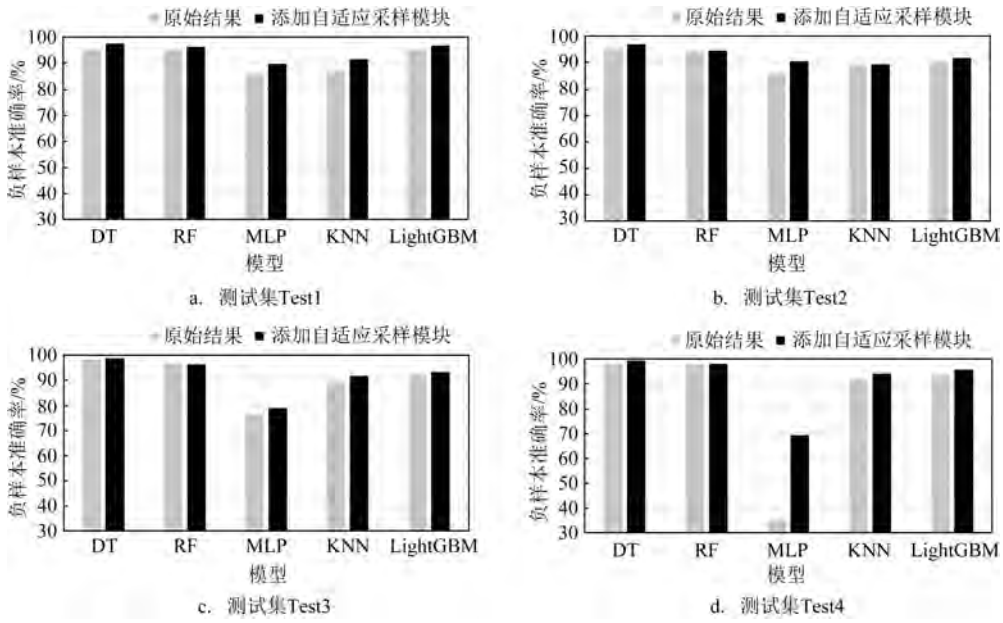

图2 不同模型在不同测试集上添加自适应下采样模块的结果

Fig.2 Results of adding adaptive down sampling modules on different Test sets for different models

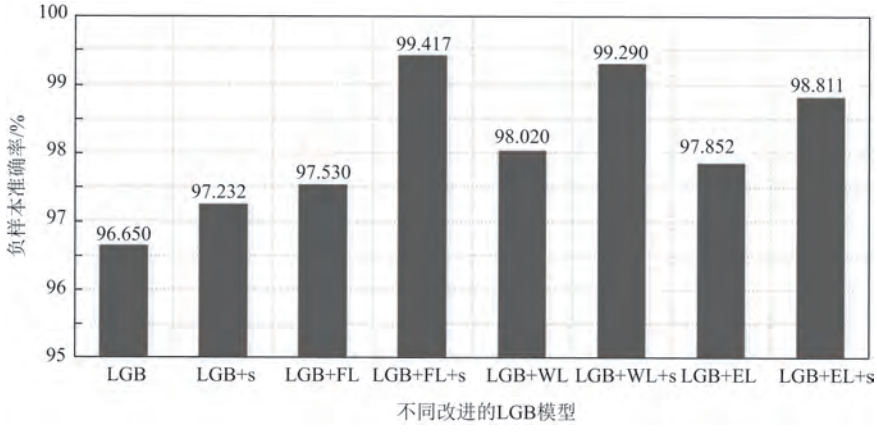

图3 不同改进LGB模型在Test1上的负样本准确率

Fig.3 Negative sample accuracy on Test1 for different variants of LGB models

率上的表现为96.650%;加入自适应采样模块后,准确率提高至97.232%;采用FL作为损失函数时,准确率又有了一定的提升,为97.530%;再加上自适应采样模块后,准确率有了显著提升,提高至99.417%。采用WL和EL作为损失函数同样取得了类似很好的效果,依次添加自适应采样模块后,模型的泛化性都得到了增强,其中,使用FL和自适应下采样模块结合的模型得到的负样本的准确率最高,为99.417%。

