doi:10.6041/j.issn.1000-1298.2025.11.059

综合光谱多特征与 XGBoost 模型的土壤高光谱遥感分类方法

侯　轶[1,2]　张　霞[1]　郑海光[3]　王一博[1,2]　王为昊[1,2]　肖　青[1]

(1. 中国科学院空天信息创新研究院, 北京 100101; 2. 中国科学院大学, 北京 100049;
3. 张家口市农业信息中心, 张家口 075000)

摘要: 土壤遥感分类对土壤快速调查及农业区划至关重要。针对单一类型光谱特征无法充分利用高光谱数据中丰富信息、土壤样本不平衡等问题,本文提出一种综合多种光谱特征与极端梯度提升算法(eXtreme gradient boosting, XGBoost)的土壤高光谱遥感分类方法。采用自适应合成过采样技术(Adaptive synthetic sampling, ADASYN)平衡土壤样本数量,提取光谱主成分、光谱吸收特征、有机质敏感波段及土壤质地敏感的光谱指数等 4 种光谱特征,并利用 XGBoost 算法构建土壤分类模型。将所建模型应用于我国北方暗棕壤、栗钙土、黑土和黑钙土 4 种典型土壤分类中。结果表明:与仅利用原始反射率光谱、光谱主成分以及光谱吸收特征相比,综合多特征的 XGBoost 模型分类精度分别提升 22.22、20.44、9.25 个百分点,表明综合多种光谱特征可显著增强土壤差异和可分性。XGBoost 模型分类结果优于支持向量机(Support vector machine, SVM)和随机森林(Random forest, RF)机器学习模型,总体分类精度和 Kappa 系数分别达 74.44% 和 0.62,其中黑钙土分类精度达 87.18%。特征重要性排名前 5 的光谱特征分别为:1 900 nm 处吸收谷深度($DP_{1\,900\,nm}$)、700 nm 附近吸收谷中心波长($L_{700\,nm}$)、700 nm 处吸收谷深度($DP_{700\,nm}$)、1 340 ~ 1 360 nm 去包络线斜率($S_{1\,340\,nm_1\,360\,nm}$)和 1 895 nm 的一阶微分($R'_{1\,895\,nm}$),这些特征与土壤中粘土矿物、有机质含量及含水量密切相关。在 4 个土壤质地敏感光谱指数中,$NDI_{665\,nm_490\,nm}$ 贡献率最高。研究结果可为土壤快速分类及精细制图提供参考。

关键词: 土壤分类; 光谱特征; 机器学习; 反射率光谱

中图分类号: TP79　　**文献标识码**: A　　**文章编号**: 1000-1298(2025)11-0612-09　　**OSID**:

Hyperspectral Remote Sensing Classification Method for Soil by Combining Spectral Multi-feature and XGBoost Model

HOU Yi[1,2]　ZHANG Xia[1]　ZHENG Haiguang[3]　WANG Yibo[1,2]　WANG Weihao[1,2]　XIAO Qing[1]

(1. *Aerospace Information Research Institute*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*
2. *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*
3. *Zhangjiakou Agricultural Information Center*, *Zhangjiakou* 075000, *China*)

Abstract: Soil classification by hyperspectral remote sensing is very important for rapid soil survey and even agricultural regionalization. To address the issues of insufficient utilization of the abundant information in hyperspectral data by single-type spectral features and unbalanced soil samples, a soil hyperspectral remote sensing classification method that integrated multiple spectral features with the eXtreme gradient boosting (XGBoost) algorithm was proposed. The adaptive synthetic sampling (ADASYN) technique was used to balance the number of soil samples. Four types of spectral features, including spectral principal component, spectral absorption characteristics, sensitive bands of organic matter and spectral index of soil texture sensitivity, were extracted. The XGBoost algorithm was used to construct the classification model. It was applied to the classification of dark brown soil, chestnut soil, black soil and chernozem soil in North China. The results showed that compared with only using the original reflectance band, spectral principal component and spectral absorption characteristics, the

收稿日期: 2024-08-08　修回日期: 2024-10-09

基金项目: 国家自然科学基金项目(42371360)和中国科学院战略性先导科技专项(XDA28080500)

作者简介: 侯轶(2001—),男,博士生,主要从事高光谱遥感信息提取研究,E-mail: houyi23@mails.ucas.ac.cn

通信作者: 张霞(1972—),女,研究员,博士,主要从事高光谱遥感多领域应用研究,E-mail: zhangxia@radi.ac.cn

classification accuracy of XGBoost was improved by 22.22, 20.44 and 9.25 percentage points by comprehensive use of spectral principal component, spectral absorption characteristics, sensitive bands of soil organic matter and spectral index of soil texture sensitivity, respectively. This indicated that combining various spectral characteristics enhanced the difference and separability of soil, and improved the classification accuracy. The classification results of XGBoost model were better than those of support vector machine (SVM) and random forest (RF) model, the overall classification accuracy and Kappa coefficient reached 74.44% and 0.62, respectively, and the classification accuracy of chernozem soil reached 87.18%. The top five spectral features of feature importance included depth of absorption valley at 1 900 nm ($DP_{1\,900\,nm}$), center wavelength near 700 nm ($L_{700\,nm}$), depth of absorption valley at 700 nm ($DP_{700\,nm}$), de-envelope slope within 1 340 ~ 1 360 nm range ($S_{1\,340\,nm_1\,360\,nm}$) as well as first-order differential at 1 895 nm ($R'_{1\,895\,nm}$); all strongly correlated with clay minerals, organic matter content and water. Among the four spectral indices of soil texture sensitivity, the contribution rate of $NDI_{665\,nm_490\,nm}$ was the highest. The research results can provide reference for rapid soil classification and mapping with a high accuracy.

