DOI:10.19951/j.cnki.1672-9331.20211130008 文章编号:1672-9331(2023)01-0065-10
引用格式:邓兴升,王清阳.基于梯度提升决策树的植被高度模型研究[J].长沙理工大学学报(自然科学版),2023,20(1):65-74.
Citation:DENG Xingsheng,WANG Qingyang.Study on vegetation height model based on the gradient boosting decision tree[J].J Changsha Univ Sci Tech (Nat Sci),2023,20(1):65-74.

基于梯度提升决策树的植被高度模型研究

邓兴升,王清阳
(长沙理工大学 交通运输工程学院,湖南 长沙 410114)

摘 要:【目的】研究以航空摄影测量的方式建立植被高度模型。**【方法】**利用数字正射影像(DOM)与数字表面模型(DSM)提取光谱特征因子和几何特征因子,采用相关性指数对植被高度与特征因子进行相关性分析,筛选出特征因子。采用梯度提升决策树算法建立植被高度模型,并通过优化参数提高模型精度。**【结果】**在默认参数下,模型精度约为2.000 m;优化参数后,模型精度达到了0.900 m;剔除部分特征因子后,模型精度可达0.840 m;通过与支持向量机算法进行对比,植被高度模型整体精度由0.893 m提高至0.758 m,运行时间由70 min缩减至10 min。**【结论】**若不考虑建模原始数据的误差,采用梯度提升决策树算法建立的植被高度模型的精度为亚米级,多次试验中模型精度较为稳定。

关键词:植被高度;梯度提升决策树;特征因子;机器学习
中图分类号:P231 **文献标志码:**A

0 引言

准确的数字高程模型(digital elevation model, DEM)数据对于林业与测绘科学是必不可少的基础数据[1]。准确获取林区植被高度是众多学者研究的热门问题。航空摄影测量在采用可见光遥感进行林区数据摄影测量时,可见光信号不能透过植被到达地面,因此,只能获得数字表面模型(digital surface model, DSM)[2]。目前,针对森林植被高度测量主要有3种方式:第一种是人工采用传统测量工具进行实地测量[3],此方式精度较高,针对性强,但代价过大且测量范围有限,易受到自然条件等因素影响。第二种是利用机载激光雷达(light detection and ranging, LiDAR)进行航飞测量,该方法便捷高效但费用较高,较适合小面积测量,如郭鹏等[4]利用机载LiDAR数据提取植被冠层结构参数,用于农田区作物的高度估算;NIE等[5]利用机载LiDAR数据对地面生物量进行估算;段祝庚等[6]利用机载激光雷达数据,采用森林冠层高度模型凹坑去除方法,实现了对森林冠层高度的估算。第三种方式为合成孔径雷达干涉测量(interferometric synthetic aperture radar, InSAR),该方法适合大面积、大尺度的森林植被高度测量。因信号未完全穿透植被,虽InSAR技术精度较高[7],但在获取植被高度的精度方面仍需进一步研究,如沈鹏等[1]提出了融合升降轨的极化干涉SAR三层模型植被高度反演方法;张兵等[8]针对植被高度反演问题,提出了几何结构法和高程精度法用来反演植被高度;解清华等[9]提出了基于极化合成孔径雷达干涉测量(polarimetric interferometric synthetic aperture radar, PolInSAR)的非线性复数最小二乘森林高度反演算法;罗环敏等[10]提出了一种基于极化相干优化和非体散射去相干补偿的森林高度反演方法;解金卫等[11]利用PolInSAR对植被区高精度DSM进行反演。随

收稿日期:2021-11-30;**修回日期:**2022-06-21;**接受日期:**2022-06-28
基金项目:湖南省自然科学基金资助项目(2020JJ4601)
通信作者:邓兴升(1971—)(ORCID:0000-0003-0158-9227),男,副教授,主要从事摄影测量数据处理方面的研究。
E-mail:383500135@qq.com

着点云边界提取算法[12]精度的不断提高,基于倾斜影像提取光谱特征与几何特征建立的随机森林分类模型对建筑、树木、低矮植被进行分类的正确率逐步提高[13],这就使得利用光谱特征与几何特征对植被高度进行分类成为可能。数字摄影测量系统的数据经空三加密后,该系统可以输出的成果包括数字正射影像(digital orthophoto map, DOM)、DSM等,例如徕卡ADS100输出的DOM、DSM以及L1级影像数据可用于制作数字线画图(digital line graphic, DLG)等。基于航空摄影测量数据,DENG等[14]提出了基于航空遥感DOM光谱特征和DSM几何特征的支持向量机植被高度模型,该模型的计算结果精度可达到亚米级。为了进一步提高植被高度模型的精度和效率,本研究在文献[14]的基础上,提出利用梯度提升决策树(gradient boosting decision tree, GBDT)算法对植被高度进行估计。试验结果显示:GBDT算法在精度和效率上均优于支持向量机算法。

