·农业信息与电气技术·

结合 Sentinel-2 影像和特征优选模型提取大豆种植区

张东彦[1]，杨玉莹[1]，黄林生[1]，杨 琦[1]，梁 栋[1]，
佘 宝[1,2※]，洪 琪[1]，姜 飞[3]

（1. 安徽大学农业生态大数据分析与应用技术国家地方联合工程研究中心，合肥 230601；
2. 安徽理工大学空间信息与测绘工程学院，淮南 232001；3. 宿州学院信息工程学院，宿州 234000）

摘 要：准确获取大豆的空间分布对于产量估计、灾害预警和农业政策调整具有重要意义，目前针对种植结构复杂地区所开展的大豆遥感识别研究鲜有报道。该研究以安徽省北部平原的典型大豆产区——龙山、青疃镇为研究区，基于 Sentinel-2 数据提出一种分层逐级提取策略的大豆识别方法。该方法首先构建决策树筛选规则，剔除研究区内非农田地物，获得田间植被的总体分布；然后生成 19 个候选特征因子，包括分辨率小于等于 20 m 的 10 个波段反射率以及 9 个植被指数。在典型地物类型样本的支持下，将 ReliefF 特征权重评估算法与随机森林（Random Forest，RF），BP 神经网络（Back-Propagation Neural Network，BPNN）和支持向量机（Support Vector Machine, SVM）相结合，分别构建 ReliefF-RF、ReliefF-BPNN、ReliefF-SVM 三种组合模型筛选出对于大豆识别最有效的特征，并基于布设在研究区内 6 个样方（大小为 1 km×1 km）的无人机影像提取得到的大豆分布来评估 3 种模型在大豆制图中的表现。结果表明，ReliefF-RF 模型表现最佳，基于该模型筛选出 7 个优选特征因子，大豆制图的总体精度介于 85.92%～91.91%，Kappa 系数在 0.72～0.81 之间，各个样方的提取效果均优于其他两种模型。此外，基于优选特征达到的提取精度明显高于原始波段反射率，虽然略低于全部 19 个特征的结果，但是数据量降低了 63.16%。该研究可以为农田景观破碎、种植结构复杂地区的大豆种植区提取相关研究提供有价值的参考和借鉴。

关键词：机器学习；模型；大豆；Sentinel-2；种植区提取；特征优选
doi：10.11975/j.issn.1002-6819.2021.09.013
中图分类号：S24 **文献标志码**：A **文章编号**：1002-6819(2021)-09-0110-10

张东彦，杨玉莹，黄林生，等. 结合 Sentinel-2 影像和特征优选模型提取大豆种植区[J]. 农业工程学报，2021，37(9)：110-119. doi：10.11975/j.issn.1002-6819.2021.09.013 http://www.tcsae.org
Zhang Dongyan, Yang Yuying, Huang Linsheng, et al. Extraction of soybean planting areas combining Sentinel-2 images and optimized feature model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021, 37(9): 110-119. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2021.09.013 http://www.tcsae.org

0 引 言

大豆是高蛋白食品、牲畜饲料的主要原料以及食用油的重要来源，在世界粮食生产中占有重要地位[1]。中国是世界大豆主产国之一，2019 年中国的大豆种植面积达 842.6 万 hm^2，位居全球第 5 位（http://www.fao.org/faostat/en/#data）。然而国内大豆产量远远无法满足生产生活需求， 2019 年大豆的进口依赖度高达 83.03%[2]，因此需要扩大种植规模，鼓励大豆生产。及时、准确地获取大豆种植面积及其空间分布对于长势监测、灾害评估等具有重要意义。以传统的农业调查方式来估算大豆种植面积通常费时费力，易受主观因素影响，结果数据亦无法提供空间分布信息。遥感技术可以更及时、高效和客观地实现大规模的农作物种植面积监测，且成本低廉[3]。

MODIS 数据具有较高的时间和光谱分辨率，适合大尺度的农作物遥感监测研究[4-5]。国内外一些研究表明基于 MODIS NDVI/EVI 时间序列数据生成的作物关键生育期的物候参数在农作物遥感识别中具有很好的表现[6-9]，如 Liu 等[10]利用随机森林方法（Random Forest，RF）提取位于美国玉米带的大豆和玉米，结果显示基于 MODIS 时间序列数据获取的 38 个物候指标在大豆和玉米的识别中具有一定优势。然而，由于不利天气及传感器工作状态等原因，无法保证时间序列数据的连续性，导致现实中的物候参数可能难以完整获取，仅采用物候信息来识别大豆具有较大的挑战。近年来，卫星传感器正逐步向高空间分辨率和多光谱的方向发展，它所增加的红边和短波红外波段在大豆种植区遥感提取中展现出了巨大潜力[11-12]。Zhong 等[13]发现在物候指标的基础上，短波红外波段（MODIS 波段 6，1 628～1 652 nm）的加入可以显著提高大豆和玉米的分离度。刘佳等[14]以黑龙江省五大连池中南部为研究区，利用最大似然方法探究红边和短波红外波段对于大豆和玉米的识别能力，结果表明引入

收稿日期：2021-02-25 修订日期：2021-04-30
基金项目：国家重点研发计划（2019YFE0115200）；农业生态大数据分析与应用技术国家地方联合工程研究中心开放课题（AE2018011）；安徽省高校优秀青年人才支持计划项目（gxyq2020001）；安徽省教育厅基金（KJ2019A0120）
作者简介：张东彦，博士，教授，研究方向为农业遥感与信息技术。
Email：zhangdy@ahu.edu.cn
※通信作者：佘宝，博士，讲师，研究方向为农业定量遥感。
Email：shebao518@aust.edu.cn

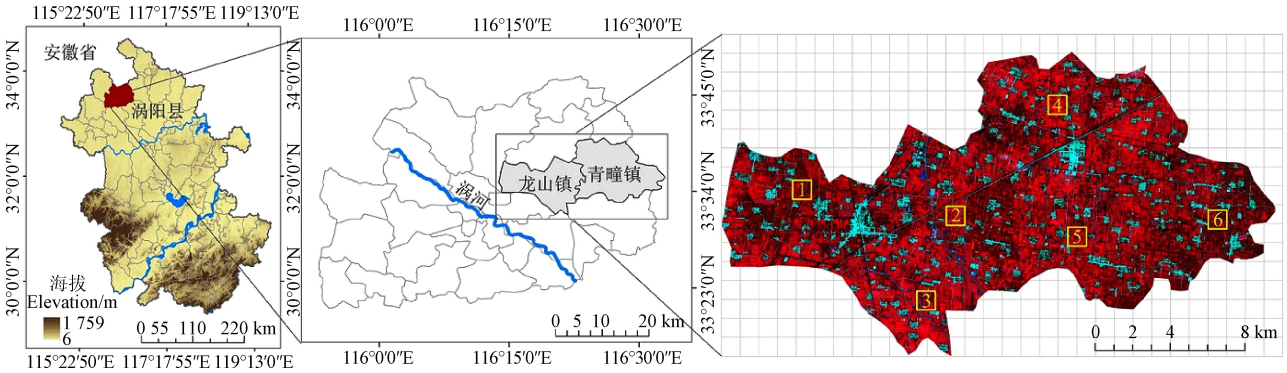
RapidEye 卫星的红边波段后，两种作物的总体识别精度提高了 7.4%；且多时相 Landsat-8 OLI 影像可以弥补因缺少短波红外波段而产生的制图精度偏低的不足。Yin 等[15]采用 RF 算法识别中国三江平原地区的大豆、玉米和水稻 3 种农作物，并提出 Sentinel-2 数据的短波红外波段可以很好地区分大豆和玉米。植被指数主要基于植被对于红光波段的强吸收和近红外波段的高反射而建立，它可以辅助光谱特征有效地提高农作物的识别精度[16]。da Silva 等[17]基于 Google Earth Engine 平台采用时间序列多光谱数据生成的物候特征和植被指数几乎可以实现巴西中西部地区内大豆种植面积的实时监测。黄健熙等[18]认为多时相 GF-1 WFV 数据生成的归一化植被指数（Normalized Difference Vegetation Index，NDVI）、归一化水分指数（Normalized Difference Water Index，NDWI）和宽动态范围植被指数（Wide Dynamic Range Vegetation Index，WDRVI）在大豆和玉米识别中表现突出，且 RF 的提取效果优于支持向量机和最大似然方法。此外，合成孔径雷达（Synthetic Aperture Radar，SAR）数据因其具有较强的云层穿透能力以及全天时、全天候的特点而备受关注[19]。光学影像和 SAR 数据的协同作用对于区分不同作物具有重要意义，如 Ajadi 等[20]基于光学影像与 Sentinel-1 数据的 VH 极化成功获取了巴西两个生长季内的大豆种植规模和空间分布。

