引文格式：李婧,杨辽,吴炜,等. 面向对象的高频时间序列乌苏市棉花分类[J]. 地理空间信息, 2025, 23(12):90-94. doi:10. 3969/j.issn. 1672-4623.2025.12.020

面向对象的高频时间序列乌苏市棉花分类

李　婧[1]，杨　辽[1,2*]，吴　炜[1,2]，邢程宏[1,3]，姜辰庚[1,3]，曾　友[1,3]
（1. 杭州市滨江区浙工大人工智能创新研究院，浙江 杭州 310056；2. 浙江工业大学地理信息学院，浙江 杭州 310014；
3. 浙江工业大学计算机科学与技术学院，浙江 杭州 310014）

摘　要：棉花种植信息的分类和提取对于农业生产和政策制定具有重要意义。现有棉花遥感提取存在影像缺失、特征冗余、耗时长等问题。基于Sentinel-2影像构建高频时间序列数据集，采用面向对象的方法采集耕地地块中心点，创建了包含分类特征的表格数据，并利用自动化机器学习方法对新疆乌苏市的棉花种植情况进行提取和分析。结果表明，该方法能有效降低“椒盐”噪声，提高分类效率和准确性。比较决策树、随机森林、LightGBM和集成模型在棉花分类中的表现，发现4种模型评价指标平均值均超过85%，其中集成模型表现最佳，准确率、精确率和召回率分别为95.14%、94.93%和98.50%。研究为基于表格数据的自动化机器学习技术在棉花种植快速识别中的应用提供了有益参考。

关键词：自动化机器学习；棉花提取；面向对象；地块中心点；kNDVI时间序列
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2025）12-0090-05

Object-oriented High-frequency Time Series Cotton Cultivation Classification in Wusu City

LI Jing[1], YANG Liao[1,2], WU Wei[1,2], XING Chenghong[1,3], JIANG Chengeng[1,3], ZENG You[1,3]
(1. Binjiang ZJUT Institute of Artificial Intelligence, Hangzhou 310056, China;
2. College of Geographical Information, Zhejiang University of Technology, Hangzhou 310014, China;
3. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310014, China)

Abstract: The classification and extraction of cotton cultivation information is of great significance for agricultural production and policy-making. Addressing issues like image loss, feature redundancy, and lengthy processing times in existing methods, we constructed a high-frequency time series dataset based on Sentinel-2 imagery. We used an object-oriented approach to collect the central points of farmland plots, created tabular data with classification features, and used automated machine learning to extract and analyze the cotton cultivation in Wusu City, Xinjiang Uygur Autonomous Region. The results show that this method effectively reduces “salt-and-pepper” noise and improves classification efficiency and accuracy. Comparing decision tree, random forest, LightGBM, and integration models, the integration model performed best with accuracy, precision, and recall rates of 95.14%, 94.93%, and 98.50%, respectively.

Key words: automated machine learning, cotton extraction, object-oriented, central point of plot, kNDVI time series

随着全球对地观测技术的飞速发展，遥感影像数据处理和农作物信息提取技术取得了显著的进步。农作物分类作为农业遥感领域的关键技术，对于提升农业生产效率、实现精准农业和可持续发展具有重大意义[1]。然而，在实际应用中，随着遥感数据量的急剧增加，传统分类方法的局限性日益明显[2-4]。因此迫切需要更高效、自动化的处理方法。

本文采用面向对象的地块中心点方法，构建自动化机器学习棉花分类模型。挖掘农作物的植被指数特征与物候特征，在新疆乌苏市开展棉花种植情况的分类研究，以期为解决传统的遥感影像农作物分类问题提供技术参考，促进农业遥感技术的应用与发展。

1　研究区概况及数据源

1.1　研究区概况

乌苏市位于新疆维吾尔自治区北部，是塔城地区下辖的一个县级市。地理位置坐标为83°40′~85°11′E，43°57′~45°29′N。地势由北向南逐渐升高，北部为准噶尔盆地，南部为天山山脉，中部则为盆地与山地的过渡带[5]。乌苏市属于干旱的温带大陆性气候，具有气温年较差和日较差大、光照充足、降雨稀少，蒸发强烈的特点。这些自然条件为农业生产活动提供了独特的机

