徐 玲,景向楠,杨 英,等.基于 SMOTE-GA-CatBoost 算法的全国地表水水质分类评价 [J]. 中国环境科学, 2023,43(7):3848-3856.
Xu L, Jing X N, Yang Y, et al. National surface water quality classification evaluation based on SMOTE-GA-CatBoost method [J]. China Environmental Science, 2023,43(7):3848-3856.

基于 SMOTE-GA-CatBoost 算法的全国地表水水质分类评价

徐 玲[1],景向楠[2],杨 英[3*],李卫华[3],刘怡心[4],严国兵[5] (1.合肥城市学院土木工程学院,安徽 合肥 238076；2.合肥城市学院经济与管理学院,安徽 合肥 238076；3.安徽建筑大学环境与能源工程学院,安徽 合肥 230009；4.中国科学技术大学地球和空间科学学院,安徽 合肥 230026；5.杭州市城建设计研究院有限公司(安徽分公司),安徽 合肥 230051)

摘要：针对地表水分类评价中水污染特征空间的高冲突性以及水质类别的不均衡性等问题,以 7 项地表水水质指标为水质评价因子,采用 SMOTE 过采样技术结合遗传算法和 CatBoost 模型对全国主要江河和重要湖库分别进行水质分类评价,并与其他 4 种改进集成算法进行对比.结果表明:SMOTE 预处理有效改善样本类别的不均衡性,提高 CatBoost 模型对少数类水质样本分类的准确性;遗传算法调参有效提高 CatBoost 模型的收敛速度和分类精度,优化了模型的分类性能;SMOTE-GA-CatBoost 模型对江河和湖库的水质分类效果均优于其他 4 种改进集成分类模型,其对江河水水质分类的准确率、精确率、召回率、F1 分别为 97.7%、97.8%、96.1%、96.9%,对湖库水水质分类的准确率、精确率、召回率、F1 分别为 96.7%、96.2%、95.4%、95.8%,该模型可以实现不同水域的水质分类评价.

关键词：地表水；水质分类评价；CatBoost；SMOTE；遗传算法
中图分类号：X824 **文献标识码**：A **文章编号**：1000-6923(2023)07-3848-09

National surface water quality classification evaluation based on SMOTE-GA-CatBoost method. XU Ling[1], JING Xiang-nan[2], YANG Ying[3*], LI Wei-hua[3], LIU Yi-xin[4], YAN Guo-bing[5] (1.School of Civil Engineering, City University of Hefei, Hefei 238076, China；2.School of Economics and Management, City University of Hefei, Hefei 238076, China；3.School of Environment and Energy Engineering, Anhui Jianzhu University, Hefei 230009, China；4.School of Earthand Space Sciences, University of Science and Technology of China, Hefei 230026, China；5.Architectural & Civil Engineering Design Institute Co.Ltd HangZhou China, Hefei 230051, China). *China Environmental Science*, 2023,43(7)：3848~3856

Abstract：Aiming at the problems such as the high conflict of water pollution feature space and the imbalance of water quality categories in surface water classification evaluation, Synthetic Minority Oversampling Technique (SMOTE) which was combined with Genetic Algorithms (GA) and CatBoost model that used seven water quality indexes of surface water as water quality evaluation factors were respectively employed to evaluate the water quality of major rivers and important lakes-reservoirs in the country. The results were compared with the other four improved ensemble algorithms, which showed that the SMOTE pretreatment could effectively enhance the imbalance of sample categories and increase the accuracy of CatBoost model for the classification of minority water quality samples. The genetic algorithm parameters could effectively improve the convergence speed and classification accuracy of CatBoost model and optimize the classification performance of the model. The SMOTE-GA-CatBoost model showed higher performance of water quality classification compared with the other four improved integrated classification models. The values of accuracy, precision, recall and F1of the SMOTE-GA-CatBoost model for river water quality classification reached 97.7%, 97.8%, 96.1% and 96.9%, respectively. The value of accuracy, precision, recall and F1for water quality classification of lakes-reservoirs water were 96.7%, 96.2%, 95.4% and 95.8%, respectively. The proposed model could be used to classify and evaluate the water quality of different water areas.

Key words：surface water；water quality classification and evaluation；CatBoost；SMOTE；genetic algorithms

根据生态环境部公布的最新数据,2022 年上半年,全国十大流域水质优良断面和水质优良湖库的占比分别为 87.3%和 76.2%,地表水环境质量得到进一步改善,但部分流域和湖库仍遭到轻度污染,地表水环境改善不平衡问题依然突出.地表水环境质量评价是解决地表水环境问题的前提条件,不仅可以实现水资源的有效管理,更有利于水体污染综合防

收稿日期：2022-12-04
基金项目：国家自然科学基金资助项目(51978003);安徽省教育厅自然科学类重点项目(2022AH052481);安徽省科技重大专项(201903a06020034);安徽省自然科学基金资助青年项目(1908085QE241);无锡市科技发展资金资助项目(G20192010)
* 责任作者, 教授, yangying5918@163.com

治方案的合理制定[1-3].

