doi:10.3969/j.issn.1003-2029.2019.05.018

随机森林模型在遥感水深反演中的应用

邱耀炜[1,2], 沈 蔚[1,2*], 纪 茜[1,2]

(1. 上海海洋大学 海洋科学学院, 上海 201306; 2. 上海河口海洋测绘工程技术研究中心, 上海 201306)

摘 要: 随着我国浅海测绘需求的日益增长, 文中利用四波段的 WorldView-2 高分辨率遥感影像, 选取我国南海西沙群岛中的甘泉岛和台湾南湾地区作为典型试验区, 开展水深反演研究。引入随机森林算法构建了随机森林水深反演模型, 并同常用的 3 种水深反演模型进行精度对比。结果表明, 在甘泉岛和南湾地区随机森林模型反演的水深值和真实水深值的 *RMSE* 分别为 0.85 m 和 1.59 m, *MRE* 分别为 8%和 12%, 均优于其他 3 种模型。

关键词: 水深遥感; Worldview-2; 随机森林; 甘泉岛; 台湾南湾

中图分类号: P237 **文献标志码**: A **文章编号**: 1003-2029(2019)05-0098-06

海洋在我国经济发展格局和对外开放中的作用日益重要, 在维护国家主权、安全、发展利益中的地位更加突出[1], 如何快速、准确地获取浅海地区的水深显得尤为重要。越来越多的卫星成功发射升空, 为水深反演提供了数据支持, 同时也促进了多光谱水深反演模型的迅速的发展。国内外学者在此基础上取得了多样的成果。例如, Lyzenga[2]根据比尔定律, 分析了电磁波在水体中的辐射传输方程, 推导出了单波段线性水深反演模型, 并首次利用多光谱航拍影像对浅海地区进行了水深估算; Polcyn 等[3]发现两波段辐射值的比值会随水深的增大而减小, 在此基础上利用波段比值算法反演了相对水深; Spitzer 等[4]从辐射传输模型出发, 分析了太阳辐射反射光谱特征, 提出了一种可以反演水深信息的双流程辐射模型; Lyzenga 等[5]在忽略水体内部散射的情况, 将参数较多的双层流模型进行了简化, 提出了一种简单的多光谱水深反演模型。

随机森林(Random Forest)是美国科学院院士 Leo Breiman 和 Adele Culter 教授在 2001 年提出的。它是一种基于 CART 决策树(Classification and Regression Tree)的机器学习算法[6]。目前, 随机森林已经在遥感图像分类、作物识别、自然语言处理、经济学分析等多个领域得到了应用[7-9], 并取得了显著的效果。研究表明, 随机森林具有更强的泛化能力, 可用于分类和回归。虽然随机森林算法应用于众多的领域, 但目前在水深反演上的应用还较少。相关研究结果表明随机森林比单棵的决策树更稳健、泛化性能好, 其在非线性回归上表现十分出众, 非常适合解决非线性的复杂问题[10-12]。随机森林算法的关键在于选择合适特征数量和决策树数量进行分类或预测[13]。因此, 本文的研究内容还包括对随机森林水深反演模型中的决策树个数以及特征数的选取, 进行分析和研究。

1 水深反演方法

1.1 随机森林

随机森林是对 Boost aggregation 集成思想和特征随机选取思想的组合, 也正是由于使用了 CART 决策树作为基础分类器, 随机森林除了能作为分类器, 同时也能完成回归分析, 这可以通过模型预测值来取代模型的分类值来实现。训练过程首先采用 bootstrap 自助抽样技术有放回地随机抽取训练样

收稿日期: 2019-04-26
基金项目: 上海市科委重点科研资助项目(14590502200)
作者简介: 邱耀炜(1994 -), 男, 硕士研究生, 主要研究方向为海洋遥感技术与应用。 E-mail : qiuyw_rs@163.com
通讯作者: 沈蔚(1977 -), 男, 博士、教授, 主要研究方向为海洋遥感与测绘。E-mail : wshen@shou.edu.cn

