

白淑英,傅志强,谢 涛,等. 基于机器学习和哨兵 2 号遥感影像的棉花种植空间分布信息提取[J]. 江苏农业科学,2024,52(20):92-104.
doi:10.15889/j.issn.1002-1302.2024.20.012

基于机器学习和哨兵 2 号遥感影像 的棉花种植空间分布信息提取

白淑英^{1,2,3},傅志强¹,谢 涛¹,张雪红¹

(1. 南京信息工程大学遥感与测绘工程学院,江苏南京 210044; 2. 自然资源部遥感导航一体化应用工程技术创新中心,江苏南京 210044;
3. 江苏省协同精密导航定位与智能应用工程研究中心,江苏南京 210044)

摘要:为快速、准确、高效地获取棉花种植空间分布信息,提高棉花信息提取精度,基于机器学习的遥感图像识别方法,是有效解决问题的途径。以新疆维吾尔自治区乌苏市为研究区,利用哨兵 2 号遥感数据,选取 6 种常用植被指数、3 种红边植被指数,基于遥感植被指数变化曲线进行棉花特征时段选择,并分别采用梯度提升决策树、随机森林、支持向量机算法,通过 RF 特征优选,进行棉花种植区空间分布信息提取,并对提取结果精度验证。结果表明,机器学习方法(GBDT、RF、SVM)的总体分类精度分别为 0.92、0.91、0.88,Kappa 系数分别为 0.91、0.89、0.85;经 RF 特征优选后的机器学习算法(RF-GBDT、RF-RF、RF-SVM)的总体分类精度分别为 0.94、0.94、0.91,Kappa 系数分别为 0.93、0.92、0.88;经 RF 特征优选后的 3 种机器学习算法(RF-GBDT、RF-RF、RF-SVM)的总体精度分别比 RF 特征优选前(GBDT、RF、SVM)的总体精度分别提高了 0.02、0.03、0.03,Kappa 系数分别提高了 0.02、0.03、0.03。GBDT 作为一种集成的机器学习算法,在地物分类与棉花提取方面有着较好的应用效果,且经过特征优选的 RF-GBDT 算法精度更高。这表明在进行机器学习分类前,通过算法对输入特征进行重要性筛选,可有效避免因特征冗余造成的分类精度下降,可实现更高精度的棉花种植区域提取。

关键词:棉花提取;哨兵 2 号;机器学习;特征优选;遥感;GBDT

中图分类号:S127;TP79 **文献标志码:**A **文章编号:**1002-1302(2024)20-0092-12

棉花作为全球重要的经济作物,其种植区域的空间分布信息对于棉花产量估算和农业经济产值预测具有至关重要的影响。在这一背景下,遥感技术凭借其广泛的覆盖范围、强大的时效性和短周期的特点,成为了快速识别棉花种植区域的有效工具,有效地弥补了传统统计数据的滞后性。机器学习方法,由于其操作便捷和高精度的特性,在提取棉花种植区空间分布信息方面展现出了巨大的潜力。结合遥感数据,这些方法能够迅速、准确、高效地捕获棉花生产管理、面积统计与产量估算等关键信息。

目前,遥感植被指数法是提取棉花种植信息的常用方法。该方法主要利用时间序列植被指数数据,通过分析棉花在盛铃期的植被指数和光谱特征

的独特性,实现与其他地物的有效区分。在这一领域,已有众多学者取得了显著的研究成果。如吕绍伦等运用遥感云计算平台和哨兵 2 号影像,利用光谱和不同物候周期作物的植被指数变化构建了高精度的棉花提取模型^[1]。魏瑞琪等使用 TIMESAT 进行棉花像元的时间序列数据分析,获得了棉花生长曲线,并提取了种植区域^[2]。王文静等利用多时相的哨兵 2 号数据、NDVI(归一化植被指数)、反射率及纹理等,经特征优选后,对石河子市的棉花种植区域进行了提取研究^[3]。刘传迹等以 MODISEVI 数据为基础,应用 Double-Logistic 滤波对棉花生长曲线进行重构,得到棉花生长阈值,由此提取了棉花种植区域^[4]。Ren 等基于 GEE 和 Sentinel-2(哨兵 2 号)数据,结合兴趣面积指数、S-G 滤波等,建立时间序列表型分析方法,筛选棉花提取关键时相数据,将面向对象的信息提取方法与光谱特征和纹理特征相结合,对棉花分布信息进行提取^[5]。此外,有学者自行提出或选用了其它的提取指标。如 Wang 等基于棉花开铃期独特的冠层特征,提出了一种新的白铃指数(WBI)进行棉花种植区域提取

收稿日期:2024-01-10

基金项目:北京空间机电研究所航天进入减速与着陆技术实验室开放基金(编号:EDL19092304)。

作者简介:白淑英(1973—),女,内蒙古宁城人,博士,教授,从事遥感与地理信息系统在生态环境中的应用研究。E-mail:001462@nuist.edu.cn。

研究^[6]。He 等基于 Sentinel-2, 利用 MERRA-2 的逐时气象数据、棉花初级生产总值 (GPP) 和叶面积指数 (LAI) 等提取了棉花信息, 并估算了棉花产量^[7]。

由于机器学习在处理地理大数据和复杂特征分类方面具有明显的优势, 因此使用机器学习算法对遥感影像进行土地利用分类及作物信息提取, 已成为研究热点。机器学习法主要包括决策树、随机森林 (RF)、支持向量机 (SVM)、 k 平均算法 (k -means) 等, 其中在棉花提取方面用得较多的是决策树、随机森林 2 种方法。Li 等利用 CBERS01 和 HJ1B 卫星图像, 使用决策树算法计算棉花种植区域面积, 并分析了棉花种植区域的时空变化规律^[8]。田野等采用支持向量机和专家知识决策树分类法, 基于不同时期的卫星数据提取了棉花种植面积等信息^[9]。荷兰提出了基于多光谱和合成孔径雷达影像的集成学习算法, 通过各种分类器和特征, 成功识别了棉花种植区域^[10]。Fei 等提出了基于光谱、植被指数、纹理等多特征选择的随机森林特征选择算法和基于不同分类器的县尺度棉花提取方法, 评价了分类时间、特征重要性和分类器对棉花提取精度的影响^[11]。王汇涵等采用随机森林 (RF)、支持向量机 (SVM)、决策树 (CART) 进行棉花种植区域提取, 利用顺序向前选择 (SFS) 和偏最小二乘算法 (PLSR) 成功预测了棉花产量^[12]。美合日阿依·莫一丁等利用哨兵 2 号数据, 构建 NDVI 和红边归一化植被指数 ($RENDVI_{783}$) 时序数据, 采用 S-G 滤波法与袋外误差法对物候特征进行特征优选; 并利用 RF 进行分类和棉花提取^[13]。Rodriguez-Sanchez 等通过使用从正交图中提取的单个地块图像, 训练具有 4 个选定特征的 SVM 分类器来识别每个地块图像中的棉花像素, 对分类后的棉花像素进行形态学图像处理, 并进行聚类及预测^[14]。Hong 等基于 Sentinel-2, 利用光谱特征、植被指数特征和纹理特征创建了 7 种分类并生成 SVM 分类器, 实现了高精度的棉花提取^[15]。王利民等基于 5 m 空间分辨率的 Rapideye 影像, 采用红边、近红外波段反射率之和构建了棉花提取指数 (CEI), 结合同期水体、裸地 (含城镇建筑) 掩模处理, 分别采用最大似然分类方法和随机森林分类方法对影像进行分类和精度验证, 实现了棉花类型的识别^[16]。

在使用机器学习算法时, 须确保所选模型具备

良好的泛化能力, 这关乎模型在不同数据环境下的稳定性和准确性。为防止模型过度拟合, 选择合适的样本数据集并进行适当的参数调整是十分必要的。基于此, 本研究首先运用遥感植被指数法来获取棉花最佳研究时期的遥感影像数据。接着, 选取样本点, 并以植被指数、红边植被指数、地形等作为输入因子。然后采用梯度提升决策树 (GBDT)、随机森林 (RF) 和支持向量机 (SVM) 3 种算法, 通过 RF 算法进行特征选择, 并进行因子相关性分析, 旨在筛选出与棉花提取最为相关的因子, 以期实现更高精度的棉花种植区域提取。