传统的单因子评价法操作简单、适用性广[4],但未考虑到各因素的相关性、水环境的复杂性,难以实现水体的综合评价.目前已有较多的学者将机器学习算法应用到水质分类评价的工作中,常见的机器学习算法主要有模糊综合评价法[5-6]、灰色聚类评价法[7-8]、神经网络(ANN)[9-10]与支持向量机(SVM)[11-12]分类算法.相比较于模糊综合评价法和灰色聚类评价法,ANN 与 SVM 算法因具有较强的自学习和自适应能力,目前已被广泛应用于水质分类评价的研究中,如 Yan 等[13]人利用全国主要流域的水质检测数据构建地表水水质分类的自适应神经网络模型,结果表明该模型分类的准确率较高且可实现数值的连续输出.黄鹤[12]等人采用粗糙集对地下水水质评价指标进行约简,并对比分析约简前后 SVM 算法的水质评价结果,表明粗糙集约简可有效消减冗余信息,增强 SVM 算法对水质等级模拟的合理性.虽然两种算法均取得较好的水质分类效果.但 ANN 算法易陷入局部最小问题,且该算法主要针对大样本水质类别的非线性拟合.与之相反, SVM 算法主要针对小样本水质数据集进行训练建模.相比较于上述单个的学习算法,集成分类算法是将多个简单的弱分类器集成一个强分类器的机器学习前沿算法,可从多个起点进行局部搜索寻找水质类别最优解,不易陷入局部最优问题[14],稳定性好、泛化能力强、分类精度高.最具代表性的集成分类算法有随机森林算法(RF)[15]、自适应提升算法(AdaBoost)[16]、极限梯度提升算法(XGBoost)[17]、分布式梯度提升算法(LightGBM)[18]、CatBoost[19-20]梯度提升算法.其中 CatBoost 算法因具备优越的分类性能备受学者们的青睐,该算法不仅可以实现类别型特征的有效处理,避免过拟合问题,而且可有效降低模型的梯度偏差,大幅度提高分类的准确性.目前国外已有学者将 CatBoost 算法应用于地表水水质分类评价中,如 Li 等[19]利用多种集成学习算法构建纽约州 3 个典型海滩的水质预测模型,结果表明,CatBoost 模型可以实现伍德兰和汉堡海滩水质的有效预测.Nasir 等[20]采用 7 种机器学习算法对印度各邦连续 9 年的地表水水质进行分类评价.其中 CatBoost 模型的水质分类准确率最高,达 94.51%.但国内关于 CatBoost 算法在全国地表水水质分类评价中的应用鲜有报道.

上述机器学习算法均以整体准确率作为水质分类评价的最终评估指标,但整体准确率难以体现分类器在少数类水质样本中的分类效果,忽视了水质样本类别的不均衡性,易导致少数类水质样本信息的丢失[21-22].合成少数类过采样技术(SMOTE)利用最邻近算法合成新的少数类样本[23],可有效均衡水质样本类别,避免随机采样中的过拟合问题,从而提高模型的水质分类性能.

针对地表水水质类别不平衡的特点,本研究以全国地表水检测断面的水质数据为基础,以高锰酸盐指数、五日生化需氧量、总磷、总氮、氨氮、溶解氧、氟化物共 7 项水质指标作为分类评价模型的输入变量,结合 SMOTE 过采样技术和遗传算法构建全国地表水水质分类评价的 CatBoost 模型,并与其他 4 种改进的集成分类模型对比分析,以期为地表水的水质分类评价提供一种新的技术思路.

1　材料与方法

1.1　实验数据

水质数据来自中国环境监测总站,2022 年 7 月份全国共设置 3579 个地表水检测断面,共计 3239 个河流检测断面和 340 个湖库检测断面,覆盖了长江、黄河、珠江等十大流域的 1686 条主要河流(下文简称江河)和太湖、滇池、巢湖三大淡水湖以及其他 204 个重要湖库(下文简称湖库).全国水质检测十大流域及三大淡水湖的空间分布见图 1.

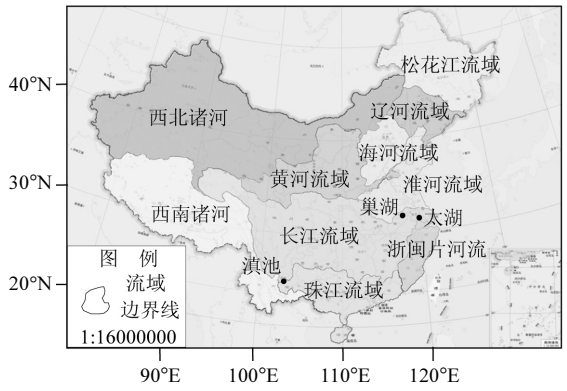

图 1　全国水质检测的十大流域及三大淡水湖的空间分布

Fig.1　The spatial distribution of the top ten river basins and three freshwater lakes in the national water quality testing

审图号:GS(2019)4345 号

依据全国地表水 7 月份水质月报,江河的主要

污染指标为 COD、高锰酸盐指数(COD_{Mn})、BOD_5、总磷(TP)、氨氮(NH_3-N).湖库的主要污染指标为 TP、COD、COD_{Mn}、氟化物(F^-)、BOD_5.基于江河和湖库的主要污染指标存在一定的差异性.采用 CatBoost 算法分别建立江河和湖库的水质分类评价模型,以探讨 CatBoost 算法在不同地表水域的水质分类评价中的可行性.

1.2 数据的预处理

1.2.1 异常值的剔除 由于检测数据中存在部分异常值和缺失值,采用 RF 算法对缺失值进行补充,采用孤立森林算法(IF)对于异常值进行处理,IF 是一种无监督的异常检测算法,通过调用 python 中 decision_function 函数对异常得分进行评估,共剔除离群数据 128 条.该方法对样本的依懒性低,可有效沉降原始信息,降低低频噪声对数据的干扰[24].