总体来看,添加损失函数和自适应采样模块都可以提高负样本的准确率,其中,相较于EL,采用FL和WL结合自适应采样模块可以取得更好的性能表现。在本文后续的多模型组合方法中,将使用添加FL的LGB模型(FLLGB)加上自适应采样模块作为模型组合的基分类器。

2.3.4　组合方式性能对比

为进一步分析组合模型Voting和Stacking在召回率上的表现,本文还以SST数据为例对比了不同模型在4个季节Test1—Test4的结果。由图4可见,RF和DT模型在4个数据集的"1"类样本上的召回率较高,但其在"0"类样本上的表现却不尽人意,而FLLGB模型则过度关注负样本,导致其正样本上的检测性能大大降低。相比之下,使用本文框架的融合模型Voting和Stacking的表现更为优秀,在几个不同季节测试集正、负样本的召回率上,两者虽然不一定达到最高精度,但综合考虑两个类别的召回率,其结果趋于稳定。对比Voting和Stacking的结果,Voting的表现更加出色,且计算复杂度更低。

2.3.5　质量控制结果

根据上述实验结果,本文选取RF、DT、FLLGB和自适应下采样模块的结合作为第一层的基学习器,通过融合多个基学习器质控结果,实现对海洋气象数据的质量控制。

为了进一步比较几种基学习器算法和融合算法的质量控制效果,本文在训练集上进行训练,将两种海洋气象要素4个测试集进行整合得到总测试集,在总测试集上进行测试,取5次实验结果的平均值。由于本文从样本不平衡问题出发,设计了采样模块和损失函数,在增大负样本的惩罚权重时,不可避免地会使模型对正样本的学习变差,因此,最终质量控制选择的评价指标为正、负类样本的精确率、召回率、F1 score以及准确率。这其中,F1 score综合考虑了精确率以及召回率,最能反映模型在不平衡海洋要素数据集上的分类表现。不同模型在SST和AT上的质量控制结果见表3和表4。

表3展示了不同分类算法在SST总测试集上各类别的分类情况。实验结果表明,所有模型在测试集上的4个指标均超过了0.95。这其中,正样本"1"类的指标较高,均在0.999以上,这是由于样本仍然不平衡,模型更倾向于学习正样本,因此,总样本的

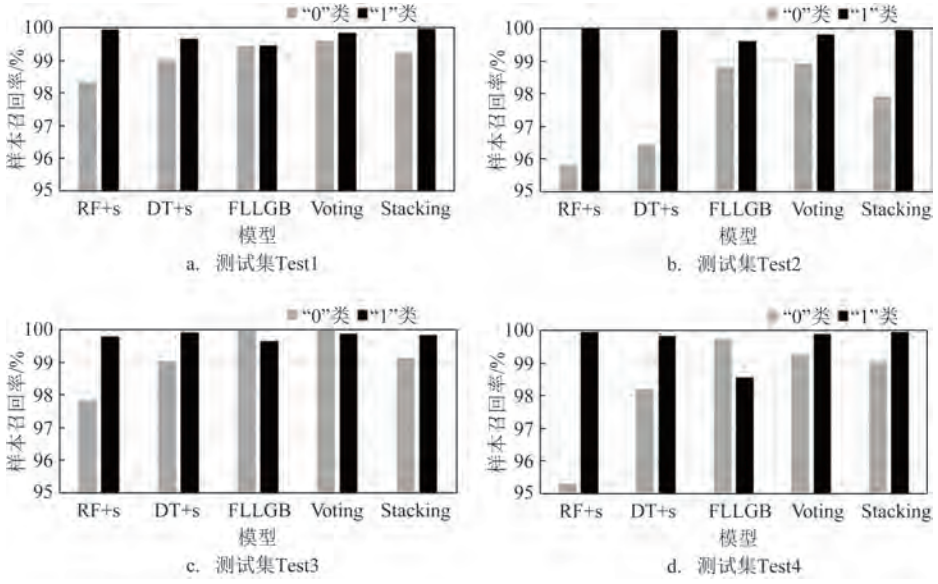

图4　不同模型在不同测试集上的正负样本召回率

Fig.4　Positive and negative sample accuracy of different models on different test sets