尽管已有不少学者开展了大豆种植区遥感提取的相关研究，但目前的研究多集中在机械化程度高、农田分布规整和大豆种植集中连片的大规模产区，例如美国、巴西、阿根廷以及中国东北地区[10-14,21]。对于以散户种植为主、种植结构较为复杂的地处中国黄淮海大豆主产区的安徽省鲜有关注。《2020 年中国农村统计年鉴》的数据显示 2019 年安徽省的大豆种植面积为 63.624 万 hm^2，仅次于黑龙江省和内蒙古自治区，位居全国第 3。该地区天气状况多变，云覆盖频繁，农田景观破碎，作物混杂种植严重，给其遥感识别带来了巨大挑战。因此，迫切需要探索出一套适合此类地区的大豆遥感识别方法。合适的卫星影像数据是应用遥感技术提取农作物种植区的基础。目前，应用广泛的 MODIS、Landsat 和 GF-1WFV 数据在空间分辨率、重访周期、工作波段设置方面具有各自的局限性。本研究先前的工作已经表明 Sentinel-2 数据适合种植结构复杂地区的大豆遥感提取，且大豆结荚早期更有利于大豆识别[22]。因此，针对目前大豆遥感识别研究存在的不足，本文基于大豆识别的优选时相（大豆结荚早期）的 Sentinel-2 影像，在田间调查数据和无人机影像的支持下，探讨 ReliefF 特征权重评估方法结合多种机器学习方法在皖北地区大豆制图中的表现，以期探索形成一套合理的大豆种植区提取方法。

1　研究区与数据

1.1　研究区概况

涡阳县位于安徽省北部（33°27′～33°47′N，115°53′～116°33′E），是中国黄淮海地区重要的大豆主产地，其大豆种植规模常年超过 7.2 万 hm^2，在安徽省所有县级行政单位中一直保持首位。该县的地形以平原为主，平均海拔为 29.5 m，属暖温带半湿润季风气候，年平均气温 15.1 ℃，年平均降雨量 851.6 mm 左右，历年平均日照时数为 2 015.7 h，适合大豆、玉米、高粱、红薯、芝麻和中药材等多种作物的生长。淮河的一级支流涡河横穿该县中部，涡河两岸呈现出截然不同的作物种植格局。涡河以北地区以大豆和玉米交错种植为主，大豆占比明显高于玉米，而涡河以南玉米规模占据绝对优势。本文选取位于涡河北部的龙山和青疃 2 个典型镇级行政单位作为研究区（图 1）。该地区的大豆通常在 6 月中下旬播种，8 月中旬开始结荚，并于当年的 9 月末至 10 月初收获（中国气象数据网 http://data.cma.cn/）。



注：1~6 代表样方编号。
Note: 1-6 represent the labels of the samples.

图 1　研究区地理位置及验证样方的空间分布

Fig.1　Geographical location of the study area and spatial distribution of verification samples

1.2　数据获取

1.2.1　Sentinel-2 数据

Sentinel-2 是多光谱成像卫星星座，拥有 2A 和 2B 两颗相同的卫星，其空间分辨率最高可达 10 m，双星协同观测可使重访周期缩短至 5 d，有利于获取作物的关键生育期图像并能展现更丰富的田间地块细节。它携带一台多光谱成像仪（Multiple Spectral Instrument，MSI），具有 13 个光谱波段，覆盖可见光、近红外到短波红外波谱范围（443～2 190 nm）[23]，可实现对地表高频次、持续和动态监测。此外，Sentinel-2 还提供了丰富的工作波段，

是唯一一个在红边范围内设置 3 个工作波段的卫星传感器，为农作物精细制图奠定了有利基础。本文通过 ESA Copernicus Open Access Hub（http://scihub.coperni cus.eu/）下载了 2019 年 8 月 18 日（大豆结荚早期）的 Sentinel-2B LIC 级数据进行后续大豆种植区提取研究。

1.2.2 UAV 图像

研究区内布设了 6 个大小为 1 km×1 km 的样方（图 1）来评估基于卫星影像的大豆种植区提取效果。样方的设置在空间上尽可能均匀分布且其内包含的人工地物占比尽可能小，且研究区内空间异质性相对较小，样方具有一定代表性。本文利用 DJI Phantom4 Pro 无人机，于 2019 年 9 月 7－9 日期间获取了 6 个样方的航拍影像。无人机平台搭载了视场角为 84°、有效像素 2 000 万的 1 英寸 CMOS 相机来获取 RGB 真彩色图像。无人机飞行期间天气良好，飞行航高均设为 200 m，航向和旁向重叠率均设为 80%，影像对应的地面分辨率约为 6 cm。此外，为了确保获取的无人机影像具有更高的地理定位精度，每个样方都布设了辨识度较高的 4 个像控点，并且采用 RTK（华测 i70）测量每个像控点的地理坐标。

1.2.3 地面调查数据

为了充分掌握样方内地物类型及典型其样本的空间位置，在获取无人机影像的同时，同步开展了地面查工作。调查时采用手持 GPS（Trimble Geo7X, USA）测量代表性地块的经纬度坐标并记录相应的植被类型。此次调查共获取地面实测点 212 个，其中大豆、玉米、高粱、裸土和其他植被的样本点个数分别为 91、79、13、5 和 24。

2 研究方法

本文首先对 Sentinel-2 卫星影像和 UAV 图像进行预处理。为使得优选得到的遥感判别特征更具有针对性，首先构建决策树筛选规则剔除非农作物分布区域，然后针对田间植被构建 ReliefF-RF、ReliefF-BPNN、ReliefF-SVM 组合模型筛选出对于大豆识别最有效的特征，并采用混淆矩阵方法评估 3 种模型在大豆制图中的表现，确定最优提取模型。

2.1 数据预处理

Sentinel-2 数据是经过辐射定标和正射校正的 Level-1C 级大气顶（Top of Atmosphere，TOA）表观反射率产品。因此，只需再对其进行大气校正，便可得到大气底部（Bottom of Atmosphere，BOA）反射率。该数据的大气校正借助 ESA 提供的 Sen2cor（http://step.esa.int/main/third-party-plugins-2/sen2cor/）来完成。本文采用空间分辨率为 10 m 的 4 个波段和 20 m 的 6 个波段开展大豆种植区提取（表 1）。为保证各波段空间分辨率的一致性，在 Sentinel Application Platform (SNAP) 平台下，使用双线性内插法将分辨率为20 m的波段重采样至10 m并输出为 ENVI 支持的 img 存储格式。最后，利用 ENVI 5.3 进行波段合成并采用涡阳县乡镇级矢量行政边界对图像进行裁剪，以获取覆盖完整研究区的影像。