1.2.2 水质类别的均衡化 江河和湖库水质类别的比例如图 2 所示,在江河的水质类别中,Ⅱ类、Ⅲ类和Ⅰ V 类水的占比均不小于 15%,Ⅰ类、V 类以及劣 V 类水的占比均小于 6%,多数类与少数类水质样本的占比约为 7.5:1;在湖库的水质类别中,Ⅰ类和劣 V 类水的占比最少,均小于 5%,多数类与少数类水质样本的占比约为 12.3:1,江河和湖库的水质样本均存在类别不均衡的现象.将江河和湖库的水质数据均按照 3:1 划分为训练集和测试集,利用 SMOTE 算法对水质数据的训练集进行采样,采样后江河和湖库的各类水质样本的占比相同,各类别样本分布均匀,且新的训练集样本数分别为 1500 和 360 组,SMOTE 采样有效地改善了水质样本类别的均衡性.

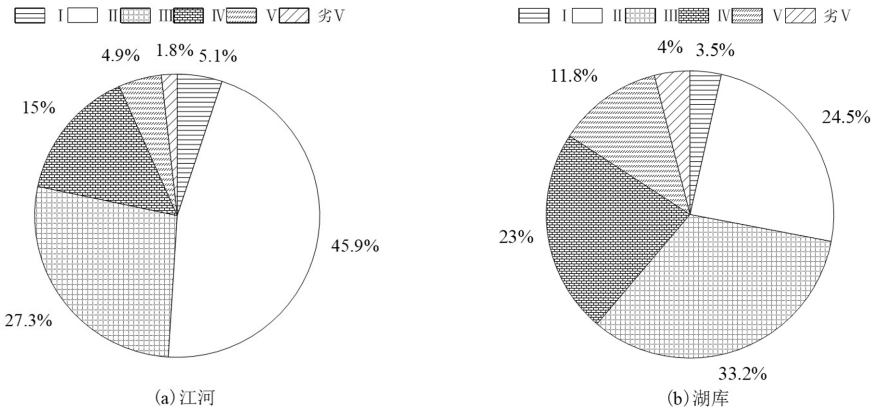

图 2 江河和湖库的水质类别比例

Fig.2 Proportion of water quality categories in rivers and lakes-reservoirs

1.2.3 水质评价指标的权重分析 层次分析法是一种通过构建判断矩阵得出各要素综合权重的主观赋权法[25-26],与之相反,变异系数法是一种客观赋权法,本实验采用层次分析—变异系数法对全国地表水水质评价指标组合赋权,以判定各水质评价指标在水质分类中的重要程度.共选取 20 项地表水水质评价指标作为特征变量,通过组合赋权法确定的权重值如表 1 所示.

表 1 地表水水质评价指标的组合权重(%)

Table 1 Combined weight of the surface water quality evaluation Index (%)

流域	COD_{Mn}	BOD_5	TP	总氮	溶解氧	NH_3-N	F^-	阴离子表面活性剂	PH 值	砷
江河	22.80	17.90	15.10	13.10	11.50	10.60	4.10	0.70	0.50	0.37
湖库	18.30	15.50	25.70	10.10	8.50	8.30	7.20	0.38	1.20	0.92

注:仅列出按权重值大小排序的前10项水质指标的权重值.

由表 1 可知,COD_{Mn} 和 TP 分别为江河和湖库水质分类评价中权重值最大的特征变量,表明 COD_{Mn} 对江河的水质分类影响最大,TP 对湖库的水质分类影响最大;权重值排序前 6 的水质指标的权重值之和均大于 86%且各指标的权重值均大于 8%,表明前 6 项指标包含了水质分类的主要信息,是江河和湖库

水质分类的重要特征变量;此外, F^-的权重值分别为4.10%和 7.20%,说明氟化物对地表水水质评价也具有一定的影响.故拟采用 COD_{Mn}、BOD_5、TP、总氮(TN)、NH_3-N、溶解氧(DO)、F^-共 7 项水质指标作为水质分类评价因子.各项水质指标的分类标准参考《地表水环境质量标准》GB3838-2002[27],如表 2 所示.

表 2　《地表水环境质量标准》基本项目的标准限值(mg/L)

Table 2　《Environmental Quality Standards for Surface Water》 Standard limits for basic items (mg/L)

基本项目	Ⅰ类	Ⅱ类	Ⅲ类	Ⅳ类	Ⅴ类
COD_{Mn}	≤2	≤4	≤6	≤10	≤15
BOD_5	≤3	≤3	≤4	≤6	≤10
NH_3-N	≤0.15	≤0.5	≤1.0	≤1.5	≤2.0
TP	≤0.02	≤0.1	≤0.2	≤0.3	≤0.4
TN	≤0.2	≤0.5	≤1.0	≤1.5	≤2.0
DO	≥7.5	≥6	≥5	≥3	≥2
氟化物(以 F^-计)	≤1.0	≤1.0	≤1.0	≤1.5	≤1.5

1.3　分析方法

1.3.1　遗传算法超参数调优　本文采用遗传算法(GA)对分类模型进行性能优化.遗传算法是通过模拟生物的进化过程来寻找最优参数,该算法调优的流程见图 3,主要操作包括选择、交叉、变异[28].

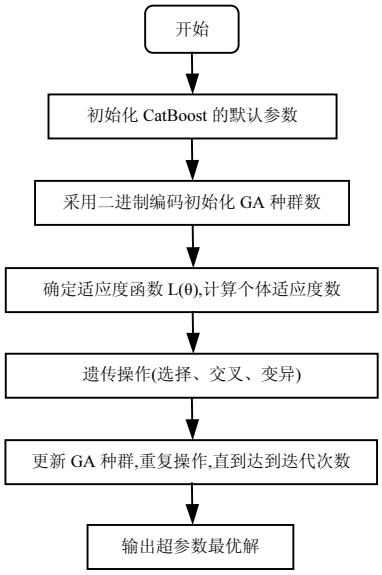

图 3　遗传算法超参数调优的流程

Fig.3　Process for hyperparameter tuning of GA

1.3.2　水质评价的分类方法　串行的 Boosting 算法与并行的 Bagging 算法是 2 种典型的集成算法,本文采用的 5 种算法均改进于这 2 种算法.