1 数据源及数据预处理

1.1 研究区概况

研究区在新疆维吾尔自治区塔城地区乌苏市, 位于新疆维吾尔自治区西北部 (如图 1 所示), 地处 $43^{\circ}34' \sim 45^{\circ}17'N$ 、 $83^{\circ}24' \sim 85^{\circ}06'E$, 全市总面积 2.07 万 km^2 。乌苏市地处北温带干旱地区, 年均气温 7.3°C , 实际日照时数可达 $2\,600 \sim 2\,800 \text{ h}$, 年均降水量为 158 mm 。乌苏市年温差较大, 光照时间长, 降水量小, 适宜棉花的生长。由于特殊的气候环境, 乌苏市不仅是全国优质棉生产基地、还是重要的粮食和水果产地。

1.2 数据源

本研究使用的遥感数据是哨兵 2 号 (Sentinel-2) 高分辨率多光谱成像卫星的 L2A 级数据, 地面分辨率有 10 、 20 、 60 m 。选用 12 个波段作为棉花提取的特征因子, 输入机器学习算法中进行棉花信息提取 (表 1)。

本研究使用谷歌地球引擎 (GEE), 获取研究区 3—10 月可用的哨兵 2 号遥感影像共 14 幅 (表 2), 分别计算不同生长时期棉花的 NDVI 与近红外 (NIR) 等植被指数。发现当棉花处于盛铃期 (7—9 月) 时, NDVI、NIR 的像元亮度 (DN) 值高于其他地物, 易与其他地类进行区分, 此时是提取棉花信息的最佳时期。

1.3 数据预处理

GEE 是由 Google 云基建提供的云平台, 用于获取和处理遥感数据。它可以处理大规模的地球科学数据集, 特别是遥感影像数据, 而且支持全球尺度的在线处理、分析和可视化^[17]。GEE 提供了 Python API 和 JavaScript API 2 种语言接口^[18]。与传统的遥感数据处理工具相比, GEE 在遥感数据处

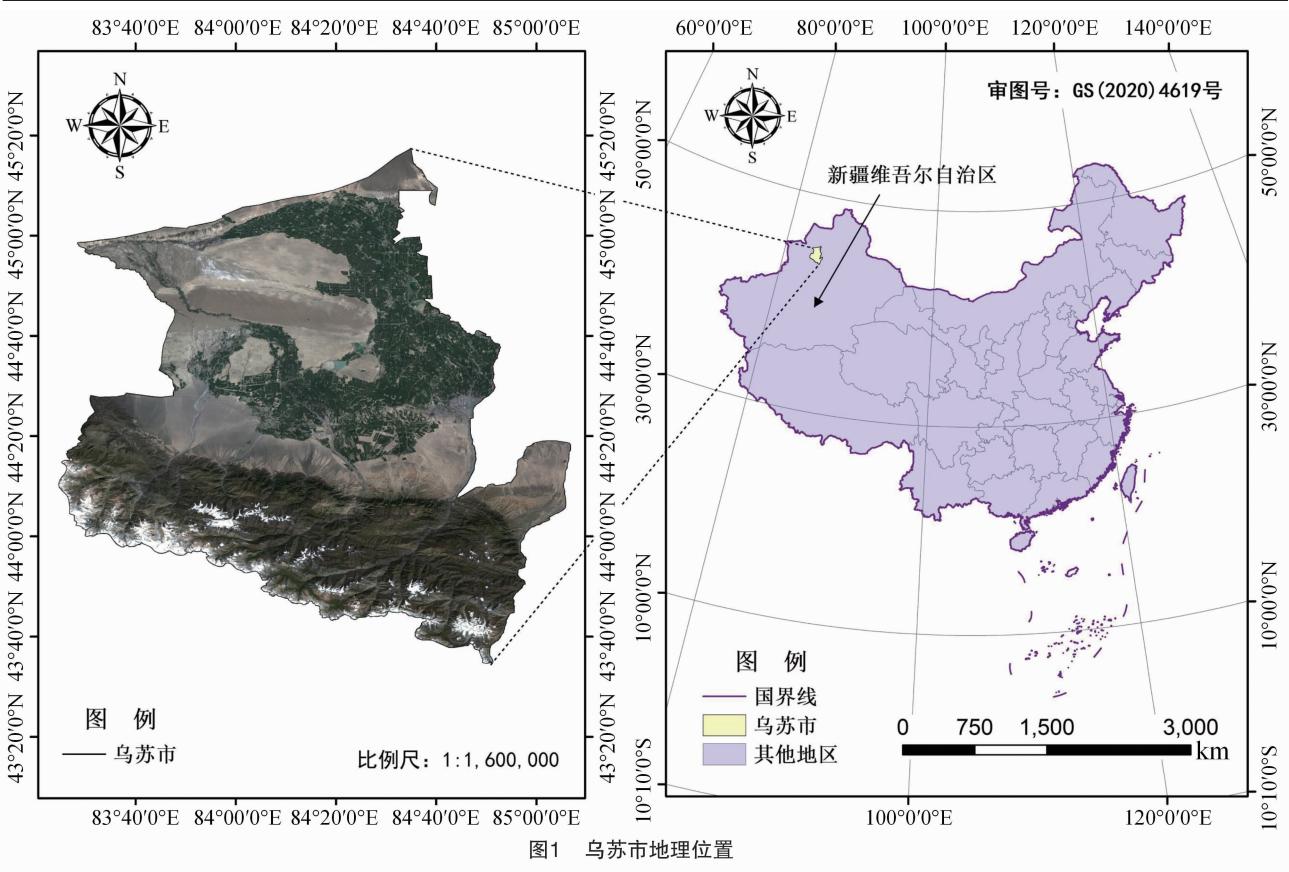


表 1 所使用的哨兵 2 号波段

编号	波段名称	波长 (nm)	空间分辨率 (m)
B1	海岸/气溶胶波段	433	60
B2	蓝光波段	490	10
B3	绿光波段	560	10
B4	红光波段	665	10
B5	红边波段	705	20
B6	红边波段	740	20
B7	红边波段	783	20
B8	近红外波段	842	10
B8A	红边波段	865	20
B9	水蒸气波段	945	60
B11	短波红外波段	1 610	20
B12	短波红外波段	2 190	20

理方面具有许多优势。

首先利用 GEE 平台进行遥感影像的下载、镶嵌、样本点的选取等,然后上传研究区范围矢量数据,利用 maskS2clouds 函数进行去云。选择哨兵 2 号数据集“COPERNICUS/S2_SR”,筛选日期与云量,并利用研究区感兴趣区域(ROI)进行裁剪操作,即可得到相应时间的遥感影像。

表 2 研究区 2021 年 3—10 月可用的 14 幅遥感影像信息

序号	日期	时间间隔 (d)
1	2021-03-17	—
2	2021-04-01	15
3	2021-04-16	15
4	2021-05-01	15
5	2021-05-16	15
6	2021-06-15	30
7	2021-06-30	15
8	2021-07-15	15
9	2021-07-30	15
10	2021-08-14	15
11	2021-08-29	15
12	2021-09-13	15
13	2021-09-28	15
14	2021-10-13	15

2 训练样本及特征选取

2.1 训练样本选取

模型训练所用的数据集也称为训练样本,是整个分类算法的基础。训练样本的质量直接决定了

分类的结果与精度。通过遥感影像目视解译,结合研究区的实际情况,利用 GEE 平台选择了 6 类训练样本,包括水体、建筑、裸地、棉花、林地和耕地(图 2),各类训练样本的数量见表 3。