本，形成各个决策树的样本子集，其次采用 CART 二元划分策略构建与样本子集对应的决策树,每个决策树的每个节点也采用 bootstrap 自助抽样技术随机抽取 F 个特征(F 小于总特征数量 P),通过计算每个特征包含的信息采用最优分裂策略进行左右分裂生长[14],最优分裂的依据如式(1)所示:

$$\frac{I}{N_l}\sum_a\left(\left|C_{a,l}\right|\right)^2+\frac{I}{N_r}\sum_a\left(\left|C_{a,r}\right|\right)^2 \tag{1}$$

式中:N_l 为左分裂的样本总数;N_r 为右分裂的样本总数;$C_{a,l}$ 为左分裂中预测值 a 的样本个数;$C_{a,r}$ 为右分裂中预测值 a 的样本个数。

对于一个输入的样本根据训练得到的各个决策数进行迭代分枝，直至每个决策树的叶节点,各个叶节点的预测结果也就是这棵树最终做出的预测值,最后利用式(2)对各棵决策树上的预测值取平均,那么就可以得出此输入据集在整个随机森林上预测结果。

$$p(c|v)=\sum_{t=i}^{T}P_t(c|v) \tag{2}$$

式中:T 为预测程中生成的树的总数;c 为某一个特定的值;P 表示的是概率函数。

随机森林中两个随机量的引入——随机选择样本和随机选择特征,使得随机森林算随着树的增加,泛化误差趋向一个上界,具有良好的防止过拟合的能力,不必担心过度拟合;在算法计算过程中,随机选取训练样本集,随机选取分裂属性集,不需要过多人工干预就能构建模型，建模过程简单方便。同时,随机森林训练的计算量是和包含的决策树的数目成正比，各个树之间又是可以并行处理,运算速度特别快,能够快速地在训练好的树上进行预测。

1.2　其他方法

遥感水深反演理论解析模型从辐射传输模型出发来求取水深值,需要多种水体内部参数。由于水体内部光学参数获取困难,模型求解复杂,没有得到广泛的应用。在此基础上发展的半理论半经验模型,是一种简化模式的理论解析模型。目前应用较广泛的有以下 3 种模型:

(1) 单波段线性回归模型:

$$Z=a\cdot\ln(L_i-L_{si})+b \tag{3}$$

(2) 双波段比值线性回归模型:

$$Z=a\cdot\frac{\ln(L_i-L_{si})}{\ln(L_j-L_{sj})}+b \tag{4}$$

(3) 多波段组合线性回归模型:

$$Z=a_0+\sum_{i=1}^{n}a_i\cdot\ln(L_i-L_{si}) \tag{5}$$

式(4)~式(6)中:Z 为水深值;$L_i(L_j)$为影像第 $i(j)$波段的辐射亮度;$L_{si}(L_j)$为影像第 $i(j)$波段深水区辐射亮度,反映了水面辐射、水体散射及大气散射等的总和,不包含底质反射;$a(a_i)$为回归系数。

1.3　技术路线

图 1 为本文的水深遥感反演技术路线图,首先对原始遥感影像进行预处理,包括辐射校正、几何校正和水陆分离。然后,利用 4 种水深反演模型进行水深反演，并利用水深检查点进行精度评价,获得最优模型的水深反演结果,最后进行制图输出。

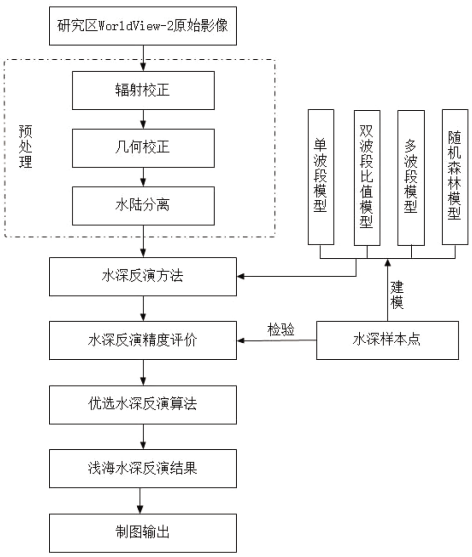

图 1　水深反演技术路线图

2　随机森林参数

2.1　随机森林性能指标

随机森林模型性能主要受到来自内外两方面因素影响。外部因素主要是指训练样本的情况,训练样本的规模,包括样本的变量个数、样本变量的类型和样本的大小；训练样本的正负类样本分布,即训练样本的平衡。因此在选择样本时,除了考虑样本空间分布均衡外，还需要设定合理的样本容量。本文采用基于多项式分布样本容量法求解验证样本的容量,公式如下:

$$N=\frac{D\sigma_i(1-\sigma_i)}{d_i^2} \tag{6}$$

式中:σ_i 为 k 个预测值中最接近 50%的第 i 个

预测值的总体比例，k 为总个数；d_i 为该预测值的期望精度；D为自由度为 1 且服从卡方分布的百分比。

内部因素主要是指随机森林中树的强度和相关性，可通过泛化误差和运行效率两方面来考虑。随机森林算法是使用 bootstrap 方法生产训练样本集的，而在生成训练样本集时，有一些样本是不能被抽取的组成了袋外数据（out of bag，OOB），这些样本未被抽中的概率为$(1-1/N)^N$，其中 N 是训练样本集的总个数。当 N 趋近于无穷大时，$(1-1/N)^N$ 将收敛于 $1/e$，约为 0.368，即有将近 36.8%的样本不会被抽中。对于每一棵数都可以利用袋外数据进行 OOB 误差估计，平均随机森林中所有的决策树的 OOB 误差估计，就可以得到整个随机森林的泛化误差，整个过程也被称为 OOB 估计。

OOB 误差估计是在整个随机森林算法决策树生成的过程计算获得的，随机森林算法可以通过并行处理同时计算出每一棵决策树生成时的 OOB 误差率，对于整个随机森林算法的泛化误差估计的计算，与交叉验证相比运行效率高，占用资源少，且其结果近似于交叉验证的结果[15]。因此本文采用 OOB 估计来评价随机森林算法的性能指标，OOB 估计越小，说明算法的性能越好。

2.2 随机森林参数优化

随机森林作为一种多决策树组合预测（分类）器，每一棵决策树在训练过程中都是随机抽取一定量的样本和特征，要想获得较好的集成预测性能，就需要对两个参数进行优化：每棵决策树上每一个节点随机抽取的特征个数 F 和森林中的决策树总个数 T。一般来说，随机森林算法对于回归预测的默认的 F 参考值为 $P/3$，而对于分类时，默认的 F 参数参考值为$\sqrt{P}$，其中 P 为训练特征集中的特征数总和。然而，对于不同的特征集，最适合模型的参数 F 值也不尽相同。而对于 T 参数，设置过低会导致预测误差较高，设置过高会增加模型复杂度，且并不一定能对最终的预测结果产生影响。

本研究通过调用 R 语言中的“RandomForest”包，并利用训练样本点的对应 4 个波段 B1，B2，B3，B4 和各个波段之间的比值 B1/B2，B1/B3，B1/B4，B2/B3，B2/B4，B3/B4 作为特征集，输入随机森林模型中进行参数优化。模型的参数优化则采用 tuneRF 函数对 F 和 T 参数进行优化，获取最优值。利用袋外样本数据进行模型内部的误差估计，产生 OOB 误差，当 OOB 误差的值到达最小时，此时 F 和 T 的组合将作为最有参数组合，基于此构建随机森林模型，最终对研究区域进行水深反演。

为了分析随机森林中的决策树数量 T 对算法性能的影响，可以控制模型中的特征数量 F 的值不变，不断调整决策树的总数量 T 的值。同时以随机森林算法的 OOB 误差为精度指标，进行 1 000 次实验，取精度指标的平均值，该值越小说明此时模型的泛化误差越小，模型的精度也就相应的越高。这里先取 F 为推荐的参考值 3，设置 T 参数的取值范围为 10 到 2 000，步长为 10，作 OOB 误差与决策树数量 T 的关系图，如图 2。

