

CHEN Zhulin, LI Tianyu, ZHANG Yaofang, XUE Wanlai, XIE Ying, WU Di, ZHAO Chenqiang, MA Li, WANG Siqi, JIA Kun. Land Cover Classification from Hyperspectral Data in the Water Ecological Space of Miyun Reservoir (in Chinese). *Chinese Journal of Space Science*, 2024, **44**(1): 103–113. DOI:10.11728/cjss2024.01.2023-0035

基于高光谱数据的密云水库水生态 空间地物精细分类*

陈珠琳¹ 李添雨² 张耀方²
薛万来² 谢 营³ 吴 迪³ 赵晨强³
马 利³ 王思棋³ 贾 坤¹

1(北京师范大学地理科学学部 遥感科学国家重点实验室 北京 100875)

2(北京市水科学技术研究院 北京 100048)

3(北京市密云水库管理处 北京 101512)

摘 要 重点水域水生态空间地物类型分布状况是其健康评估以及生态规划的重要基础。基于高分五号 (GF-5) 卫星高光谱数据, 采用混合式特征选择算法开展北京市密云水库水生态空间地物精细分类研究。采用随机森林算法获取波段重要性排序, 经过特征降维将总体分类精度最高的模型对应的特征集作为初始特征子集。利用后向序列选择算法搜索地物精细分类的最佳特征子集, 进而开展密云水库水生态空间的地物精细分类。结果表明, 高光谱数据可以实现较高精度的地物分类 (总体分类精度为 93.61%, Kappa 系数为 91.71%), 相比于哨兵二号 (S-2) 卫星多光谱数据, 在精细树种分类方面具有明显的优势。

关键词 遥感分类, 高光谱数据, 随机森林, 水生态空间

中图分类号 P2

Land Cover Classification from Hyperspectral Data in the Water Ecological Space of Miyun Reservoir

CHEN Zhulin¹ LI Tianyu² ZHANG Yaofang²
XUE Wanlai² XIE Ying³ WU Di³ ZHAO Chenqiang³
MA Li³ WANG Siqi³ JIA Kun¹

1(State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875)

2(Beijing Water Science and Technology Institute, Beijing 100048)

3(Beijing Miyun Reservoir Management Office, Beijing 101512)

* 国家自然科学基金项目 (42192581, 42171318) 和北京市科技计划课题项目 (Z221100005222013) 共同资助

2023-03-07 收到原稿, 2023-05-10 收到修定稿

©The Author(s) 2024. This is an open access article under the CC-BY 4.0 License
(<https://creativecommons.org/licenses/by/4.0/>)

Abstract With the acceleration of China's urbanization process, the problem of the structure and function of water ecological space has become increasingly severe. Monitoring the detailed distribution of land cover types in the key water ecological space is critical for their health assessment and future ecological planning. This study investigated a hybrid feature selection algorithm and GF-5 hyperspectral data (with a spatial resolution of 30 m) to generate a fine land cover classification method for the water ecological space of Miyun Reservoir in Beijing. Firstly, the feature importance ranking was determined using the Random Forest (RF) algorithm and several feature subsets were generated with feature amount gradually carried out in a step size of 10. Then, the classification model was generated based on each subset using the RF algorithm. The feature subset that achieved the highest overall classification accuracy was determined as the initial feature subset. Next, the backward sequential selection algorithm was used to this initial subset to search for the best feature subset. Finally, the classification model of the water ecological space of Miyun Reservoir was generated based on the best feature subset and RF algorithm. To validate the advance of GF-5 hyperspectral data, this study also developed a classification model using Sentinel-2 multispectral data (with a spatial resolution of 10 m) for comparison. The results indicated that hyperspectral data achieved high classification accuracy (overall classification accuracy of 93.61%, and Kappa coefficient of 91.71%), especially in the accurate recognition of tree species (The producer's accuracy and user's accuracy of the chestnut forest are 81.25% and 73.03%, respectively). The reflectance of shortwave infrared bands of GF-5 data has increased the differentiation between chestnut forests and other tree species. By contrast, Sentinel-2 data-based model achieved lower classification accuracy with an overall accuracy of 85.91% and a Kappa coefficient of 82.00%. This result indicated that although Sentinel-2 data has higher spatial resolution than GF-5 data, it still has difficulty identifying chestnut forests due to a lack of fine band information. The classification algorithm proposed in this study can provide accurate basic data for supporting the rational planning and management of water ecological spaces.

Key words Remote sensing classification, Hyperspectral data, Random forest, Water ecological space

0 引言

水生态空间是指河流、湖泊、水库、湿地、蓄滞洪涝区的管理和保护范围,承担着维持水生态系统健康稳定、保障水安全等多重功能,是为人类提供水生态系统服务的重要生态空间.近年来,由于多重因素的影响,水生态空间的结构和功能受到威胁和破坏^[1].水生态空间范围内的地物组成是影响水生态空间健康状况的重要因素,因此了解和掌握水生态空间的重要地物类型分布可以为其健康评估以及未来生态规划起到关键性作用.遥感具有快速、实时和大范围的动态监测能力和优势^[2],发展基于遥感技术的水生态空间范围内地物类型精细分类方法可以为其综合性治理与规划提供技术支持.

光学遥感影像是地物分类的主要数据源,可分为多光谱遥感和高光谱遥感分类.多光谱遥感数据种类和数量多,空间分辨率高,广泛应用于地物类型分类

研究.例如 Zhu 等^[3]基于 Landsat-8 卫星数据实现了海南岛热带天然林类别的识别. Wang 等^[4]利用资源三号 (ZY-3)、哨兵二号 (S-2) 和高分一号 (GF-1) 卫星多源数据实现了安徽省合肥市关键生育期的冬小麦和油菜的分类识别.但由于多光谱遥感数据获取到的波段数较少,对异物同谱的地物识别精度有限.相比之下,高光谱遥感数据能够获取到可见光到短波红外范围内数百个窄波段的反射率数据,在精细地物类型分类方面具有巨大的应用潜力^[5].例如, Han 等^[6]基于多时相 CHRIS 高光谱卫星数据实现了吉林省汪清试验区的 7 种优势树种的识别,总体分类精度达到 92.7%. Guo 等^[7]提出一种三维卷积神经网络和超像素分类的算法实现了高光谱分类,在 Salinas 数据集和 Pavia University 数据集进行了测试,总体分类精度分别为 97.53% 和 98.48%. Bhosle 等^[8]基于卷积神经网络对 Hyperion 高光谱数据进行精细地物类型分类,最终总分类精度达到 79.43%.但由于高光谱数据

种类和数量较少且空间分辨率一般低于多光谱遥感数据(例如 GF-1 和 S-2 卫星), 其应用相对匮乏. 近年来, 随着中国多颗高光谱卫星的发射, 例如高分五号 (GF-5)、资源一号 02 D (ZY-02 D)、珠海一号系列卫星 (OHS-1/2/3/4) 等, 极大提高了高光谱数据的获取能力. 目前, 上述高光谱数据在水生态、水环境等方面的应用较少, 因此亟需评估高光谱数据用于水生态空间范围内地物精细分类的可行性并探索适宜的分类方法.