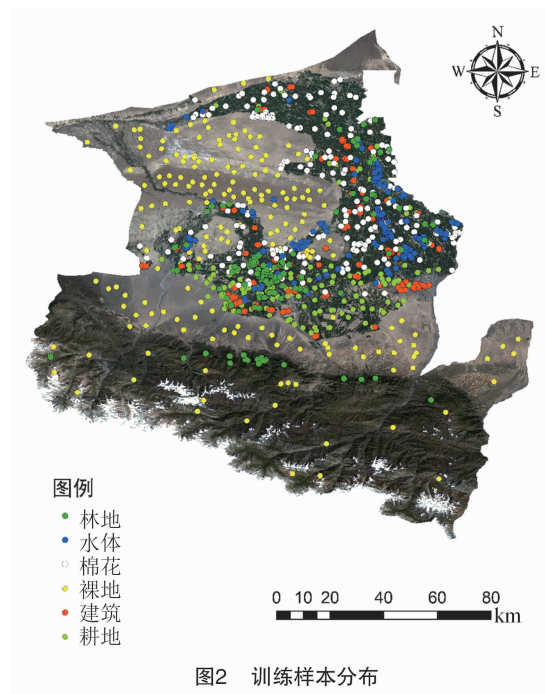


图2 训练样本分布

表 3 训练样本的类别与数量

类别	数量(个)
水体	183
建筑	202
裸地	193
棉花	211
林地	195
耕地	173
总计	1 157

为了量化各类样本间的可分离性,本研究采用了转换分离性和 JM 距离作为关键指标。转换分离性是基于马氏距离概念的统计度量,用于评估多变量分布之间的差异,特别适用于模式识别和图像处理领域。而 JM 距离是用于度量 2 个概率分布之间差异的统计量,广泛应用于模式识别和遥感影像分类。这 2 个参数的值在 0~2 之间^[19],大于 1.9 则样本之间的可分离性较好;小于 1.8 则可分离性较差,需重新选取。在本研究中,所选训练样本的可分离性见表 4。经分析,各样本类别间的 JM 距离均大于 1.8,这表明所选样本在特征空间中具有较好的区分度。这一结果为后续的遥感影像分类提供了坚实的基础,确保了分类过程的有效性和准确性。

表 4 6 种地类样本的可分离性

地类	JM 距离				
	建筑	裸地	棉花	林地	耕地
水体	1.925	2.000	1.997	1.947	1.943
建筑		1.998	1.954	1.946	1.921
裸地			2.000	2.000	1.999
棉花				1.960	1.923
林地					1.823

2.2 特征因子选取

特征因子中,植被指数包括:归一化植被指数(NDVI)、比值植被指数(RVI)、差值植被指数(DVI)、增强型植被指数(EVI)、归一化水体指数(NDWI)、土壤调节植被指数(SAVI)6 种。Sentinel-2 的优势在于其拥有 3 个红边波段,使其在识别植被信息方面非常有效,因此,本研究选取了红边植被指数(REP)、地面叶绿素指数(MTCI)、归一化差值红边指数(NDRE1)3 种红边植被指数。

纹理特征包括 7 类:均值、协同性、对比度、相异性、熵、角二阶距和相关性。采用灰度共生矩阵(GLCM)方法提取纹理特征,它通过描述像元对之间的空间结构特征及其相关性来定量描述遥感影像^[20],是应用最广泛的一种纹理特征提取方法。通过主成分分析方法,选取前 2 个主分量的 7 种纹理特征(共 14 个)作为纹理因子。选取了坡度、坡向、海拔 3 个地形因子。

本研究共选取了哨兵 2 号的 12 个波段、6 种植被指数、3 种红边植被指数、前 2 个主成分的 7 个纹理因子,以及 3 个地形因子,共 38 个特征因子作为机器学习算法的输入参数(表 5)。

3 研究方法

3.1 技术路线

由图 3 可知,首先,从 GEE 平台获取哨兵 2 号遥感数据,选取训练样本点和特征因子;其次,根据棉花夏季在 NDVI、NIR 上的特殊光谱曲线特征,进行棉花特征时段选择;随后,将所有特征因子输入,分别利用 GBDT、RF、SVM 机器学习算法,进行棉花信息提取;然后,利用 RF 进行特征优选,并将经优选的所有特征因子,再次输入 3 种机器学习算法,进行棉花提取;最后,比较几种方法的提取结果和精度,评价 RF 特征优选对于棉花提取效果和分类精度的影响。

表 5 特征因子及其计算公式

类别	名称	公式
植被指数	NDVI	$NDVI = \frac{B8 - B4}{B8 + B4}$
	DVI	$DVI = B8 - B4$
	RVI	$RVI = \frac{B8}{B4}$
	EVI	$EVI = 2.5 \times \frac{B8 - B4}{B8 + 6 \times B4 - 7.5 \times B2 + 1}$
	NDWI	$NDWI = \frac{B3 - B8}{B3 + B8}$
	SAVI	$SAVI = \frac{B8 - B4}{(B8 + B4 + L)} \times (1 + L), L = 0.5$
红边植被指数	REP	$REP = 705 + 35 \times \frac{\frac{B4 + B7}{2} - B5}{B6 - B5}$
	MTCI	$MTCI = \frac{B6 - B5}{B5 - B4}$
	NDREI	$NDREI = \frac{B6 - B5}{B6 + B5}$
纹理特征	均值	$\sum_i \sum_j p(i, j) \cdot i$
	协同性	$\sum_i \sum_j p(i, j) \cdot \frac{1}{1 + (i - j)^2}$
	对比度	$\sum_i \sum_j p(i, j) \cdot (i - j)^2$
	相异性	$\sum_i \sum_j p(i, j) i - j $
	熵	$\sum_i \sum_j p(i, j) \cdot \ln p(i - j)$
	角二阶距	$\sum_i \sum_j p(i, j)^2$
	相关性	$\sum_i \sum_j \frac{[i - \sum_i \sum_j p(i, j) \cdot i] \cdot [j - \sum_i \sum_j p(i, j) \cdot i] \cdot p(i, j)^2}{\sum_i \sum_j p(i, j) \cdot [i - \sum_i \sum_j p(i, j) \cdot i]^2}$
地形因子	坡度	
	坡向	
	海拔	

注: B1 为海岸/气溶胶波段反射率; B2 为蓝光波段反射率; B3 为绿光波段反射率; B4 为红光波段反射率; B5、B6、B7 均为红边波段反射率; B8 为近红外波段反射率; i 表示矩阵的行; j 表示矩阵的列; $p(i, j)$ 代表灰度级之间联合条件概率密度, 即在给定空间距离和方向时, 灰度以 i 为始点, 出现灰度级为 j 的概率。

3.2 棉花信息提取方法

3.2.1 基于遥感植被指数变化曲线的棉花特征时段选择 由图 4 可知, 棉花的生长周期主要包括 5 个阶段, 分别是出苗期、苗期、蕾期、盛铃期和吐絮期。当棉花处于盛铃期时, NIR 的 DN 值会大幅上升, 此时棉花 NIR 的 DN 值会明显高于其他农作物与森林、灌木等植被。所以, NIR 可有效地将棉花与各类作物、植被进行区分。因此, 可通过 NDVI 与 NIR 结合设定光谱阈值的方法, 获取棉花的特征时段。

3.2.2 基于梯度提升决策树的棉花信息提取方法 1999 年 Freiman 提出了梯度提升决策树 (GBDT), 该算法是一种迭代的决策树算法, 主要是由多棵 CART 树组成^[21]。GBDT 的主要思想是, 每

次建立的新模型均以上一个模型损失函数的负梯度为基础, 通过多个弱学习器合成为强学习器^[22]。GBDT 属于 Boosting 算法家族, 核心在于迭代地训练决策树, 以便每一棵树都能修正前一棵树的错误, 从而逐渐减少模型在训练集上的损失, GBDT 不仅可用于分类, 还可用于回归^[23]。GBDT 算法的流程如下。

对弱分类器进行初始化:

$$f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)。$$

式中: L 表示损失函数; γ 表示使损失函数最小化的值, 为常数。

对每次迭代 $m = 1, 2, \dots, M$ 。计算第 i 个样本第 m 轮的残差:

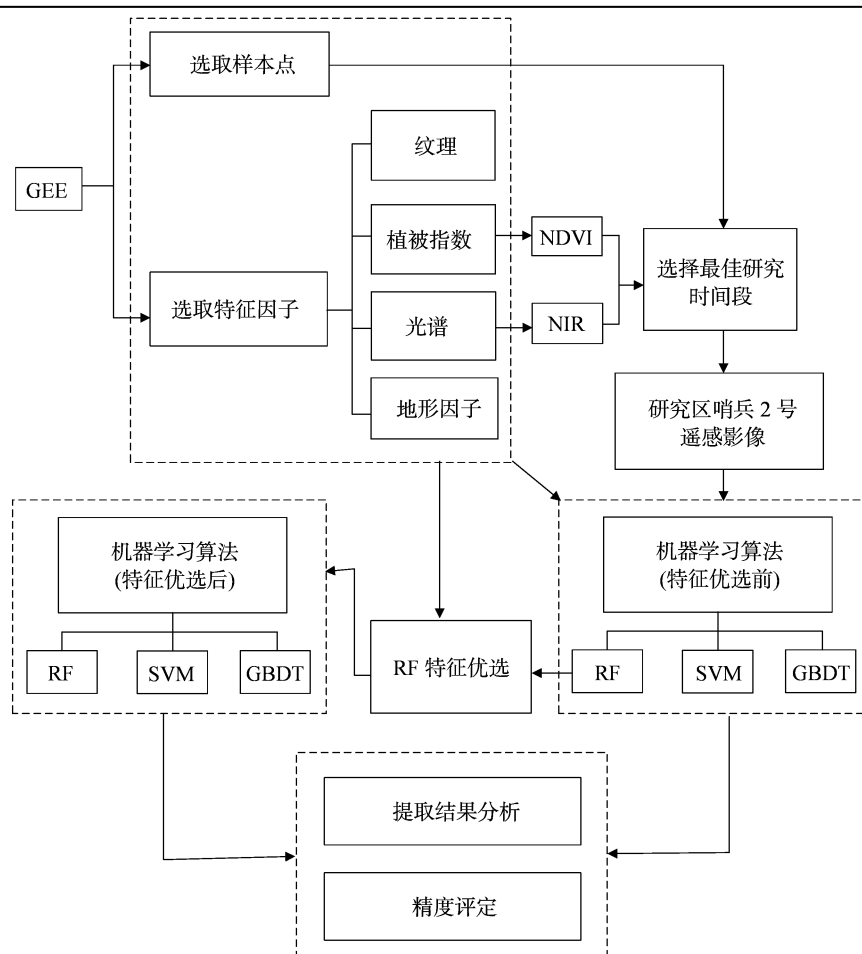


图3 技术路线

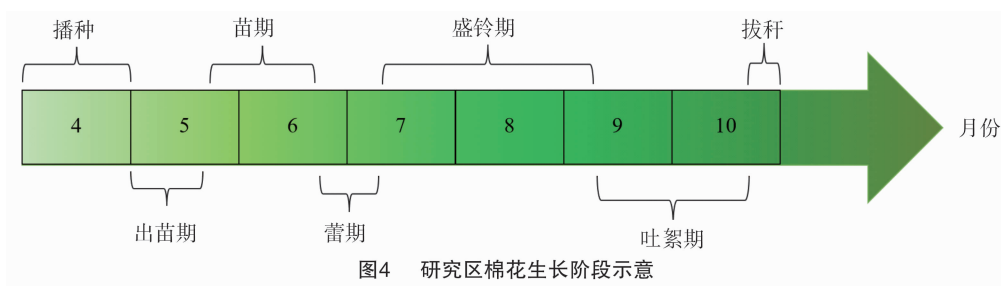


图4 研究区棉花生长阶段示意

$$\gamma_{im} = - \left\{ \frac{\partial L[y_i, f(x_i)]}{\partial f(x_i)} \right\}_{f(x) = f_{m-1}(x)}.$$

利用 (x_i, r_{im}) , 对 CART 回归树进行拟合, 其叶子节点将空间划分为独立区域。该回归树的叶子区域为 $R_{jm}, j=1, 2, \dots, J$ (节点个数)。

对于 $j=1, 2, \dots, J$, 计算最佳拟合值:

$$\gamma_{jm} = \arg \min_m \sum_{x \in R_{jm}} L[y_i - f_{m-1}(x) + \gamma].$$

持续更新强学习器:

$$f_m(x) = f_{m-1} + \sum_{j=1}^J \gamma_{jm} I.$$

根据上述流程得最终的强分类器为

$$f(x) = f_0 + \sum_{m=1}^M \sum_{j=1}^J \gamma_{jm} I.$$

式中: j 表示叶子区域; J 为叶子节点个数; 若公式为真则 I 是 1, 为假则 I 为 0。

3.2.3 基于随机森林的棉花信息提取方法 随机森林(RF)是由 Breiman 提出的一种基于决策树组合的方法, 是一种在样本空间、特征空间同时进行的集成学习算法。RF 中的每棵决策树都依赖于由训练确定的参数组成的随机向量, 每棵树在特征集中选择部分特征, 进行决策树的构造并贡献一票, 随后通过 Bagging 算法形成独立分布的训练样本集

进行训练,通过投票的方式获得最终的分类或预测结果^[24]。RF 分类的原理见图 5。

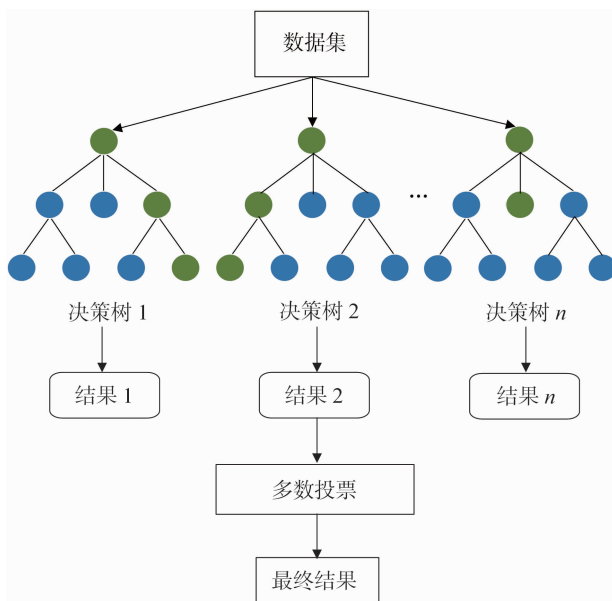


图5 RF 算法路线示意

RF 是一种非参数化的机器学习算法,它具有精确度高、不需要降维、训练速度快、无需剪枝、较少出现过拟合现象、能容忍一定的干扰和异常值,且能处理具有高维特性的输入样本的优点^[25]。因此,随机森林可用于各种数据类型的分类,并在性能上超越了传统统计方法及许多机器学习算法^[26]。

3.2.4 基于支持向量机的棉花信息提取方法 支持向量机(SVM)是 Vapnik 团队基于统计学 VC 维理论和结构风险最小化原理,开发的一种基于统计学习理论的机器学习算法^[27]。SVM 的特点是同时最小化经验误差和最大化分类间隔,其具有强大的非线性和高维数据处理能力,特别适用于小样本、非线性和高维模式识别问题^[28],同时也有效解决了“维数灾难”和“过度学习”等问题。

SVM 常采用的核函数有 3 种:线性核函数、多项式核函数以及径向基核函数,其表达式见表 6。本研究选用线性核函数作为 SVM 的核函数。

表 6 3 种核函数的表达式

核函数名称	表达式
线性核函数	$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
多项式核函数	$K(\mathbf{x}_i, \mathbf{x}_j) = (g\mathbf{x}_i^T \mathbf{x}_j + \gamma)^2, g > 0$
径向基核函数	$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\ \mathbf{x}_i - \mathbf{x}_j\ ^2}{\sigma^2}}$

式中: $\mathbf{x}_i, \mathbf{x}_j$ 表示输入空间的向量; g 表示常数; e 为自然常数; σ 为函数的宽度参数,控制了函数的径向作用范围。

此外,SVM 中可以设置惩罚系数 C ,其默认值为 1。 C 的取值影响了对分错样本的惩罚程度,较大的 C 值会导致在训练样本中获得更高的准确率,但可能会降低对测试数据的分类准确率,泛化能力较低。相反,减小 C 允许训练样本中存在一些误分类的样本,但可以提高模型的泛化能力。

3.2.5 基于 RF 特征优选的棉花信息提取方法

一般情况下,一份数据集有几十上百种特征,由于各特征的重要性不同,所以为了保证训练模型的精确度,应尽量降低复杂程度,筛选出最优特征以进行进一步研究。常见的特征优选方法有主成分分析、LASSO、RF 等。

随机森林(RF)用袋外数据(OOB)做预测。在训练过程中,约 1/3 的样本不被抽取,在每次重抽样建立决策树时,都会有一些样本未被选中,则可用这些样本进行交叉验证,这也是用 RF 进行特征优选的优点之一^[29]。这些袋外数据可用于计算特征重要性指标,进而进行特征选择。该方法无需做交叉验证,直接用 oob score 对模型性能进行评估。其基本原理为:

(1) 每棵决策树的袋外数据误差,记为 err_{OOB1} ;

(2) 然后随机对 OOB 所有样本的特征 i 加入噪声干扰,再次计算袋外数据误差,记为 err_{OOB2} ;