对于无人机所获得的航拍影像，首先对数据进行质量检查，剔除成像质量相对略差的影像。将筛选后的影像导入 Context Capture Center（version 4.4.9）中自动完成影像匹配、空中三角测量和不规则三角网模型的构建进而生成密集点云。为保证图像的空间定位精度，需要导入像控点坐标，然后基于点云数据生成三维模型，并通过该模型获取数字正射影像（Digital Orthophoto Map，DOM）。最后，使用 Global Mapper 14 对 DOM 影像进行拼接。

表 1 文中所采用的 Sentinel-2 的 10 个光谱波段描述

Table 1 Description of the 10 spectral bands of Sentinel-2 employed in this study

波段编号 Band number	波段名称 Band name	中心波长 Central wavelength/nm	空间分辨率 Spatial resolution/m
B2	蓝波段	490	10
B3	绿波段	560	10
B4	红波段	665	10
B5	红边波段 1	705	20
B6	红边波段 2	740	20
B7	红边波段 3	783	20
B8	近红外波段	842	10
B8a	窄带近红外	865	20
B11	短波红外 1	1 610	20
B12	短波红外 2	2 190	20

2.2 非农作物地物类型的剔除

基于 Sentinel-2 影像，本文采用分层逐级提取策略，首先借助归一化建筑指数（Normalized Difference Building Index，NDBI）[24]、改进的归一化水体指数（Modified Normalized Difference Water Index，MNDWI）[25]以及近红外波段反射率构建决策树筛选规则剔除人工地物（如建筑、道路等）、水体、裸土和林地等非农作物分布区域。

$$NDBI=(R_{11}-R_8)/(R_{11}+R_8) \quad (1)$$

$$MNDWI=(R_3-R_{11})/(R_3+R_{11}) \quad (2)$$

式中 R_3、R_8 和 R_{11} 分别代表绿波段（B3）、近红外波段（B8）和短波红外波段（B11）的反射率值。为了进一步增强结果的可信度，本文借助 2017 年 FROM-GLC10 全球土地利用产品[26]（http://data.ess.tsinghua.edu.cn/）提供的耕地分布（类型编号 10，空间分辨率 10 m）作为决策树的附加判别条件，以进一步筛除结果中可能存在的部分非耕地像元。最后基于生成的掩膜文件，对研究区影像进行掩膜处理得到农田植被的总体分布，再执行后续的大豆种植区提取。

2.3 大豆遥感识别模型的构建

2.3.1 机器学习方法

RF 算法在遥感制图领域应用广泛，其抗噪能力强，运算速度快，预测准确率高，且能有效抑制过拟合。研究表明，通常情况下 RF 算法仅需要设置 2 个关键的用户参数，并且在默认参数下即可取得令人满意的结果[27]。鉴于此，本文的参数保持默认设置即分支节点的特征数为参与分类的特征总数的平方根，决策树的数量为 100。

BP 神经网络（Back-Propagation Neural Network，BPNN）具有较强的非线性映射能力和良好的网络容错性，可以很好地解决现实场景中非线性建模问题[28]。在

农业遥感领域，BPNN 被广泛用于土地覆盖分类和植被理化参数定量反演模型的构建。为获得较好的大豆提取效果，通过多次试验探究，本文采用单隐含层 BPNN，迭代次数设置为 1 000，学习率设置为 0.02，训练目标的最小误差为 0.001。

支持向量机（Support Vector Machine，SVM）的基本原理是通过构造最优分割超平面，以此实现训练样本分类。已有研究表明径向基核函数（Radial Basis Function，RBF）更适用于区分不同类型的农作物[29]。因此，本研究选择 RBF 作为分类模型中的核函数来提取大豆种植区。该核函数的 Gamma 取所用卫星影像波段数的倒数；分类阈值设为 0，其他参数保持默认。

2.3.2　候选特征变量

传统意义上通常采用波段反射率作为指定地物的遥感判别特征，然而现实中可能并非所有工作波段对于大豆识别均足够有效，因此本文考虑加入一些扩展特征如多种植被指数参与大豆种植区提取，评估各个扩展特征在大豆识别中的表现。基于 Sentinel-2 影像，选取了包括 9 个植被指数（表 2）和原始 10 个波段反射率在内的共 19 个候选特征因子来执行大豆遥感识别，并且在此基础上对候选特征进行优选。本文将这些植被指数和原始波段一起，统称为“特征”。

表 2　本文所选用的遥感植被指数

Table 2　Vegetation indices employed in this paper

植被指数 Vegetation index	计算公式 Expression formula	参考文献 Reference
增强型植被指数 Enhanced Vegetation Index (EVI)	$EVI=2.5(R_8-R_4)/(R_8+6R_4-7.5R_2+1)$	[30]
土壤调节植被指数 Soil Adjusted Vegetation Index (SAVI)	$SAVI=1.5(R_8-R_4)/(R_8+R_4+0.5)$	[31]
MERIS 陆地叶绿素指数 MERIS Terrestrial Chlorophyll Index (MTCI)	$MTCI=(R_6-R_5)/(R_5-R_4)$	[32]
红边位置指数 Red-Edge Position (REP)	$REP=705+35[0.5(R_4+R_7)-R_5]/(R_6-R_5)$	[33]
归一化植被指数 Normalized Difference Vegetation Index (NDVI)	$NDVI=(R_8-R_4)/(R_8+R_4)$	[34]
红边归一化植被指数 1 Red-edge NDVI 1 (NDVIre1)	$NDVIre1=(R_8-R_5)/(R_8+R_5)$	[35]
红边归一化植被指数 2 Red-edge NDVI 2 (NDVIre2)	$NDVIre2=(R_8-R_6)/(R_8+R_6)$	[35]
绿色归一化植被指数 Green Normalized Difference Vegetation Index (GNDVI)	$GNDVI=(R_8-R_3)/(R_8+R_3)$	[35]
宽动态范围植被指数 Wide Dynamic Range Vegetation Index (WDRVI)	$WDRVI=(0.2R_8-R_4)/(0.2R_8+R_4)$	[36]

注：R_a 表示 a 波段的反射率，a = 2~8。

Note: R_a represents the reflectance of band a, a = 2-8.

2.3.3　特征变量重要性评估

ReliefF 算法的核心思想是通过计算类别之间的假设间隔对候选特征因子进行分类贡献度评价。若特征集 A 中的某个特征使得异类样本间的距离大于同类样本，说明此特征有利于分类，故增加其权重值；反之则降低其权重。最后将 n 次计算结果的均值作为每个特征的最终权重，计算公式如下[37]：

$$\omega(A_i)=\omega(A_i)-\frac{1}{nk}\sum_{h\in H}\left|R_i-h_i\right|+\frac{1}{nk}\sum_{m\in M}\left|R_i-h_i\right| \quad (3)$$

式中 $\omega(A_i)$表示特征 i 的权重值，$\sum_{h\in H}\left|R_i-h_i\right|$ 为 k 个同类最近邻样本与 R 样本在特征上的距离之和，$\sum_{m\in M}\left|R_i-h_i\right|$ 代表 k 个异类最近邻样本与样本 R 在特征 i 上的距离之和。

本文基于掩膜后的图像，从 4 种田间作物类型（大豆、玉米、高粱、其他）中选取近 2 000 个样本进行特征敏感性分析。由于该算法固有的随机性可能导致权重评估结果具有一定的不确定性，本文取 20 次运算结果的平均值作为各个特征的最终权重值。