高光谱遥感数据波谱信息丰富, 但同时也不可避免地造成了数据冗余. 所以在构建分类模型之前需要对其进行特征降维处理, 寻找地物精细分类的敏感特征波段. 目前, 常用的降维方法可分为特征提取和特征选择两类^[9]. 典型的特征提取算法包括主成分分析法、独立成分分析、流形学习等算法, 其通过建立高维空间到低维空间的映射实现降维, 但由于对原始空间进行了变换, 导致最终特征提取结果可解释性较差. 特征选择算法则是在原始的特征空间内, 按照某种原则分析因变量与自变量之间的关系, 从而选择出代表性较高的特征, 因此该类算法的可解释性较强. 根据算法策略的不同, 特征选择算法包括过滤法、包装法和嵌入式算法^[10]. 三种算法各有优劣, 其中过滤

法依赖数据特征, 运算速度快, 但无法做到目标最优; 包装法引入了学习算法, 可以搜索最佳子集, 但在高维空间的搜索效率较低; 嵌入式算法虽然对运算效率和精度进行了权衡, 但仍然难以实现最优子集的搜索. 近年来, 混合式特征选择算法逐渐被推广应用, 通过将不同类型的算法进行结合, 可以实现降低模型复杂度, 提高精度等目标^[11].

本文基于混合式特征选择算法, 研究基于 GF-5 卫星高光谱数据的北京市密云水库水生态空间地物类型精细分类方法, 确定水生态空间重要地物识别的最优特征子集, 实现水生态空间地物精细分类. 此外, 通过对高光谱遥感数据与 S-2 多光谱数据的分类结果进行比较, 分析了高光谱遥感数据在水生态空间地物精细分类的优势.

1 研究区与数据源

1.1 研究区与地面特征

选择北京市密云水库水生态空间范围作为研究区 ($117^{\circ}47'31.20''$ — $117^{\circ}9'46.80''$ E, $40^{\circ}26'2.4''$ N— $40^{\circ}38'49.20''$ N), 如图 1 所示. 根据密云水库水生态空间典型地类的分布特点, 可以将地物精细分类的类别

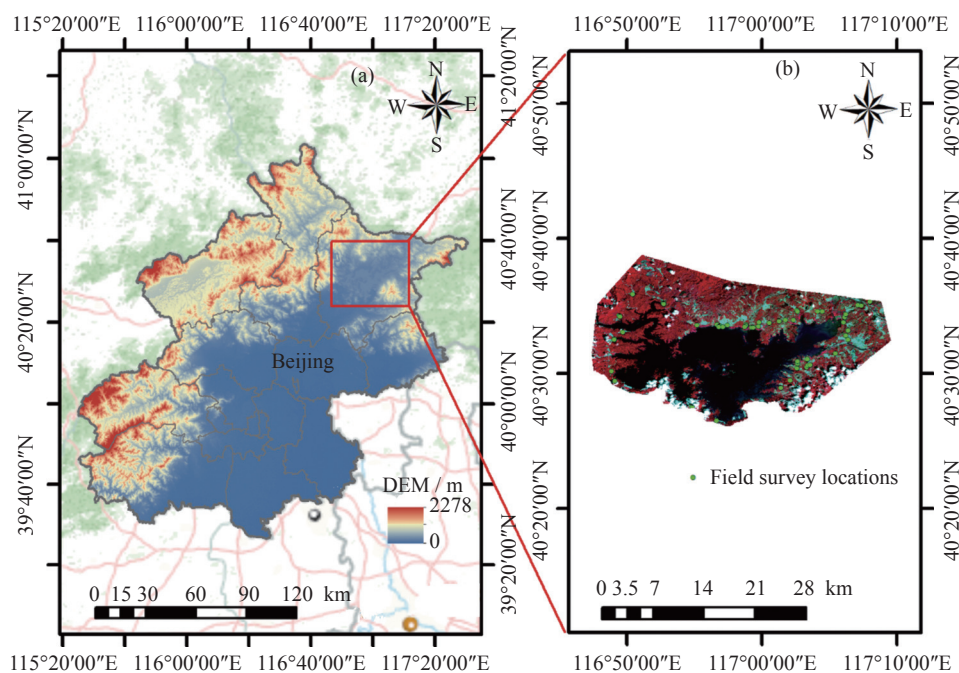


图 1 研究区地理位置以及 GF-5 卫星伪彩色合成影像 (红为波段 107, 绿为波段 65, 蓝为波段 41)

Fig. 1 Study area location and false color image of GF-5 satellite hyperspectral data

(Red: band 107. Green: band 65. Blue: band 41)

分为 5 类: 水体、不透水面及裸土、板栗林、农田和草地、其他林地. 为了获取研究区地物类型的实际分布数据, 于 2022 年 8 月 1 日至 4 日开展了野外调查工作, 在调查同期获取的北京二号 (BJ-2) 多光谱影像 (空间分辨率为 0.8 m) 基础上进行地物类型的样本选择, 各类别在 GF-5 卫星高光谱和 S-2 多光谱影像上获取的训练集和验证集样本数列于表 1. 由于各类型的包含样本量不同, 所以本研究对每个类型单独进行随机划分, 训练集和测试集的样本比例约为 2:1.