(3) 特征 i 的重要性为 $\frac{SUM(err_{OOB2} - err_{OOB1})}{N}$

(N 为树的棵数);若加入随机噪声后,袋外数据准确率大幅下降,则说明这个特征对预测结果有很大的影响,进而说明其重要程度比较高。

本研究将所有特征(38 个)输入 GBDT、RF、SVM 算法后,使用 RF 进行特征重要性的排序选择,再将经特征优选后的特征再次输入 3 种机器学习算法中,以探究 RF 特征优选前后,3 种机器学习算法的分类结果及精度变化。

3.3 精度评价方法

遥感影像分类结果的精度评价至关重要,本研究将 70% 的样本作为训练集,30% 的样本作为测试集,并采用混淆矩阵进行精度评估;混淆矩阵的列为参考数据,行为遥感数据的分类结果^[30]。评估分类效果的指标包括用户精度(UA)和生产者精度(PA),评价分类效果的指标包括总体分类精度(OA)和 Kappa 系数^[31]。这些精度指标从不同角度反映了分类的准确性。

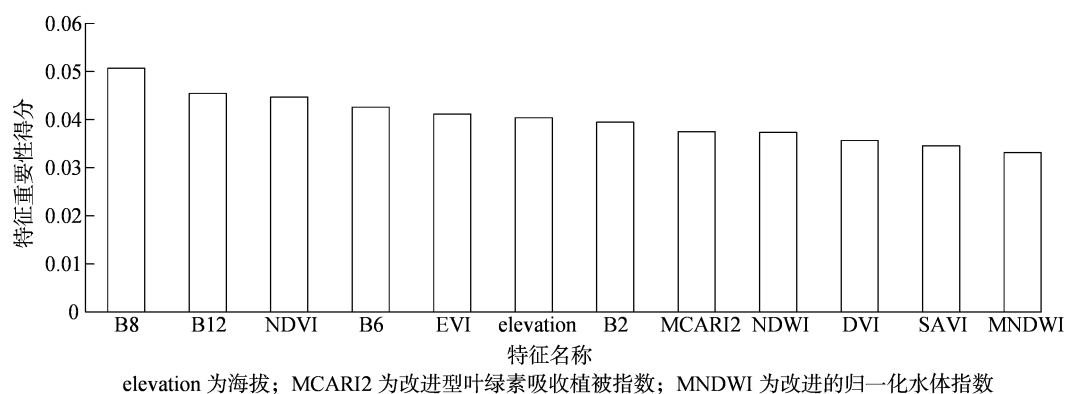
4 提取结果与分析

4.1 乌苏市棉花特征时段提取结果

在 GEE 平台中,通过选取适量各类典型地物的样本点进行 NDVI、NIR 时序分析,并结合棉花生长时序,可得到在棉花生长期(3 月 17 日至 10 月 13 日)各类典型地物的 NDVI、NIR 时序光谱曲线图。如图 6、图 7 所示,在 5 月中旬(出苗期),棉花的 NDVI 值从 0.1 大幅上升,且 NIR 值也开始逐渐上升;在 7 月中旬至 9 月中旬(盛铃期),棉花的 NDVI、NIR 值均在 0.6 左右,远高于其他地物。这也说明,7 月中旬至 9 月中旬的盛铃期,是观测棉花生长、获取棉田信息的最佳时期,也是进行棉花种植区域提取的最佳时期。

4.2 特征重要性排序结果

将 38 个特征因子,通过 RF 算法进行特征排序后发现,当特征数量达到 12 时,分类精度达到最高,大于 12 后呈下降趋势。因此,选择排名前 12 个特征构建训练模型的输入因子。这 12 个特征的重要性排名见图 8,B8、B12、NDVI 位列前三。



elevation 为海拔; MCARI2 为改进型叶绿素吸收植被指数; MNDWI 为改进的归一化水体指数

图8 特征重要性排名示意

4.3 GBDT、RF、SVM 3 种方法及特征优选分类结果比较

将特征优选前的 38 个因子,以及特征优选后的 12 个因子(图 8),分别利用 3 类机器学习算法,对棉花种植区域进行提取,即可得到各方法的棉花种植区域提取结果(图 9),以及各分类方法的精度(表 7、表 8)。

由表 7、表 8 可知,上述几种机器学习算法中,尽管部分地类的分类精度略低,但棉花的 UA、PA 始终在 0.90 以上,上述机器学习方法的棉花提取精度均达到优良水平。

RF-GBDT 和 RF-RF 方法的棉花提取精度较

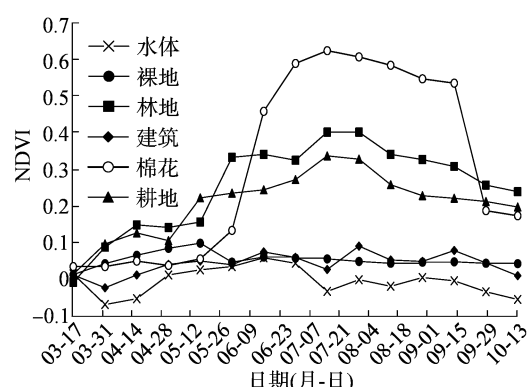


图6 各类地物NDVI 时序变化曲线

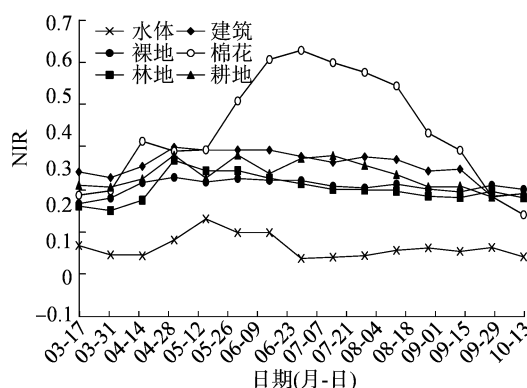


图7 各类地物 NIR 时序变化曲线

高,总体精度达到了 0.94。GBDT 的优势在于它通过每一次的残差计算增加了分错样本的权重,从而提高了泛化性能。然而,GBDT 对异常值较为敏感,而且由于分类器之间存在依赖关系,难以实现并行计算。但总体而言,GBDT 算法通常在一些方面优于 RF 算法。

使用 RF 分类器需要设置训练棵数。RF 特征优选前,棵数为 80~90 时(图 10),精度最高。随着棵数的不断增加,总体精度出现波动,当棵数为 100 时精度开始保持稳定。RF 特征优选后,棵数为 50 时,精度达到最大。随着棵数的不断增加,总体精度也出现波动,且总体呈下降趋势,当棵数为 140 时