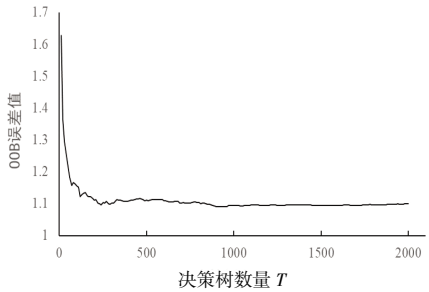

图 2 OOB 误差与决策树数量 T 的关系图

从图 2 可以看出，当特征数 F=3 时，当决策树的数量为 1 000~2 000 时，OOB 误差值基本趋于稳定，不再变化，说明当决策树数量达到一定程度后，随机森林算法的精度趋向于稳定，其性能也相对更加的稳定。同理，取 T 为 1 000，设置 F 参数的取值范围为 1 到 10，步长为 1，作 OOB 误差与特征数 F 的关系图，如图 3 所示。

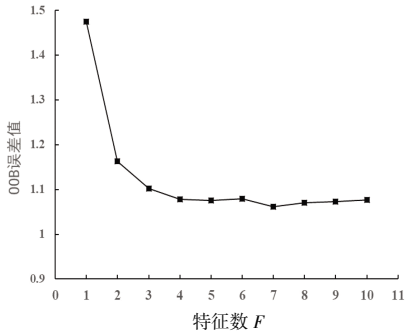

图 3 OOB 误差与特征数 F 的关系图

从图 3 可以看出，当随机抽取的特征数量在 1~4 范围时，OOB 误差值随着特征数量增加而逐渐降低，说明随机森林算法的精度随着模型中的特征数量增大而提升。此时，F 参数值与随机森林算法的预测性能是正相关的；当随机抽取的特征数量在 4~10 时，OOB 误差值整体趋于稳定，其中当 F=7 时，OOB 的误差最小。综上考虑，选定随机森林模型最优参数组合为 T=1 000，F=7。

3　实验与分析

3.1　研究区域与数据

本文采用的研究区域为甘泉岛和南湾。甘泉岛位于我国西沙群岛永乐群岛的永乐环礁西部,南湾是我国台湾省最南端两陆岬的行界点。使用的原始影像为高分辨率的 WorldView-2 卫星影像,多光谱分辨率为 2 m,全色分辨率为 0.5 m,有蓝(Blue)、绿(Green)、红(Red)、近红外(NIR)4 个波段。

使用的实测数据主要来自于由机载雷达测深系统获得的点云数据以及购买的官方电子海图。台湾南湾采用的电子海图为 S57 标准格式,编号为 C1514940,比例尺为 1:700 000,将海图上的声呐水深点矢量化后获得 203 个水深点,均匀且随机的取建模点 141 个,检验点 62 个。甘泉岛点云的数据机载雷达测深系统型号为加拿大 Optech 公司生产的 SHOALS-3000,采集时间是 2013 年 1 月 29 日,精度较高,可代替真实水深值进行建模和验证。其水平精度 2.0 m,测深精度 25 cm,可测深范围为 0.2~50 m。对点云数据进行抽稀,本试验一共采用 494 点实测水深点,选取其中 353 个作为建模点,141 个作为检验点。

3.2　预处理

本文使用的 WorldView-2 数据属于 Digital Global 公司的高分辨率商业卫星数据,Digital Global 给出了 QuickBird、WorldView 系列卫星影像的辐射亮度转换公式(7),通过该公式可以直接将无量纲的灰度值转换为大气层顶 (Top-of-Atmosphere, TOA)光谱辐射亮度值 $L_{\mathrm{TOA}}(\lambda)$。

$$L_{\mathrm{TOA}}(\lambda)=\frac{q(\lambda)*absCalFactor(\lambda)}{effetiveBandwidth(\lambda)} \tag{7}$$

式中:$absCalFactor\ (\lambda)$ 为波段 λ 的绝对定标因子;$effetiveBandwidth(\lambda)$为波段 λ 的波段有效宽度,该元数据信息存储在.IMD 文件中。

大气校正使用 ENVI 中 FLAASH 模块进行。由于缺少 870~1 020 nm 范围的短波红外波段,不能用 K-T 模型反演气溶胶,因此气溶胶反演选 NONE。接着进行几何校正。为了更有效地获取遥感影像上水体部分的水深信息,增强海体部分信息,减少非水体部分对水深反演的影响,采用面向对象的方法对研究区的遥感影像进行水陆分离。