1.2 遥感数据与预处理

采用 GF-5 卫星高光谱数据开展密云水库水生态空间的典型地物精细分类研究. GF-5 卫星是中国高分专项中实现高光谱分辨率对地观测的卫星, 也是国家高分辨率对地观测能力的重要标志^[12]. GF-5 号 02 星于 2021 年 9 月 7 日发射升空, 搭载了一台幅宽为 60 km、空间分辨率为 30 m 的可见光短波红外高光谱相机 (AHSI) 可获取可见光到短波红外范围内 (400~2500 nm) 的 330 个谱段数据, 其中可见光近红外波谱分辨率为 5 nm, 短波红外波谱分辨率为 10 nm, 数据获取重访周期为 51 天^[13]. 采用的高光谱数据是 GF-5 号 02 星于 2022 年 7 月 10 日获取到的密云水库水生态空间范围的影像 (见图 1). GF-5 卫星高光谱数据预处理主要包括以下步骤: 首先进行辐射定标、大气校正以及 RPC 正射校正, 输出数据为大气校正后的正射影像. 然后根据相近获取时相的 S-2 多光谱数据进行几何精校正, 最后进行云掩膜及研究区范围提取处理. 为分析高光谱数据地物精细分类的优势, 采用 2022 年 7 月 23 日获取的 S-2 多光谱数据分类作为对比, 该数据包括 13 个波段, 去除 60 m 空间分辨率的波段 1, 9 和 10, 采用蓝 (496.6 nm)、绿 (560.0 nm)、红 (664.5 nm)、3 个红边波段 (703.9 nm, 740.2 nm, 782.5 nm)、2 个近红外波段 (835.1

nm 和 864.8 nm) 和 2 个短波红外波段 (1613.7 nm 和 2202.4 nm) 共 10 个波段作为分类模型的输入. 上述 S-2 多光谱数据波段中除蓝、绿、红和近红外 1 波段的空间分辨率为 10 m, 其他波段的空间分辨率为 20 m, 因此本研究将其进行重采样生成空间分辨率为 10 m 的反射率数据.

2 研究方法

2.1 J-M 距离分析

J-M 距离 (Jeffreys-Matusita Distance) 是一种判断类别可分性的重要指标, 研究表明相比于其他指标 (如离散度、Bhatta-Charyya Distance), J-M 距离更适合用于判断类别可分性^[14]. J-M 距离的取值在 0~2 之间, 值越大代表可分性越好, 其中 0 代表完全不可分, 2 代表完全可分, 当 J-M 距离大于 1.9 时说明可分性较大^[15]. 主要采用 J-M 距离判断类别之间的可分性, 计算公式如下:

$$J_{ij} = 2(1 - e^{-B_{ij}}),$$

$$B_{ij} = \frac{1}{8}(m_i - m_j)^T \left\{ \frac{C_i + C_j}{2} \right\}^{-1} (m_i + m_j) + \frac{1}{2} \ln \left\{ \frac{|(C_i + C_j)/2|}{|C_i C_j|} \right\}. \quad (1)$$

其中, m_i , m_j , C_i , C_j 分别表示 2 个类别的均值和方差.

2.2 随机森林分类算法

采用随机森林分类算法 (Random Forest, RF) 构建水生态空间地物精细分类模型 (见图 2). RF 算法是由多棵决策树组成的集成算法, 与神经网络、支持向量机等机器学习算法相比, RF 算法鲁棒性较强, 运算效率高, 具有处理高维数据和不均衡数据的优势^[16]. RF 算法利用自助采样策略随机且有放

表 1 训练集和验证集中的样本数量
Table 1 Number of samples in training and validation set

数据集	类别				
	水体	不透水面及裸土	板栗林	农田和草地	其他林地
训练集(GF-5)	400	600	165	400	610
测试集(GF-5)	200	296	80	195	308
训练集(S-2)	1000	1000	600	1000	1000
测试集(S-2)	500	500	300	500	500

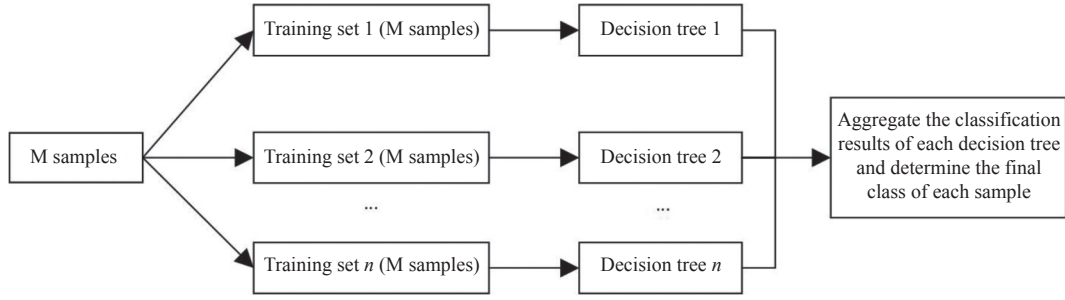


图 2 随机森林分类算法

Fig. 2 Random forest classification algorithm

回地从原始数据集中抽取一定数量的样本, 对每个训练集建立决策树分类模型^[17]. RF 算法中, 特征选择的随机性也使得每个特征子空间存在一定差异. 因此, 特征选择和样本集自助采样过程共同增加了决策树构造的随机性. 而 RF 的整体分类结果是每棵决策树分类结果的综合, 泛化误差低^[18]. 根据已有研究可知, RF 算法中有两个重要参数需要设置, 分别为特征子集的个数和决策树的数量^[19]. 两个参数的值通过交叉验证确定. 其中特征子集个数的取值为 1~5, 步长为 1, 决策树的数量取值为 100~600, 步长为 100. 通过对特征子集的个数和决策树的数量进行组合, 分别构建模型, 对比模型精度确定最佳的取值. 最终确定特征子集的个数和决策树的数量值分别为 2 和 300.

RF 算法不仅可以输出分类结果, 还能够输出分类特征的重要性排序. 在 RF 算法中, 每棵决策树的袋外 (Out of Bag, OOB) 数据可以得到一个误差率, 用于评价分类特征变量的重要性, 具体步骤如下^[20].

步骤 1 选择相应的 OOB 数据, 计算每棵决策树的 OOB 数据误差 E_{OOB1} .