Key words: soil classification; spectroscopy feature; machine learning; reflectance spectroscopy

0　引言

精确的土壤分类有助于了解土壤质量、物化性质以及空间分布规律,为土壤调查、土地评价提供科学依据,进而为农业区划、国土整治规划以及改土培肥等提供重要基础[1-2]。传统土壤分类依赖于野外调查和实验室分析得到的土壤理化性质(如质地、有机质含量、铁氧化物含量等),耗时费力且成本较高[3],而高光谱遥感技术可快速低成本精细地获取土壤反射率光谱,在土壤分类中具有显著优势[4]。目前国内外学者已广泛使用高光谱特征结合机器学习算法实现土壤分类。

常用的光谱特征有光谱主成分和光谱吸收特征等,其能准确反映土壤理化性质,剔除高光谱遥感数据中的无关冗余信息,对于实现土壤分类具有重要意义。光谱主成分因子是原始光谱的线性组合,旨在实现数据降维并尽可能地保留原始数据中的信息,例如 DE OLIVEIRA 等[3]使用反射率光谱主成分,在土壤亚纲级别进行了分类研究。光谱吸收特征是基于光谱去包络线提取的吸收谷中心波长、吸收谷深度、吸收谷面积等参数,能显著增强不同土壤光谱间的差异,相较于主成分具有明确的物理意义[5]。ZHANG 等[6]提取土壤反射率光谱的吸收谷特征参数,实现松嫩平原黑土、黑钙土、风砂土和草甸土分类;WANG 等[7]使用光谱吸收特征参数,在土属水平上实现松嫩平原 4 种典型土壤准确分类。但现有研究大多仅使用单一光谱特征作为分类标准,单一光谱输入量会忽略不同光谱输入量间的互补性[8-9],提供的信息量有限。MARQUES 等[10]证明使用光谱强度、形状等与土壤特性密切相关的特征,结合主成分等定量方法有助于提升土壤分类精度。此外,研究表明土壤光谱微分、光谱指数能有效提升光谱与土壤有机质含量[11-13]、质地[14-15]等土壤理化特性间的相关性。若综合运用多样的土壤反射率光谱特征,有望提升土壤间差异性,同时规避原始反射率光谱的冗余问题,进而提升分类准确度。

在土壤高光谱分类领域,常用支持向量机(SVM)[16-17]和随机森林(RF)[18-21]等机器学习模型,可以在一定程度上解决土壤样本量较少、数据不平衡的问题。邓永鹏等[16]使用 SVM 等机器学习算法对 96 个不同退耕年限的土壤样本进行分类;XIE 等[18]结合光谱主成分因子和 RF 模型实现类别分布不平衡的土壤分类。然而,SVM 对噪声敏感并且核函数的高维映射解释性差;RF 在噪声较多的情况下,每棵决策树都会学习噪声从而容易导致过拟合。极端梯度提升算法(XGBoost)具有并行处理运算、引入正则化项避免过拟合以及内置交叉运算等优势[16]。相较于没有自我校正能力的 SVM 和 RF,XGBoost 能在每次迭代中增加对错误分类数据点的权重,在不平衡数据上有更好表现[22],因此较传统的 SVM 和 RF 机器学习模型具有更高的精度且不易过拟合[23-24]。尽管 XGBoost 在土壤属性反演[25-26]以及高光谱植被分类领域[27-28]被证明是一个强大的模型,但目前尚缺乏利用多种光谱特征结合 XGBoost 模型进行土壤高光谱遥感分类的应用。

因此,基于我国北方 4 种典型土壤反射率光谱,本文综合光谱多特征和 XGBoost 模型进行土壤精确分类。使用自适应合成过采样技术(ADASYN)扩充样本,分别进行主成分分析、包络线去除、一阶微分变换处理,构建土壤反射率光谱主成分、光谱吸收特征、有机质敏感波段及土壤质地敏感的光谱指数等 4 类光谱特征,综合利用这 4 类特征并结合 XGBoost 构建土壤高光谱遥感分类模型,以期提高土壤高光谱遥感分类准确性。

1 材料与方法

1.1 土壤数据与光谱测试

实验数据来自《2017 年 8 月中国东北土壤室内测量反射率数据集》[29],该数据集采集于黑龙江省、吉林省和内蒙古自治区,采集方法为根据地区土壤类型空间位置和面积占比进行土壤样点位置设计,按照五点混合法进行表层 0 ~ 20 cm 土壤样品采集。选取数据集中样本数量大于等于 10 的暗棕壤、栗钙土、黑土和黑钙土 4 种北方典型土壤样本,共 74 个样本,其中暗棕壤 11 个,栗钙土 10 个,黑土 22 个,黑钙土 31 个,采样点分布如图 1 所示,研究区经纬度范围为 40.995° ~ 50.474°N,116.283° ~ 130.717°E。

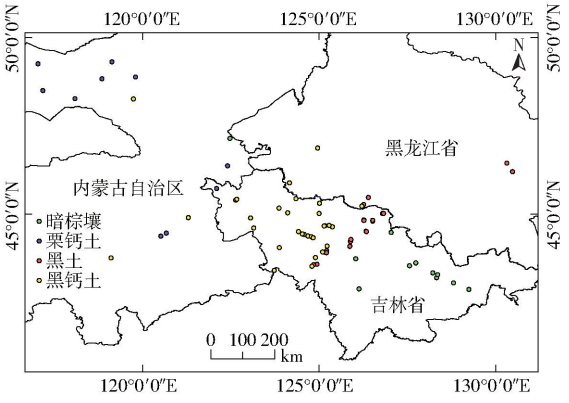
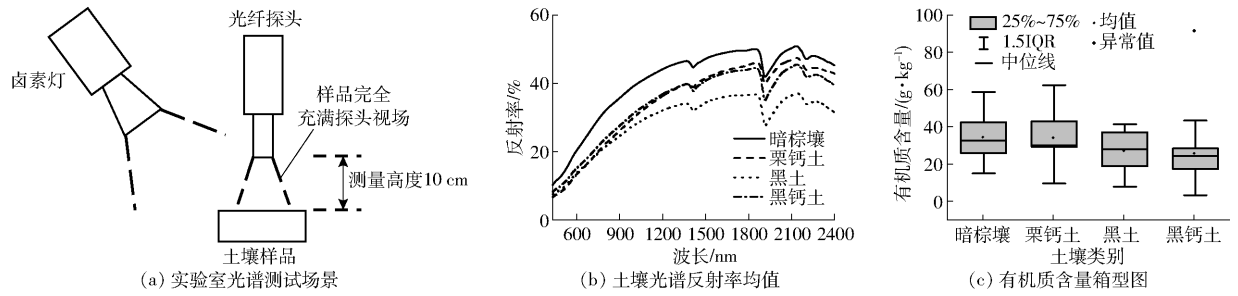