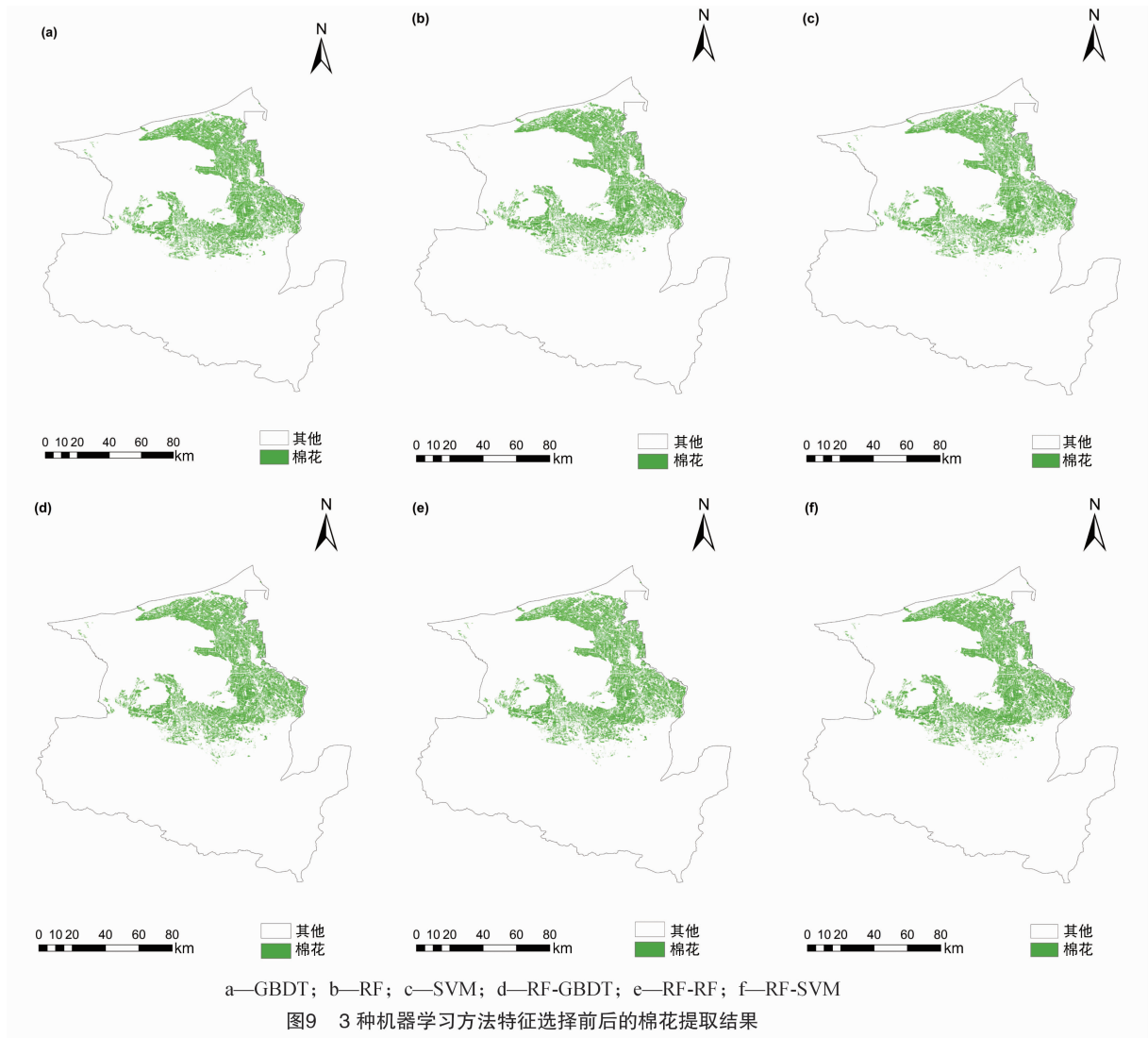


表 7 3 种方法在特征选择前的精度

地类	GBDT				RF				SVM			
	UA	PA	OA	Kappa 系数	UA	PA	OA	Kappa 系数	UA	PA	OA	Kappa 系数
水体	1.00	0.90	0.92	0.91	0.97	0.97	0.91	0.89	0.96	1.00	0.88	0.85
建筑	0.83	0.87			0.84	0.87			0.96	0.70		
裸地	0.96	1.00			0.97	0.93			0.77	0.96		
棉花	0.96	1.00			0.93	1.00			0.95	0.91		
林地	0.96	0.89			0.78	0.88			0.89	0.89		
耕地	0.85	0.88			0.90	0.75			0.77	0.86		

精度开始保持稳定。由图 10 可知,经 RF 特征优选的提取精度始终高于未经 RF 特征优选的提取精度。

此外,使用 SVM 设置惩罚参数 C 。RF 特征优选前, C 为 20 时精度最高。随着 C 的不断增加,总体精度总体呈下降趋势,并最终稳定在 0.8 左右。RF 特征优选后, C 为 15 时,精度最高,随后总体精度总体呈下降趋势,并最终也稳定在 0.8 左右(图 11)。

在经过特征优选之后,3 种机器学习算法(GBDT、RF、SVM)的分类精度均得到了提升,且在提取棉花种植区域方面表现出更高的准确性,极少出现漏提取现象。然而,在使用相同的遥感影像和训练样本的情况下,RF-SVM 方法在某些情况下将田间道路误识别为棉田,同时在部分区域出现了较为明显的“椒盐现象”(图 12),这导致其分类精度

表 8 3 种方法在特征选择后的精度

地类	RF - GBDT				RF - RF				RF - SVM			
	UA	PA	OA	Kappa 系数	UA	PA	OA	Kappa 系数	UA	PA	OA	Kappa 系数
水体	0.96	1.00	0.94	0.93	0.89	1.00	0.94	0.92	1.00	1.00	0.91	0.88
建筑	0.95	0.75			0.96	0.89			0.81	0.69		
裸地	0.88	1.00			0.96	1.00			0.90	0.94		
棉花	0.92	0.96			1.00	1.00			1.00	0.95		
林地	1.00	0.94			0.96	0.87			0.90	0.97		
耕地	1.00	1.00			0.81	0.89			0.77	0.82		

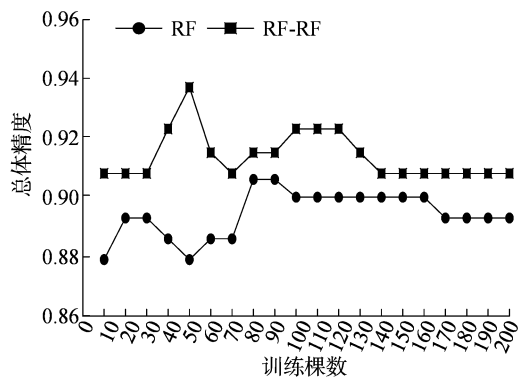


图10 RF 和 RF-RF 的训练棵数与精度关系示意

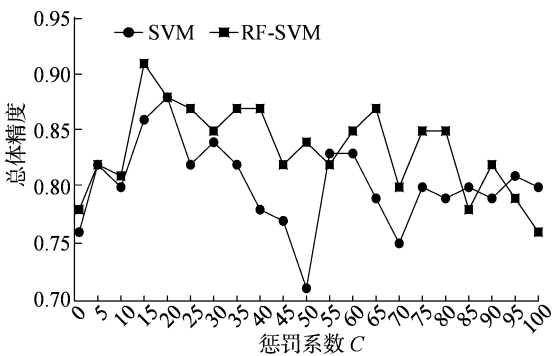


图11 SVM 和 RF-SVM 的惩罚系数 C 与精度关系示意

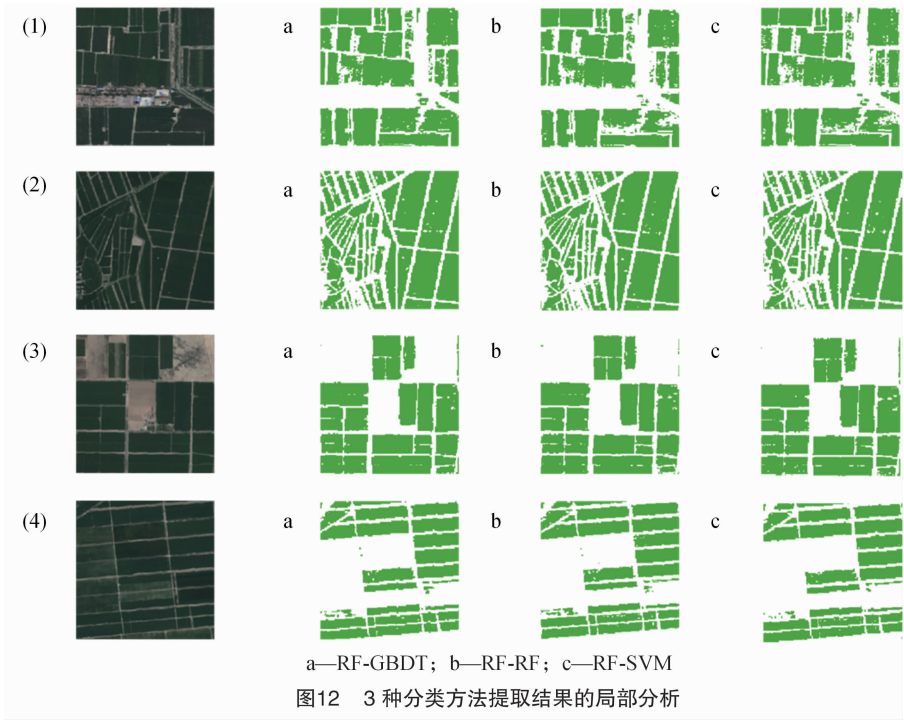


图12 3 种分类方法提取结果的局部分析

相较于其他 2 种算法略有下降。

RF 特征优化的主要目标是剔除冗余或不相关的特征,这不仅有效减少了特征的数量,而且提高了模型的精确度。特别是在处理如细窄田间道路等复杂地物特征时,结合 RF 特征优选和 GBDT 算法能够有效减少将道路误识别为棉田的情况。这

