# 基于特征优选随机森林算法的 GF-2 影像分类

杨迎港[1] 刘培[*2] 张合兵[1] 张文志[1]
（1 河南理工大学测绘与国土信息工程学院，焦作 454003）
（2 海南省海洋与渔业科学院，海口 570100）

**摘 要** 基于随机森林算法（RF，Random Forest）对“高分二号”（GF-2）卫星遥感数据进行面向对象地表信息提取时存在如下不足：1）有限的光谱波段导致随机森林可选特征变量受限，影响分类器性能；2）面向对象影像分割尺度以经验判别为主，缺少定量化的判定标准。为了克服上述问题，文章提出了一种优化特征空间的随机森林分类算法。首先根据面向对象分割的理论方法，引入方差变化率，获取研究区影像的最优分割尺度；然后利用随机森林–平均精度减少模型（RF-MDA，Random Forest-Mean Decrease in Accuracy）与 K 折交叉验证算法（K-CV，K-Cross Validation），进行特征重要性排序并优化特征空间；最后，基于不同特征组合的随机森林分类算法进行面向对象分类，并对分类结果进行对比分析。结果表明，改进的基于特征优选随机森林分类算法的总体精度和 Kappa 系数分别为 93.44%和 0.928，优于原始 RF 算法。该方法能够有效提高 GF-2 卫星遥感影像在土地利用分类方面的精度，可为国土监测和管理提供技术支持和理论指导。

**关键词** “高分二号”卫星遥感影像 特征优选 随机森林 面向对象分类 最优分割尺度
**中图分类号:** N37; N39 **文献标志码:** A **文章编号**: 1009-8518(2022)02-0115-12
DOI: 10.3969/j.issn.1009-8518.2022.02.012

# Research on GF-2 Image Classification Based on Feature Optimization Random Forest Algorithm

YANG Yinggang[1] LIU Pei[*2] ZHANG Hebing[1] ZHANG Wenzhi[1]

（1 School of Surveying and Mapping Land Information Engineering, Henan Polytechnic University, Jiaozuo 454003, China）
（2 Hainan Academy of marine and Fishery Sciences, Haikou 570100, China）

**Abstract** To overcome the following boundaries of random forest object-based classification for high-resolution remote sensing images, that 1) limited spectral bands of high spatial resolution remotely sensed data has restricted the performance of random forest; 2) Segmentation scale of object-oriented method is based on empirical discrimination, which lacks quantitative criteria. In this paper, a random forest classification algorithm with optimized feature space is proposed. Firstly, according to the theory and method of object-oriented segmentation, the variance change rate is introduced to obtain the optimal segmentation scale of the image in the study area. Then, the Random Forest -Mean Decrease in Accuracy (RF-MDA) model and

收稿日期：2021-12-11
基金项目：国家自然科学基金（41601450，U1810203）；河南理工大学杰出青年基金（J2021-3）；江苏省水利科技基金（2020002）
引用格式：杨迎港, 刘培, 张合兵, 等. 基于特征优选随机森林算法的 GF-2 影像分类[J]. 航天返回与遥感, 2022, 43(2): 115-126.
YANG Yinggang, LIU Pei, ZHANG Hebing, et al. Research on GF-2 Image Classification Based on Feature Optimization Random Forest Algorithm[J]. Spacecraft Recovery & Remote Sensing, 2022, 43(2): 115-126. (in Chinese)

K-Cross Validation (K-CV) are used to rank the feature importance and optimize the feature space. Finally, the random forest classification algorithm based on different feature combinations is used for object-oriented classification, and the classification results are compared and analyzed. The results show that the overall accuracy and kappa coefficient of the improved random forest classification algorithm based on feature optimization are 93.44% and 0.928 respectively, which are better than the original RF algorithm. This method can effectively improve the accuracy of GF-2 remote sensing image in land use classification, and can provide technical support and theoretical guidance for land monitoring and management.

**Keywords** GF-2 satellite remote sensing images; feature optimization; random forest; object-based classification; optimal segmentation scale

## 0 引言

土地利用分类在土地动态监测、城市规划与管理、区域合理开发与保护等方面具有重要作用，是当前全球环境变化研究领域的重要内容之一[1]。随着我国社会经济快速发展，城镇化、工业化进程不断加快，人类活动范围不断扩大，建筑物用地规模、不透水地表比例大幅度增长，地表覆盖类型变化剧烈[2-3]。及时精准地获取市区土地利用分类信息，对于促进城乡转型、新型城市化、城乡一体化和区域社会经济的可持续发展具有重要意义[4-5]。高分辨率遥感影像具有空间分辨率高、覆盖范围大、更新速度快等优点，被广泛应用于土地利用信息提取[6-7]。随机森林（Random Forest，RF）作为集成统计机器学习的杰出代表，具有运算速率快、分类精度高、稳定性强、处理多维数据变量能力强等特点，广泛用于数据挖掘与分类。

由于随机森林的这些特性，在土地利用和地表覆盖分类的遥感应用领域也具有较强的适用性[8-9]。例如：文献[10]将随机森林应用于 Landsat ETM+影像的土地利用信息提取，通过实验证明随机森林非常适合于土地利用分类；文献[11]比较了基于航空高光谱图像对生态区分类的三种分类方法，得出随机森林在训练中比自适应增强（Adaboost）算法更快更稳定；文献[12]采用 121 个 UCI 数据集，通过一系列实验对比分析了 179 种分类算法的分类能力，结果表明随机森林算法具有最佳的分类性能；文献[13]在研究洪河湿地影像的分类时，将随机森林算法与最大似然分类法（MLC）和分类回归树（CART）算法进行了对比分析，结果显示随机森林的分类效果相对于后两者有巨大改善；文献[14]将随机森林算法运用于 Landsat 8 遥感影像数据的土地利用分类研究，并将分类结果与其他面向对象的机器学习分类算法进行对比研究，结果表明随机森林算法的分类精度和分类效率更具优势，更适于拥有复杂地类影像的地物信息提取。