步骤 2 随机地对 OOB 数据所有样本特征 X 加入噪声进行干扰, 再次计算 OOB 数据误差 E_{OOB2} .

步骤 3 若 RF 模型中共有 N 棵决策树, 则特征 X 的重要性值为

$$X_i = \left(\sum_{i=1}^N E_{OOB2} - E_{OOB1} \right) / N. \quad (2)$$

2.3 后向序列选择算法

后向序列选择算法 (Sequential Backward Selection, SBS) 是序列特征选择算法的一种典型应用, 其基于贪婪搜索算法^[21], 用于将原始的 n 维特征空间缩减到一个 k 维特征子空间 ($k < n$). 该算法的基本思想

是在给定特征空间的基础上, 在性能衰减最小的约束下, 通过试探法逐次剔除一个特征, 确保精度保持上升 (或不变) 趋势, 直到达到最高精度^[22]. SBS 算法通过缩减不相关特征降低模型的泛化误差, 提高预测精度. SBS 算法的步骤如下.

步骤 1 将所有特征 (n 维) 作为原始特征集合, 构建模型, 并使用某评价指标对模型进行评估.

步骤 2 删除原始特征集合中的其中 1 个特征 ($n-1$ 维), 重新构建模型, 使用相同的评价指标进行评估, 找到 $n-1$ 维中精度最高的集合进行后续降维.

步骤 3 重复上述步骤, 直到模型的性能不再得到明显的提升为止, 即得到最优特征子集.

2.4 混合式特征选择算法

本研究综合利用 RF 算法特征变量重要性排序和后向序列选择算法构建混合式特征选择算法进行高光谱遥感地物精细分类研究^[23], 模型构建的流程图如图 3 所示. 该算法需要进行两步特征选择处理, 步骤如下.

步骤 1 采用高光谱数据的全部波段和 RF 算法构建初始的地物精细分类模型, 并输出特征波段的重要性排序.

步骤 2 根据 RF 算法计算的特征变量重要性排序结果, 按照步长为 10, 进行降维处理, 根据分类精度选择合适数量的波段构建初始的特征集合. 由于 GF-5 高光谱数据去除无效波段 (波段 192~波段 200、波段 246~波段 264 和波段 325~波段 330) 后得到 296 个波段, 波段数并非 10 的倍数, 因此首先将 296 维降到 290 维, 然后按照 10 为步长进行逐步递减.

步骤 3 基于 SBS 算法寻找最优特征子集, 以提高地物分类的精度, 减小模型的复杂度. 对比每次遍

历得到的分类结果, 探索分类精度最高的组合, 最终构建水生态空间地物精细分类的最佳特征子集, 作为 RF 分类算法的输入特征变量.

2.5 精度评价

选择总体分类精度 (Overall Accuracy, OA)、Kappa 系数、用户精度 (User's Accuracy, UA) 和生产者精度 (Producer's Accuracy, PA) 对分类结果进行评价. OA(定义符号 A) 和 Kappa 系数 (定义符号 κ) 的计算公式如下:

$$A = \frac{1}{N} \sum_{i=1}^r x_{ii},$$

$$\kappa = \frac{N \sum_{i=1}^r x_{ii} - \sum_{i=1}^r (N x_{ii})}{N^2 - \sum_{i=1}^r (N x_{ii})}. \quad (3)$$

式中, r 为类别总数量, N 为样本总数量, x_{ii} 为第 i 类被识别为第 i 类的样本数.

OA 和 Kappa 系数是针对整个分类系统的评价指标, 而 UA 和 PA 则是针对每一类别的评价指标, 由分类混淆矩阵计算, 其中 UA 为被正确识别的第 i 类样本与第 i 类真实样本量之比; PA 则为被正确识别的第 i 类样本与被识别为第 i 类的所有样本之比.

3 结果与分析

3.1 典型地物高光谱反射特征

图 4(a) 为 5 种地物类型的高光谱反射特征 (样本平均值), 可以发现每种地物类型的光谱反射曲线均符合其典型的反射光谱特性. 其中林地 (板栗林和其他林地)、农田和草地在近红外的混淆程度较高, 但在可见光、红边波段和短波红外波段的区分度较大. 图 4(b) 给出了 5 种地物类型的 NDVI 分布, 从图 4(b)

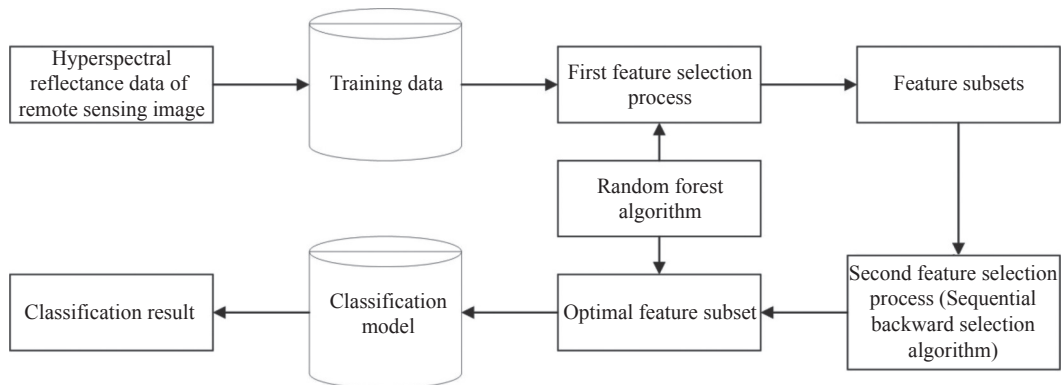


图 3 基于混合特征选择算法的分类模型构建流程

Fig. 3 Classification model construction process based on hybrid feature selection algorithm