3.3　实验结果与分析

为了探究各个模型反演水深的能力,对随机森林模型构建中进行参数优化,确定随机森林水深反演算法中最佳的模型参数,即特征数量 F 和决策树数量 T,并同其余 3 种模型进行比较。本文采用均方根误差($RMSE$)和平均相对误差(MRE)来对模型的反演精度进行定量的评价,各个参数的计算公式可表达为:

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y})^2} \tag{8}$$

$$MRE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}}{y_i}\right|\cdot 100\% \tag{9}$$

式中:y_i 表示样本点的真实水深值;$\hat{y}$表示的是反演的水深值;n 为样本点的个数。其中,MRE 和 $RMSE$ 越小,则反演结果精度越高。

为了定量地比较随机森林水深反演模型和其余 3 种常用水深反演模型的精度情况,计算了各个模型的水深反演值和实测水深值之间的 $RMSE$ 和 MRE,计算结果见表 1。

表 1　多光谱水深反演模型比较

	甘泉岛		南湾地区	
	$RMSE$/m	MRE/%	$RMSE$/m	MRE/%
单波段模型	2.10	18	2.77	25
双波段模型	2.52	17	6.08	51
多波段模型	1.10	10	6.12	52
随机森林模型	0.85	8	1.59	12

由表 1 可知,在甘泉岛和南湾地区随机森林模型相对于其余 3 种模型的 $RMSE$ 和 MRE 均为最小;甘泉岛地区 4 个模型的反演精度整体上优于南湾地区,推测可能与甘泉岛地区采用的是机载 LiDAR 获取的实测水深数据有关,而南湾地区的实测提取自海图,两者精度上有差异;从 3 个半理论半经验水深反演模型的精度对比来看,在甘泉岛地区多波段模型反演精度较优,而在南湾地区则是单波段模型较优,推测可能是两个研究区的水质情况导致的差异,甘泉岛地区水体清澈,符合理论上推导的前提假设,多波段模型结合了多个波段的信息所以反演精度较优,而南湾地区水质较为浑浊,具有一定的悬浮物质,不同波段在水体中的衰减系数的差值是稳定不变的理论假设并不成立,导致采用两个或多个波段进行水深反演时反而精度较低。然而,无论是水体清澈区域的甘泉岛还是水体较为浑

浊的南湾地区，随机森林模型都具有较好的反演精度，因为随机森林模型综合多个波段特征信息，且随机选择随机取用，同时也得益于其强大的非线性回归预测能力。

图 4~图 5 利用随机森林水深反演模型对研究区进行水深反演，计算获得水深栅格图。

4 结束语

本文提出了一种随机森林水深反演方法，并利用 WorldView-2 多光谱卫星影像在甘泉岛和台湾南湾地区开展水深反演试验，确定了模型在水深反演中过程的重要参数，并同常用的 3 种水深反演模型进行了精度对比。结果表明，随机森林水深反演模型精度可靠，同时也为大范围遥感获取水深信息提供了新的思路和方法。

本文的研究工作仍存在不足之处，在水深反演领域，随机森林模型的原始输入特征没有相关的先验知识，本文选取了各个波段和各个波段的比值共 10 个特征作为输入特征，具有一定的人为性和经验性，下一步考虑增加相关的特征进行研究，如水体指数等。