不仅提高了分类精度,也增强了模型在处理复杂地表特征时的鲁棒性。

在使用 RF 进行特征分析及通过特征重要性排序筛选和模型优化之后,分类精度提高了 2~3 百分点。RF 特征优化的主要目的是剔除多余或不相关的特征,这不仅减少了特征数量,而且还提升了模

型的精确度。尽管特征优选能够在一定程度上减少像素级别的分类错误,从而视觉上缓解了“椒盐现象”,但这并不代表可完全消除“椒盐现象”(图 13)。影响分类结果的因素不仅局限于特征选择,

还包括模型参数的配置、训练样本的选择,以及影像数据本身的质量和特性等,这些因素共同作用,最终决定了分类结果的准确性和可靠性。

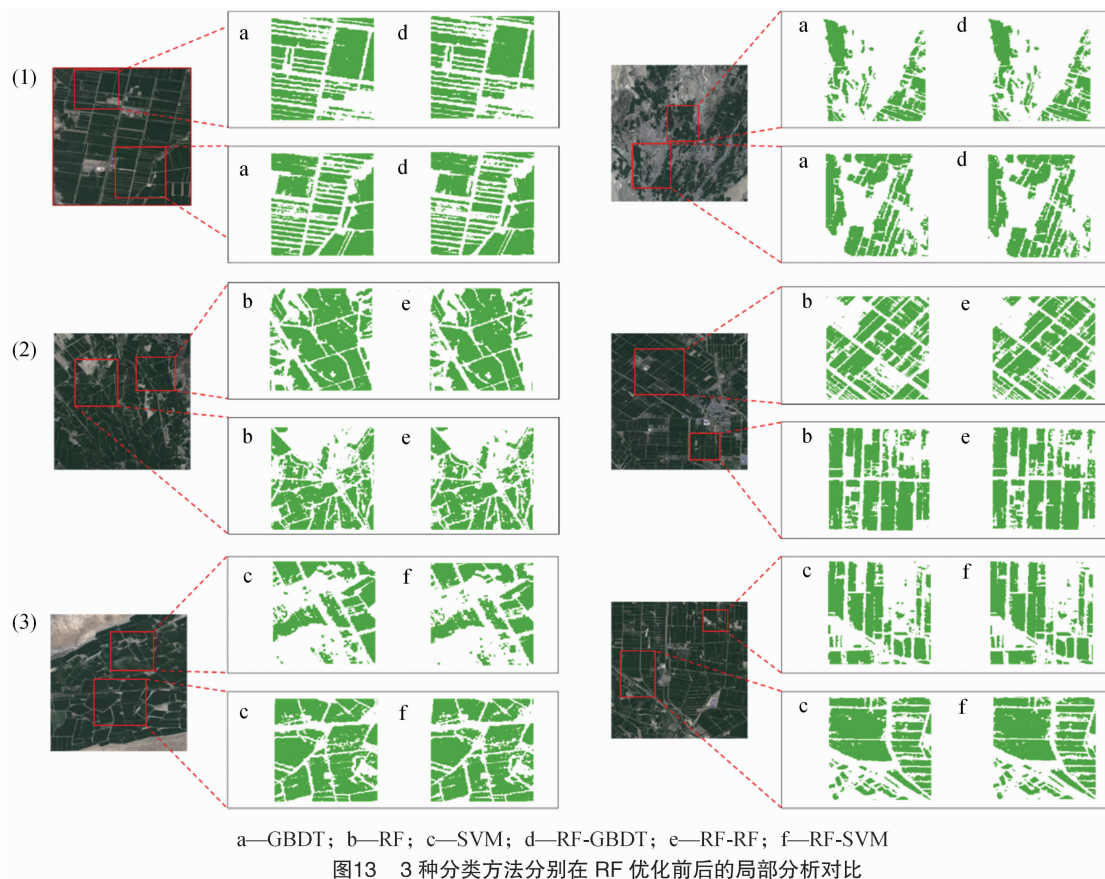


图13 3种分类方法分别在 RF 优化前后的局部分析对比

5 结论与讨论

本研究将遥感和机器学习方法相结合,基于哨兵 2 号影像,利用 GEE、python、ENVI、ArcGIS 等软件平台,对研究区的棉花种植区域进行了提取研究,并将几种方法的分类结果进行了对比。结果表明,RF-GBDT、RF-RF 在乌苏市的棉花信息提取中取得了较好的效果,GBDT、RF 次之,RF-SVM 与 SVM 的精度较低。研究结果表明:(1)根据研究区内典型地物的遥感植被指数变化曲线可知,7—8 月棉花的 NDVI、NIR 值远高于其他地物,此时是提取棉花信息的最佳时期。(2)通过算法发现,B8、B12、NDVI 等波段与棉花的相关性最高,说明这些波段特征对于棉花提取、估产等有重要意义。(3)经 RF 特征优选后的 3 种机器学习算法(RF-GBDT、RF-RF、RF-SVM)的总体精度分别比 RF 特征优选前(GBDT、RF、SVM)的总体精度分别提高了 0.02、

0.03、0.03, Kappa 系数分别提高了 0.02、0.03、0.03。由此可见,在进行机器学习分类前,通过算法对输入特征进行重要性筛选,可有效避免因特征冗余造成的分类精度下降,可实现更高精度的棉花种植区域提取。(4)使用多种机器学习方法对棉花种植区域进行提取,均取得了较好的分类效果。其中,RF-GBDT 算法的分类精度最高,其 Kappa 系数比 RF-RF 方法还提高了 0.01;由此可见,GBDT 算法作为一种集成的机器学习算法,在地物分类与棉花提取方面有着较好的应用效果。

本研究选取新疆乌苏市作为研究区域,运用 RF 算法对各类特征进行重要性排序,并最终筛选出前 12 个关键特征,输入至 3 种机器学习算法中,以提高分类精度并减少特征冗余。此外,本研究首次尝试将 GBDT 算法应用于棉花种植区域的提取,并取得了显著的成效。GBDT 在分类精度上高于 RF 和 SVM 的主要原因,在于其采用了增强学习策略,通

过迭代构建决策树并逐步减少残差,从而提升模型的准确性。相较之下,RF 采用多个决策树的平均或多数投票机制进行预测,而 SVM 则在特征空间中寻找最优分割面以区分不同类别。GBDT 专注于每轮迭代中减少分类误差,因此在某些情况下能够提供更精确的分类结果。然而,GBDT 的逐步优化策略也可能导致其在处理大规模数据或高维特征时出现过拟合的风险。

尽管本研究在提取精度上取得了一定成果,但仍存在提升空间。主要原因包括:(1)遥感图像获取条件的复杂性,包括光照变化、大气条件和传感器角度等,这些都影响遥感影像的质量,从而影响分类结果;(2)训练样本选取中混合像元的存在,导致建筑、水体、裸地等区域与棉花种植区域无法完全分离,影响分类精度;(3)尽管机器学习算法在遥感图像分类中表现出色,但它们在处理大规模数据或高维特征时,无法避免地会存在过拟合或泛化能力不足的问题。此外,本研究方法的选取部分基于前人在其他研究区的成果和文献经验,与前人所选训练样本的差异可能导致试验结果的误差。后续研究将致力于提高训练样本的准确性,结合遥感和野外实地考察选取棉花样本,避免因样本选取误差导致的精度下降,并尝试应用更多机器学习算法及神经网络(如 U-Net)算法,以进一步优化提取结果,提升精度。

本研究表明,通过使用 GEE 平台获取高分辨率遥感影像,选取训练样本,并应用机器学习方法提取棉花种植区域,能够有效提升提取精度。这为棉花种植区域的提取提供了新的解决方案和技术路径,为棉花面积估算研究提供了重要参考。

参考文献:

- [1] 吕绍伦,赵阳,陈万基,等. 基于遥感云计算的阿拉尔市棉花种植面积提取[J]. 棉花科学,2022,44(4):19-25.
- [2] 魏瑞琪,李林峰,仙巍,等. 利用 TIMESAT 软件和时间序列卫星影像提取新疆石河子棉花种植区域[J]. 湖北农业科学,2018,57(4):105-112.
- [3] 王文静,张霞,赵银娣,等. 综合多特征的 Landsat 8 时序遥感图像棉花分类方法[J]. 遥感学报,2017,21(1):115-124.
- [4] 刘传迹,金晓斌,徐伟义,等. 2000—2020 年南疆地区棉花种植空间格局及其变化特征分析[J]. 农业工程学报,2021,37(16):223-232.
- [5] Ren B Y, Zhou H Z, Shen H, et al. Research on cotton information extraction based on Sentinel-2 time series analysis[C]//2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics). Turkey: IEEE, 2019:1-6.
- [6] Wang N, Zhai Y G, Zhang L F. Automatic cotton mapping using time series of Sentinel-2 images[J]. Remote Sensing, 2021, 13(7):1355.
- [7] He L M, Mostovoy G. Cotton yield estimate using Sentinel-2 data and an ecosystem model over the southern US[J]. Remote Sensing, 2019, 11(17):2000.
- [8] Li M, Zhao G X, Qin Y W. Extraction and monitoring of cotton area and growth information using remote sensing at small scale: a case study in dingzhuang town of Guangrao County, China [C]//2011 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring. Changsha: IEEE, 2011:816-823.
- [9] 田野,张清,李希灿,等. 基于多时相影像的棉花种植信息提取方法研究[J]. 干旱区研究,2017,34(2):423-430.
- [10] 荀兰. 基于 Sentinel-1/2 卫星影像的棉花种植区识别方法研究[D]. 北京:中国科学院大学(中国科学院空天信息创新研究院),2022.
- [11] Fei H, Fan Z H, Wang C K, et al. Cotton classification method at the county scale based on multi-features and random forest feature selection algorithm and classifier[J]. Remote Sensing, 2022, 14(4):829.
- [12] 王汇涵,张泽,康孝岩,等. 基于 Sentinel-2A 的棉花种植面积提取及产量预测[J]. 农业工程学报,2022,38(9):205-214.
- [13] 美合日阿依·莫一丁,买买提·沙吾提,李金朝. 基于 Sentinel-2 时间序列数据及物候特征的棉花种植区提取[J]. 干旱区地理,2022,45(6):1847-1859.
- [14] Rodriguez-Sanchez J, Li C Y, Paterson A H. Cotton yield estimation from aerial imagery using machine learning approaches[J]. Frontiers in Plant Science, 2022, 13:870181.
- [15] Hong Y, Li D R, Wang M, et al. Cotton cultivated area extraction based on multi-feature combination and CSSDI under spatial constraint[J]. Remote Sensing, 2022, 14(6):1392.
- [16] 王利民,刘佳,姚保民,等. 基于 Rapideye 数据的棉花特征光谱指数构建及类型识别[J]. 中国农业信息,2019,31(5):25-37.
- [17] Gorelick N, Hancher M, Dixon M, et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone[J]. Remote Sensing of Environment, 2017, 202:18-27.
- [18] 郝斌飞,韩旭军,马明国,等. Google Earth Engine 在地球科学与环境科学中的应用研究进展[J]. 遥感技术与应用,2018,33(4):600-611.
- [19] Bruzzone L, Roli F, Serpico S B. An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection[J]. IEEE Transactions on Geoscience and Remote Sensing, 1995, 33(6):1318-1321.
- [20] Iqbal N, Mumtaz R, Shafi U, et al. Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms[J]. PeerJ Computer Science, 2021, 7:e536.
- [21] Friedman J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis, 2002, 38(4):367-378.

刘珊珊,刀 剑,张连刚,等. 基于随机森林的水稻信息提取与空间格局分析[J]. 江苏农业科学,2024,52(20):104-112.
doi:10.15889/j.issn.1002-1302.2024.20.013

基于随机森林的水稻信息提取与空间格局分析

刘珊珊¹,刀 剑^{2,3},张连刚¹,付 伟¹

(1. 西南林业大学经济管理学院,云南昆明 650224; 2. 云南农业大学植保学院,云南昆明 650500;
3. 云南省植物病理重点实验室,云南昆明 650500)

摘要:为准确了解岭南丘陵平原区水稻种植空间格局,以 Sentinel-2A 影像数据及耕地类型矢量数据为基础,采用随机森林(random forest,RF)对研究区水田范围内覆被地物进行分类,进而提取研究区水稻种植信息,以乡镇为空间单元尺度,分别从区域分布特征、空间破碎度、地形分布指数(P)3 个方面统计分析其种植空间格局。结果表明:(1)基于 RF 结合 Sentinel-2A 数据获得的组合植被指数(NDVI 和 NDRE₇₀₅)能够较好地对研究区水田掩膜后的影像进行覆被地物分类识别,分类的总体精度、Kappa 系数分别为 95.238%、0.926,其中水稻的用户精度最高,为 98.703%;根据提取结果得到水稻种植面积为 12 529.797 hm²,占比为 64.281%。(2)水稻种植区主要分布在石滩镇和中新镇,占比分别为 21.149%、16.982%;增江街道水稻种植面积最少,仅占 5.451%。(3)研究区水稻田块的破碎度在空间上的差异较为明显,破碎度高的水稻种植区域主要集中在研究区西部,而东部地区整体较低,在北部派潭镇、中部朱村街道、正果镇、荔城街道和增江街道以及南部的石滩镇,水稻种植地块破碎度相对较低,而中新镇、小楼镇和新塘镇反之。(4)水稻种植区分别在坡度 0°~8°、高程 0~32 m、半阳坡和阳坡(112.6°~247.5°)范围内处于优势水平, P 值远大于 1。研究成果可为制定区域国土管理制度和农业科学决策提供参考,对调整和优化水稻结构布局具有积极作用。

关键词:水稻;种植信息;哨兵 2 号影像;随机森林;空间格局;地形分布指数

中图分类号:S127 **文献标志码:**A **文章编号:**1002-1302(2024)20-0104-09

耕地是社会发展过程中最重要的土地覆盖类型之一^[1],它作为粮食、燃料等生产的主要载体,对人类至关重要。然而,伴随着人口持续增长、社会

经济快速发展等原因导致的城市扩张,土地不可持续利用、城乡迁移、耕地撂荒和土地边际化可能是对耕地影响最大的因素,进而决定了农作物的景观转换和破碎化^[2-4],致使农作物的空间格局受到影响。农作物的空间格局反映人类农业生产在空间范围内利用农业生产资源的状况,是了解农作物种类、结构、分布特征的重要信息,也是进行作物结构调整和优化的重要依据^[5]。

收稿日期:2023-11-03

基金项目:云南省科技计划基础研究专项(编号:202401CF070082);
校级科研启动专项(编号:110223010);校级人文社科科研专项
(编号:WKQN2309)。

作者简介:刘珊珊(1992—),女,山西大同人,博士,讲师,研究方向为
农业遥感。E-mail:274881086@qq.com。

[22]张海洋,张 瑶,田泽众,等. 基于 GBDT 和 Google Earth Engine 的冬小麦种植结构提取[J]. 光谱学与光谱分析,2023,43(2): 597-607.

[23]卓越,严海军. 基于梯度提升树算法的玉米施肥模型构建[J]. 水资源与水工程学报,2020,31(4):223-228,237.

[24]林志坚,姚俊萌,苏校平,等. 基于 MODIS 指数和随机森林的江西省早稻种植信息提取[J]. 农业工程学报,2022,38(11): 197-205.

[25]李旭青,刘世盟,李 龙,等. 基于 RF 算法优选多时相特征的冬小麦空间分布自动解译[J]. 农业机械学报,2019,50(6):218-225.

[26]Belgiu M, Drăguț L. Random forest in remote sensing: a review of applications and future directions [J]. ISPRS Journal of

Photogrammetry and Remote Sensing,2016,114:24-31.

[27]肖博林. 基于支持向量机的高光谱遥感影像分类[J]. 科技创新与应用,2020,10(4):22-24.

[28]费 浩. 综合多特征的县域尺度棉花种植面积遥感提取方法[D]. 阿拉尔:塔里木大学,2021:29-30.

[29]刘浩然,刘秀清,王春乐. 基于随机森林和超像素的极化 SAR 图像分类[J]. 国外电子测量技术,2021,40(9):29-35.

[30]黄鹏程,张明明,王新宇,等. 基于 Landsat-8 OLI 的西安市土地利用类型遥感分类研究[J]. 测绘与空间地理信息,2020,43(1):85-88,92.

[31]张 群. 基于高分遥感的黑方台滑坡识别[D]. 西安:长安大学,2017:28-30.