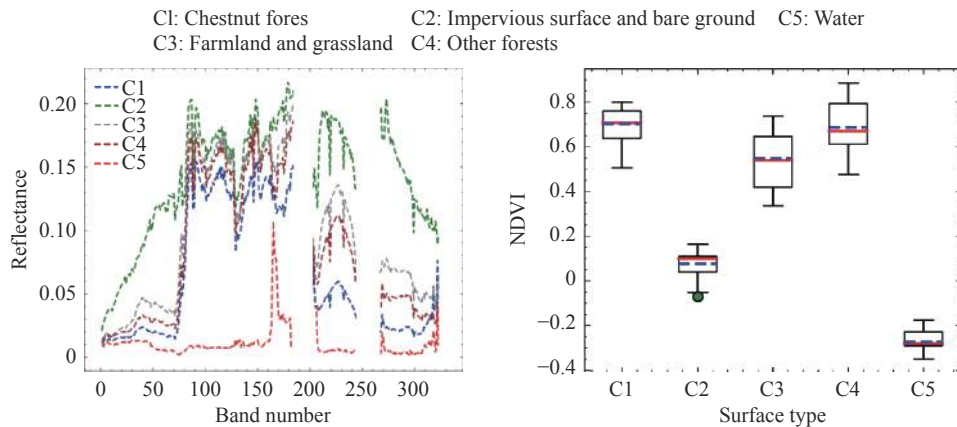


图 4 典型地物光谱曲线 (a) 及其 NDVI 分布箱形图 (b)

Fig. 4 Spectral curve of typical surface features (a) and NDVI distribution boxplot (b)

可以发现,不同地物类型 NDVI 分布区间有一定的差别. 农田和草地的 NDVI 分布值低于林地 (板栗林和其他林地), 但板栗林和其他林地的 NDVI 混淆程度较严重, 说明仅利用红和近红外波段无法实现板栗林的精细识别.

3.2 J-M 距离可分性

表 2 列出了利用 GF-5 和 S-2 全部波段计算的各地物类别之间的 J-M 距离. 根据 GF-5 高光谱数据得到的结果可知, 不透水面及裸土和水体与其他类别具有明显的可分析, J-M 距离大于 1.9. 板栗林、农田和草地、林地三者之间 J-M 距离在 1.5~1.8 范围内, 说明三者之间存在一定的光谱可分性, 但也存在一定程度的混淆. 其中, 农田和草地与板栗林和其他林地之间的 J-M 距离分别为 1.7260 和 1.7470, 而板栗林与其他林地的 J-M 距离为 1.5003, 可分性弱于其他地物类型之间的可分性. 由 S-2 多光谱数据计算的 J-M 距离可以发现, 农田和草地、不透水面及裸土、其他林地和水体之间的 J-M 距离要略高于 GF-5 高光谱数据, 这可能是由于 S-2 多光谱数据具有较高的空间分辨率, 以及 GF-5 高光谱数据在计算 J-M 距离时采用了所有波段作为输入, 没有去除冗余波段和低相

关波段等的影响. 此外, S-2 多光谱数据计算的板栗林和其他林地之间的 J-M 距离仅为 0.8491, 说明二者之间可分性较差.

3.3 混合式特征选择

图 5 给出了 OA 随波段数量的减少而产生的变化. 可以发现, 当波段数从 296 降低到 10 时, OA 的值由 0.9171 降低到了 0.9101, 表明较少数量的窄波段特征集合便可表征大部分地物类型之间的区别. 当输入特征维度从 290 降低到 160 时, OA 的变化幅度较小, 其中最大值为 0.9184, 最小值为 0.9180, 说明上述波段对分类模型的精度提升贡献度较小. 当输入维度由 160 降低到 140 时, OA 由 0.9184 降低到 0.9171, 说明上述 20 个波段对分类精度具有较大的影响. 当输入维度由 140 降低到 80 的过程中, OA 的变化比较平稳. 但随后下降到 50 维的过程中, OA 出现了上升趋势, 由 0.9169 上升到 0.9179, 说明上述 30 个波段影响分类精度的提升. 当维度由 50 继续下降时, OA 出现大幅度降低, 因此, 本研究选择 50 作为初始的特征子集维度.

确定了初始特征子集后, 采用 SBS 算法进行搜索, 最终构建了由 31 个波段组成的最佳特征子集 (见

表 2 基于 GF-5 和 S-2 数据的地物类型间 J-M 距离
Table 2 J-M distance between different surface features based on GF-5 data

类别	板栗林(GF-5/S-2)	农田和草地(GF-5/S-2)	不透水面(GF-5/S-2)	其他林地(GF-5/S-2)	水体(GF-5/S-2)
板栗林(GF-5/S-2)	0/0	1.7260/1.7861	1.9405/1.9501	1.5003/0.8491	1.9667/1.9999
农田及草地(GF-5/ S-2)	1.7260/1.7861	0/0	1.9022/1.9855	1.7470/1.7667	1.9662/1.9999
不透水面及裸土 (GF-5/ S-2)	1.9405/1.9501	1.9022/1.9855	0/0	1.9484/1.9807	1.9644/1.9999
其他林地(GF-5/ S-2)	1.5003/0.8491	1.7470/1.7667	1.9484/1.9807	0/0	1.9679/1.9999
水体(GF-5/ S-2)	1.9667/1.9999	1.9662/1.9999	1.9644/1.9999	1.9679/1.9999	0/0

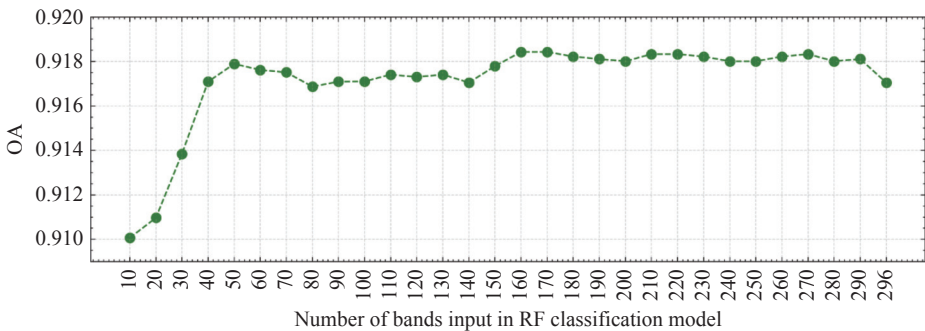


图 5 RF 算法排序下的总体分类精度 (OA) 随变量个数递减的变化趋势

Fig. 5 Trend of Overall Accuracy (OA) with decreasing number of variables under RF algorithm ranking