表 3 不同模型在SST总测试集上的最终结果

Tab. 3 Final results of different models on the SST total Test set

模型	“0”类			“1”类		
	精确率	召回率	F1 score	精确率	召回率	F1 score
RF+s	0.982 4	0.970 4	0.978 9	0.999 6	**0.999 8**	0.999 7
DT+s	0.950 0	**0.987 4**	0.969 6	**0.999 8**	0.998 8	0.999 4
FLLGB+s	0.980 4	0.974 2	0.977 6	0.999 6	0.999 6	0.999 6
Voting	**0.984 0**	0.977 6	0.980 5	0.999 7	**0.999 8**	**0.999 8**
Stacking	0.983 6	0.978 7	**0.981 4**	0.999 6	**0.999 8**	0.999 7

注:粗体表示最高的精度

表 4 不同模型在AT总测试集上的最终结果

Tab. 4 Final results of different models on the AT total Test set

模型	“0”类			“1”类		
	精确率	召回率	F1 score	精确率	召回率	F1 score
RF+s	0.988 7	0.969 1	0.978 8	0.999 3	**0.999 8**	0.999 6
DT+s	0.997 6	0.993 2	0.995 9	0.999 6	0.999 5	0.999 6
FLLGB+s	0.849 9	0.972 0	0.906 9	0.985 1	0.882 8	0.942 6
Voting	**0.999 7**	0.997 4	**0.998 5**	**0.999 7**	0.999 7	0.999 7
Stacking	0.998 2	**0.998 4**	0.998 3	**0.999 7**	**0.999 8**	**0.999 8**

注:粗体表示最高的精度

准确率也能达到0.999以上;而各类模型在负样本“0”类上的分类表现不同。

在负类“0”样本上,DT+s模型的召回率最高,达到了0.987 4,但是其精确率和F1 score最低,分别只有0.950 0和0.969 6,并且其在正样本“1”上的召回率也最低,为0.998 8。由于正样本的数据量庞大,即使0.999的召回率也会导致大量的数据被漏判,因此,我们认为DT+s模型独立检测并不能兼顾正、负两类样本。其他两个基模型RF+s和FLLGB+s在负样本上的F1 score略高于DT+s模型,分别达到了0.978 9和0.977 6,主要是这两个模型的精确率优于DT+s。使用本文框架的组合模型Voting和Stacking的F1 score比各独立基模型更高,但在负样本上的召回率低于DT+s。

表4展示了不同分类算法在AT总测试集上各类别的分类情况。由表可知,其整体趋势与SST数据类似,本文提出的组合结果在“0”类样本上的F1 score最高,分别达到了0.998 5和0.998 3。

由以上实验分析可知,各个基模型在处理SST和AT数据都存在精度不高或者精度两极化的情况,本文提出的两层模型集成融合的方法可以缓解不同模型的局限性,在两种不平衡的海洋气象数据集上正、负样本的分类结果更加趋于稳定。

3 结论与展望

真实的海洋要素数据集往往是不平衡的,其负样本远远小于正常类数据,而这些负样本恰恰又是我们更加关心的。传统的机器学习算法难以准确地对少数类进行分类。本研究针对海洋要素数据集不平衡的问题,提出了一种基于多模型组合的两层框架,在国际综合海洋大气数据集上进行质量控制,这种方法可以有效提高负样本的分类效果。通过对常见分类算法的实验比较,在SVM、DT、RF、MLP、KNN以及LGB中,SVM、MLP和KNN算法在实验数据集上的分类效果远不如其他分类算法。