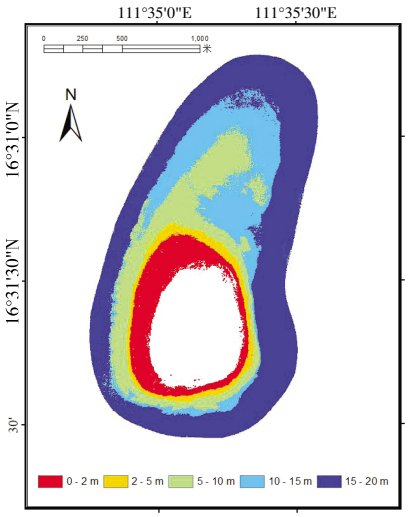

图 4 甘泉岛随机森林模型水深反演结果图

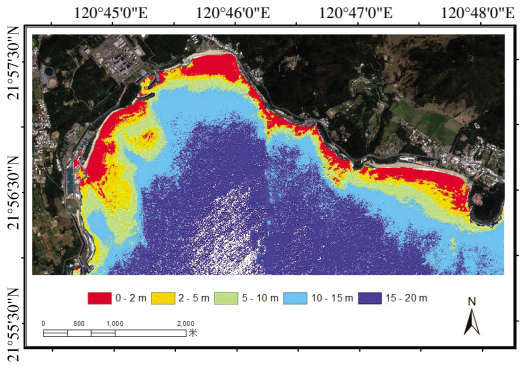

图 5 南湾随机森林模型水深反演结果图

参考文献：

[1] 徐胜. 走中国特色的海洋强国之路[J]. 求是，2013(21): 41-42.

[2] Lyzenga D R. Passive remote sensing techniques for mapping water depth and bottom features[J]. Applied Optics, 1978, 17(3): 379.

[3] Polcyn F C, Lyzenga D R. Calculation of water depth from ERTS-MSS data [C]// Proceedings Symposium on Significant Results Obtained from ERTS-1, New Carrollton, Maryland, 1973: 1433-1436.

[4] Spitzer D, Dirks R W J. Bottom influence on the reflectance of the sea [J]. International Journal of Remote Sensing, 1987, 8(3): 279-308.

[5] Lyzenga D R, Malinas N P, Tanis F J. Multispectral bathymetry using a simple physically based algorithm [J]. IEEE Transactions on Geoscience and Remote Sensing, 2006, 44(8): 2251-2259.

[6] Breiman L. Using iterated bagging to debias regressions[J]. Machine Learning. 2001.

[7] 张磊,宫兆宁,王启为,等.Sentinel-2 影像多特征优选的黄河三角洲湿地信息提取[J].遥感学报, 2019,23(02):313-326.

[8] 郭玉宝,池天河,彭玲,等. 利用随机森林的高分一号遥感数据进行城市用地分类[J]. 测绘通报, 2016(05): 73-76.

[9] 李高玲,帖云,齐林. 基于随机森林分类优化的多特征语音情感识别[J]. 微电子学与计算机, 2019, 36(01): 70-73.

[10] Ham J, Chen Y, Crawford M M, et al. Investigation of the random forest framework for classification of hyperspectral data [J]. IEEE Transactions on Geoscience & Remote Sensing, 2005, 43(3): 492-501.

[11] Isaac E, Easwarakumar K S, Isaac J. Urban landcover classification from multispectral image data using optimized AdaBoosted random forests[J]. Remote Sensing Letters, 2017, 8(4):350-359.

[12] Stumpf A, Kerle N. Object-oriented mapping of landslides using Random Forests[J]. Remote Sensing of Environment, 2011, 115(10): 2564-2577.

[13] 王宇恒. 推荐系统中随机森林算法的优化与应用[D]. 浙江: 浙江大学, 2016.
[14] 顾海燕,闫利,李海涛,等. 基于随机森林的地理要素面向对象自动解译方法[J]. 武汉大学学报(信息科学版), 2016, 41(02): 228-234.
[15] 方匡南,吴见彬,朱建平,等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(03): 32-38.

Satellite-Derived Bathymetry Using Random Forest Model

QIU Yao-wei[1,2], SHEN Wei[1,2], JI Qian[1,2]

1. *College of Marine Science, Shanghai Ocean University, Shanghai* 201306, *China*;
2. *Shanghai Engineering Research Center of Estuarine and Oceanographic Mapping, Shanghai* 201306, *China*

Abstract: With the growing demand for shallow sea surveying and mapping in China, this paper uses four-bands high resolution WorldView-2 images to conduct satellite-derived bathymetry experiments in the Ganquan Island in Xisha Islands and the Nanwan area in Taiwan. The random forest satellite-derived bathymetry model is constructed by using the random forest algorithm, and the accuracy is compared with three classic satellite-derived bathymetry model. The results show that the bathymetry accuracy of the random forest model is optimal in the Ganquan Island and Nanwan area, with *RMSE* of 0.85 m and 1.59 m, and *MRE* of 8% and 12%, respectively.

Key words: satellite-derived bathymetry; Worldview-2; random forest; Ganquan Island; Nanwan of Taiwan Island