2.3.4　不同模型下的特征子集优选

在特征权重评估的基础上，判断特征子集的最佳维度是实现特征优选的关键。鉴于传统的针对特征权重的阈值判定方法存在强烈的主观性，本文提出一种与分类器相耦合的顺序前向选择判定方法。该方法首先将权重最大的特征因子输入某个分类器，得到初始分类精度，紧接着按照特征权重从高到低的顺序依次加入下一个权重略低的特征，与前面已加入的特征组合成新的输入数据，并计算相应的总体分类精度（Overall Accuracy, OA）；每次添加一个特征并执行精度评估，直到 19 个特征全部输入完毕。若某个特征使 OA 数值降低，则剔除该特征，模型对应的优选特征集合依据 OA 确定。考虑到机器学习算法的随机性，本文取 50 次 OA 的均值作为特征子集的优选结果，最后将不同分类器各自优选出的特征组合作为相应模型的输入数据来提取大豆种植区。本文基于 MATLAB 2018 实现特征变量优选。

2.4　最优模型的判定

基于研究区内 6 个样方的无人机影像提取得到的大豆分布来评估不同模型的提取效果。由于航拍工作开展时间稍迟导致不同大豆田的物候期存在差异，部分地块已进入黄熟期，不利于计算机自动提取。而经过预处理后的无人机影像具有足够精细的纹理信息，可以较为容易地对地物进行目视解译和类型归属判断。因此，本文在 ArcGIS 10.4 软件平台的支持下，基于拼接后的无人机影像采用数字化方式描绘大豆种植地块的边界，并保存为矢量图层，以此作为真值来检验不同模型的提取效果。

借助地面真实数据，通过构建混淆矩阵可以对不同模型的分类结果进行精度评价。由该矩阵派生出的评价指标主要包括 4 个，即制图精度、用户精度、总体精度和 Kappa 系数。与其他 3 个指标相比，Kappa 系数是根据所有待评估地类的漏分和错分情况给出的一种更为全面、更权威的分类准确性评估指标，其计算公式如下[38]：

$$\text{Kappa}=\frac{N\sum_{i=1}^{m}x_{ii}-\sum_{i=1}^{m}x_{i+}\cdot x_{+i}}{N^2-\sum_{i=1}^{m}x_{i+}\cdot x_{+i}} \quad (4)$$

式中 N 表示像元总数，m 是类别数，x_{ii} 是混淆矩阵对角线上的像元个数，x_{i+}和 x_{+i} 分别是第 i 行和第 i 列的像元总数。

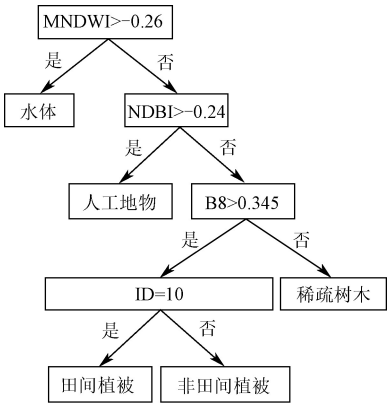
2.5 最佳提取模型的效果评价

为了进一步考查优选模型在大豆制图中的表现，本文设计了 3 种大豆提取方案。方案 A 所用的特征为 Sentinel-2 原始 10 个波段反射率；方案 B 包含未经过特征选择的全部 19 个特征；方案 C 为上一节得到的优选模型。将不同特征组合形式作为输入，采用优选模型所对应的机器学习算法，基于相同的训练样本和检验样本，评估 3 种方案各自的分类精度，据此考查优选指标在大豆提取中的表现，分析该工作的实际意义。

3 结果与分析

3.1 非农作物像元的剔除

通过多次对比试验发现，MNDWI 指标上建筑与水体的差异更为显著，更容易实现水体的分离。需要指出的是，研究区内树木多沿道路和房屋周围呈零星分布，植被指数（如 NDVI 或 EVI 等）数值处于中等水平，简单利用植被指数难以将其与农作物进行区分。而树木在近红外波段 B8（中心波长 842 nm）和农作物具有明显差异，因此可基于该指标构建判别规则。具体的决策树筛选规则如图 2 所示。



注：NDBI 为归一化建筑指数；MNDWI 为改进的归一化水体指数；ID 为类型编号。

Note：NDBI is normalized difference building index; MNDWI is modified normalized difference water index; ID is type number.

图 2 决策树筛选规则

Fig.2 Filtering rules of decision tree

3.2 不同模型的最佳特征子集

基于剔除非农作物像元的 Sentienl-2 影像，采用 ReliefF 算法评估了 19 个候选特征因子在大豆识别中的重要性（图 3）。B8 权重最高，表明近红外波段对大豆提取的贡献度最大。REP、NDVIre2 是有红边波段参与生成的特征因子；B5、B6 是 Sentinel-2 的 2 个红边波段反射率，从特征权重评估结果来看，这些与红边波段相关的特征因子重要性排序比较靠前，意味着红边波段对于大豆遥感识别具有重要意义；此外，短波红外反射率 B12 和 B11 对于实现大豆与其他田间植被之间的分离也十分有效；SAVI 和 EVI 相比其他常用植被指数更有利于此研究区内的大豆识别。

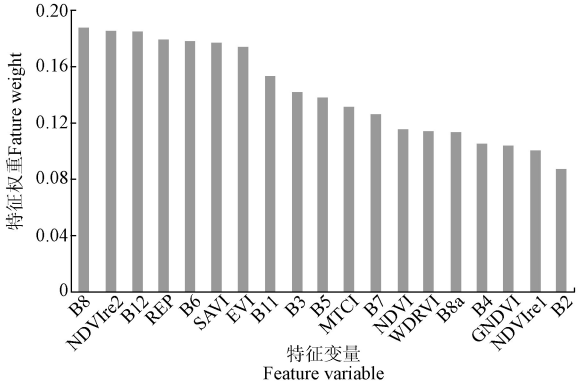


图 3 各候选特征的重要性排序

Fig.3 Ranking of the importance of each candidate feature

根据前文所述的优选特征变量的最佳维度判定方法，随着特征变量的逐步加入，不同模型的分类精度如图 4 所示。图 4a 结果显示当特征个数达到 9 时，分类精度达到局部最优，随着后续特征的加入，精度略有下降且在一个小范围内波动；当 19 个特征因子全部参与分类时，分类精度达到最大值，但仅比前 10 个特征所得精度高 0.31 个百分点，因此首先舍弃排名在第 9 位以后的特征；此外，排在第 6 位（SAVI）和第 8 位（B11）的特征在加入后未能提升分类精度，同样予以舍弃。最终 ReliefF-RF 模型选取了特征权重排名前 9 位的 7 个特征因子作为该模型的优选特征子集。同理，ReliefF-BPNN 模型在特征个数达到 9 时，精度达到最大值（图 4b），舍弃排名在第 7 位（EVI）和 8 位（B11）对应的特征以及第 9 位以后的 11 个特征，该模型的最佳特征维度为 7；ReliefF-SVM 模型的最佳特征因子的个数为 5（图 4c）。表 3 给出了 3 种模型的特征变量优选结果。