RF 算法是 Bagging 算法与随机子空间算法的有效融合[29-30],该算法通过 CART 函数构建决策树形成基分类器,各基分类器相互独立且决策树形成过程中无需采取剪枝处理,模型的鲁棒性强.

AdaBoost 算法改进于 Boosting 算法[31],主要体现了重赋权思想,通过迭代训练获得最优分类器,然后将每个弱分类器的分类结果加权融合得到强分类器的最终决策结果.

LightGBM、XGBoost、CatBoost 3 种算法都是对梯度提升决策树(GBDT)算法的优化,GBDT 算法是一种分阶段学习的叠加模型,有效结合了 Boosting 与回归决策树,GBDT 多分类算法的原理如图 4 所示,即通过前一个基学习器的学习误差率来调整下一个样本训练集的权重分布进行迭代训练,在梯度降低的方向不断建立 CART 决策树[32],最终形成一个强学习器;训练过程中,采用梯度下降技术对参数进行迭代优化,不断减小残差,直到损失函数的期望值达到最小,以获得最佳分类结果.

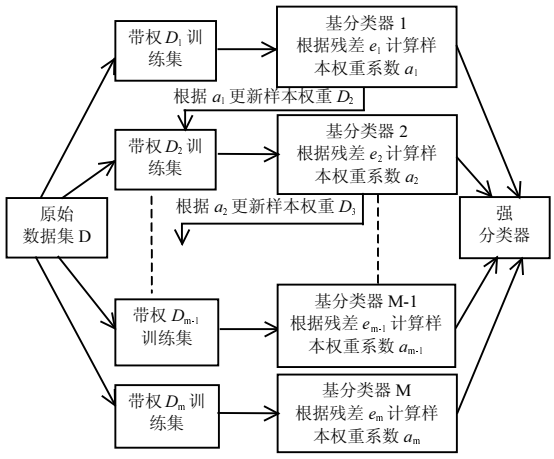

图 4　GBDT 算法的框架

Fig.4　Frame of GBDT multi-classification model

相比较于 GBDT 算法,XGBoost 算法通过二阶泰勒公式展开式对残差进行逼近;且将正则化项引入目标函数中,增加了参数的平滑度,避免过拟合现象的发生[33].

LightGBM 是对 XGBoost 的进一步优化,LightGBM 的优势在于采用点分裂的方法提高学习精度,并结合了单边采样(GOSS)和专有特征捆绑(EFB) 2 种技术,其中,GOSS 有效地保留了小梯度样

本的信息,EFB 通过对稀疏特征空间中的特征进行合并,降低了模型的特征维度,该技术在确保模型精度的同时加快了模型的训练速度[34].

CatBoost 采用对称的二叉树作为基分类器[35],以二进制的形式存储每个叶子节点的索引,并基于二进制特征计算模型输出值,模型的评分速率较 LightGBM、XGBoost 大幅度提高[36].

上述 5 种集成分类算法与传统的单因子评价法均可实现水质的分类评价,各方法的优缺点分析见表 3.经比较可知,5 种集成分类算法均考虑了多种水质指标对水质分类评价的综合影响,不易出现单因子评价法中的过保护问题.此外,相较于其他 4 种集成分类算法,CatBoost 在处理非线性的地表水水质分类评价问题时优势较明显,不仅可以处理高维度数据,实现类别型特征的有效处理,而且克服了梯度偏差避免过拟合.

表 3 不同水质分类评价方法的分析与比较

Table 3 Analysis and comparison of different water quality classification evaluation methods

评价方法	单因子评价法	AdaBoost	RF	XGBoost	LightGBM	CatBoost
优点	适用性广、操作简单;数据原理直观可视化	客观赋权摈弃主观影响;弱分类器的构建可基于多种分类算法	客观赋权摈弃主观影响;可处理高维度数据;抗噪能力强,不易过拟合	客观赋权摈弃主观影响;内存消耗低;增加了参数的平滑度,不易过拟合	客观赋权摈弃主观影响;支持类别特征,可高效并行;内存消耗低,运行速度快	客观赋权摈弃主观影响;对超参数要求低;可处理高维度数据;可实现类别型特征的有效处理,梯度偏差低;不易过拟合
缺点	易出现过保护问题,不能反映水质指标的综合影响	不易确定迭代次数;训练较耗时;不平衡数据分类效果差	不易处理高维度稀疏数据;训练较耗时	不易处理高维度数据;对超参数要求高	抗噪能力弱;需增设深度限制以防过拟合	分类准确性受随机数设定的影响

1.3.3 模型的评价指标 采用准确率(accuracy)、精确率(precision)、召回率(recall)、F1 得分作为分类算法的性能评价指标.

二分类问题:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{1}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{2}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

$$\text{F1}=\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{4}$$

式中:TP(True positives)、FN(False negatives)分别表示将正类样本预测为正样本和负样本的数量;FP(False positives)、TN(True negatives)分别表示将负类样本预测为正样本和负样本的数量.