进一步针对不平衡样本采用自适应下采样模块和设计损失函数,最后通过融合模型Stacking和Voting进行质量控制,融合模型在测试集上的分类效果同样优于单独模型分类效果。

本文所提出的以DT、RF、LGB作为基模型的融合模型Stacking和Voting,在ICOADS要素数据集的质量控制方面有着较为优秀的表现。未来可以深入研究在更多海洋要素的加入下,如何针对实际应用场景对该框架进行不断修正和优化,以此来提高海洋数据质量控制的准确性,为海洋领域的科学研究、渔业生产等提供更为准确的数据服务。

参考文献:

[1] FREEMAN E, WOODRUFF S D, WORLEY S J, et al. ICOADS release 3.0: a major update to the historical marine climate record [J]. International Journal of Climatology, 2017, 37(5): 2211-2232.

[2] WU G K, ZHANG B P, XU J. Numerical computation of ocean HABs image enhancement based on empirical mode decomposition and wavelet fusion[J]. Applied Intelligence, 2023, 53(16): 19338-19355.

[3] 谭哲韬, 张斌, 吴晓芬, 等. 海洋观测数据质量控制技术研究现状及展望[J]. 中国科学: 地球科学, 2022, 52(3): 418-437.
TAN Z T, ZHANG B, WU X F, et al. Quality control for ocean observations: from present to future[J]. Science China Earth Sciences, 2022, 65(2): 215-233.

[4] GOURETSKI V. World ocean circulation experiment-Argo global hydrographic climatology[J]. Ocean Science, 2018, 14(5): 1127-1146.

[5] SCAVIA D, RABALAIS N N, EUGENE TURNER R, et al. Predicting the response of Gulf of Mexico hypoxia to variations in Mississippi River nitrogen load[J]. Limnology and Oceanography, 2003, 48(3): 951-956.

[6] 任焕萍, 张斌, 谭哲韬, 等. 一种精细化的海洋浮标数据质量控制方法[J]. 海洋科学, 2021, 45(10): 93-103.
REN H P, ZHANG B, TAN Z T, et al. A new quality control scheme for marine buoy temperature and salinity data[J]. Marine Sciences, 2021, 45(10): 93-103.

[7] 刘首华, 陈满春, 董明媚, 等. 一种实用海洋浮标数据异常值质控方法[J]. 海洋通报, 2016, 35(3): 264-270.
LIU S H, CHEN M C, DONG M M, et al. A quality control method for the outlier detection of buoy observations[J]. Marine Science Bulletin, 2016, 35(3): 264-270.

[8] 王辉赞, 张韧, 王桂华, 等. Argo浮标温盐剖面观测资料的质量控制技术[J]. 地球物理学报, 2012, 55(2): 577-588.
WANG H Z, ZHANG R, WANG G H, et al. Quality control of Argo temperature and salinity observation profiles[J]. Chinese Journal of Geophysics, 2012, 55(2): 577-588.

[9] WONG A, KEELEY R, CARVAL T. Argo quality control manual for CTD and trajectory data[R]. ARGO, 2024.

[10] 许立兵, 王安喜, 汪纯阳, 等. 基于机器学习的海洋环境预报订正方法研究[J]. 海洋通报, 2020, 39(6): 695-704.
XU L B, WANG A X, WANG C Y, et al. Research on correction method of marine environment prediction based on machine learning[J]. Marine Science Bulletin, 2020, 39(6): 695-704.

[11] TIMMS G P, DE SOUZA JR P A, REZNIK L, et al. Automated data quality assessment of marine sensors[J]. Sensors, 2011, 11 (10): 9589-9602.

[12] ZHOU Y S, QIN R F, XU H P, et al. A data quality control method for seafloor observatories: the application of observed time series data in the East China Sea[J]. Sensors, 2018, 18(8): 2628.

[13] LE GUEN R. Machine Learning applied to Argo floats temperature and salinity Delayed-Mode Quality Control (Core-Argo DMQC)[R]. ARGO, 2019: 71-100.