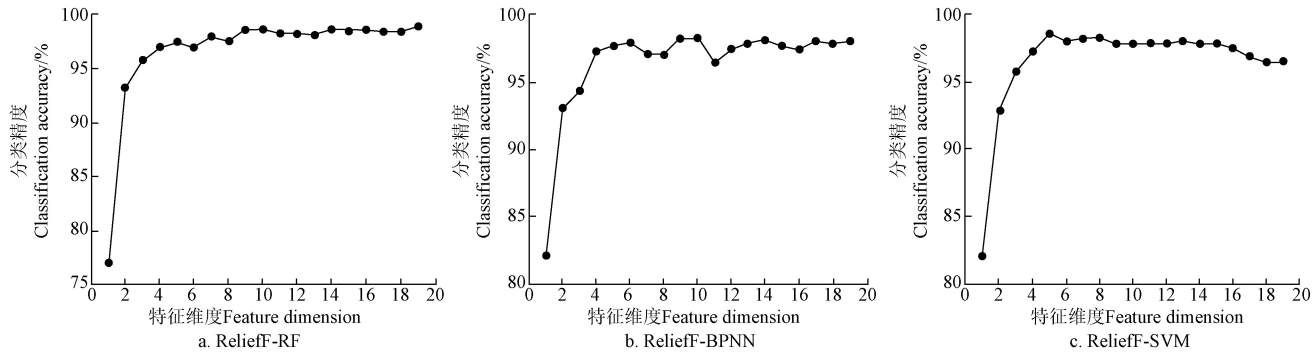


图 4 3 种模型的分类精度与特征维度的关系

Fig.4 Relationship between classification accuracy and feature dimension of the three models

表 3　不同模型的优选特征子集

Table 3　Optimum feature-subsets of different models

模型 Model	特征因子 Feature factor
ReliefF-RF	B8，NDVIre2，B12，REP，B6，EVI，B3
ReliefF-BPNN	B8，NDVIre2，B12，REP，B6，SAVI，B3
ReliefF-SVM	B8，NDVIre2，B12，REP，B6

3.3　大豆种植区的最佳提取模型

本文将各模型对应的优选特征子集作为输入执行分类得到大豆种植区，并基于各验证样方内 UAV 影像的大豆种植区分布对不同模型的大豆提取效果进行评估（表 4）。结果表明，基于 ReliefF-RF 模型得出的 6 个样方的总体精度和 Kappa 系数均高于其他 2 种模型，且在样方 3 上优势最为明显；ReliefF-BPNN 模型的大豆制图精度较高但用户精度较低，说明该模型将较多其他地类错分为大豆；ReliefF-SVM 模型的制图精度和用户精度在样方 2、3、5、6 均低于 ReliefF-RF，而在样方 1 内的用户精度比 ReliefF-RF 高 0.06 个百分点，但制图精度明显低于 ReliefF-RF，这表明该模型在样方 1 内大豆的漏分情况相对更为严重。3 种模型在不同样方内的提取效果具有差异，但得出的大豆空间分布格局总体较为一致（图 5）。研究区内大豆的种植规模占据绝对优势且空间分布较为均衡，然而作物间交错混杂种植现象普遍存在，大豆田块集中程度低、分布分散，ReliefF-RF、ReliefF-BPNN、ReliefF-SVM 模型提取得到的大豆种植区总面积分别为 9 291.16、10 277.70、9 451.24 hm^2。

表 4　大豆种植区的提取精度

Table 4　Extraction accuracy of soybean planting areas

样方编号 Sample No.	ReliefF-RF				ReliefF-BPNN				ReliefF-SVM				A	B
	PA/%	UA/%	OA/%	Kappa	PA/%	UA/%	OA/%	Kappa	PA/%	UA/%	OA/%	Kappa	Kappa	Kappa
1	86.63	84.46	90.20	0.78	91.64	78.56	88.84	0.76	81.32	84.52	88.77	0.75	0.71	0.79
2	86.15	84.60	85.92	0.72	88.40	80.79	84.45	0.69	85.65	84.23	85.51	0.71	0.69	0.73
3	81.84	87.24	89.50	0.76	85.16	77.95	86.44	0.71	79.63	81.96	86.80	0.71	0.72	0.78
4	92.58	90.78	88.25	0.72	90.48	90.93	87.05	0.69	92.76	90.05	87.79	0.70	0.68	0.74
5	89.59	87.52	88.49	0.77	92.35	84.07	87.51	0.75	87.28	87.09	87.26	0.75	0.73	0.78
6	89.70	85.29	91.91	0.81	92.77	79.79	90.36	0.79	85.67	84.73	90.66	0.78	0.75	0.81

注：PA、UA、OA 分别表示制图精度、用户精度和总体精度；A、B 代表方案 A 和 B，方案 A 所用的特征是 Sentinel-2 原始 10 个波段反射率；方案 B 包含未经过特征选择的全部 19 个特征。

Note: PA, UA, OA represent the producer's accuracy, user's accuracy and overall accuracy, respectively; A, B represent the scheme A and B, the features employed in scheme A were reflectance of original 10 bands of Sentinel-2; scheme B contained all 19 features without feature selection.

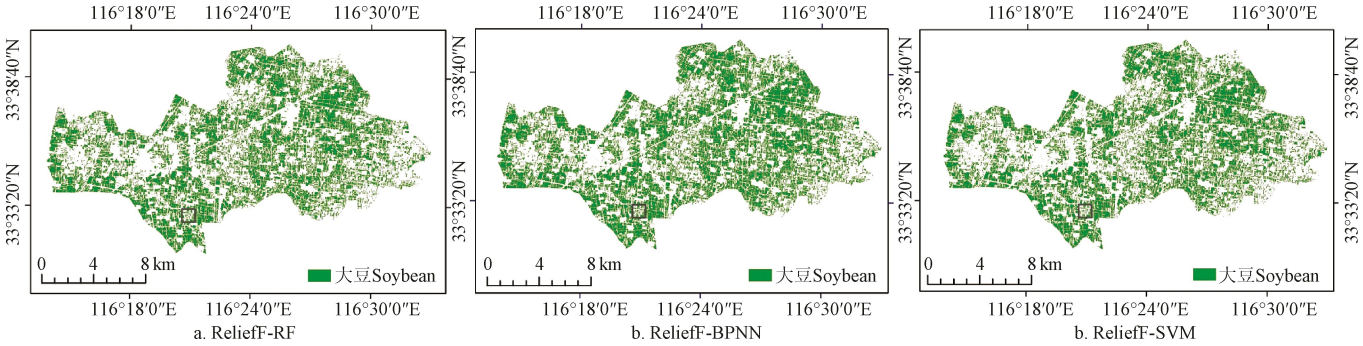


图 5　基于 3 种模型提取的大豆种植区空间分布

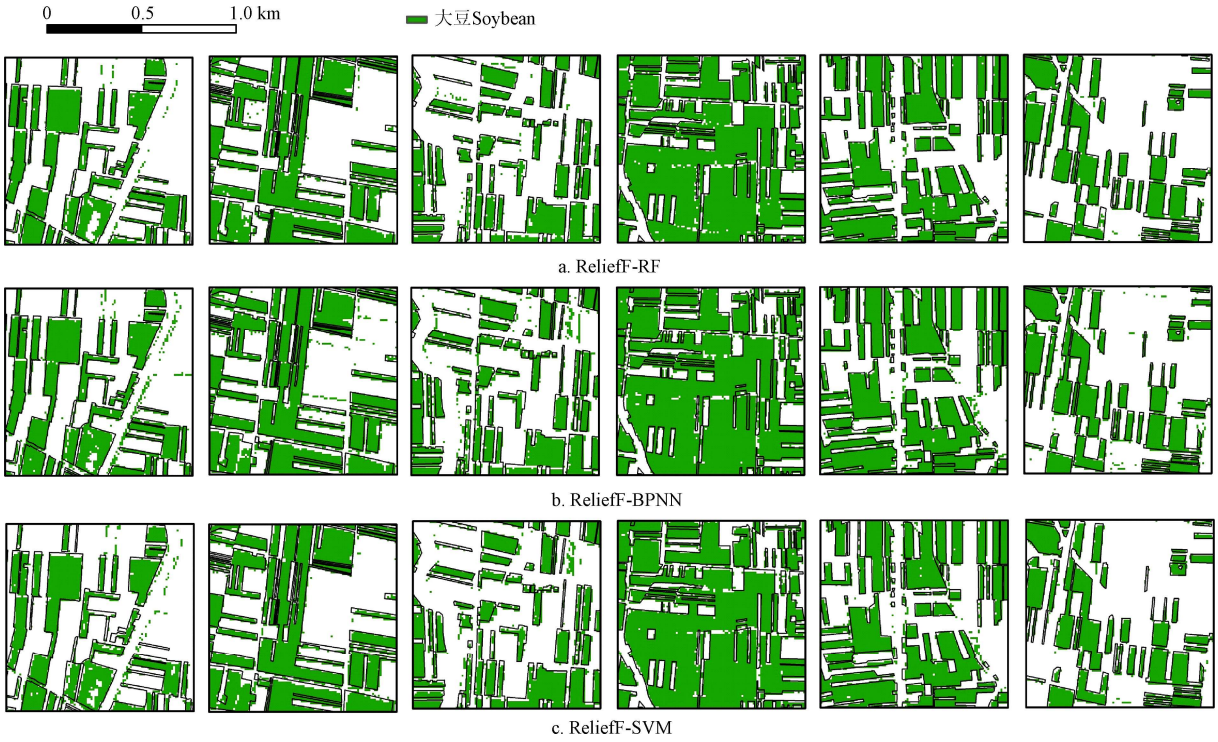
Fig.5　Spatial distribution of soybean planting areas extracted by using three models