随机森林的优势在于样本的多样性和随机特征的多样性，而主流高分辨率遥感影像通常只有蓝、绿、红、近红外 4 个有限的光谱波段，这就导致随机森林可选特征变量受限，影响分类器性能。比如文献[15]基于“哨兵 2 号”卫星多光谱成像仪的光谱特征，采用随机森林算法对印度农耕区进行土地分类信息提取，但是由于采用特征变量单一，对休耕地、甘蔗地等类型的分类精度较低；文献[16]采用“哨兵 2 号”影像数据，利用随机森林分类算法将托斯卡纳东部地区森林资源分为 4 类，由于以像元为研究尺度，单纯采用光谱特征而未考虑纹理特征、几何特征，造成分类精度较低；文献[17]利用 Landsat 8 OLI 光学遥感影像，基于随机森林算法进行土地覆盖监督分类，由于只采用光谱特征，导致农田分类精度较低。

为解决上述问题，本研究提出了基于特征优选的随机森林分类算法，通过多特征提取和特征优选在提高随机森林可用特征的基础上，同时降低特征冗余可能造成的干扰，达到提高分类器性能和地物提取精度的效果。研究以新郑市“高分二号”（GF-2）卫星影像数据为例，首先，通过统计影像同质性的局部方差 LV（Local Variance）及其变化率值 ROC（Rate Of Change）来确定最优分割尺度；然后通过 RF-MDA

模型度量各特征变量的重要性程度，并经过 K 折交叉验证获得最佳特征子集；最后应用优化的随机森林算法获取高精度的分类结果，并与只基于光谱特征及未经特征优选的随机森林分类结果进行对比，分析评价特征优选随机森林算法在 GF-2 卫星遥感影像土地利用分类中的优势。

## 1　研究区域和数据源

研究区位于河南省新郑市，经度范围为 113°46′50"E~113°48′50"E，纬度范围为 34°46′30"N~34°47′20"N，属暖温带大陆性季风气候，夏季炎热、冬季寒冷，季节温差变化明显，整体而言气候比较温和、适宜居住。研究区为城市区域，地势平坦，地物类型复杂多样，主要地类有建筑物、道路、阴影、林地、草地、裸地及水体。

影像数据源为 GF-2 卫星影像，这是我国自主研发的高分辨率民用卫星。研究采用新郑市 2020 年 12 月 21 日的 GF-2 影像，包含分辨率为 1m 的全色影像和分辨率为 4m 的多光谱影像[18]。影像预处理包括辐射定标、大气校正、正射校正和影像增强等，使用 Gram-Schmidt 影像融合方法获取分辨率为 1 m 的高品质多光谱影像，并通过影像裁剪获得最终的研究区范围。研究区影像如图 1 所示。

图 1　研究区影像

Fig.1　Image of the study area

## 2　研究方法

本文对预处理后的高品质 GF-2 卫星影像进行多尺度分割，并利用 ESP 工具获取最优分割尺度；然后基于影像分割对象提取光谱特征、几何特征、纹理特征及自定义特征，组成初始特征空间；接着利用特征优选算法对特征空间中所有特征进行优化选择，确定最佳特征子集；最后，以面向对象的随机森林分类算法为分类器，基于三组不同特征空间组合进行影像地物分类实验，并对分类结果进行精度验证与对比分析。

### 2.1　影像分割

#### 2.1.1　分割算法

影像分割是高分辨率遥感影像面向对象分类中至关重要的一步，是特征表达的前提，分割效果的好坏会影响最终的分类精度。研究应用分形网络演化方法（Fractal Net Evolution Approach，FNEA）对高分辨率遥感影像进行分割。该方法是一种多尺度分割算法，以异质性最小准则对像素进行合并，综合考虑光谱和形状信息，最终得到面向对象分类的最小单元——对象[19]。对一幅影像而言，当分割后对象内的异质性最低，对象间的异质性最大，且光谱、几何、纹理等信息表达完整时，即为最佳分割尺度。为快速获取最佳分割尺度，提高分类效率，采用 Dragut 等[20]提出的 ESP（Estimation Scale Parameter）算法计

算每个分割尺度对应分割层次上对象标准差的均值，引入方差变化率来表述层次间的变化，计算公式为

$$\mathrm{ROC}=\frac{L-L_{1-}}{L_{1-}}\times 100\% \quad (1)$$

式中 ROC 为方差变化率；$L$ 为目标层次的局部方差；$L_{1-}$表示相邻较低层次的方差。本研究利用 eCongnition Developer 软件中的 ESP 工具，生成局部方差（LV）和方差变化率（ROC）关系曲线。在构建了对象的 LV-ROC 曲线后，方差变化率的峰值即表示最优的分割尺度。

2.1.2 最优分割尺度的确定

分割尺度是一个描述同质对象异质度的阈值，选择的分割尺度越大，分割得到的对象数量就越少，易将不同类型的地物分割为一个同质对象，降低分类精度。反之，选择的分割尺度越小，分割得到的对象数量就越多，大大降低了运算效率。因此，对多尺度分割结果进行定量评价，找到最优分割尺度尤为重要[21]。

本文在进行最佳分割尺度选择实验时，采用自下而上多层次流程，起始分割尺度为 80（依据经验选取），形状因子和紧致度因子分别为 0.7 和 0.5，得到 LV 和 ROC 曲线如图 2 所示，其中蓝色曲线表示方差变化率，其峰值对应的即为各个层次下的最优分割尺度。由于影像的复杂性，最优分割尺度并不唯一，需要结合目视解译进行选择。本文通过观察 LV 和 ROC 曲线的变化，发现图中 ROC 的曲线变化率出现了大小不一的多个峰值，如 112、190、285 等。