多分类问题是将多分类问题拆分为多个二分类问题,从而获得每个类别下的精确率、召回率,通过计算各类别下的精确率、召回率的算术平均值得到宏精确率(Macro-precision)、宏召回率(Macro-recall).

$$\text{Macro-precision}=\frac{1}{n}\sum_{i=1}^{n}\text{precision}(i) \tag{5}$$

$$\text{Macro-recall}=\frac{1}{n}\sum_{i=1}^{n}\text{recall}(i) \tag{6}$$

$$\text{F1}=\frac{2\cdot\text{Macro-precision}\cdot\text{Macro-recall}}{\text{Macro-precision}+\text{Macro-recall}} \tag{7}$$

由于精确率与召回率呈负相关性,选取 F1 作为模型的综合评价指标,F1 是精确率和召回率的加权调和平均值.

2 结果与分析

2.1 输入特征变量的共线性检测

地表水环境是一个多样性的生态系统,水中微生物复杂的生化反应搭建了各水质指标之间的相关性.为了避免上述拟输入特征变量之间产生多重共线性问题,采用皮尔逊(Pearson)相关系数表征 7 个输入变量的相关性.结果见表 4.

表 4 水质特征变量的相关性

Table 4 Correlation of water quality characteristic variables

水质特征变量	DO	COD_{Mn}	BOD_5	NH_3-N	TP	TN	F^-
F^-	0.02	0.02	0.03	0.06	−0.02	0.02	1
TN	−0.32	0.38	0.26	0.61	0.11	1	0.02
TP	−0.32	0.36	0.21	0.10	1	0.11	−0.02
NH_3-N	0.31	0.37	0.27	1	0.10	0.61	0.06
BOD_5	−0.26	0.67	1	0.27	0.21	0.26	0.03
COD_{Mn}	−0.32	1	0.67	0.37	0.36	0.38	0.02
DO	1	−0.32	−0.26	0.31	−0.32	−0.32	0.02

由表 4 可得, NH_3-N 和 TN 以及 BOD_5 和 COD_{Mn} 的相关系数都大于 0.6,其他各水质指标的相关性较低;由于 NH_3-N 代表水中以有机化合物(NH_3 和 NH_4^+)形式存在的氮,仅是 TN 的组成成分之一,不能代表水中氮元素的全部信息;COD_{Mn} 可以表征水体中大部分有机物的含量,而 BOD_5 仅能表征水体中可生化降解的有机物含量,无法表征难生物降解的有机物含量;此外,NH_3-N 和 BOD_5 的权重值均比 TN 和 COD_{Mn} 低,对水质分类的影响较 TN 和 COD_{Mn} 小,故剔除 NH_3-N 和 BOD_5 两项水质指标,保留其他 5 项水质指标作为建模的输入特征变量.

2.2　模型的遗传算法调参

基于系统默认的数据库,采用遗传算法对模型进行超参数调参,调参后 CatBoost 模型的迭代次数为 50 次,分类精度为 97.80%,比手动调参提高 2.2%;训练时间为 30s,比手动调参缩短 8s,时间效率进一步提高.遗传算法调参有效地提高了模型的收敛速度和分类精度,优化了模型的分类性能.经过遗传算法调优后,5 种集成分类模型的最优参数见表 5.

表 5　遗传算法调参后 5 种集成分类模型的主要参数的最优值

Table 5　The optimal values of the main parameters of the five ensemble classification models after parameter tuning by genetic algorithm

模型	超参数名称	含义	最优值
GA-CatBoost	iterations	最大树数	100
	learning_ rate	学习率	0.04
	depth	树深	25
	Od_wait	迭代次数	100
GA-RF	N_estimators	树的数量	1600
	Max_depth	树的最大深度	20
	Min_sample_leaf	叶子节点最小样本	4
	Min_samples_split	限制子树	10
GA-XGBoost	Max_depth	树的最大深度	7
	Min_chlid_weight	最小叶子节点权重	2
	Ggamma	最小损失函数下降值	0.5
	Learning_rate	学习率	0.1
GA-LightGBM	Learning_rate	学习率	0.02
	Max_depth	树的最大深度	5
	Lambda_l1	L1 正则化	0.2
	Lambda_l2	L2 正则化	0.3
GA-AdaBoost	N_estimators	树的数量	100
	Learning_rate	学习率	0.2
	Max_depth	树的最大深度	3
	Min_sample_split	限制子树	5

2.3　水质分类评价结果

2.3.1　江河的水质分类评价

采用 8 种水质分类模型对江河的水质进行分类评价,根据每种分类模型的混淆矩阵计算各模型的性能评价指标.不同模型对江河的水质分类结果见图 5.

由图 5 可得, GA-CatBoost 模型的准确率、精确率、召回率、F1 得分较标准 CatBoost 模型分别提高了 0.5%、1.2%、2.2%、2.6%,说明遗传算法调参有效优化了 CatBoost 模型的水质分类性能;SMOTE-GA-CatBoost 模型的各项性能评价指标均达到 96%以上,准确率、精确率、召回率、F1 得分分别为 97.7%、97.8%、96.1%、96.9%,相比较于 SMOTE-GA-RF、SMOTE-GA-XGBoost、SMOTE-GA-LightGBM 、SMOTE-GA-AdaBoost 模型,SMOTE-GA-CatBoost 模型的精确率分别提高了 1.1%、1.9%、3.1%、3.8%;召回率分别提高了 1.6%、2.0%、2.8%、3.0%,结果表明,SMOTE-GA-CatBoost 模型对江河水质的分类评价具有可行性,且较其他 4 种基于 SMOTE-GA 改进的集成分类模型,该模型对江河水质的分类表现出更好的适应性.