第一作者简介：李婧（1994—），硕士研究生，工程师，主要研究方向为地理信息系统工程与应用，E-mail：lijing181@mails.ucas.ac.cn。
通信作者：杨辽（1972—），本科，正高级工程师，主要研究方向为遥感与地理信息应用，E-mail: yangliao@yeah.net。

遇与挑战。

1.2　数据源

1.2.1　遥感数据

以欧空局提供的Sentinel-2多光谱遥感数据为主要数据源，获取乌苏市所在位置的Sentinel-2格网影像，如图1所示。本文选取蓝、绿、红、近红外共4个波段数据，用于提取棉花分类的核心特征。

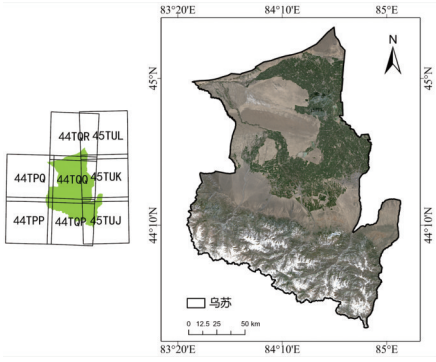


图1　研究区位置及Sentinel-2格网编号

审图号：GS(2024)0650号

考虑到5-10月为该地区棉花的关键生长期，本文以2020年1-12月共96期Sentinel-2多光谱遥感数据为基础，重构了5-10月等时间间隔的遥感影像数据集，作为棉花分类的数据源。

1.2.2　地面调查数据

地面调查数据主要使用2020年在乌苏地区的实地调绘数据，共采集了1 153个棉花样本斑块。实验样本包含782个参与分类的训练样本，371个用于精度评价的验证样本。其中，前者包括565个正样本和217个负样本。后者包括266个正样本和105个负样本。

2　研究方法

本文基于Sentinel-2卫星数据进行高频时间序列重构，并结合实地调查数据，运用决策树、随机森林、LightGBM以及集成模型4种算法，构建了基于表格数据的自动化机器学习棉花分类模型，提取乌苏市棉花种植信息并进行精度评价。研究的目的是探究自动化机器学习方法在面向对象的高频时间序列数据棉花信息提取中的应用，以及不同模型对提升棉花提取精度的作用。技术路线如图2所示。

2.1　植被指数提取

植被指数是通过线性或非线性计算方法处理地表在特定波谱范围内的反射率，以高亮显示植物的关键特性。在遥感技术中，归一化植被指数（normalize difference vegetation index，NDVI）通过标准化处理减轻了大气和传感器的影响，从而显著提高了对植被反

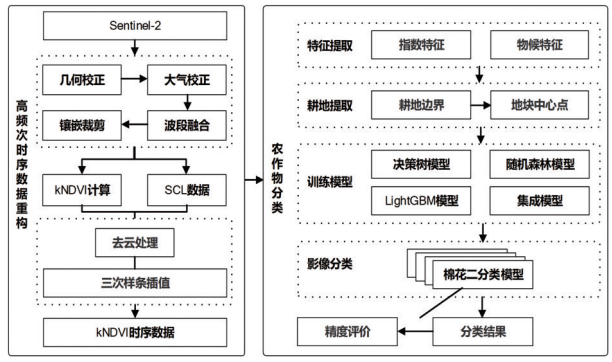


图2　技术路线

应的灵敏度和准确性，成为提取农作物信息的常用指标[6]。而kNDVI（kernel normalized difference vegetation index）是对NDVI的一种非线性扩展，它利用核方法原理和机器学习技术以提高分析的准确性[7]。与NDVI相比，kNDVI在处理信号饱和、应对复杂的植被周期和季节性变化以及解决混合像素问题等方面具有更明显优势。这些特点使得kNDVI与生态系统生产力之间存在更强的相关性，为遥感领域提供了一种更为高效和精确的植被分析工具。因此，本文选取kNDVI构建棉花的分类特征集，其计算公式如下：