为了更为直观地展现 3 种模型在 6 个验证样方内大豆提取效果的差异，图 6 给出了各个样方的大豆种植区空间分布。3 种模型在局部地块，尤其是在大豆田和玉米田邻接的地块仍存在一定差异。将无人机影像解译得到的大豆种植区作为地面真值（用实线多边形表示），可以看出 ReliefF-RF 模型的提取结果与真值差距相对较小。除了样方 4 以外，ReliefF-BPNN 模型的大豆高估情况要比其他 2 种模型更为严重，该模型将更多其他地类错分为大豆，导致用户精度偏低。与 ReliefF-SVM 与 ReliefF-BPNN 模型相比，ReliefF-RF 模型的提取结果更接近大豆的真实分布，提取效果明显优于其他 2 种模型，因此，本研究文将该模型作为大豆种植区提取的最佳模型。

3.4　最佳提取模型在大豆识别中的表现

为了进一步评估最佳提取模型 ReliefF-RF 在大豆识别中的表现，基于 RF 算法的方案 A 与 B 的大豆提取精度如表 4 所示。与 Sentinel-2 原始 10 个波段（方案 A）相比，基于优选特征子集的 ReliefF-RF 模型的 Kappa 系数在样方 1、6 分别提高了 0.07、0.06，大豆的提取效果有显著改善；在样方 2 提高了 0.03；在样方 3、4、5 提高了 0.04，该模型在所有样方中均实现了精度提升。与

未经优选的 19 个特征的提取结果（方案 B）相比，优选模型所得的 Kappa 系数仅比前者低 0.01 或 0.02。结果表明，基于优选特征建立的 ReliefF-RF 模型在保障提取精度的同时，相比将所有候选特征作为输入能够减少 63.16%的数据量。因此，本文提出的最佳模型 ReliefF-RF 在大豆种植区提取中具有较为明显的方法优势。



注：每个子图中从左到右依次表示样方 1~6。样方内黑色多边形所圈定的区域对应大豆的真实分布。
Note: Each sub-figure shows sample 1-6 from left to right. The area delineated by black polygon in each sample corresponds to the ground truth distribution of soybean.

图 6 3 种模型在 6 个样方内的大豆种植区提取结果
Fig.6 Extraction results of soybean planting areas extracted by using three models in the six samples

4 讨 论

先前的有关大豆识别的研究多以 MODIS、Landsat 和高分系列卫星影像为主要数据源，研究区多集中在作物类型简单的美国、巴西等国家[5,12,15,17]，且有利于大豆识别的最佳特征的相关研究较少。本文基于单时相 Sentinel-2 影像采取分层逐级提取策略实现了种植结构复杂地区的大豆种植区提取。分层逐级提取策略可以实现较高精度的大豆提取，该提取策略的研究对象聚焦田间植被，筛选得到的特征针对性更强，理论上所得结果的适用性和推广性更优且不受其他非农地物占比的影响，在大豆卫星遥感提取中具有重要意义。该策略在构建决策树筛选规则时，借助 2017 年 FROM-GLC10 全球土地利用产品来修正决策树提取结果，在剔除非农地物类型之后，它所提供的耕地类型的空间分布尽可能多的包含了可耕作地块。经对比分析后发现耕地面积高于实际的农田分布区，因此该产品有助于非农地物像元的剔除来提高大豆识别精度，并且不会对作物提取结果产生不利影响。

本文的特征优选是采用与分类器相耦合的方式自行筛选出与之相匹配的最佳判别特征，这能够在最大程度上兼顾不同分类算法的特异性，并且在一定程度上降低了利用传统阈值方法[39]执行最佳维度判断时所带来的主观性。3 种模型的优选特征因子均包含了权重排名前 5 的特征（B8, NDVIre2, B12, REP, B6），表明红边、近红外和短波红外波段在大豆识别中具有显著优势，与王利民等[12-16]的研究结论一致，同时也说明了植被指数与光谱波段相结合的特征集能有效区分不同农作物。不同模型的最佳特征子集中所删除的特征可能是冗余信息，但是本研究未单独对候选特征以及优选特征子集进行冗余性分析，后续研究将进一步探讨各个特征之间可能存在的信息冗余问题。

本研究采用布设于研究区内的 6 个样方来检验大豆提取精度，样方的设置同时兼顾了空间地理位置的均匀性、样方内的作物类型以及大豆的占比等情况，且研究区为相邻的 2 个乡镇，其空间异质性较小，因此验证样方具有一定代表性。由于权威的乡镇级作物播种面积统计数据在现实中难以获取（统计年鉴通常只能提供县区级以上的统计数据），因此本研究未进行种植面积估算精度的检验。最佳提取模型 ReliefF-RF 在样方 1、3、5、6 的 Kappa 系数均大于等于 0.76，而在样方 2 和 4 仅有 0.72，其他 2 种模型的提取精度和该模型具有类似结果，提取精度尚未达到较高水平。经田间实地调查发现，在散户耕种模式下，研究区内田间种植结构复杂，夏季作物类型多样，大豆的品种、播种时间以及管理方式的不同，导致大豆作物类内差异较为明显，比如肉眼可见的

植株高度与叶片颜色的不一致，田间存在杂草和斑秃地块等，这些因素均给大豆遥感提取带来了巨大挑战，尤其是在大豆田和玉米田的邻接地块。由于样方 2 和样方 4 内的大豆田和玉米田更为细小和狭长，田间破碎程度相对更高，而所用 Sentinel-2 波段的像元大小为 10 m，某些地块宽度可能不及一个像元，导致“混合像元效应”显著，增加了遥感提取的难度和不确定性，在一定程度上降低了提取精度，今后将探讨混合像元分解方法在该地区内大豆种植区的提取效果。此外，作为验证数据的无人机影像的空间分辨率为 6 cm，Sentinel-2 影像的空间分辨率为 10 m，二者的空间尺度差异十分明显，结果之间不容易实现匹配，必然会对检验精度产生影响。后续研究中考虑采用高分辨率卫星影像如 GF-2 PMS, Superview-1, Pleiades 等来评估提取效果。