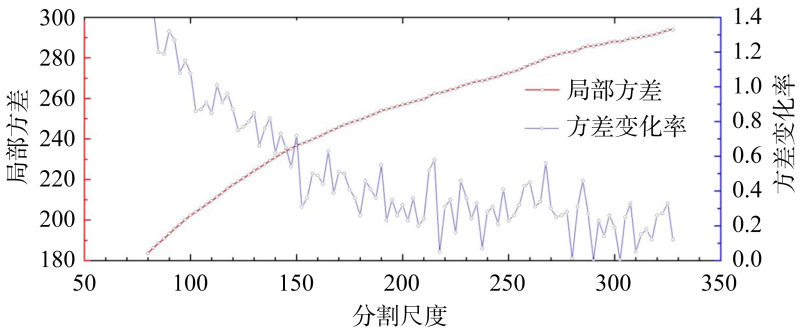
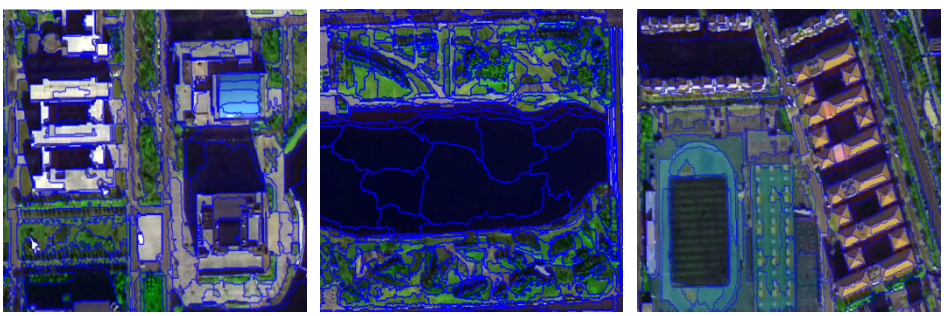



图 2 方差变化率的计算

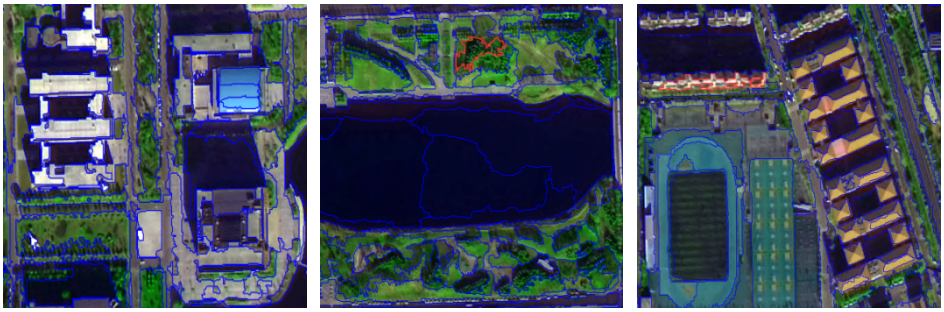
Fig.2 Calculation of variance change rate

对各峰值对应分割尺度进行分割实验，其分割效果局部对比如图 3 所示。由图 3 可知，当分割尺度为 112 时，影像分割的较为细碎，表现为过分割现象；当分割尺度为 285 时，不同地物（如建筑物和道路）之间分割不完全，同一对象内包含不同地物类型，表现为欠分割现象；当分割尺度为 190 时，建筑物和道路分割效果较好，各地物（如林地、草地、水体等）分割边界清晰，分割对象符合实际地物边界特征，满足实验需求。因此，本文选取分割尺度 190 为最优分割尺度。



（a）分割尺度 112

（a）Segmentation scale values are 112



（b）分割尺度 190

（b）Segmentation scale values are 190

（c）分割尺度 285

（c）Segmentation scale values are 285

图 3　各尺度下影像分割效果

Fig.3　Image segmentation effect at each scale

## 2.2　RF-MDA 和 K-CV 相结合的特征优选

### 2.2.1　RF-MDA-CV 特征优选模型

本文构建了一种随机森林–平均精度减少模型（RF-MDA）与 K 折交叉验证（K-CV）相结合的特征优选算法 RF-MDA-CV。首先通过 RF-MDA 算法对初始特征空间内的所有特征进行重要性排序，然后基于 K-CV 算法直接利用训练样本对不同特征组合进行循环交叉验证，根据交叉验证得分获取最佳特征子集。该方法克服了基于经验构建特征空间的盲目性，可以在保证随机森林可用特征基础的同时，降低特征冗余可能造成的干扰，达到提高分类器性能和地物提取精度的效果，在 GF-2 遥感影像的地物信息分类提取中表现出较好的适用性。RF-MDA-CV 特征优选算法的流程如图 4 所示。

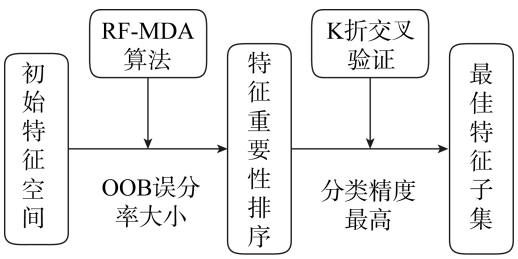



图 4　RF-MDA-CV 算法流程

Fig.4　Technical process of RF-MDA-CV algorithm

（1）特征提取

根据研究区影像的特点，在最优分割的基础上，提取光谱特征、纹理特征、几何特征和自定义特征等 63 个特征组成初始特征空间。具体如表 1 所示。

**表 1 63 种对象特征**

Tab.1 Characteristics of 63 objects

| 特征类别 | 特征名称 | 个数 | 特征类别 | 特征名称 | 个数 |
|---|---|---|---|---|---|
| 光谱特征 | 各波段均值 | 4 | 几何特征 | 面积 | 1 |
| | 波段比率 | 4 | | 长度 | 1 |
| | 各波段标准差 | 4 | | 长宽比 | 1 |
| | 亮度 | 1 | | 宽度 | 1 |
| | 最大差 | 1 | | 像元个数 | 1 |
| 纹理特征 | 相关性 | 5 | | 边界指数 | 1 |
| | 同质性 | 5 | | 非对称性 | 1 |
| | 标准差 | 5 | | 紧致度 | 1 |
| | 对比度 | 5 | | 密度 | 1 |
| | 平均值 | 5 | | 主方向 | 1 |
| | 差异性 | 5 | | 形状指数 | 1 |
| | 信息熵 | 5 | | 椭圆拟合 | 1 |
| | | | 自定义特征 | NDVI、NDWI | 2 |