$$\mathrm{NDVI}=\frac{(\mathrm{NIR}-\mathrm{Red})}{(\mathrm{NIR}+\mathrm{Red})}$$

$$\mathrm{kNDVI}=\tanh(\mathrm{NDVI}^2)$$

式中，NIR为近红外波段（B8）的反射率；Red为红波段（B4）的反射率。

2.2　高频时间序列重构与评估方法

三次样条插值技术是一种分段多项式插值方法，能够有效地拟合原始数据，创建平滑连续的曲线，填补时间序列中缺失的数据。该方法将数据点分割成多个区间，并在每个区间上定义一个三次多项式函数，构造连续、平滑的插值函数，确保这些函数在数据点上连续，而且一阶和二阶导数也连续[8]，通过这些不等间隔的时间点观测值，可推测出相等间隔时间点的预测值。因此，针对数据中云层遮挡引起的图像空缺以及某些区域数据覆盖不全的问题，本文将三次样条插值运用到植被指数时间序列数据重构中[9]，见图3。

时间序列影像重构的评估可以通过定性和定量2种方法。定性评估主要依赖于视觉判断；而定量评估则从误差和数据结构的相似性2个维度进行。本文采用的定量评估指标包括均方根误差（root mean square error，RMSE）、决定系数（R^2）和峰值信噪比（peak signal-to-noise ratio，PSNR）[10]。计算公式如下：

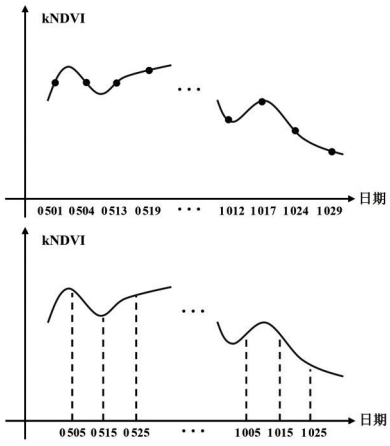


图3 三次样条插值原理图

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$\mathrm{PSNR}=10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right)$$

式中，n 为数据点的数量；MAX 为图像像素的最大可能值；MSE 为均方误差。

2.3 面向对象的地块中心点分类特征集构建

通过引入面向对象的思想，选取耕地地块，提取地块中心纯净像素点，创建了一个面向对象的农田图斑数据集，为了减少噪声干扰问题，本文采用基于地块中心点的分类方法，利用农田图斑的中心点代替耕地地块，提取高频时间序列的植被指数，构建面向对象的数据集参与棉花的分类过程。通过对乌苏市所有耕地进行编码，获得面向对象的地块中心点集。

地块中心点的计算采用地块数据的重心，重心定义为图形的几何中心，权重是对象的面积，计算公式如下：

$$\bar{x}=\frac{1}{6A}\sum_{i=1}^{n}(x_i+x_{i+1})(x_iy_{i+1}-x_{i+1}y_i)$$

$$\bar{y}=\frac{1}{6A}\sum_{i=1}^{n}(y_i+y_{i+1})(x_iy_{i+1}-x_{i+1}y_i)$$

$$A=\frac{1}{2}\sum_{i=1}^{n}(x_iy_{i+1}-x_{i+1}y_i)$$

式中，图形顶点坐标定义为（x_1，y_1），（x_2，y_2），…，（x_n，y_n），$x_{n+1}=x_1$，$y_{n+1}=y_1$，图形闭合，重心为（$\bar{x}$，$\bar{y}$），A 为图形面积。

2.4 农作物分类模型构建

分类实验以样本点为输入数据，采用相同的样本集，基于 AutoML 工具训练了决策树、随机森林、LightGBM 以及集成模型4种分类模型，并将其分别应用于预测数据集以获取分类结果，同时对结果进行精度评价。其中，集成模型（Ensemble）是模型的集合，每个模型的权重由 AutoML 工具根据不同模型在验证数据上的性能，使用加权平均方法自动计算得出。本文所采用的集成模型集合了决策树、随机森林、LightGBM 这3种模型。该方法能够有效地融合来自不同单个模型的预测，以此来减少过拟合、增强模型的泛化能力，并在复杂的数据集上达到更高的精确度。