图 1 研究区及采样点分布

Fig. 1 Study area and distribution of soil sampling points

土壤样本室内光谱测量在暗室进行,如图 2a 所示,光源为 1 个 50 W 卤素灯,由 ASD Fieldspec 3 型光谱仪测量,将光纤探头固定在土壤样本上方 10 cm,保证光纤探头视场内充满样品。为减少实验台的背景反射光,在实验台上铺 1 块黑布。将过 100 目筛的土壤样本以培养皿的形状依次、等距排列在黑布上。测量过程中使用 1 块 $BaSO_4$ 板作为定标板,用于土壤反射率定标。每个土壤样本测量 10 条光谱,取平均光谱记为样本最终光谱,并对水汽吸收带进行平滑处理。样本光谱波长为 350 ~ 2 500 nm,光谱采样间隔为 1 nm。去除噪声强烈的 350 ~ 429 nm 和 2 401 ~ 2 500 nm,保留 430 ~ 2 400 nm 区间内的波段用于后续研究。有机质含量采用重铬酸钾法测定。以土壤实验室光谱反射率及有机质含量作为实验数据(图 2b、2c),其中有机质含量主要用于辅助选取敏感波段。



(a) 实验室光谱测试场景 (b) 土壤光谱反射率均值 (c) 有机质含量箱型图

图 2 实验室光谱测试场景、光谱反射率均值曲线和有机质含量箱型图

Fig. 2 Laboratory spectrum test scenario, mean spectral reflectance curve, and box plot of organic matter content

1.2 分类特征提取

1.2.1 光谱主成分

高光谱数据包含海量信息并且波段间相关性较强,主成分是原始反射率数据的线性组合,保留高光谱包含的信息,并且主成分间互不相关,因此可有效实现高光谱数据降维。本文选取贡献率大于 95% 的主成分作为分类特征[3]。

1.2.2 光谱吸收特征

光谱吸收特征基于去包络线计算,表示光谱曲线的吸收特征描述。本文提取的光谱吸收特征包括吸收谷最小值对应的波长(L)、吸收谷深度(DP)、吸收谷面积(A)、2 个波长间去包络线斜率(S)。DP、A、S 计算公式为

$$DP = 1 - C_\lambda \tag{1}$$

$$A = \sum_{\lambda=\lambda_L}^{\lambda_R} (1 - C_\lambda) \tag{2}$$

$$S = \frac{C_{\lambda_1} - C_{\lambda_2}}{\lambda_1 - \lambda_2} \tag{3}$$

式中 C_λ——波长 λ 处去包络线值

C_{λ_1}——波长 λ_1 处去包络线值

C_{λ_2}——波长 λ_2 处去包络线值

λ_R——吸收谷右侧边界波长

λ_L——吸收谷左侧边界波长

1.2.3 土壤有机质敏感波段

相对于原始光谱反射率,一阶微分转换后的反射率光谱与有机质含量相关性会显著增强。使用一阶微分与有机质含量进行相关性分析获取有机质敏感波段作为分类特征,一阶微分计算公式为

$$R'_{\lambda_i} = \frac{R_{\lambda_{i+1}} - R_{\lambda_i}}{\lambda_{i+1} - \lambda_i} \tag{4}$$

式中 λ_i——第 i 波段波长

λ_{i+1}——第 $i+1$ 波段波长

R_{λ_i}——波长 λ_i 处反射率

$R_{\lambda_{i+1}}$——波长 λ_{i+1}处反射率

1.2.4　土壤质地敏感的光谱指数特征

土壤质地与土壤光谱有一定相关性，一般情况下，土壤颗粒越小，光谱反射率越大；土壤颗粒越大，表面越粗糙，光谱反射率越小。参照刘琼等[14]的黑土区相关研究，提取与土壤质地相关性较强的 4 个光谱指数：$DI_{1\,610\ nm_490\ nm}$、$DI_{2\,190\ nm_490\ nm}$、$DI_{2\,190\ nm_560\ nm}$、$NDI_{665\ nm_490\ nm}$，计算公式为

$$DI_{1\,610\ nm_490\ nm} = R_{1\,610\ nm} - R_{490\ nm} \tag{5}$$

$$DI_{2\,190\ nm_490\ nm} = R_{2\,190\ nm} - R_{490\ nm} \tag{6}$$

$$DI_{2\,190\ nm_560\ nm} = R_{2\,190\ nm} - R_{560\ nm} \tag{7}$$

$$NDI_{665\ nm_490\ nm} = \frac{R_{665\ nm} - R_{490\ nm}}{R_{665\ nm} + R_{490\ nm}} \tag{8}$$

式中 R 为反射率，下角标为波长。

1.3　自适应合成过采样方法

数据集存在数据不平衡的问题，暗棕壤、栗钙土和黑土样本数量相较于黑钙土数量较少，这会导致模型倾向于学习数量多的样本，解决这类问题的常用方法是对原始数据进行过采样或欠采样[30]。ADASYN 是一种基于合成少数过采样技术(SMOTE)的过采样办法，即使用不同少数类的加权分布决定少数类需要新合成样本的数量[31]。本文采用 ADASYN 方法对训练集进行过采样处理。

1.4　极端梯度提升分类算法

XGBoost 是基于梯度提升决策树的机器学习算法，其核心是通过添加新树拟合上一次预测残差以减少损失，并采用 Boosting 算法组合弱分类器[22]。XGBoost 具有数据压缩、分片和缓存访问模式，这些机制使得 XGBoost 在速度与准确性上都优于大多数机器学习算法。

1.5　模型建立与精度评价

本文技术路线如图 3 所示。首先去除 350 ~ 429 nm 和 2 401 ~ 2 500 nm 噪声波段，样本采用分层抽样方法，按比例 2∶1 划分训练集与测试集，由于数据集较小，重复 10 次实验。采用 ADASYN 对暗棕壤、栗钙土和黑土训练集进行扩充，黑钙土不作处理。由于合成样本所需的邻近样本数量设置过大会导致生成的样本难以与其他样本区分，过小会导致生成的样本过于集中，因此样本数取 5[32]。然后提取光谱主成分、光谱吸收特征、土壤有机质敏感波段和土壤质地敏感的光谱指数 4 种光谱特征，有机质含量用于筛选土壤有机质敏感波段。将原始反射率、光谱主成分、光谱吸收参数与 4 种光谱特征分别作为模型输入，利用 XGBoost 模型进行土壤分类。取所有实验均值作为最终结果，采用总体分类精度(f_{OA})和 Kappa 系数来评估模型分类精度。