（2）随机森林–平均精度减少模型（RF-MDA）

所有特征参与分类必然会产生信息冗余，大量冗余的特征会增加计算机运算负担，并造成“休斯（Hughes）”现象，通过特征优选排除冗余特征十分必要[23-24]。研究基于 RF-MDA 模型进行特征重要性评估，该模型通过袋外数据（Out of Bag，OOB）误分率对特征重要性进行度量。随机森林模型在构建 CART 决策树时，通过重复抽样得到一个数据集用于训练决策树，这时还有大约 1/3 的数据没有被利用，没有参与决策树的建立，这部分数据即为袋外数据。袋外数据可以用于评估决策树的性能，计算模型的预测错误率，称为袋外数据误差。RF-MDA 的思想就是打乱每个特征的特征值顺序，并且度量顺序变动对随机森林模型分类精度的影响，对于不重要的特征变量来说，打乱顺序对随机森林模型的精度影响不会太大，但对于重要的特征变量来说，打乱顺序就会降低模型的分类精度。

具体计算流程为：1）训练随机森林模型，选取袋外样本数据计算每棵树的 OOB 误差，记为 $\varepsilon_1$；2）将噪声随机添加到 OOB 数据样本的特征 $X$ 中，然后重新计算每棵树的 OOB 误差，并记录为 $\varepsilon_2$；3）如果随机森林中存在 $N$ 棵 CART 决策树，则特征 $X$ 的重要性 MDA($X$)可以表示为

$$\mathrm{MDA}(X)=\frac{1}{N}\sum(\varepsilon_1-\varepsilon_2) \tag{2}$$

（3）K 折交叉验证（K-CV）

根据特征的重要性排序不能直接获得最佳的特征组合，需要对不同特征组合的分类结果进行定量化对比分析。基于训练样本数据，采用 K-CV 算法进行不同特征组合的性能评定。K-CV 分类器的原理是将训练样本分成 $k$ 份，将每份数据作为一个测试集，剩余的 $k-1$ 份作为训练集，从而得到 $k$ 个评估模型。K-CV 算法以 $k$ 个模型的平均分类准确率作为性能指标。K 折交叉验证能充分利用所有样本数据，最终获得的结论也更具说服力。

K 折交叉验证的步骤如下：1）将全部训练集 $S$ 分成 $k$ 个不相交的子集，假设 $S$ 中的训练样本个数为 $n$，那么每一个子集有 $n/k$ 个训练样本，相应的子集称作{$S_1$，$S_2$，…，$S_k$}；2）每次从分好的子集中拿出一个作为测试集，其他 $k-1$ 个作为训练集；3）根据训练集训练出模型或者假设函数；4）把该模型应用于测试集，得到分类率；5）计算 $k$ 次所得分类率的平均值，作为模型或者假设函数的真实分类率。

2.2.2 特征变量重要性排序及最佳特征子集的确定

为提高分类算法的效率和精度，对初始高维特征空间进行优选。首先，通过 RF-MDA 模型获得各特征变量的重要性程度，并将特征变量按照重要性从大到小进行排序。重要性程度数值大代表特征的分类

能力较强，可以较好地区分不同的地物类型；重要性程度数值小的特征则与分类无关，属于冗余特征，过多的参与分类会导致模型的复杂度上升，降低分类效率和分类精度。根据 RF-MDA 特征重要性的计算原理，在 Jupyter 开发环境中采用 Python 编程得出 63 个特征的重要性得分，其中，Mean_Layer3 的重要性程度最高为 6.43%，这表示该特征对各地类的区分度较高，对模型预测结果准确度的贡献最大；GLCM_Correlation4 的重要性最低为 0.12%，这表示该特征对地物分类的作用微乎其微，在分类模型中无法有效区分不同地物类型，反而会增加模型运算负担；GLCM_StdDev3 的重要性为 1.81%，位列第 20；GLCM_Homogeneity2 的重要性为 1.03%，位列第 25，排序在 25 以后的特征重要性程度则均在 1%以下，这说明存在大量冗余特征，对初始特征空间进行优化很有必要。

K-CV 算法通过逐次减少重要性最小的特征进行交互运算，可快速获得 63 个不同特征组合下模型的分类性能得分，得分最高代表模型的分类效果最好，对应的特征组合即为最佳特征子集。考虑到交叉验证次数 $k$ 对随机森林模型分类精度的影响，本文基于训练样本数据设计了 $k$={2，3，…，10}共 9 种交叉验证模型，提取出与之对应的最高交叉验证得分进行对比，评估 9 种模型在最优条件下的分类性能高低，进而选出适用于研究区训练样本的最优交互次数。9 种交互次数与对应模型的最高性能得分如图 5 所示。由图 5 可知，不同交互次数对应的模型最高得分不同，当交互次数为 4 时，随机森林分类模型的性能最优，样本的预测效果最好，因此，研究选取 $k$=4 作为特征子集优选的交叉验证次数。

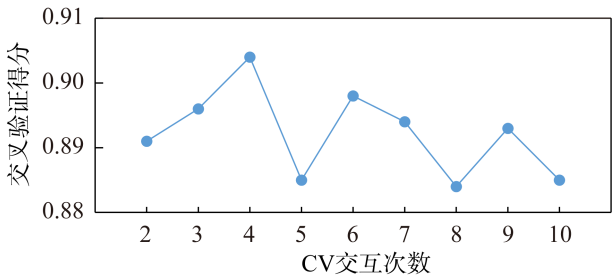



图 5　交互次数与对应的最高交叉验证得分

Fig.5　Interaction times and corresponding highest cross validation score

当交叉验证次数设为 4 时，特征个数与交叉验证得分关系如图 6 所示。由图 6 可以看出，在前 10 个特征中，随着特征数量的增多，模型交叉验证分数迅速提高，因为前 10 个特征对影像分类有很大的贡献率，均在 4%以上。随着特征个数的继续增加，分类精度并不是一味的提高，而是存在一定波动，这说明特征的选取并非越多越好，而是要恰到好处。特征太少，信息量不足，分类精度低；特征过多，则会产生冗余，不利于模型运算。根据研究的实验结果，当特征个数为 20 时，模型交叉验证分数最高，因此选取前 20 个重要特征作为最佳特征子集。