2.5 精度评价

在机器学习和统计分类领域，真正例（true positive，*TP*）、假正例（false positive，*FP*）、真负例（true negative，*TN*）、假负例（false negative，*FN*）是评估分类模型性能的4个基本指标[11]，见表1。其中，*TP* 代表模型正确预测正类的实例数；*FP* 代表模型将负类误判为正类的实例数；*TN* 代表模型正确预测负类的实例数；*FN* 代表模型将正类误判为负类的实例数[11]。

表1 混淆矩阵结构

混淆矩阵		真实类别	
		Positive	Negative
预测类别	Positive	*TP*	*FP*
	Negative	*FN*	*TN*

准确率（overall accuracy，*OA*）、精确率（precision，*P*）、召回率（recall，*R*）是基于 *TP*、*FP*、*TN*、*FN* 的概念，用于评估分类模型性能的3个关键指标[12]。本文利用地面调查数据集中的验证样本，采用准确率、精确率和召回率这3个指标对决策树、随机森林、LightGBM 和集成模型4个分类模型的分类结果进行精度评价，通过计算这些指标，可以全面地评估不同模型在分类任务中的表现，了解模型在识别棉花方面的准确性、效率和完整性。3种评价指标的计算方式如下：

$$OA=\frac{TP+TN}{TP+FP+TN+FN}\times 100\%$$

$$P=\frac{TP}{TP+FP}\times 100\%$$

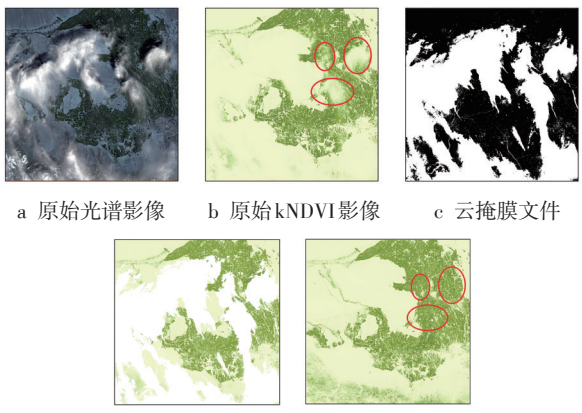
$$R=\frac{TP}{TP+FN}\times 100\%$$

3 结果与分析

3.1 高频时间序列重构结果及评价

本文按3次样条插值方法最终获取了2020年5月5日–10月25日的相等时间间隔的新疆维吾尔自治区

北部乌苏地区时间序列遥感数据，每月3期，共18期。以2020年7月的影像为例，在图5展示时序数据重构的步骤的效果。其中，图5a为原始光谱影像，显示为真彩色图像，大块白色区域代表云层覆盖。图5b为原始kNDVI影像，是根据原始光谱影像数据计算得到的kNDVI植被指数，红色圈出的区域为受云层干扰的耕地区域。图5c为云掩膜文件，显示云层检测结果，白色区域为云层和云影。图5d为kNDVI影像去云效果图，应用云掩膜移除影像中的云层影像。图5e为kNDVI影像重构效果图，通过时序分析和云去除技术，重构出清晰的植被指数图，红圈区域展示了经过云去除和数据重构后的耕地效果，呈现了无云状态下植被的真实情况。从视觉效果上看，时间序列重构有效地解决了原始影像中云层污染的干扰的问题，为后续的遥感分析提供了高质量的植被指数数据。



a 原始光谱影像　b 原始kNDVI影像　c 云掩膜文件

d kNDVI影像去云　e kNDVI影像重构

图5　时间序列重构结果

进一步地，以峰值信噪比（PSNR）、均方根误差（RMSE）和决定系数（R^2）作为为评价因子，定量探究时序重构影像的准确性。表2展示了原始影像和重构影像，同参考影像的PSNR、RMSE和R^2值的对比分析。结果表明，重构影像在所有评估指标均优于原始图像。具体而言，原始影像与参考影像的PSNR为33.38，而重构影像与参考影像的PSNR提高到40.75，这表明重构影像与参考影像更为相似；重构影像与参考影像的RMSE值为2.20，低于原始影像的2.24，这表明重构结果的误差较小；重构影像与参考影像的R^2值达到0.96，高于原始影像的0.75，说明重构后的影像与参考影像相关程度更大。综上所述，时序重构技术不仅有效解决了原始影像中的云层干扰问题，而且显著提升了影像数据的质量，能够为后续的遥感分析和农作物监测提供更为可靠的数据支持。