研究区地处南北方过渡地带，天气变化较为剧烈，阴雨天气出现的频率较高，导致可用光学影像的覆盖频率受限，多个物候期内的光学影像不一定有条件获取。因此，基于单时相数据实现大豆种植区提取在现实中更为可行且更具有实际意义，后续研究将进一步探讨多时相数据在皖北地区大豆精细遥感提取中的应用效果。此外，本研究的田间试验以及地面调查工作还不够完善，尤其是调查样点数量不够充足，覆盖的地物类型不够广泛，对于大豆异谱现象的形成机理尚未进行深入探讨。后续工作需要进行更为系统和全面的田间实地调查。本文的研究区相对较小，仅覆盖 2 个镇，且只关注了 2019 年一个生长季，后续将在更大尺度上针对多个生长季开展大豆种植区提取研究，来检验该项研究所得结论的适用性和鲁棒性。

5　结　论

本文基于 Sentinel-2 影像，以安徽省皖北典型大豆主产区为例，针对种植结构复杂、晴空观测有限以及田间景观破碎的客观实际，提出了一种基于分层逐级提取策略的大豆识别方法。研究结果表明，ReliefF-RF 的 Kappa 系数介于 0.72～0.81，相比其他 2 种模型表现出了更好的大豆识别能力（ReliefF-BPNN 和 ReliefF-SVM 模型的 Kappa 系数分别在 0.69～0.79 和 0.70～78 之间）；此外，优选特征子集的提取精度明显高于 Sentinel-2 原始 10 个波段参与提取得到的结果（后者 Kappa 系数在 0.68～0.75 范围内），尽管略低于全部 19 个特征因子的结果（Kappa 系数相差 0.01～0.02），但是降低了 63.16%的数据量。因此，ReliefF-RF 模型是本研究中大豆提取的最佳模型，该模型筛选出的红边波段 B6（740 nm）、近红外波段 B8（842 nm）、短波红外波段 B12（2 190 nm）和绿波段 B3（560 nm）可以有效地识别大豆，且红边归一化植被植数 2、红边位置指数和增强型植被指数在大豆识别中相比其他常用遥感植被指数更有优势。该研究弥补了复杂种植条件下大豆提取相关研究的不足，文中所提出的研究思路可以为相似种植条件下的大豆遥感识别相关研究提供有益参考，研究成果可以为当地农业部门开展农情调查、长势评估等工作提供有价值的依据。

［参　考　文　献］

[1] Liu X, Rahman T, Song C, et al. Changes in light environment, morphology, growth and yield of soybean in maize-soybean intercropping systems[J]. Field Crop Research, 2017, 200: 38-46.

[2] 李锁强. 中国农村统计年鉴[M]. 北京：中国统计出版社，2020，140-242.

[3] Weiss M, Jacob F, Duveiller G. Remote sensing for agricultural applications: A meta-review[J]. Remote Sensing of Environment, 2020, 236: 111402.

[4] 王鹏新，荀兰，李俐，等. 基于时间序列叶面积指数稀疏表示的作物种植区域提取[J]. 遥感学报，2019，23(5)：959-970.
Wang Pengxin, Xun Lan, Li Li, et al. Extraction of planting areas of main crops based on sparse representation of time-series leaf area index[J]. Journal of Remote Sensing, 2019, 23(5): 959-970. (in Chinese with English abstract)

[5] Zhong L H, Yu L, Li X C, et al. Rapid corn and soybean mapping in US Corn Belt and neighboring areas[J]. Science Report, 2016, 6(1): 1-14.

[6] 张莎，张佳华，白雲，等. 基于 MODIS-EVI 及物候差异免阈值提取黄淮海平原冬小麦面积[J]. 农业工程学报，2018，34(11)：150-158
Zhang Sha, Zhang Jiahua, Bai Yun, et al.Extracting winter wheat area in Huanghuaihai Plain using MODIS-EVI data and phenology difference avoiding threshold[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(11): 150-158. (in Chinese with English abstract)

[7] Li R Y, Xu M Q, Chen Z Y, et al. Phenology-based classification of crop species and rotation types using fused MODIS and Landsat data: The comparison of a random-forest-based model and a decision-rule-based model[J]. Soil and Tillage Research, 2020, 206: 104838.

[8] Grzegozewski D M, Johann J A, Uribe-Opazo M A, et al. Mapping soya bean and corn crops in the State of Parana, Brazil, using EVI images from the MODIS sensor[J]. International Journal of Remote Sensing, 2016, 37: 1257-1275.

[9] Chen Y L, Song X D, Wang S S, et al. Impacts of spatial heterogeneity on crop area mapping in Canada using MODIS data[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016,119: 451-461.

[10] Liu X X, Yu L, Zhong L H, et al. Spatial-temporal patterns of features selected using random forests: A case study of corn and soybeans mapping in the US[J]. International Journal of Remote Sensing, 2019, 40: 269-283.

[11] 田富有，吴炳方，曾红伟，等. 基于多层神经网络与 Sentinel-2 数据的大豆种植区识别方法[J]. 地球信息科学学报，2019，21(6)：918-927.
Tian Fuyou, Wu Bingfang, Zeng Hongwei, et al. Identifying Soybean cropped area with Sentinel-2 data and multi-Layer neural network[J]. Journal of Geo-information Science, 2019, 21(6): 918-927. (in Chinese with English abstract)

[12] 王利民，刘佳，杨玲波，等. 短波红外波段对玉米大豆种植面积识别精度的影响[J]. 农业工程学报，2016，32(19)：169-178.

Wang Limin, Liu Jia, Yang Lingbo, et al. Impact of short infrared wave band on identification accuracy of corn and soybean area[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2016, 32(19): 169-178. (in Chinese with English abstract)

[13] Zhong L H, Hu L N, Yu L, et al. Automated mapping of soybean and corn using phenology[J]. ISPRS J. Photogramm. Remote Sensing, 2016, 119: 151-164.

[14] 刘佳，王利民，杨福刚，等. 红边与短波红外谱段的玉米大豆识别能力研究[J]. 中国农学通报，2018，34(35)：120-129.
Liu Jia, Wang Limin, Yang Fugang, et al. Recognition ability of red edge and short wave infrared spectrum on maize and soybean[J]. Chinese Agricultural Science Bulletin, 2018, 34(35): 120-129. (in Chinese with English abstract)

[15] Yin L K, You N S, Zhang G L, et al. Optimizing feature selection of individual crop types for improved crop mapping[J]. Remote Sensing, 2020, 12: 1-20.

[16] 王利民，刘佳，杨玲波，等. 随机森林方法在玉米-大豆精细识别中的应用[J]. 作物学报，2018，44(4)：569-580.
Wang Limin, Liu Jia, Yang Lingbo, et al. Application of random forest method in maize-soybean accurate identification[J]. Acta Agronomica Sinica, 2018, 44(4): 569-580. (in Chinese with English abstract)

[17] da Silva C A, Leonel A H S, Rossi F S, et al. Mapping soybean planting area in midwest Brazil with remotely sensed images and phenology-based algorithm using the Google Earth Engine platform[J]. Computers and Electronics in Agriculture, 2020, 169: 105194

[18] 黄健熙，侯裔焯，苏伟，等. 基于 GF-1 WFV 数据的玉米与大豆种植面积提取方法[J]. 农业工程学报，2017，33(7)：164-170.
Huang Jianxi, Hou Yuzhuo, Su Wei, et al. Mapping corn and soybean cropped area with GF-1 WFV data[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2017, 33(7): 164-170. (in Chinese with English abstract)

[19] Liu C A, Chen Z X, Shao Y, et al. Research advances of SAR remote sensing for agriculture applications: A review[J]. Journal of Integrative Agriculture, 2019, 18(3): 506-525.

[20] Ajadi O A, Barr J, Liang S Z, et al. Large-scale crop type and crop area mapping across Brazil using synthetic aperture radar and optical imagery[J]. International Journal of Applied Earth Observation and Geoinformation, 2021, 97: 102294.