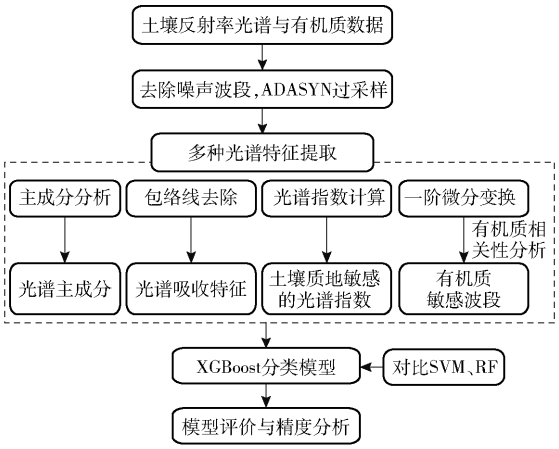


图 3　技术路线

Fig. 3　Technical route of proposed method

2　结果与分析

2.1　特征提取结果

2.1.1　光谱主成分

所有光谱数据的第一主成分(PC1)贡献率为 94.42%，第二主成分(PC2)贡献率为 4.79%，累计贡献率为 99.21%，表明第一、第二主成分几乎包含光谱中所有信息，省略剩余主成分不会丢失数据中的信息[3]。图 4a 为主成分散点图，暗棕壤和黑土在一、二象限的样本点明显可分，在三、四象限 4 种土壤存在重叠。图 4b 为主成分特征值，第一主成分在各波长处特征值随波长增加而升高，与土壤光谱反射率存在直接关系，第二主成分在 600 ~ 900 nm 可

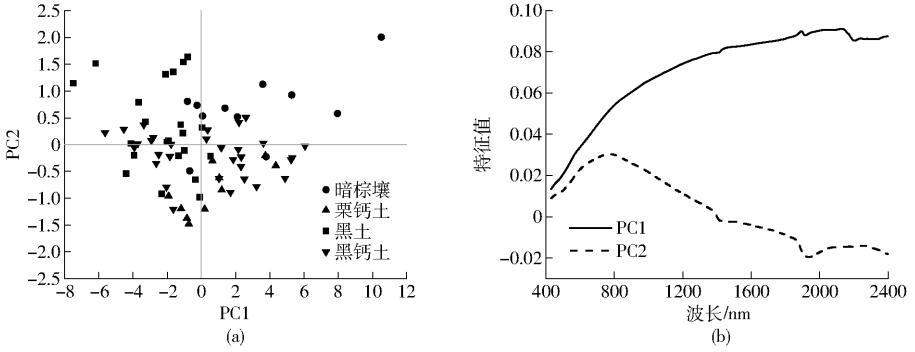


图 4　光谱主成分散点图和主成分特征值变化曲线

Fig. 4　Spectral principal component scatter diagram and load factor curve

见光区域存在峰值,1 900 nm 处特征值最小。

2.1.2 光谱吸收特征

图 5 为 4 类土壤不同类型光谱去包络线反射率,在 500、700、1 400、1 900、2 200 nm 共有 5 个吸收谷。500、700 nm 处吸收谷与土壤有机质、氧化铁和粘土矿物有关[33-34],1 400、1 900 nm 处吸收谷主要是由土壤表面吸附水、黏土矿物 O—H 晶格结构水引起,此外,1 900 nm 处吸收谷还受碳酸盐矿物影响[35]。2 200 nm 处吸收谷与土壤中有机质含量以及黏土矿物分子羟基(—OH)伸缩振动的双频有关[10,36]。为了能有效区分不同土壤,选择特征有:500 nm 处吸收谷面积($A_{500\ nm}$)、吸收谷中心波长($L_{500\ nm}$),700 nm 处吸收谷中心波长($L_{700\ nm}$),500 nm 和 700 nm 处吸收谷面积之和($A_{500\ nm_700\ nm}$),4 个吸收谷吸收深度($DP_{500\ nm}$、$DP_{700\ nm}$、$DP_{1\ 400\ nm}$、$DP_{1\ 900\ nm}$)以及 650 ~ 810 nm、1 340 ~ 1 360 nm、1 420 ~ 1 470 nm 3 个波段的去包络线斜率($S_{650\ nm_810\ nm}$、$S_{1\ 340\ nm_1\ 360\ nm}$、$S_{1\ 420\ nm_1\ 470\ nm}$)。

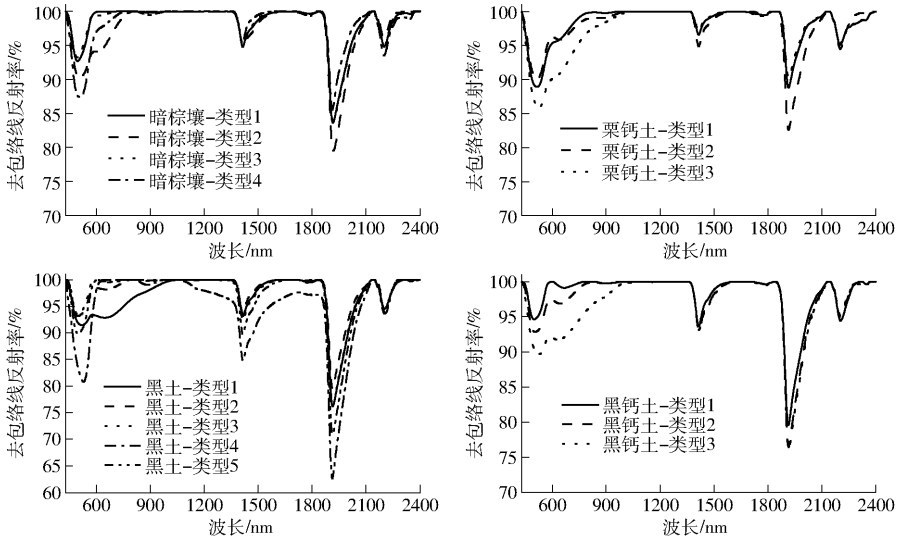


图 5 4 类土壤光谱去包络线反射率

Fig. 5 Continuum removal of spectral reflectance of four types of soil

2.1.3 土壤有机质敏感波段

图 6 为一阶微分光谱与土壤有机质含量的相关系数。有机质含量与 1 400、1 900、2 200 nm 3 个波长有较高相关性是由于土壤吸湿作用,有机质含量越高土壤中的吸附水含量则越高。在 2 300 nm 附近还存在一个与有机质相关的 C—H 特征峰[11]。选择提取的 573、1 388、1 427、1 895、1 928、2 194、2 238、2 337 nm 等 8 个波长的反射率作为有机质敏感波段分类特征。