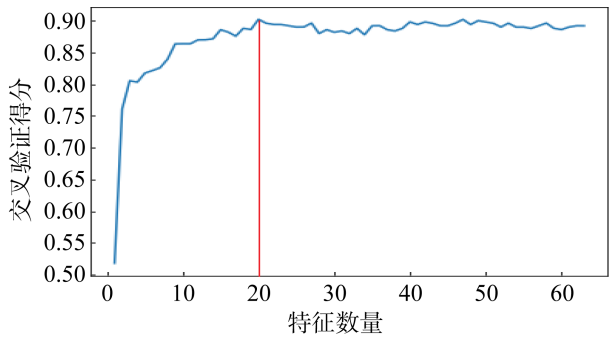



图 6　特征数与交叉验证得分的关系

Fig.6　Relationship between feature number and cross validation score

### 2.3 样本数据选取

样本数据的选取是影像分类的关键，样本的好坏直接影响最终的分类结果。考虑到样本数量对分类结果的影响，采用分层抽样的方法进行样本点选取，保证各地物拥有相同的样本量。根据研究区实际情况选择 7 类地物，分别为建筑物、道路、阴影、林地、草地、裸地及水体。结合研究区 2020 年土地变更调查数据和当地 Google Earth 高清影像及预处理后的 GF-2 影像，选取上述具有代表性的 7 类地物，形成样本集。采用分层抽样的原则抽取一部分作为训练样本，进行特征优选和模型训练，剩下的作为验证样本评价分类的精度。

### 2.4 随机森林分类

随机森林分类算法由 Leo Breiman[22]于 2001 年提出，其原理是将 Bagging 集成学习思想与随机子空间方法相结合。随机森林以决策树为基础学习器构建 Bagging 集成，并在决策树的训练历程中添加了随机属性的选取，以此加大分类模型之间的相异性，进而增强该模型的泛化能力以及预测能力。

随机森林分类算法可以概括为两个主要步骤：1）模型训练，首先利用自助法重采样技术从训练样本集中有放回的反复随机抽取 $N$ 个子集，相应的构造 $N$ 棵 CART 决策树，然后在每个决策树的全部节点处有放回的选择 $m$（$m$≤样本子集中总特征数）个特征变量，最后统计各变量所包含的信息量实现完全分裂；2）决策分类，对每一棵决策树的决策结果使用简单多数投票法进行归纳，最终分类结果由各决策树投票形成的分数而定。

### 2.5 精度评定指标

混淆矩阵又称误差矩阵，在精度评价中，主要用于比较分类结果和实测值之间的混淆程度[25]。本研究通过验证样本与分类结果生成混淆矩阵，选择目前普遍采用的总体精度、Kappa 系数、制图精度和用户精度作为评价指标对分类结果进行评价。

1）总体精度：表示所有被正确分类的对象数占所有验证样本数的比例，为混淆矩阵对角线上所有数值之和与样本总数之比。该指标计算公式为

$$A_{\mathrm{o}}=\frac{\sum_{i=1}^{t}P_{i,i}}{T} \tag{3}$$

式中 $A_{\mathrm{o}}$ 为总体精度；$T$ 为样本总数；$t$ 为总类别数；$P_{i,i}$ 为某一类正确分类的样本数。

2）Kappa 系数：表示分类结果与实际验证样本的匹配程度，取值范围为[0,1]，Kappa 系数越大，表示影像分类精度越高。计算公式为

$$K=\frac{T\sum_{i=1}^{t}P_{i,i}-\sum_{i=1}^{t}(P_{i+}\cdot P_{+i})}{T^{2}-\sum_{i=1}^{t}(P_{i+}\cdot P_{+i})} \tag{4}$$

式中 $K$ 代表 Kappa 系数；$P_{i+}$ 和 $P_{+i}$ 分别表示 $i$ 行和 $i$ 列的样本总数。

3）制图精度：表示某类样本对象在实际分类中被正确分类的可能性，由该类别所在列的对角线上的对象个数 $P_{i,i}$ 除以该列总的对象个数 $P_{+i}$ 而得到。制图精度 $A_{\mathrm{p}}$ 的计算公式为

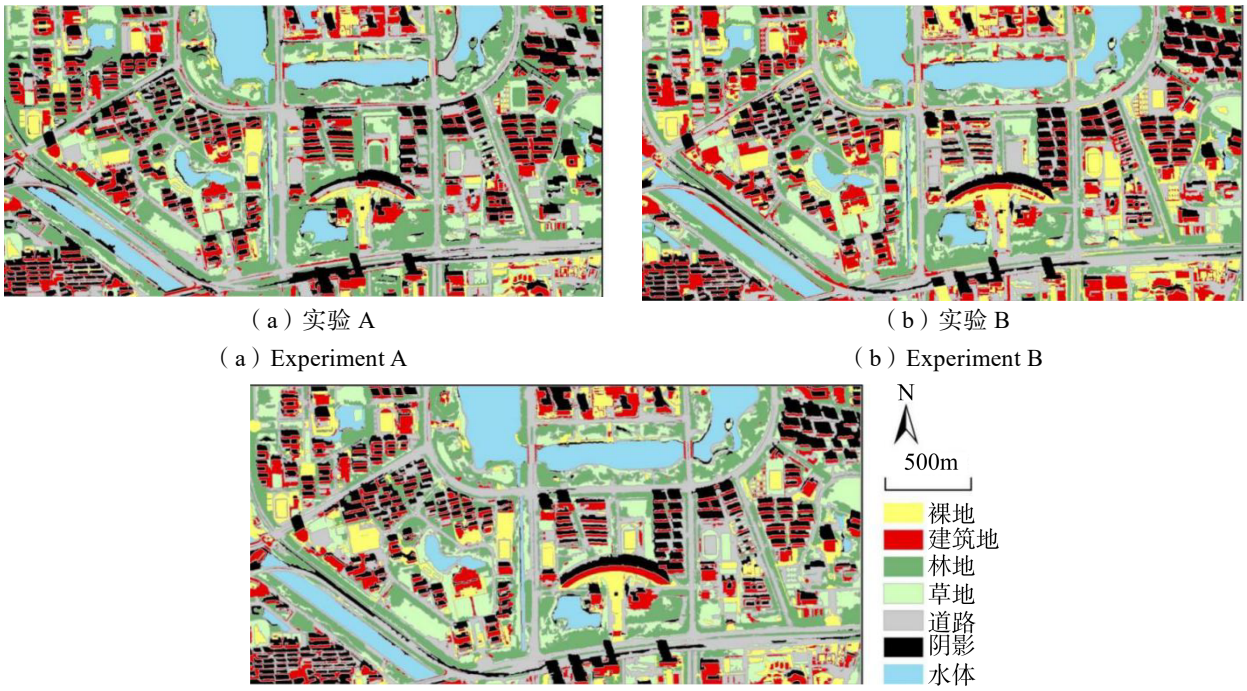
$$A_{\mathrm{p}}=\frac{P_{i,i}}{P_{+i}} \tag{5}$$