图 6), 波段号和中心波长分别为 50 (600.030 nm), 77 (715.430 nm), 89 (766.765 nm), 90 (771.043 nm), 107 (843.768 nm), 109(852.324 nm), 110(856.602 nm), 111 (860.880 nm), 114(873.713 nm), 118 (890.825 nm), 145(1006.880 nm), 151(1004.77 nm), 154(1030.050 nm), 160(1080.620 nm), 180(1249.210 nm), 202(1435.060 nm), 211(1511.090 nm), 215 (1544.880 nm), 216(1553.330 nm), 218(1569.230 nm), 219(1577.610 nm), 224(1620.210 nm), 226(1637.050 nm), 227(1645.470 nm), 229(1662.310 nm), 231 (1679.15 nm), 236(1721.250 nm), 237(1729.670 nm), 239(1746.500 nm), 242(1771.760 nm), 270(2007.500 nm). 从上述结果可以发现, 特征波段多集中在短波红外范围内. 表 3 为最终的验证集分类混淆矩阵, 分类模型的 OA 为 93.61%, Kappa 系数为 91.71%, 具

有较好的分类效果. 由表 3 可知, 该区域水体、不透水面及裸土和其他林地的识别精度较高, UA 和 PA 均达到 90% 以上. 板栗林与农田和草地的识别精度略低于上述三种类型, 其中板栗林的生产者精度为 81.25%, 用户精度为 73.03%. 这是由于高光谱数据受云的影响, 导致一些板栗林样本点无法使用. 此外, GF-5 高光谱数据的空间分辨率较低, 板栗林样本点的数量有限, 与其他林地的数量差距较大, 模型学习到的规律多集中于大样本的类别, 导致板栗林的识别精度受限. 尽管如此, 高光谱数据仍然表现出了在精细树种识别方面的有效性.

3.4 分类模型对比

表 4 为利用 S-2 多光谱数据的分类混淆矩阵, 其 OA 和 Kappa 系数分别为 85.91% 和 82%. 由表 4 可知, 水体、不透水面及裸土、农田和草地 3 类的分类

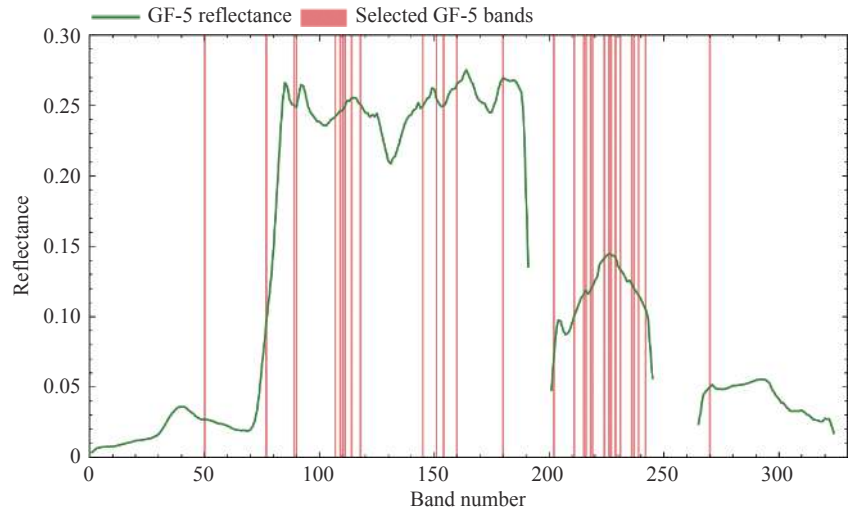


图 6 GF-5 波段选择结果
Fig. 6 GF-5 bands selection result

表 3 GF-5 验证集分类混淆矩阵
Table 3 Confusion matrix of GF-5 verification set

类别	水体	不透水面和裸土	板栗林	农田和草地	其他林地	PA/(%)
水体	200	0	0	0	0	100
不透水面及裸土	0	290	0	6	0	97.97
板栗林	0	0	65	2	13	81.25
农田和草地	0	10	3	175	7	89.74
其他林地	0	0	21	7	280	90.91
UA/(%)	100	96.67	73.03	92.11	93.33	—

注 UA代表用户精度, PA代表生产者精度.

精度较高, 且农田和草地对应的 UA 和 PA 显著高于 GF-5 高光谱数据. 这是由于密云水库水生态空间范围多以林地为主, 农田和草地的面积小且分布较破碎, 增加了 GF-5 样本选择的难度. 相比之下, S-2 多光谱数据的空间分辨率较高, 混合像元数量相对较少. 本研究中, GF-5 训练数据中农田和草地的样本个数为 400, S-2 训练数据中农田和草地的样本个数为 1000. 因此, 训练数据的数量和质量影响了 GF-5 的分类精度, 导致 GF-5 数据在农田和草地类型上的分

类精度显著低于 S-2 数据. 但 S-2 多光谱数据无法区分板栗林和其他林地, 使得其他林地的 PA 仅为 61.57%. 根据 GF-5 高光谱数据构建的最佳分类特征子集可知, 大部分波段集中在短波红外, 而 S-2 多光谱数据仅有 2 个短波红外波段, 不足以实现精细树种分类. 因此, 上述结果综合表明 GF-5 高光谱数据不仅可以实现 5 种地物类型的精细区分, 且能够实现板栗林的可靠识别. 最终, 密云水库水生态空间典型地物类型的空间分布如图 7 所示.

表 4 S-2 验证集分类混淆矩阵
Table 4 Confusion matrix of S-2 verification set

类别	水体	不透水面和裸土	板栗林	农田和草地	其他林地	PA/(%)
水体	500	0	0	0	0	100.00
不透水面及裸土	0	497	0	3	0	98.61
板栗林	0	0	0	0	300	0.00
农田和草地	0	7	0	484	9	98.37
其他林地	0	0	0	5	495	61.57
UA/(%)	100	99.40	0	96.80	99.00	—