[14] 刘玉龙, 王国松, 侯敏, 等. 基于深度学习的海温观测数据质量控制应用研究[J]. 海洋通报, 2021, 40(3): 283-291.
LIU Y L, WANG G S, HOU M, et al. Quality control of sea temperature observation data using deep learning neural networks [J]. Marine Science Bulletin, 2021, 40(3): 283-291.

[15] MIERUCH S, DEMIREL S, SIMONCELLI S, et al. SalaciaML: a deep learning approach for supporting ocean data quality control [J]. Frontiers in Marine Science, 2021, 8: 611742.

[16] 向先全, 路文海, 杨翼, 等. 海洋环境监测数据集质量控制方法研究[J]. 海洋开发与管理, 2015, 32(1): 88-91.
XIANG X Q, LU W H, YANG Y, et al. Research on quality control methods of marine environmental monitoring datasets[J]. Ocean Development and Management, 2015, 32(1): 88-91.

[17] LIU Z N, CAO W, GAO Z F, et al. Self-paced ensemble for highly imbalanced massive data classification[C]//Proceedings of the 2020 IEEE 36th International Conference on Data Engineering. Dallas: IEEE, 2020: 841-852.

[18] 李颖. 基于决策树算法的信息系统数据挖掘研究[J]. 信息技术, 2022(2): 116-120.
LI Y. Research on information system data mining based on decision tree algorithm[J]. Information Technology, 2022(2): 116-120.

[19] 耿丹, 刘婷婷, 李超. 结合FY-4A卫星及随机森林的日间沿海海雾识别模型的研究[J]. 海洋预报, 2022, 39(3): 83-93.
GENG D, LIU T T, LI C. Research on a daytime sea fog identification model based on FY-4A satellite data and random forest algorithm[J]. Marine Forecasts, 2022, 39(3): 83-93.

[20] 王丹, 李林, 赵丹. 基于LightGBM的企业财务风险预测[J]. 信息科学, 2022, 602: 259-268.
WANG D, LI L, ZHAO D. Corporate finance risk prediction based on LightGBM[J]. Information Sciences, 2022, 602: 259-268.

[21] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of 2017 IEEE International

Conference on Computer Vision. Venice: IEEE, 2017: 2999-3007.
[22] 孙昭, 李云, 江毓武, 等. 基于Stacking机器学习模型的南海北部海温预报[J]. 海洋预报, 2023, 40(1): 39-45.
SUN Z, LI Y, JIANG Y W, et al. Sea temperature forecast in the northern South China Sea base on Stacking machine learning model[J]. Marine Forecasts, 2023, 40(1): 39-45.
[23] TAN J R, WANG C B, LI B Y, et al. Equalization loss for long-tailed object recognition[C]//Proceedings of 2020 IEEE / CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11659-11668.

Quality control method for class-imbalanced oceanographic data based on multi-model combination

SONG Wei[1], ZHANG Guiqing[1], XIE Jingrong[1], DONG Mingmei[2], YUE Xinyang[2], YANG Yang[2]
(1. School of Information, Shanghai Ocean University, Shanghai 201306, China; 2. National Marine Information Center, Tianjin 300171, China)

Abstract: This paper proposes a two-layer framework for ocean data quality control based on the combination of multiple models. Various common classification algorithms are chosen as base learners to predict the primary quality labels of ocean data, and a Voting or Stacking strategy is used to identify the quality of the data. To address the issue of class imbalance, an adaptive undersampling strategy is combined with the Focal loss function to enhance the model's ability to recognize difficult samples. To verify the performance of the proposed method, we apply it to the quality control of sea surface temperature and air temperature data that are from ICOADS (International Comprehensive Ocean-Atmosphere Data Set). The results show that the F1 score (the weighted harmonic mean of precision and recall) of rare anomaly samples by the Voting or Stacking methods can reach 0.980 6 and 0.981 2 for sea surface temperature data, and 0.998 5 and 0.998 3 for air temperature data.

Key words: quality control; ocean-atmosphere data; ensemble learning; class imbalance