4）用户精度：表示某一个被分为某类地物的对象确实属于该地物的可能性，由该类别所在行的对角线上的对象个数 $P_{i,i}$ 除以该行总的对象个数 $P_{i+}$ 而得到。用户精度 $A_{\mathrm{U}}$ 的计算公式为

$$A_{\mathrm{U}}=\frac{P_{i,i}}{P_{i+}} \tag{6}$$

## 3　分类实验与结果分析

本文在研究区的 GF-2 卫星遥感影像上提取了 4 类特征变量共 63 个特征，其中：光谱特征 14 个，几何特征 12 个，纹理特征 35 个，自定义特征 2 个。为了验证 RF-MDA-CV 特征优选模型的优劣，设计了三种分类实验，在其他条件相同的情况下，选取三种不同的特征子集组合方式。其中，实验 A 采用研究区影像的红波段、绿波段、蓝波段、近红外波段共 14 个光谱特征变量作为特征子集；实验 B 采用研究区影像的光谱特征、几何特征、纹理特征、自定义特征共 63 个特征变量作为特征子集；实验 C 采用经过 RF-MDA-CV 模型优选的 20 个特征变量作为特征子集。三组实验都采用随机森林分类器进行面向对象分类，分类结果如图 7 所示。结合研究区影像进行目视判读可知，图 7（a）中建筑物和道路，道路和裸地，林地和草地之间均存在明显的误分现象；图 7（b）中裸地和道路存在着较多误分的情况,且有部分建筑物被误分为裸地；图 7（c）的分类结果相对较好，各地类之间没有明显的误分现象。




（a）实验 A
（a）Experiment A
（b）实验 B
（b）Experiment B
（c）实验 C
（c）Experiment C

图 7　三组实验的分类结果

Fig.7　Classification results of three experiments

利用验证样本对三组实验分类结果统计混淆矩阵，计算出用户精度、制图精度、总体精度、Kappa 系数作为分类精度评价指标。三组实验中 A 组实验的分类总体精度为 91.17%，Kappa 系数为为 0.8945；B 组实验的分类总体精度为 92.51%，Kappa 系数为 0.9105；C 组实验的分类总体精度为 93.44%，Kappa 系数为 0.928。

由上述结果可知，只采用遥感影像光谱信息作为特征子集的分类实验 A 和采用全部 63 个特征变量的实验 B，基于随机森林分类算法得到的分类精度均低于实验 C。实验 C 的总体精度数值比实验 A 和 B 分别高了 2.27%和 0.93%，Kappa 系数值提高了 0.0335 和 0.0175。说明相较于前两种特征组合的随机森

林分类，以 RF-MDA-CV 模型优选的 20 个特征变量作为特征子集的随机森林面向对象分类整体分类精度更优，经过特征优选的随机森林分类算法在 GF-2 遥感影像上的地物分类中表现出了最佳的分类性能。

三组实验中各地类的用户精度和制图精度统计结果如表 2 所示。

**表 2 精度统计表**

Tab.2 User accuracy statistics

| 实验方案 | 用户精度/% | | | | | | | 制图精度/% | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 草地 | 林地 | 道路 | 阴影 | 裸地 | 水体 | 建筑物 | 草地 | 林地 | 道路 | 阴影 | 裸地 | 水体 | 建筑物 |
| A | 83.86 | 91.50 | 83.20 | 100 | 96.22 | 100 | 78.27 | 87.45 | 87.50 | 94.03 | 100 | 77.91 | 100 | 82.25 |
| B | 87.34 | 91.83 | 83.33 | 100 | 96.12 | 99.54 | 92.30 | 87.44 | 91.46 | 93.52 | 98.38 | 78.33 | 100 | 94.97 |
| C | 81.86 | 93.53 | 88.32 | 100 | 97.86 | 100 | 93.49 | 89.82 | 87.86 | 93.52 | 100 | 87.80 | 100 | 95.75 |

由表 2 可知：在用户精度方面，实验 C 中的草地精度相对前两组实验有所降低，这是因为大部分草地和林地接壤。相对于林地而言，草地的色调更为均匀，对象轮廓边界更为平滑，区分两者的主要特征为各波段的均值、辐射率以及部分纹理特征，而优选特征子集中相关特征变量并不完整，由此导致部分草地被错分为林地，致使草地用户精度有所下降。除草地外，实验 C 中各地类的用户精度均大于实验 A 和 B。在林地方面，实验 C 的用户精度为 93.53%，相比实验 A 和 B 分别提高了 2.03%和 1.7%。在道路地类中，实验 C 的用户精度为 88.32%，相比实验 A 和 B 分别提高了 5%左右；在建筑物提取方面，依然是实验 C 的用户精度最高，为 93.49%，仅靠光谱信息作为数据源的实验 A 明显差于实验 B 和实验 C。从用户精度来看，实验 C 的整体分类效果最好，实验 B 次之，实验 A 最差。