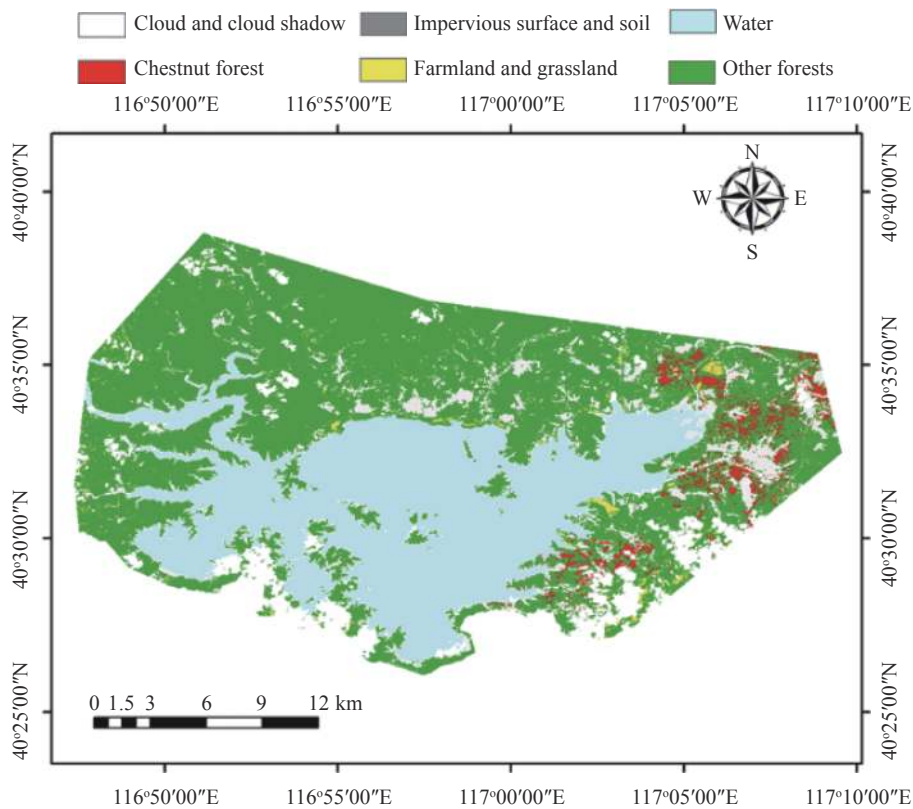


图 7 密云水库水生态空间典型地物类型分类结果
Fig. 7 Classification results of typical surface features in Miyun reservoir

4 结论

分析了 GF-5 号卫星高光谱数据对于北京市密云水库水生态空间范围内典型地物类型的精细识别能力. 研究表明, GF-5 高光谱数据可以较好区分水体、林地 (其他林地、板栗林)、农田和草地、不透水面及裸地, OA 达到 0.9361. 其中 GF-5 数据在短波红外范围内成功捕捉到了板栗林与其他林地反射特征的区别, 实现了二者的区分. 但由于空间分辨率较低 (30 m) 且受云覆盖影响导致板栗林的样本较少, 混合像元数量多, 但仍可以实现 81.25% 的生产者精度和 73.03% 的用户精度, 因此在提高影像质量 (无云影像) 和增加样本量的情况下, GF-5 高光谱数据可以实现更高精度的板栗林识别. 而具有 10 m 空间分辨率的 S-2 多光谱数据无法实现板栗林的识别, 也证明了高光谱数据在树种分类方面的有效性和优越性. 本研究提出的分类方法可在密云水库水生态空间范围内实现较为准确的精细地物类型分类, 为相关部门提供可靠的基础数据, 并且可结合相关数据实现该地区的水生态空间健康评估 (例如根据板栗林的面积等信息评估水土保持情况), 为水生态空间的合理规划和综合治理提供技术支持.

参考文献

- [1] ZHOU Guangjin, TONG Yali, WANG Lingqing, *et al.* Multi-dimensional identification technology and application of water ecological space and protection line in the territorial spatial planning[J]. *Journal of Natural Resources*, 2022, **37**(12): 3102-3117 (周广金, 童亚莉, 王凌青, 等. 国土空间规划中水生态空间及保护线的多维识别技术与应用 [J]. *自然资源学报*, 2022, **37**(12): 3102-3117)
- [2] ZHOU Qigang, LI Jian, MENG Haobin, *et al.* Evaluation on quality of important ecological space in Chongqing city based on remote sensing interpretation[J]. *Research of Soil and Water Conservation*, 2021, **28**(6): 292-300 (周启刚, 李剑, 孟浩斌, 等. 基于遥感解释的重庆市重要生态空间质量评价 [J]. *水土保持研究*, 2021, **28**(6): 292-300)
- [3] ZHU Qi, GUO Huadong, ZHANG Lu, *et al.* Classification of tropical natural forests in Hainan Island based on multi-temporal Landsat8 remote sensing images[J]. *Remote Sensing for Natural Resources*, 2022, **34**(2): 215-223 (朱琦, 郭华东, 张露, 等. 基于多时相 Landsat8 影像的海南岛热带天然林类型遥感分类 [J]. *自然资源遥感*, 2022, **34**(2): 215-223)
- [4] WANG Linjiang, WU Bingfang, ZHANG Miao, *et al.* Winter wheat and rapeseed classification during key growth period by integrating multi-source remote sensing data[J]. *Journal of Geo-Information Science*, 2019, **21**(7): 1121-1131 (王林江, 吴炳方, 张森, 等. 关键生育期冬小麦和油菜遥感分类方法 [J]. *地球信息科学学报*, 2019, **21**(7): 1121-1131)
- [5] SU Hongjun, LIU Hao. A novel dynamic classifier selection algorithm using spatial-spectral information for hyperspectral classification[J]. *Remote Sensing for Land & Resources*, 2017, **29**(2): 15-21 (苏红军, 刘浩. 一种利用空间和光谱信息的高光谱遥感多分类器动态集成算法 [J]. *国土资源遥感*, 2017, **29**(2): 15-21)
- [6] HAN Wenjun, ZHANG Su, JIAO Quanjun, *et al.* Dominant tree species mapping based on multi-temporal CHRIS hyperspectral satellite data[J]. *Forest Inventory and Planning*, 2019, **44**(2): 1-6 (韩文军, 张苏, 焦全军, 等. 基于多时相 CHRIS 高光谱卫星数据的优势树种分类研究 [J]. *林业调查规划*, 2019, **44**(2): 1-6)
- [7] GUO Qiang, PENG Long. Hyperspectral classification based on 3D convolutional neural network and super pixel segmentation[J]. *Acta Optica Sinica*, 2021, **41**(22): 2210001 (国强, 彭龙. 基于三维卷积神经网络与超像素分割的高光谱分类 [J]. *光学学报*, 2021, **41**(22): 2210001)
- [8] BHOSLE K, MUSANDE V. Evaluation of deep learning CNN model for land use land cover classification and crop identification using hyperspectral remote sensing images[J]. *Journal of the Indian Society of Remote Sensing*, 2019, **47**(11): 1949-1958
- [9] DING X H, ZHANG S Q, LI H P, *et al.* A restrictive polymorphic ant colony algorithm for the optimal band selection of hyperspectral remote sensing images[J]. *International Journal of Remote Sensing*, 2020, **41**(3): 1093-1117
- [10] LI J D, CHENG K W, WANG S H, *et al.* Feature selection: A data perspective[J]. *ACM Computing Surveys*, 2017, **50**(6): 94
- [11] WEI G F, ZHAO J, FENG Y L, *et al.* A novel hybrid feature selection method based on dynamic feature importance[J]. *Applied Soft Computing*, 2020, **93**: 106337
- [12] CHEN Z L, JIA K, XIAO C C, *et al.* Leaf area index estimation algorithm for GF-5 hyperspectral data based on different feature selection and machine learning methods[J]. *Remote Sensing*, 2020, **12**(13): 2110
- [13] ZHANG Lifu, ZHAO Xiaoyang, SUN Xuejian, *et al.* Comparison of fusion methods on GF-5 hyperspectral data[J]. *National Remote Sensing Bulletin*, 2022, **26**(4): 632-645 (张立福, 赵晓阳, 孙雪剑, 等. 高分五号高光谱数据融合方法比较 [J]. *遥感学报*, 2022, **26**(4): 632-645)
- [14] HAN Bingbing, CHEN Shengbo, ZENG Qinghong, *et al.* Time-series classification of Sentinel-1 data based on J-M distance[J]. *Science Technology and Engineering*, 2020, **20**(17): 6977-6982 (韩冰冰, 陈圣波, 曾庆鸿, 等. 基于 J-M 距