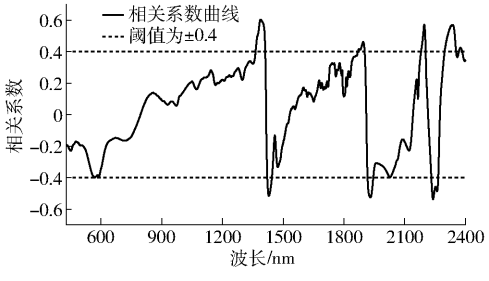


图 6 土壤光谱一阶微分转换数据与有机质含量相关系数变化曲线

Fig. 6 Correlation coefficient curve between soil first derivative spectrum and organic matter

2.2 分类精度对比

为证明本文提出模型的有效性,选取 SVM 和 RF 模型作为基准进行对比。SVM 模型通过优化决策边界,使得不同类别之间的间隔最大化,从而实现精准分类。RF 模型采用 Booststraping 生成决策树,通过构建多个决策树并对其预测结果进行多数投票,以提高分类准确性和鲁棒性。本文模型采用贝叶斯优化(BO)方法自动搜寻最优超参数,贝叶斯优化迭代速度更快,且在非凸、多峰等问题上有较好的鲁棒性[37]。参数调整后,XGBoost 中树数量为 185、最大深度为 10、学习率为 0.134 6、子采样率为 0.632 7、列采样率为 0.579 1。SVM 模型惩罚系数为 10,核函数为 rbf,核函数系数为 1。RF 模型学习器数量为 31,最大深度为 7,最大特征数为 4。

由表 1 可以看出,使用原始光谱反射率进行分类时,由于存在数据冗余,3 类模型分类精度均很低。使用主成分进行分类,RF 和 XGBoost 模型分类精度略微提升,但由主成分散点图(图 4a)可以看出 4 类土壤光谱样本之间存在重叠,因此分类效果不好。相较于原始反射率和光谱主成分,使用光谱吸收特征能降低数据冗余,增强不同土壤间差异,因此 3 类模型分类精度均有所提升,其中 RF 和 XGBoost 整体分类精度均超过 60%。尽管单一特征能降低

表 1　分类精度

Tab. 1　Classification accuracy

分类特征	模型	总体精度/%				平均总体精度/%	Kappa 系数
		暗棕壤	栗钙土	黑土	黑钙土		
原始反射率	SVM	47.22	52.78	58.02	49.57	52.22	0.33
	RF	47.22	69.44	56.79	44.44	51.58	0.33
	XGBoost	75.00	72.22	60.49	33.33	52.22	0.35
光谱主成分	SVM	60.00	60.00	25.56	31.54	43.33	0.26
	RF	67.86	64.29	46.03	41.76	52.67	0.35
	XGBoost	67.86	64.29	53.97	39.56	54.00	0.37
光谱吸收特征	SVM	55.56	50.00	69.14	57.26	59.63	0.43
	RF	55.56	52.78	60.49	70.94	63.33	0.48
	XGBoost	55.56	50.00	67.90	70.94	65.19	0.50
4 类光谱特征	SVM	62.50	47.50	56.67	56.92	56.33	0.38
	RF	69.44	58.33	62.96	79.49	70.37	0.57
	XGBoost	69.44	47.22	70.37	87.18	74.44	0.62

冗余，但并未充分利用高光谱包含的丰富信息，因此分类精度不能令人满意。综合利用主成分、光谱吸收特征、土壤有机质敏感波段以及土壤质地敏感的光谱指数等 4 类光谱特征进行土壤分类，SVM、RF、XGBoost 总体分类精度分别达 56.33%、70.37%、74.44%，与仅利用原始反射率进行分类相比，分别提高 4.11、18.79、22.22 个百分点。黑土与黑钙土分类精度提升明显，黑土在 RF 和 XGBoost 模型中分类精度分别提升 6.17、9.88 个百分点，黑钙土分类精度在 3 类模型中分别提升 7.35、35.05、53.85 个百分点。暗棕壤和栗钙土分类精度在不同模型和分类特征下变化较大。通过对比 SVM、RF、XGBoost 3 个模型分类表现，RF 和 XGBoost 在综合利用 4 种光谱特征进行土壤分类上的表现优于 SVM，XGBoost 分类精度最高。以上结果表明，综合利用提取的 4 种光谱特征进行分类有助于增加各土壤间的光谱差异性和可分性，剔除冗余信息，提升土壤分类精度。

图 7 为 XGBoost 各特征重要性排序。前 5 个特征分别为 1 900 nm 处吸收谷深度（$DP_{1\,900\ nm}$）、700 nm 附近吸收谷中心波长（$L_{700\ nm}$）、700 nm 处吸收谷深度（$DP_{700\ nm}$）、1 340 ~ 1 360 nm 去包络线斜率（$S_{1\,340\ nm_1\,360\ nm}$）和 1 895 nm 的一阶微分数据（$R'_{1\,895\ nm}$）。第二主成分比第一主成分贡献率高。在 4 个土壤质地敏感光谱指数中，$NDI_{665\ nm_490\ nm}$ 贡献率最高。