在制图精度方面：在草地信息的分类提取中，实验 C 的制图精度最高，为 89.82%，比实验 A 和 B 分别提高了 2.37%和 2.38%；在林地方面，实验 A 和实验 C 中林地制图精度差别不大，但均低于实验 B，这是因为相对于其他地物而言，林地具有较为独特的群落结构，基于其顶部和侧面的形状及纹理特征可以较好的进行分类提取，而在实验 A 与实验 C 中相关特征变量存在缺失，导致部分林地被漏分，进而制图精度有所下降；在道路、阴影及水体的分类提取中，三组实验的制图精度差别不大；在裸地方面，实验 A 和 B 的制图精度相差不大，均明显低于实验 C；在建筑物提取方面，实验 A 的建筑物制图精度明显低于实验 B 和 C，说明仅仅具有光谱特征不能较好的提取建筑物信息，而实验 B 的制图精度低于实验 C，说明特征冗余也会降低建筑物分类提取的精度，经过特征优选的随机森林分类实验 C 对高分遥感影像中建筑物提取的制图精度最高。总体来看，实验 C 中各地物的整体分类提取效果最好，实验 B 次之，实验 A 最差。

通过三组实验的总体分类效果以及分类提取精度的对比分析可知，以 RF-MDA-CV 模型对 63 个初始特征变量进行优选得出的优选特征子集作为数据源，获取的分类结果要优于只以 14 个光谱特征和以 63 个特征变量为数据源的分类结果。改进的基于特征优选随机森林分类算法能够有效提高 GF-2 卫星遥感影像在土地利用分类提取方面的精度。

## 4 结束语

本研究以 GF-2 影像为数据源，通过计算影像方差变化率确定出最优分割尺度，利用 RF-MDA 和 K-CV 相结合的特征优选算法获取最佳特征子集，所提方法在研究区 GF-2 影像土地利用分类实验中取得了理想的结果。将经过特征优选的随机森林分类结果与原始随机森林算法分类结果进行对比分析，主要得到如下结论：1）通过计算方差变化率来确定影像最优分割尺度，可以减小分割尺度选择的盲目性，进而提高遥感影像面向对象分类的效率和精度。2）在遥感影像的分类过程中，分类精度和特征个数并不是

简单的正相关关系。依靠经验所选的特征集往往不是最佳的特征组合，通过特征优选找到最佳特征子集，可以排除人为因素所带来的冗余特征，进而提高分类精度。3 ) 基于特征优选随机森林算法的分类精度较高，总体精度和 Kappa 系数分别为 93.44%和 0.928，满足高分遥感影像地物信息分类提取的精度要求。

本文提出的 RF-MDA-CV 特征优选模型，可以在提高随机森林可用特征基础上同时降低特征冗余造成的干扰，提高了分类器性能和地物提取精度，找到了一条提高 GF-2 影像市区土地利用分类精度的新途径。但由于分类精度会受到研究区大小、样本数据、地物类型、影像数据源等诸多因素的影响，后续还需寻找更多的研究区域来证实这种方法的普适性。

## 参考文献(References)


[1] 马玥, 姜琦刚, 孟治国, 等. 基于随机森林算法的农耕区土地利用分类研究[J]. 农业机械学报, 2016, 47(1): 297-303.
MA Yue, JIANG Qigang, MENG Zhiguo, et al. Study on Land Use Classification in Agricultural Areas Based on Random Forest Algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(1): 297-303. (in Chinese)

[2] 谭莞怡. 江浙地区土地利用变化及驱动因素分析[J]. 安徽农学通报, 2021, 27(14): 118-122.
TAN Wuanyi. Analysis on Land Use Change and Driving Factors in Jiangsu and Zhejiang[J]. Anhui Agricultural Science Bulletin, 2021, 27(14): 118-122. (in Chinese)

[3] 宁全可, 谢世成, 仲臣, 等. 土地利用/覆盖变化的遥感分析[J]. 北京测绘, 2021, 35(7): 921-925.
NING Quanke, XIE Shicheng, ZHONG Chen, et al. Remote Sensing Analysis of Land Use/Cover Change[J]. Beijing Surveying and Mapping, 2021, 35(7): 921-925. (in Chinese)

[4] 黄天能, 张建中, 庞艳展. 边境口岸城市土地利用转型及其社会经济关联因子分析[J]. 广西财经学院学报, 2020, 33(4): 68-78.
HUANG Tianneng, ZHANG Jianzhong, PANG Yanzhan. Analysis of Land Use Transformation and Its Socio-economic Correlation Factors at Border Ports[J]. Journal of Guangxi University of Finance and Economics, 2020, 33(4): 68-78. (in Chinese)

[5] 李智礼, 匡文慧, 张澍. 近 70a 天津主城区城市土地利用/覆盖变化遥感监测与时空分析[J]. 遥感技术与应用, 2020, 35(3): 527-536.
LI Zhili, KUANG Wenhui, ZHANG Shu. Remote Sensing Monitoring and Temporal and Spatial Analysis of Urban Land Use/Cover Change in the Main Urban Area of Tianjin in Recent 70 years[J]. Remote Sensing Technology and Application, 2020, 35(3): 527-536. (in Chinese)

[6] WANG L J, ZHANG G M, WAMG Z Y, et al. Bibliometric Analysis of Remote Sensing Research Trend in Crop Growth Monitoring: A Case Study in China[J]. Remote Sensing, 2019, 11(7): 809-820.