3.2　分类特征评价

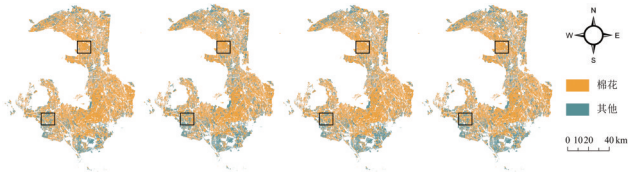
本文以kNDVI指数为分类特征，结合时间序列信息建立时序kNDVI特征集，分别对4种分类模型时序kNDVI特征的重要性进行了优选和评估，见图6，其中颜色越深代表分类特征对于模型的贡献度越高。结果表明，6月5日，10月25日和10月15日kNDVI值在所有模型中都显示出较高的重要性。这可能是因为6月上中旬棉花处于苗期，其快速生长导致kNDVI值迅速上升；而10月中下旬则是棉花的吐絮期，kNDVI值随之下降。由此可见，变化明显的kNDVI值是分类的关键分类特征，对应的日期则为棉花分类的关键物候期。此外，相比其他模型，集成模型中多数时序特征重要性都较高，这说明集成模型能够最大限度地运用时序特征来确保分类的准确度，并适当地采取了特征优化策略来提高分类效率。这进一步验证了集成模型在处理复杂数据和提高分类性能方面的优势。

表2　时间序列重构评价结果

影像	PSNR	RMSE	R^2
原始影像	33.38	2.24	0.75
重构影像	40.75	2.20	0.96

3.3　分类结果对比

基于4种分类模型获得的结果，本文绘制了研究区内棉花和其他作物的空间分布图（图6）。从4种分类模型得到的结果来看，棉花的种植分布显示出较高的一致性。图上地块显示为大片黄色，代表种植的农作物为棉花，这是因为乌苏地区以棉花为主要农作物，在耕地区域的大部分地区均有种植。而研究区的南部地区地块则显示为绿色，代表种植了棉花以外的其他作物，这与实地考察的结果对应，即，该区域种植了部分玉米、番茄和蔬菜等其他作物。



a 决策树模型　b 随机森林模型　c LightGBM模型　d 集成模型

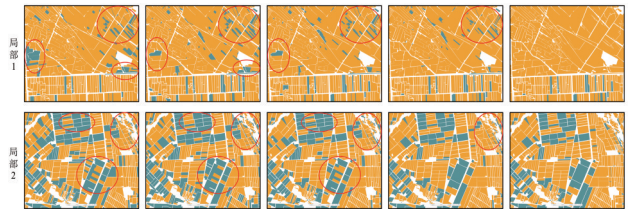
图6　基于4种模型的棉花分类结果图（总体）

审图号：GS(2024)0650号

通过比较模型分类面积和实地调查面积，发现集成模型是表现最好的模型，其分类的棉花面积和实际棉花面积更为接近。

同时图7展示了研究区内2个局部地区的结果图（对应于图6中黑框标注位置），具体比较了4种模型分类结果（a~d）与实地调查结果（e）的差异。红色圆圈标注了各个模型错误分类的细节。结果显示，利用集成模型分类，错误分类地块相对较少，分类结