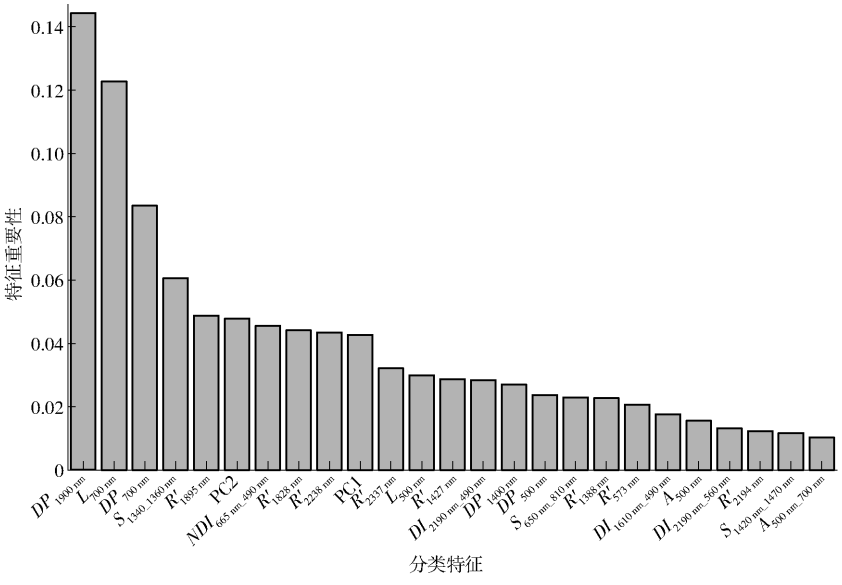


图 7　特征重要性排序

Fig. 7　Feature importance ranking

3　讨论

单一光谱特征在描述土壤理化性质和内在结构方面存在局限，因此需构建多种光谱特征以增强不同土壤间差异。相较于单独使用原始反射率（平均总体精度为 52.22%）、光谱主成分（平均总体精度为 54.00%）以及光谱吸收特征（平均总体精度为 65.19%）方法，本研究提出的综合 4 种光谱特征与 XGBoost 模型的土壤分类框架，平均总体精度最高（74.44%），表明多种光谱特征可以更加准确地捕捉土壤性质，与 XGBoost 模型的结合进一步提升了土壤分类准确性。目前尚未有综合多种光谱特征与 XGBoost 模型实现土壤分类的研究。但值得注意的是，本研究使用光谱吸收特征相较于仅依靠原始反射率、光谱主成分土壤分类精度显著提升，与现有研究结果一致[5-7]。而综合 4 种光谱特征作为输入，XGBoost 模型分类精度比仅使用光谱吸收特征提升 9.25 个百分点，这在一定程度上验证了本研究提出的模型有效性。

以不同特征作为输入，XGBoost 模型均能获得

比 SVM 和 RF 模型更高的分类精度(表 1)。由于土壤类型和数量不同,现有研究分类精度难以对比,DE OLIVEIRA 等[3]结合光谱主成分与 SVM 和 RF 模型,使用 5 600 个土壤样本训练,在土壤亚纲级别总体分类精度均达 95%,然而,XIE 等[18]的研究中,由于土壤数量少和类别偏分布,采用相同方法在土壤亚纲级别分类精度仅为 63%。XGBoost 模型在迭代过程中赋予错误预测点更多权重,并且采用修剪树机制防止过拟合多数类,因此相较于 SVM 和 RF 模型可以更有效地处理类别不平衡数据[24]。此外,本研究还使用 ADASYN 方法进行平衡和扩充样本数量,可以为少数类样本赋予不同权重,迫使模型专注于少数类样本,以补偿偏态分布[31]。

最重要的光谱特征主要与 700、1 900 nm 两个波长有关。这两个波长敏感的土壤理化性质主要是粘土矿物、有机质含量及含水量。暗棕壤中粘土矿物以水云母、高岭石为主,发育地区降水量少蒸发量大,水分少。栗钙土和黑钙土中粘土矿物相似,但由于分别发育于阳坡和阴坡,因此栗钙土含水量更少,反射率相对于黑土更高。黑土中粘土矿物以蒙脱石、伊利石为主,且含有少量绿泥石、赤铁矿、褐铁矿,且黑土属于典型湿润土壤,含水量高[38]。因此,4 类土壤的 $L_{700\ nm}$ 和 $DP_{700\ nm}$ 差异明显。含水量差异导致暗棕壤和栗钙土在 1 400 nm 和 1 900 nm 的水分吸收带吸收深度浅,而黑土和黑钙土吸收深度更深。刘焕军等[39]研究结果表明不同土壤在 500 nm 和 700 nm 两个波段处吸收谷差异均较为明显,但本文中 $L_{500\ nm}$ 和 $DP_{500\ nm}$ 对分类的贡献率不高,一方面是由于土壤类型不同,另一方面土壤光谱曲线受多种因素影响,文献[39]中黑土和黑钙土光谱去包络线与本文中也不完全相同。

成土因素(气候、母质、植被、地形)影响和决定土壤形成方向、发育程度和理化性质,这些因素对土壤分类具有一定贡献,但如何将其参量化并提取有效分类特征还有待深入研究;小波变换具有强大的时频分析能力,是增强光谱特征的有效方法,能够表征光谱信号的局部特征;遥感影像还可获取土壤纹理等空间特征,因此未来有必要探讨引入成土因素、小波变换、空间纹理特征提升基于高光谱图像的土壤分类精度。

4 结论

(1)与仅利用原始反射率波段、光谱主成分与光谱吸收特征相比,综合利用光谱主成分、光谱吸收特征、土壤有机质敏感波段以及土壤质地敏感的光谱指数,XGBoost 模型分类精度分别提升 22.22、20.44、9.25 个百分点,表明综合多种光谱特征显著增强了土壤差异和可分性。

(2)XGBoost 模型分类结果优于 SVM 和 RF 模型,最高总体分类精度及 Kappa 系数分别达 74.44%、0.62,其中黑钙土分类精度达 87.18%。

(3)在 XGBoost 模型中特征重要性排名前 5 的特征分别为 1 900 nm 处吸收谷吸收深度($DP_{1\ 900\ nm}$)、700 nm 附近吸收谷中心波长($L_{700\ nm}$)、700 nm 处吸收谷吸收深度($DP_{700\ nm}$)、1 340 ~ 1 360 nm 去包络线斜率($S_{1\ 340\ nm_1\ 360\ nm}$)和 1 895 nm 的一阶微分($R'_{1\ 895\ nm}$),这些特征与土壤中的粘土矿物、有机质及水分密切相关。在 4 个土壤质地敏感光谱指数中,$NDI_{665\ nm_490\ nm}$ 贡献率最高。

参 考 文 献

[1] LI X Y, FAN P P, LI Z M, et al. Soil classification based on deep learning algorithm and visible near-infrared spectroscopy [J]. Journal of Spectroscopy, 2021, 2021:1 - 11.