上述各分类模型在不同水质类别下的精确率和召回率见表 6.由表 6 可知,经 SMOTE 水质类别均衡化处理后,5 种 SMOTE-GA 改进的集成分类模型对各类别水质分类的精确率和召回率均较高;未经过 SMOTE 预处理的 CatBoost 模型对Ⅰ类、Ⅴ类和劣Ⅴ类水的分类精确率和召回率较多数类水质样本低,经过 SMOTE 平滑后,CatBoost 模型对Ⅰ类、Ⅴ类和劣Ⅴ类江河水质分类的精确率较标准模型分别提高了 6.7%、4.3%、8.1%,召回率较标准模型分别提高了 6.9%、7.8%、7.6%,且对江河水质分类的总体精确率、总体召回率分别提高了 3.3%、3.8%,结果表明,水质类别的均衡化处理有效提高了 CatBoost 模型对少数水质样本的学习效果,降低了模型在江河水质分类评价中的偏向性,显著提高了模型对江河水质样本的总体分类精度.

2.3.2　湖库的水质分类评价

基于不同模型对江河水质分类的结果可知,遗传算法调参和 SMOTE 过采样技术均可提高模型的分类效果,故采用 5 种 SMOTE-GA 改进的集成分类算法对 50 个湖库断面的水质进行分类评价.湖库的水质分类结果见图 6.

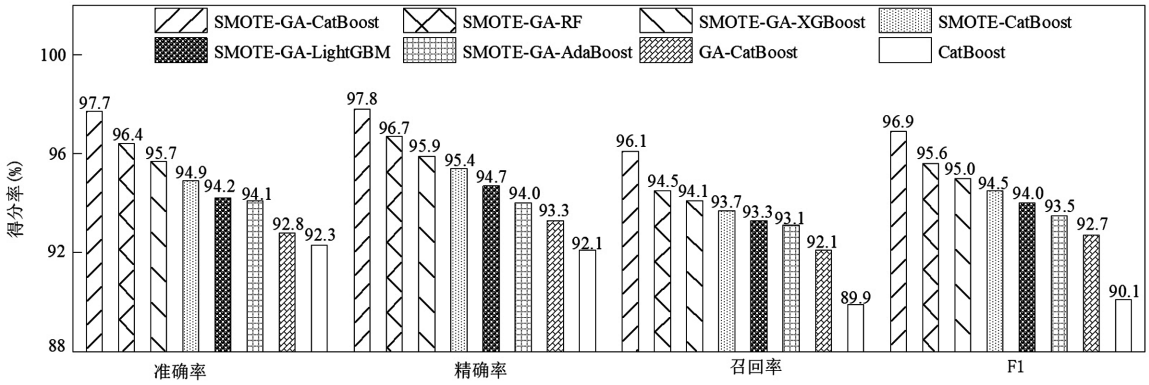

图 5 不同模型对江河的水质分类结果的比较

Fig.5 Comparison of water quality classification results with different models in rivers

精确率与召回率分别代表各模型的宏精确率、宏召回率

表 6 各分类模型在不同水质类别下的精确率和召回率(%)

Table 6 Precision and recall of each classification model under different water quality categories (%)

分类算法	Ⅰ类		Ⅱ类		Ⅲ类		Ⅳ类		Ⅴ类		劣Ⅴ类		算术平均	
	精确率	召回率	精确率	召回率	精确率	召回率	精确率	召回率	精确率	召回率	精确率	召回率	精确率	召回率
SMOTE−GA−CatBoost	97.5	95.8	99.7	97.1	98.2	96.9	97.0	95.3	96.9	95.6	97.7	95.7	97.8	96.1
SMOTE−GA−RF	96.2	94.2	97.5	95.2	97.4	94.9	96.9	94.6	95.9	94.0	96.5	94.3	96.7	94.5
SMOTE−GA−XGBoost	95.2	93.8	96.2	94.9	97.2	94.5	95.9	94.1	95.0	93.5	96.0	94.0	95.9	94.1
SMOTE−CatBoost	95.0	93.3	95.8	94.0	95.3	93.5	96.2	94.6	94.8	93.2	95.6	93.6	95.4	93.7
SMOTE−GA−LightGBM	94.7	93.3	95.6	94.2	95.1	93.4	94.5	93.1	94.2	92.7	95.1	93.2	94.7	93.3
SMOTE−GA−AdaBoost	92.9	92.5	95.3	93.7	94.7	93.6	93.7	93.2	92.6	92.1	94.3	93.3	94.0	93.1
GA−CatBoost	90.1	88.0	96.9	96.6	97.2	97.0	96.6	96.5	89.7	87.0	89.7	87.2	93.3	92.1
CatBoost	88.3	86.4	95.1	94.0	95.5	93.0	95.6	94.5	90.5	85.4	87.5	86.0	92.1	89.9

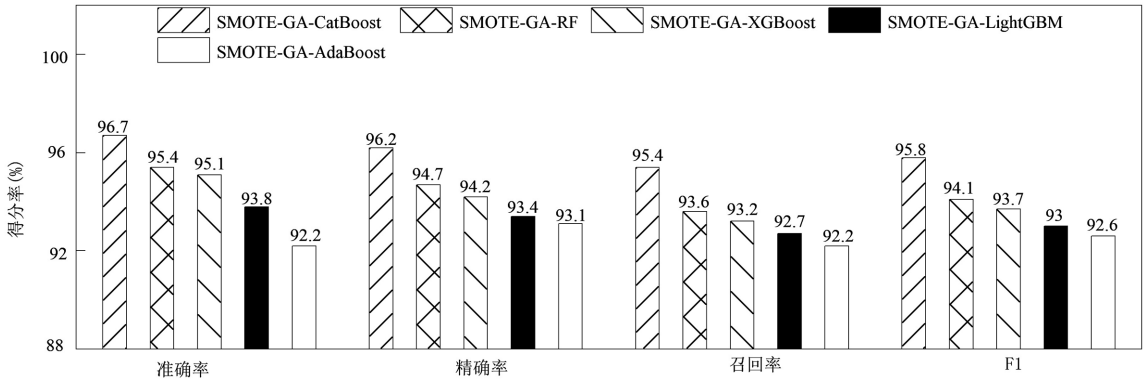

图 6 不同模型对湖库的水质分类结果的比较

Fig.6 Comparison of water quality classification results with different models in lakes−reservoirs