[21] You N S, Dong J W, Huang J X, et al. The 10-m crop type maps in Northeast China during 2017-2019[J]. Scientific Data, 2021, 8(1): 41.

[22] She B, Yang Y Y, Zhao Z G, et al. Identification and mapping of soybean and maize crops based on Sentinel-2 data[J]. International Journal of Agricultural and Biological Engineering, 2020, 13(6): 171-182.

[23] Drusch M, Del Bello U, Carlier S, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services[J]. Remote Sensing of Environment, 2012, 120: 25-36.

[24] 查勇，倪绍祥，杨山. 一种利用 TM 图像自动提取城镇用地信息的有效方法[J]. 遥感学报，2003，7(1)：38-40.
Zha Yong, Ni Shaoxiang, Yang Shan. An effective approach to automatically extract urban land-use from TM imagery[J]. Journal of Remote Sensing, 2003, 7(1): 38-40. (in Chinese with English abstract)

[25] 徐涵秋. 利用改进的归一化差异水体指数（MNDWI）提取水体信息的研究[J]. 遥感学报，2005，9(5)：590-595.
Xu Hanqiu. A study on information extraction of water body with the modified normalized difference water index (MNDWI)[J]. Journal of Remote Sensing, 2005, 9(5): 590-595. (in Chinese with English abstract)

[26] Gong P, Liu H, Zhang M N, et al. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017[J]. Science Bulletin, 2019, 64(6): 370-373.

[27] Immitzer M, Atzberger C, Koukal T. Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data[J]. Remote Sensing, 2012, 4: 2661-2693.

[28] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back propagating errors[J]. Nature, 1986, 323: 533-536.

[29] 王霞，王占岐，金贵，等. 基于核函数支持向量回归机的耕地面积预测[J]. 农业工程学报，2014，30(4)：204-211.
Wang Xia, Wang Zhanqi, Jin Gui, et al. Land reserve prediction using different kernel based support vector regression[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2014, 30(4): 204-211. (in Chinese with English abstract)

[30] Huete A R, Didan K, Miura T, et al. Overview of the radiometric and biophysical performance of the MODIS vegetation indices[J]. Remote Sensing of Environment, 2002, 83: 195-213.

[31] Huete A R. A soil-adjusted vegetation index (SAVI)[J]. Remote Sensing of Environment, 1988, 25: 295-309.

[32] Dash J, Curran P. Evaluation of the MERIS terrestrial chlorophyll index (MTCI)[J]. Advances in Space Research, 2007, 39: 100-104.

[33] Guyot G, Baret F, Major D. High spectral resolution: determination of spectral shifts between the red and the near infrared[J]. International Archives of Photogrammetry and Remote Sensing, 1988, 11: 750-760.

[34] Rouse J W, Hass R H, Scheel J A, et al. Monitoring vegetation systems in the Great Plain with ERTS[C]. Washington, DC, USA: NASA Special Publication, 1973, 309-317.

[35] Gitelson A A, Merzlyak M. Remote estimation of chlorophyll content in higher plant leaves[J]. International Journal of Remote Sensing, 1997, 18: 2691-2697.

[36] Gitelson A A. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation[J]. Journal of Plant Physiology, 2004, 161(2): 165-173.

[37] 韦娜，王涛. 结合 ReliefF 与支持向量机的特征选择方法研究[J]. 计算机应用与软件，2008，25(1)：283-285.
Wei Na, Wang Tao. Research into the feature selection method by combining ReliefF and support vector machine[J]. Computer Applications and Software, 2008, 25(1): 283-285. (in Chinese with English abstract)

[38] Congalton R G. A review of assessing the accuracy of classifications of remotely sensed data[J]. Remote Sensing of Environment. 1991, 37: 35-46.

[39] 黄林生，阮超，黄文江，等.基于 GF-1 遥感影像和 relief-mRMR-GASVM 模型的小麦白粉病监测[J]. 农业工程学报，2018，34(15)：167-175.
Huang Linsheng, Ruan Chao, Huang Wenjiang, et al. Wheat powdery mildew monitoring based on GF-1 remote sensing image and relief-mRMR-GASVM model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(15): 167-175. (in Chinese with English abstract)

Extraction of soybean planting areas combining Sentinel-2 images and optimized feature model

Zhang Dongyan[1], Yang Yuying[1], Huang Linsheng[1], Yang Qi[1], Liang Dong[1], She Bao[1,2※], Hong Qi[1], Jiang Fei[3]

(1. *National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University, Hefei* 230601*, China*; 2. *School of Geomatics, Anhui University of Science & Technology, Huainan* 232001*, China*; 3. *School of Information Engineering, Suzhou University, Suzhou* 234000*, China*)

Abstract: Accurate mapping of the soybean planting area is greatly significant to yield estimation, crop-damage warning, and structural adjustments in modern agriculture. But there are only a few reports on the remote sensing technology in soybean identification, particularly in view of the high frequency of cloud cover, diverse types of summer crops, and complex planting structure of fields. In this study, taking Longshan and Qingtuan towns situated in typical soybean producing areas in North Anhui plain as the study area, a hierarchical extraction was proposed to obtain the spatial distribution of soybean planting area in the 2019 growing season. The Sentinel-2 image was acquired at the early pod-setting stage of soybean (August 18, 2019). A series of filtering rules for decision trees were first established to eliminate non-agricultural cover types, such as water, sparse trees, bare soil, and artificial objects (buildings, roads). As such, the overall distribution of field vegetation was obtained. The Sentinel-2 image was then utilized to generate 19 candidate features containing the reflectance of 10 spectral bands with a resolution of less than or equal to 20 m and 9 vegetation indices. ReliefF algorithm was used to evaluate the significance of each candidate feature in typical ground-feature samples. The ReliefF algorithm was combined with three machine learning, including Random Forest (RF), BP Neural Network (BPNN), and Support Vector Machine (SVM). Three models were established, including ReliefF-RF, ReliefF-BPNN, and ReliefF-SVM. The most effective features were screened out for the soybean identification, thereby evaluating the performance of three models in soybean mapping. The UAV images covering six ground samples (each was 1 km×1 km in size) were selected to evaluate the extraction. Results showed that the best performance was achieved in the ReliefF-RF model with the Kappa coefficient ranging from 0.72-0.81, and the overall accuracy of 85.92%-91.91%. The Kappa coefficient of the present model was higher than that of another two models in each ground sample, where 0.69-0.79 and 0.70-0.78 for ReliefF-BPNN and ReliefF-SVM, respectively. The ReliefF-RF was used to single out the near-infrared B8 (842 nm), red-edge normalized difference vegetation index (NDVIre2) that derived from B8 and B6, short-wave infrared B12 (2190 nm), red-edge position (REP), red-edge B6 (740 nm), green B3 (560 nm), and enhanced vegetation index (EVI). It indicated that these seven optimum features were more advantageous than other commonly-used spectral bands and remote-sensing vegetation indices in soybean identification, where the red edge-related variables were particularly highlighted. In addition, the mapping data derived from the optimum features significantly outperformed that generated from the 10 spectral bands. Since the performance of the optimum feature subset was slightly inferior to total 19 features, ReliefF-RF that contained only seven optimum features showed obvious advantages in terms of time and computation cost, as well as data volume. Consequently, the optimum features were more targeted without any inference from the proportion of non-agricultural land cover types, due mainly to the hierarchical extraction focused only on the field vegetation. Better applicability and generalization were gained in theory. The findings can provide a valuable reference for the extraction of soybean areas under complex planting conditions.

Keywords: machine learning; models; soybean; Sentinel-2; planting area extraction; feature optimization