[2] 赵华甫,吴克宁. 试论土壤调查、土地类型与耕地评价——兼论倪绍祥先生的土地评价思想[J]. 中国农业资源与区划, 2021, 42(10):245 - 252.
ZHAO Huafu, WU Kening. Discussion of soil survey, land type, and cultivated land evluation—based on academic thoughtofmr NI Shaoxiang's land evaluation [J]. Chinese Journal of Agricultural Resources and Regional Planning, 2021, 42(10): 245 - 252. (in Chinese)

[3] DE OLIVEIRA K M, FALCIONI R, GONÇALVES J V F, et al. Rapid determination of soil horizons and suborders based on VIS - NIR - SWIR spectroscopy and machine learning models[J]. Remote Sensing, 2023, 15(19):4859.

[4] ZENG R, ZHANG G L, LI D C, et al. How well can VNIR spectroscopy distinguish soil classes? [J]. Biosystems Engineering, 2016, 152:117 - 125.

[5] 刘焕军,孟祥添,王翔,等. 反射光谱特征的土壤分类模型[J]. 光谱学与光谱分析, 2019, 39(8):2481 - 2485.
LIU Huanjun, MENG Xiangtian, WANG Xiang, et al. Soil classification model based on the characteristics of soil reflectance spectrum[J]. Spectroscopy and Spectral Analysis, 2019, 39(8):2481 - 2485. (in Chinese)

[6] ZHANG X, LIU H, ZHANG X, et al. Allocate soil individuals to soil classes with topsoil spectral characteristics and decision trees[J]. Geoderma, 2018, 320:12 - 22.

[7] WANG X, ZHANG X, LI H, et al. The minimum level for soil allocation using topsoil reflectance spectra: genus or species?

[J]. CATENA, 2019, 174:36 - 47.
[8] 张笑寒,孟祥添,唐海涛,等. 优化光谱输入量的土壤有机质随机森林预测模型[J]. 农业工程学报, 2023, 39(2):90 - 99.
ZHANG Xiaohan, MENG Xiangtian, TANG Haitao, et al. Random forest prediction model for the soil organic matter with optimized spectral inputs[J]. Transactions of the CSAE, 2023, 39(2): 90 - 99. (in Chinese)
[9] 李硕,王雅晋,覃卫林,等. 基于近地传感光谱协同的土壤重金属含量空间分布预测方法[J]. 农业机械学报, 2025, 56(3):451 - 457.
LI Shuo, WANG Yajin, QIN Weilin, et al. Prediction method for soil heavy metal content based on covariates over proximally sensed spectra[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(3):451 - 457. (in Chinese)
[10] MARQUES K P P, RIZZO R, CARNIELETTO DOTTO A, et al. How qualitative spectral information can improve soil profile classification? [J]. Journal of Near Infrared Spectroscopy, 2019, 27(2):156 - 174.
[11] SHI Z, WANG Q, PENG J, et al. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations[J]. Science China Earth Sciences, 2014, 57(7):1671 - 1680.
[12] ZHANG X, XUE J, CHEN S, et al. Digital mapping of soil organic carbon with machine learning in dryland of northeast and north plain China[J]. Remote Sensing, 2022, 14(10):2504.
[13] 张霞,孙友鑫,尚坤,等. 基于有机质特征谱段的土壤 Cd 含量高光谱遥感反演[J]. 农业机械学报,2024,55(1):186 - 195.
ZHANG Xia, SUN Youxin, SHANG Kun, et al. Soil Cd content retrieval from hyperspectral remote sensing data based on organic matter characteristic spectral bands[J]. Transactions of the Chinese Society for Agricultural Machinery,2024,55(1): 186 - 195. (in Chinese)
[14] 刘琼,罗冲,孟祥添,等. 典型黑土区耕作土壤质地遥感时间窗口及影响因素分析[J]. 农业工程学报, 2022, 38(18): 122 - 129.
LIU Qiong, LUO Chong, MENG Xiangtian, et al. Time window and influencing factors analysis of tillage soil texture remote sensing in the typical black soil region[J]. Transactions of the CSAE, 2022, 38(18): 122 - 129. (in Chinese)
[15] YOUSIF B S, MUSTAFA Y T, FAYYADH M A. Digital mapping of soil-texture classes in Batifa, Kurdistan Region of Iraq, using machine-learning models[J]. Earth Science Informatics, 2023, 16(2):1687 - 1700.
[16] 邓永鹏,朱洪芬,丁皓希,等. 基于高光谱技术的退耕还林地年限判别[J]. 农业工程学报, 2022, 38(3):66 - 74.
DENG Yongpeng, ZHU Hongfen, DING Haoxi, et al. Identification of the years of returning farmland to forest land using hyperspectral technology[J]. Transactions of the CSAE, 2022, 38(3): 66 - 74. (in Chinese)
[17] CHEN S, LI S, MA W, et al. Rapid determination of soil classes in soil profiles using vis-NIR spectroscopy and multiple objectives mixed support vector classification[J]. European Journal of Soil Science, 2019, 70(1):42 - 53.
[18] XIE X L, LI A B. Identification of soil profile classes using depth-weighted visible-near-infrared spectral reflectance[J]. Geoderma, 2018, 325:90 - 101.
[19] DOTTO A C, DEMATTE J A M, ROSSEL R V, et al. Soil environment grouping system based on spectral, climate, and terrain data: a quantitative branch of soil series[J]. Soil, 2020, 6:163 - 177.
[20] 孟祥添,鲍依临,刘焕军,等. 基于高分 5 号影像的东北典型黑土区土壤分类[J]. 农业工程学报, 2020, 36(16):231 - 238.
MENG Xiangtian, BAO Yilin, LIU Huanjun, et al. Soil classification in typical black soil areas of Northeast China based on Gaofen - 5 images[J]. Transactions of the CSAE, 2020, 36(16): 231 - 238. (in Chinese)
[21] MENG X, BAO Y, LIU H, et al. A new digital soil mapping method with temporal-spatial-spectral information derived from multi-source satellite images[J]. Geoderma, 2022, 425:116065.
[22] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016:785 - 794.
[23] SAMAT A, LI E, DU P, et al. GPU-accelerated CatBoost-forest for hyperspectral image classification via parallelized mRMR ensemble subspace feature selection[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14:3200 - 3214.
[24] SHARIFIFAR A, SARMADIAN F. Coping with imbalanced data problem in digital mapping of soil classes[J]. European Journal of Soil Science, 2023, 74(3):e13368.
[25] 邱士其,赵明松,芦圆圆,等. 基于 XGBoost 模型的安徽省土壤 pH 空间建模[J]. 科学技术与工程, 2023, 23(4):1472 - 1480.
QIU Shiqi, ZHAO Mingsong, LU Yuanyuan, et al. Spatial modeling of soil pH in Anhui Province based on XGBoost model [J]. Science Technology and Engineering, 2023, 23(4):1472 - 1480. (in Chinese)
[26] 韩文霆,崔家伟,崔欣,等. 基于特征优选与机器学习的农田土壤含盐量估算研究[J]. 农业机械学报,2023,54(3): 328 - 337.
HAN Wenting, CUI Jiawei, CUI Xin, et al. Estimation of farmland soil salinity content based on feature optimization and machine learning algorithms[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(3):328 - 337. (in Chinese)
[27] ZHANG H, EZIZ A, XIAO J, et al. High-resolution vegetation mapping using eXtreme gradient boosting based on extensive features[J]. Remote Sensing, 2019, 11(12):1505.
[28] ZHEN J, MAO D, SHEN Z, et al. Performance of XGBoost ensemble learning algorithm for mangrove species classification