[7] 杨贵军, 李长春, 于海洋, 等. 农用无人机多传感器遥感辅助小麦育种信息获取[J]. 农业工程学报, 2015, 31(21): 184-190.
YANG Guijun, LI Changchun, YU Haiyang, et al. Agricultural UAV Multi-sensor Remote Sensing Assisted Wheat Breeding Information Acquisition[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(21): 184-190. (in Chinese)

[8] 王李娟, 孔钰如, 杨小冬, 等. 基于特征优选随机森林算法的农耕区土地利用分类[J]. 农业工程学报, 2020, 36(4): 244-250.
WANG Lijuan, KONG Yuru, YANG Xiaodong, et al. Land Use Classification in Agricultural Areas Based on Feature Optimization Random Forest Algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2020, 36(4): 244-250. (in Chinese)

[9] 张红华, 赵威成, 刘强凯. 特征优选随机森林的土地利用分类[J]. 黑龙江科技大学学报, 2020, 30(5): 490-494.
ZHANG Honghua, ZHAO Weicheng, LIU Qiangkai. Land Use Classification of Random Forest With Feature Optimization[J]. Journal of Heilomgjiang University of Science and Technology, 2020, 30(5): 490-494. ( in Chinese )

[10] PAL M. Random Forests for Land Cover Classification[C]// Proceedings of 2003 IEEE International Geoscience and Remote Sensing Symposium. [S.l.]: IEEE, 2004: 3510-3512.

[11] CHAN C W, DESIRE P. Evaluation of Random Forest and Adaboost Tree-based Ensemble Classification and Spectral Band Selection for Ecotope Mapping Using Airborne Hyperspectral Imagery[J]. Remote Sensing of Environment, 2008, 112(6): 2999-3011.

[12] FERNANDEZ D M, CERNADAS E, BARRO S, et al. Do We Need Hundreds of Classifiers to Solve Real World Classification Problem[J]. Journal of Machine Learning Research, 2014, 15: 3133-3181.

[13] 王书玉, 张羽威, 于振华. 基于随机森林的洪河湿地遥感影像分类研究[J]. 测绘与空间地理信息, 2014, 37(4): 83-85, 93.
WANG Shuyu, ZHANG Yuwei, YU Zhenhua. Research on Remote Sensing Image Classification of Honghe Wetland Based on Random Forest[J]. Geomatics & Spatial Information Technology, 2014, 37(4): 83-85, 93. (in Chinese)
[14] 谷晓天. 基于机器学习的湟水流域土地利用/土地覆被分类研究[D]. 西宁: 青海师范大学, 2018.
GU Xiaotian. Research on Land Use/Land Cover Classification in Huangshui Basin Based on Machine Learning[D]. Xining: Qinghai Normal University, 2018. ( in Chinese)
[15] SAINI R, GHOSH S K. Exploring Capabilities of Sentinel-2 for Vegetation Mapping Using Random Forest[C]//ISPRS TC III Mid-term Symposium: Developments, Technologies and Applications in Remote Sensing. [S.l.: s.n.], 2018: 1499-1502.
[16] PULETTIN N, CHIANUCCI F, CASTALDI C. Use of Sentinel-2 for Forest Classification in Mediterranean Environments[J]. Annals of Silvicultural Research, 2018, 42(1): 32-38.
[17] 王笑影, 周玉科, 温日红. 基于 Landsat-8 影像和随机森林方法的土地分类研究[J]. 测绘与空间地理信息, 2020, 43(11): 1-3.
WANG Xiaoying, ZHOU Yuke, WEN Rihong. Research on Land Classification Based on Landsat-8 Image and Random Forest Method[J]. Geomatics & Spatial Information Technology, 2020, 43(11): 1-3. (in Chinese)
[18] 彭力恒, 刘凯, 朱远辉, 等. 旋转森林算法在 GF-2 卫星影像土地利用分类中的应用[J]. 航天返回与遥感, 2019, 40(1): 112-122.
PENG Liheng, LIU Kai, ZHU Yuanhui, et al. Application of Rotating Forest Algorithm in Land Use Classification of GF-2 Satellite Image[J]. Spacecraft Recovery & Remote Sensing, 2019, 40(1): 112-122. (in Chinese)
[19] 张华, 张改改. 面向对象的 GF-1 遥感影像多尺度分割研究[J]. 甘肃农业大学学报, 2018, 53(4): 116-123.
ZHANG Hua, ZHANG Gaigai. Object Oriented Multi-scale Segmentation of GF-1 Remote Sensing Image[J]. Journal of Gansu Agricultural University, 2018, 53(4): 116-123. (in Chinese )
[20] DRAGUT L, EISANK C, STRASSER T. Local Variance for Multi-scale Analysis in Geomorphometry[J]. Netherlands Geomor-phology, 2011, 130(3): 162-172.
[21] 王猛, 张新长, 王家耀, 等. 结合随机森林面向对象的森林资源分类[J]. 测绘学报, 2020, 49(2): 235-244.
WANG Meng, ZHANG Xinchang, WANG Jiayao, et al. Object Oriented Classification of Forest Resources Combined With Random Forest[J]. Acta Geodaetica et Cartographica Sinica, 2020, 49(2): 235-244. (in Chinese)
[22] BREIMAN L. Random Forests Machine Learning[J]. Journal of Clinical Microbiology, 2001(2): 199-228.
[23] 何云, 黄翀, 李贺, 等. 基于 Sentinel-2A 影像特征优选的随机森林土地覆盖分类[J]. 资源科学, 2019, 41(5): 992-1001.
HE Yun, HUANG Chong, LI He, et al. Random Forest Land Cover Classification Based on Sentinel-2A Image Feature Optimization[J]. Resource Science, 2019, 41(5): 992-1001. (in Chinese)
[24] WANG L J, DONG T F, ZHANG G M, et al. LAI Retrieval Using PROSAIL Model and Optimal Angle Combination of Multi-angular Data in Wheat[J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 2013, 6(3): 1730-1736.
[25] 赵存秀. 基于混淆矩阵的分类器性能评价指标比较[J]. 电子技术与软件工程, 2020(13): 146-147.
ZHAO Cunxiu. Comparison of Classifier Performance Evaluation Indexes Based on Confusion Matrix[J]. Electronics and Software Engineering, 2020(13): 146-147. (in Chinese)

**作者简介**

杨迎港，男，1997 年生，河南理工大学在读硕士研究生，主要研究方向为摄影测量与遥感。E-mail：764814256@qq.com。

**通讯作者**

刘培，男，1985 年生，河南许昌人，副教授，主要研究方向为遥感信息挖掘与信息处理。E-mail：liupei@hpu.edu.cn。

（编辑：夏淑密）