果基本与实地调查结果一致。此外，图8还显示了决策树、随机森林和LightGBM模型分类结果中的棉花面积，相比于集成模型偏少。经过与实地调查数据对比，发现这是由“棉花-玉米”的混分现象造成的。玉米作为乌苏地区种植量仅次于棉花的作物，其分布相对分散，在图7中大多数显示为绿色的耕地地块事实上是玉米的种植区域。从对验证样本的进一步分析结果来看，由于玉米和棉花在物候特征上的相似性，决策树、随机森林和LightGBM模型都存在不同程度的棉花与玉米间的分类混淆，这可以解释3种模型在预测棉花种植面积时面积偏少的现象。



a 决策树模型　b 随机森林模型　c LightGBM模型　d 集成模型　e 实地调查结果

图7　基于4种模型的棉花分类结果图（局部）

3.4　分类特征评价

如表2所示，模型精度降低的主要原因在假正例（*FP*）和假负例（*FN*）的较高值，这主要是由棉花与其他作物分类混淆所致。在4种模型中，集成模型的*FP*和*FN*最小，仅占4.9 %，这表明其分类混淆程度较低，从而使得集成模型取得了最优的性能。

表2　混淆矩阵基本指标

分类模型	*TP*	*FP*	*TN*	*FN*
决策树	253	35	70	13
随机森林	251	20	85	15
LightGBM	254	16	89	12
集成模型	262	14	91	4

表3的分类精度也表明集成模型取得了最优的分类性能，取得了95.14 %的准确率、94.93 %的精确率以及98.50 %的召回率，在所有指标上均高于其他3种模型，表明其减少了误分类情况，整体性能最优。

表3　分类精度对比

分类模型	*OA*	*P*	*R*
决策树	87.06	87.85	95.11
随机森林	90.57	92.62	94.36
LightGBM	92.45	94.07	95.49
集成模型	95.14	94.93	98.50

综上所述，不同模型的分类结果存在一定差别，除了决策树模型之外，其他模型的准确率均在90 %以上，达到了较高的分类精度。这也验证了面向对象的高频时间序列分类方法可以快速和准确地完成棉花种植提取。此外，研究结果还验证了基于表格数据的自动化机器学习技术在快速识别棉花种植信息的可行性。

4　结　语

本文针对遥感影像棉花种植提取，采用面向对象的地块中心点集和kNDVI植被指数，结合时间序列重构与AutoML技术，实现了自动化分类。决策树、随机森林、LightGBM及集成模型平均精度超85%，有效抑制噪声并提升特征区分力。未来将着力解决作物混分问题，进一步提高准确性。

参考文献

[1] 李思佳,孙艳楠,李蒙,等. 国内外农作物遥感估产的研究进展[J]. 世界农业,2013,34(5):125-127

[2] 金梦婷,徐权,郭鹏,等. 基于面向对象多特征学习的无人机影像农作物精细分类方法[J]. 遥感技术与应用,2023,38(3):588-598

[3] 余晓敏,湛飞并,廖明生,等. 利用改进SEaTH算法的面向对象分类特征选择方法[J]. 武汉大学学报(信息科学版),2012,37(8):921-924

[4] 黄双燕,杨辽,陈曦,等. 机器学习法的干旱区典型农作物分类[J]. 光谱学与光谱分析,2018,38(10):3 169-3 176

[5] 姜洁. 新疆玛纳斯河中下游流域湿地维管植物研究[D]. 石河子:石河子大学,2018

[6] 李杭燕. 时间序列NDVI数据集重建方法研究[D]. 兰州:兰州大学,2010

[7] Camps-valls G, Campos-taberner M, Moreno-martínez Á, et al. A Unified Vegetation Index for Quantifying the Terrestrial Biosphere[J]. Science Advances,2021,7(9):eabc7 447

[8] 蒋祖华,徐韶勇. 高阶多段凸轮型线的拟合方法研究及软件开发[J]. 机械,2003,41(3):1-3

[9] 郭庭威,黄红莲,孙晓兵,等. 环境二号卫星多光谱图像的薄云检测及去除[J]. 大气与环境光学学报,2023,18(4):383-400

[10] 李瑄,李大成,陈金勇,等. 基于时空融合的Landsat反射率数据时序重建与分类质量评价[J]. 太原理工大学学报,2020,51(6):889-899

[11] 钱骥. 基于机器学习的KPI异常检测技术研究[D]. 长沙:国防科技大学,2020

[12] 赵岩. 基于机器学习的白龙江流域潜在低频泥石流沟识别[D]. 兰州:兰州大学,2020