with multisource spaceborne remote sensingdata[J]. Journal of Remote Sensing, 2024, 4:01406.
[29] 郑兴明,马明国,赵凯,等. 2017 年 8 月中国东北土壤室内测量反射率数据集[DB/OL]. 中国科学院东北地理与农业生态研究所, 中国科学院寒区旱区环境与工程研究所, 2019. http://doi.org/10.11878/db.202005.000015.000001.
[30] TAGHIZADEH-MEHRJARDI R, SCHMIDT K, EFTEKHARI K, et al. Synthetic resampling strategies and machine learning for digital soil mapping in Iran[J]. European Journal of Soil Science, 2020, 71(3):352 - 368.
[31] HE H B, BAI Y, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning[C]//2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008:1322 - 1328.
[32] DATTA D, MALLICK P K, REDDY A V N, et al. A hybrid classification of imbalanced hyperspectral images using ADASYN and enhanced deep subsampled multi-grained cascaded forest[J]. Remote Sensing, 2022, 14(19):4853.
[33] TAN K, CHEN L, WANG H, et al. Estimation of the distribution patterns of heavy metal in soil from airborne hyperspectral imagery based on spectral absorption characteristics[J]. Journal of Environmental Management, 2023, 347:119196.
[34] STONER E R, BAUMGARDNER M F. Characteristic variations in reflectance of surface soils[J]. Soil Science Society of America Journal, 1981, 45(6):1161 - 1165.
[35] VISCARRA R R A, WEBSTER R. Discrimination of Australian soil horizons and classes from their visible-near infrared spectra[J]. European Journal of Soil Science, 2011, 62(4):637 - 647.
[36] CHEN L, LAI J, TAN K, et al. Development of a soil heavy metal estimation method based on a spectral index: combining fractional-order derivative pretreatment and the absorption mechanism[J]. Science of the Total Environment, 2022, 813: 151882.
[37] 丁昌伟,王新,陈同俊,等. 贝叶斯优化的 XGBoost 在小断层地震解释中的应用[J]. 煤炭学报, 2023, 48(6):2530 - 2539.
DING Changwei, WANG Xin, CHEN Tongjun, et al. Application of Bayesian optimized XGBoost in seismic in terpretation of small-scale faults[J]. Journal of China Coal Society, 2023, 48(6):2530 - 2539. (in Chinese)
[38] 朱鹤健,陈健飞,陈松林. 土壤地理学[M]. 3 版. 北京: 高等教育出版社,2019.
[39] 刘焕军,张小康,张新乐,等. 面向土壤分类的高光谱反射特征参数模型[J]. 遥感学报, 2017, 21(1):105 - 114.
LIU Huanjun, ZHANG Xiaokang, ZHANG Xinle, et al. Hyperspectral reflectance characteristics paramter extraction for soil classification model[J]. Journal of Remote Sensing, 2017, 21(1):105 - 114. (in Chinese)

(上接第 598 页)

[23] ZHOU X H, WANG J P, CAO X K, et al. Simulation of future dissolved oxygen distribution in pond culture based on sliding window-temporal convolutional network and trend surface analysis[J]. Aquacultural Engineering, 2021, 95: 102200.
[24] DUPUIS A, DADOUCHI C, AGARD B. Performances of a Seq2Seq - LSTM methodology to predict crop rotations in Québec [J]. Smart Agricultural Technology, 2023, 4: 100180.
[25] JOBARTEH B, NEETHIRAJAN S. Leveraging satellite data for greenhouse gas mitigation in Canadian poultry farming[J]. Smart Agricultural Technology, 2025, 10: 100704.
[26] BANDARU L, IRIGIREDDY B C, KOUTILYA P V N R, et al. DeepQC: a deep learning system for automatic quality control of in-situ soil moisture sensor time series data[J]. Smart Agricultural Technology, 2024, 8: 100514.
[27] 张观山, 丁小明, 何芬, 等. 基于 LSTM - AT 的温室空气温度预测模型构建[J]. 农业工程学报, 2024, 40(18): 194 - 201.
ZHANG Guanshan, DING Xiaoming, HE Fen, et al. Predicting greenhouse air temperature using LSTM - AT[J]. Transactions of the CSAE, 2024, 40(18): 194 - 201. (in Chinese)
[28] MISRA I, SHRIVASTAVA A, GUPTA A, et al. Cross-stitch networks for multi-task learning[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016: 3994 - 4003.
[29] ZHANG Q, WANG H, DONG J, et al. Prediction of sea surface temperature using long short-term memory[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(10): 1745 - 1749.
[30] SONG J, GUO J, CAI Y, et al. Prediction and analysis of sea surface temperature based on LSTM-transformer model[J]. Regional Studies in Marine Science, 2024, 78: 103726.
[31] JIANG Y, ZHAO M, ZHAO W, et al. Prediction of sea temperature using temporal convolutional network and LSTM - GRU network[J]. Complex Engineering Systems, 2021, 1: 6.
[32] QIU R, WANG Y, RHOADS B, et al. River water temperature forecasting using a deep learning method[J]. Journal of Hydrology, 2021, 595: 126016.
[33] DENG J, CHEN X, JIANG R, et al. A multi-view multi-task learning framework for multi-variate time series forecasting[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(6): 6099 - 6112.
[34] CHOI H, KANG P. Multi-task self-supervised time-series representation learning[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023: 2613 - 2623.