精确率与召回率分别代表各模型的宏精确率、宏召回率

由图6可知, SMOTE−GA−CatBoost模型在湖库的水质分类评价中表现最好, SMOTE−GA−RF 模型仅次之, SMOTE−GA−AdaBoost 模型的表现相对最差; SMOTE−GA−CatBoost 模型的准确率、精确率、召回率、F1得分分别为96.7%、96.2%、95.4%、95.8%, 分类的可信度较高,可更好地实现湖库水质的分类评价.

综上所述, 采用 SMOTE−GA−CatBoost 模型对全国地表水进行水质分类评价具有可行性.

2.4 讨论

本研究采用的分类评价方法训练速度快,预测精度高,目前已广泛应用于解决水质指标的非线性

回归问题,但在水质分类评价方面应用较少.本文是国内较早采用 CatBoost 算法对全国地表水水质进行分类的理论研究,通过模型性能指标的对比分析,验证了该方法在水质分类评价中的优越性,值得推广应用.但该方法可解释性不高,能否通过模型的融合或引入新的解释机制如局部可解释的模型无关解释(LIME)、SHapley 加成解释(SHAP)等提高模型的可解释性有待进一步研究.

3　结论

3.1　基于全国地表水检测断面的水质数据,结合层次分析—变异系数法对水质指标组合赋权,结果表明,COD_{Mn}、BOD_5、TP、TN、NH_3-N、DO 6 项水质指标在江河和湖库水质分类评价中的权重值之和均大于 86%且各指标的权重值均大于 8%,表明该 6 项指标是地表水水质分类的重要影响因子.

3.2　SMOTE-CatBoost 模型对少数类江河水质样本分类的精确率和召回率均较标准模型显著提高,且总体精确率和召回率较标准模型分别提高了 3.3%、3.8%,表明 SMOTE 平滑处理可有效降低模型的偏向性,提高少数类水质样本分类的准确性,改善模型总体的水质分类效果.

3.3　经 SMOTE-GA 改进后, CatBoost 模型在江河和湖库的水质分类评价中表现均最好, RF、XGBoost 模型仅次之, LightGBM、AdaBoost 模型的分类效果相对较差, SMOTE-GA-CatBoost 模型对江河水水质分类的准确率、精确率、召回率、F1 分别为 97.7%、97.8%、96.1%、96.9%;对湖库水水质分类的准确率、精确率、召回率、F1 分别为 96.7%、96.2%、95.4%、95.8%,可实现对全国地表水水质的有效分类.

参考文献：

[1] 嵇晓燕,侯欢欢,王姗姗,等.近年全国地表水水质变化特征 [J]. 环境科学, 2022,43(10):4419-4429.
Ji X Y, Hou H H, Wang S S, et al. Variation characteristics of surface water quality in China in recent years [J]. Environmental Science, 2022,43(10):4419-4429.

[2] Behmel S, Damour M, Ludwig R, et al. Water quality monitoring strategies—A review and future perspectives [J]. Science of the Total Environment, 2016,571:1312-1329.

[3] Gitau M W, Chen J Q, Ma Z. Water quality indices as tools for decision making and management [J]. Water Resources Management, 2016,30(8):2591-2610.

[4] 孙　悦,李再兴,张艺冉,等.雄安新区—白洋淀冰封期水体污染特征及水质评价 [J]. 湖泊科学, 2020,32(4):952-963.
Sun Y , Li Z X, Zhang Y R, et al. Water pollution characteristics and water quality evaluation during the freezing period in Lake Baiyangdian of Xiong'an New Area [J]. Journal of Lake Sciences, 2020,32(4):952-963.

[5] 韩晓刚,黄廷林,陈秀珍.改进的模糊综合评价法及在给水厂原水水质评价中的应用 [J]. 环境科学学报, 2013,33(5):1513-1518.
Han X G, Huang T L, Chen X Z. Improved fuzzy synthetic evaluation method and its application in raw water quality evaluation of water supply plant [J]. Acta Scientiae Circumstantiae, 2013,33(5):1513-1518.

[6] Hu G J, Mian H R, Abedin Z, et al. Integrated probabilistic-fuzzy synthetic evaluation of drinking water quality in rural and remote communities [J]. Journal of Environmental Management, 2022,301: 113937.

[7] 江　敏,刘金金,卢　柳,等.灰色聚类法综合评价滴水湖水系环境质量 [J]. 生态环境学报, 2012,21(2):346-352.
Jiang M, Liu J J, Lu L, et al. Synthetical evaluation of water quality of Dishui Lake water system by gray clustering method [J]. Ecology and Environmental Sciences, 2012,21(2):346-352.

[8] Ip W C, Hu B Q, Wong H, et al. Applications of grey relational method to river environment quality evaluation in China [J]. Journal of Hydrology, 2009,379(3/4):284-290.

[9] Garcia-Alba J, Barcena J F, Ugarteburu C, et al. Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries [J]. Water research, 2019,150:283-295.

[10] Zhang J, Qiu H, Li X Y, et al. Real-Time Nowcasting of microbiological water quality at recreational beaches: A wavelet and artificial neural network based hybrid modeling approach [J]. Environmental Science & Technology, 2018,52(15):8446-8455.