- 离的多时相 Sentinel-1 农作物分类 [J]. *科学技术与工程*, 2020, **20**(17): 6977-6982)
- [15] LIU Yongmei, GE Xinghua, DONG Xingzhi, *et al.* Spectral discrimination of flowers and leaves of major species in the degraded alpine meadow[J]. *Journal of Northwest University (Natural Science Edition)*, 2022, **52**(2): 159-168 (刘咏梅, 盖星华, 董幸枝, 等. 退化高寒草甸主要物种花与叶片的光谱识别方法 [J]. *西北大学学报 (自然科学版)*, 2022, **52**(2): 159-168)
- [16] NIU Quanfu, FU Jiankai, LU Ming, *et al.* Comparison of GF-6 WFV and Landsat8 OLI remote sensing image classification based on random forest[J]. *Geospatial Information*, 2022, **20**(8): 49-54 (牛全福, 傅键恺, 陆铭, 等. 基于随机森林的 GF-6 WFV 和 Landsat8 OLI 遥感影像分类比较 [J]. *地理空间信息*, 2022, **20**(8): 49-54)
- [17] BIAU G, SCORNET E. A random forest guided tour[J]. *Test*, 2016, **25**(2): 197-227
- [18] CAO Zetao, FANG Zidong, YAO Jin, *et al.* Loess landform classification based on random forest[J]. *Journal of Geo-information Science*, 2020, **22**(3): 452-463 (曹泽涛, 方子东, 姚瑾, 等. 基于随机森林的黄土地貌分类研究 [J]. *地球信息科学学报*, 2020, **22**(3): 452-463)
- [19] YANG Hongyan, DU Jianmin, RUAN Peiying, *et al.* Vegetation classification of desert steppe based on unmanned aerial vehicle remote sensing and random forest[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2021, **52**(6): 186-194 (杨红艳, 杜健民, 阮培英, 等. 基于无人机遥感与随机森林的荒漠草原植被分类方法 [J]. *农业机械学报*, 2021, **52**(6): 186-194)
- [20] AHMED K A, ALJAHDALI S, HUSSAIN S N. Comparative prediction performance with support vector machine and random forest classification techniques[J]. *International Journal of Computer Applications*, 2013, **69**(11): 12-16
- [21] RUAN F Q, QI J, YAN C H, *et al.* Quantitative detection of harmful elements in alloy steel by LIBS technique and sequential backward selection-random forest (SBS-RF)[J]. *Journal of Analytical Atomic Spectrometry*, 2017, **32**(11): 2194-2199
- [22] GUNASUNDARI S, JANAKIRAMAN S. A hybrid PSO-SFS-SBS algorithm in feature selection for liver cancer data[M]//*Power Electronics and Renewable Energy Systems*. New Delhi: Springer, 2015: 1369-1376
- [23] CHEN Zhulin, JIA Kun, LI Qiangzi, *et al.* Hybrid feature selection for cropland identification using GF-5 satellite image[J]. *National Remote Sensing Bulletin*, 2022, **26**(7): 1383-1394 (陈珠琳, 贾坤, 李强子, 等. 基于混合式特征选择的高分五号影像农田识别 [J]. *遥感学报*, 2022, **26**(7): 1383-1394)

作者简介



陈珠琳 女, 1994 年 7 月出生于山东省德州市, 现为中国林业科学研究院资源信息研究所助理研究员, 主要研究方向为资源环境遥感。
E-mail: chenzl@ifrit.ac.cn



李添雨 (共同第一作者) 男, 1991 年 8 月出生于北京市, 现为北京市水科学技术研究院流域生态环境研究所工程师, 主要研究方向为水生态监测评价、水生态保护修复技术等。
E-mail: lt0554@sina.com



张耀方 (共同第一作者) 女, 1991 年 2 月出生于四川省宜宾市, 现为北京市水科学技术研究院高级工程师, 主要研究方向为水生态系统评价、水生态保护修复技术研究、流域面源污染研究等。
E-mail: ucaszyf@163.com



贾 坤 (通信作者) 男, 1983 年 9 月出生于河北省唐山市, 现为北京师范大学地理科学学部教授, 博士生导师, 主要研究方向为生态环境遥感、地理信息分析、植被遥感等。
E-mail: jiakun@bnu.edu.cn