[11] Xu T T , Coco G , Neale M. A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning [J]. Water Research, 2020,177:115788.

[12] 黄　鹤,梁秀娟,肖　霄,等.基于粗糙集的支持向量机地下水质量评价模型 [J]. 中国环境科学, 2016,36(2):619-625.
Huang H, Liang X J, Xiao X, et al. Model of groundwater quality assessment with support vector machine based on rough set [J]. China Environment Science, 2016,36(2):619-625.

[13] Yan H, Zou Z H, Wang H W. Adaptive neuro fuzzy inference system for classification of water quality status [J]. Journal of Environmental Sciences, 2010,22(12):1891-1896.

[14] 王　清.集成学习中若干关键问题的研究 [D]. 上海:复旦大学, 2011.
Wang Q. Research on several key problems of ensemble learning algorithms [D]. Shanghai: Fudan University, 2011.

[15] 吴　敏,温小虎,冯　起,等.基于随机森林模型的干旱绿洲区张掖盆地地下水水质评价 [J]. 中国沙漠, 2018,38(3):657-663.
Wu M, Wen X H, Feng Q, et al. Assessment of groundwater quality based on random forest model in arid oasis area [J]. Journal of Desert Research, 2018,38(3):657-663.

[16] Yao J Q, Sun S Y, Zhai H R, et al. Dynamic monitoring of the largest

reservoir in North China based on multi-source satellite remote sensing from 2013 to 2022: Water area, water level, water storage and water quality [J]. Ecological Indicators, 2022,144:109470.

[17] Yusri H I H, Ab Rahim A A, Hassan S L M, et al. Water quality classification using SVM and XGBoost method [C]. New York: IEEE, 2022.

[18] Narita K, Matsui Y, Matsushita T, et al. Screening priority pesticides for drinking water quality regulation and monitoring by machine learning: Analysis of factors affecting detectability [J]. Journal of Environmental Management, 2023,326:116738.

[19] Li L B, Qiao J D, Yu G, et al. Interpretable tree-based ensemble model for predicting beach water quality [J]. Water Research, 2022,211: 118078.

[20] Nasir N, Kansal A, Alshaltone O, et al. Water quality classification using machine learning algorithms [J]. Journal of Water Process Engineering, 2022,48:102920.

[21] Chen X G, Liu H T, Liu F R, et al. Two novelty learning models developed based on deep cascade forest to address the environmental imbalanced issues: A case study of drinking water quality prediction [J]. Environmental Pollution, 2021,291:118153.

[22] Xu T, Coco G, Neale M. A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning [J]. Water Research, 2020,177:115788.

[23] Luengo J, Fernández A, García S, et al. Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling [J]. Soft Computing, 2011,15(10):1909-1936.

[24] Liu F T, Ting K M, Zhou Z H. Isolation-based anomaly detection [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2012,6(1):1-39.

[25] 张亚青,王 相,孟凡荣,等.基于熵权和层次分析法的 VOCs 处理技术综合评价 [J]. 中国环境科学, 2021,41(6):2946-2955.
Zhang Y Q, Wang X, Meng F R, et al. Comprehensive evaluation of VOCs processing technology based on entropy weight method and analytic hierarchy process [J]. China Environment Science, 2021,41(6): 2946-2955.

[26] Zyoud S H, Fuchs-Hanusch D. A bibliometric-based survey on AHP and TOPSIS techniques [J]. Expert Systems with Applications, 2017, 78(7):158-181.

[27] GB 3838-2002 地表水环境质量标准 [S].
GB 3838-2002 Environmental quality standards for surface water [S].

[28] 梁 泽,王玥瑶,岳远紊,等.耦合遗传算法与 RBF 神经网络的 $PM_{2.5}$ 浓度预测模型 [J]. 中国环境科学, 2020,40(2):523-529.
Liang Z, Wang Y Y, Yue Y W. A coupling model of genetic algorithm and RBF neural network for the prediction of $PM_{2.5}$ concentration [J]. China Environment Science, 2020,40(2):523-529.

[29] Ho T K. The random subspace method for constructing decision forests [J]. IEEE transactions on pattern analysis and machine intelligence, 1998,20(8):832-844.

[30] Speiser J L, Miller M E, Tooze J, et al. A comparison of random forest variable selection methods for classification prediction modeling [J]. Expert Systems with Applications, 2019,134:93-101.

[31] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to Boosting [J]. Journal of computer and system sciences, 1997,55(1):119-139.

[32] Zhang C S, Zhang Y, Shi X J, et al. On incremental learning for Gradient Boosting Decision Trees [J]. Neural Processing Letters, 2019, 50(1):957-987.

[33] 连克强.基于 Boosting 的集成树算法研究与分析 [D]. 北京:中国地质大学, 2018.
Lian K Q. The study and application of ensemble of trees based on Boosting [D]. Beijing: China University of Geosciences, 2018.

[34] Zhang D Y, Gong Y C. The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure [J]. IEEE Access, 2020,8:220990-221003.

[35] Hancock J T, Khoshgoftaar T M. CatBoost for big data: an interdisciplinary review [J]. Journal of Big Data, 2020,7(1):94-139.

[36] 苗丰顺,李 岩,高 岑,等.基于 CatBoost 算法的糖尿病预测方法 [J]. 计算机系统应用, 2019,28(9):215-218.
Miao F S, Li Y, Gao C, et al. Diabetes prediction method based on CatBoost algorithm [J]. Computer Systems & Applications, 2019,28 (9):215-218.

作者简介：徐 玲(1990-),女,安徽庐江人,讲师,硕士,主要从事水处理理论与技术研究.发表论文 4 篇.2757743398@qq.com.