1 基于梯度提升决策树的植被高度模型提取方法

1.1 梯度提升决策树算法

梯度提升决策树(GBDT)[15]是一种应用于回归预测的机器学习方法,是以分类回归树(classification and regression tree, CART)为基学习器的Boosting集成学习算法,其通过迭代方式将较弱的学习器组合成一个较强的学习器[16]。在GBDT的每次迭代过程中,都在残差减小的梯度方向新建一棵CART决策树,经多次迭代,使残差趋近于0,最后将所有决策树的结果进行累加,获得最终的预测结果[17-18]。GBDT在大样本数据的预测中具有很好的效果,能够灵活地处理复杂的非线性关系[19],即使样本数据的类型不同,也可以较好地处理。相对于文献[14]采用的支持向量机算法,GBDT在进行数据处理时,运行速度更快,精度更高。本研究利用GBDT的这一优势,对特征因子与植被高度之间复杂的函数关系进行处理。植被高度GBDT预测模型的建立过程如下:

1)建立初始弱学习器:

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

式中:$L(y_i,\gamma)$为损失函数;γ为损失函数达到最小值时的常数。对于回归问题,GBDT采用平方误差损失函数,即:

$$L(y,f(x))=(y-f(x^2))^2 \tag{2}$$

式中:$f(x)$为机器学习植被高度模型预测值。

2)迭代次数m取1,2,…,M,将此损失函数的负梯度值r_{mi}作为残差的估计值,即:

$$r_{mi}=-\left(\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right)_{f(x)=f_{x-1}(x)}= -2(y_i-f(x_i)) \tag{3}$$

根据所有样本x_i的负梯度值r_{mi},得到由J个叶节点组成的决策树,定义其对应的叶节点区域为R_{mj},其中,$j=1,2,\cdots,J$。各个叶节点的最佳残差拟合值γ_{mj}为:

$$\gamma_{mj}=\arg\min_{\gamma_m}\sum_{x_i\in R_{mj}}[y_i-(f_{m-1}(x_i)+\gamma_m)]^2 \tag{4}$$

式中:γ_{mj}为迭代次数为m时的节点残差拟合值。

3)更新学习器$f_m(x)$,得到:

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}\gamma_{mj}I,x\in R_{mj} \tag{5}$$

$$I=\begin{cases}1,x\in R_{mj}\\0,x\notin R_{mj}\end{cases} \tag{6}$$

4)经过M次迭代,得到最终的植被高度预测模型为:

$$f(x)=f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}\gamma_{mj}I,x\in R_{tj} \tag{7}$$

利用GBDT算法建立植被高度模型,其建模过程包括数据预处理、特征提取、梯度提升决策树建模三部分。主要技术流程如图1所示。

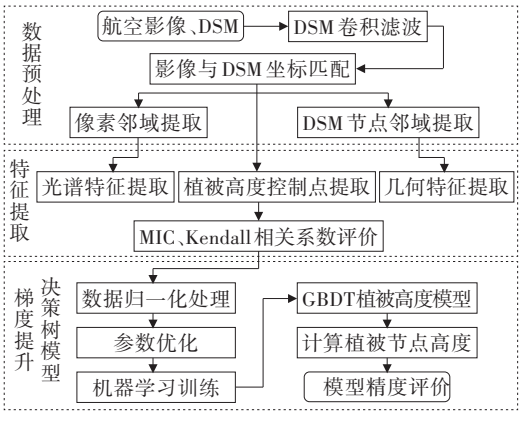

图1 植被高度梯度提升决策树建模流程图

Fig. 1 Modeling flow chart of vegetation height using the gradient boosting decision tree

1.2　特征因子提取

根据可见光植被指数,DSM节点可分为植被点、地面点两类。植被高度控制点,即DSM格网节点处植被高度的提取方法见文献[14]。

光谱特征因子主要包括颜色特征和植被指数。Lab颜色特征由两个颜色通道a和b、亮度L三个要素组成。植被指数选取可见光植被指数,有可见光波段差异植被指数(visible-band difference vegetation index, VDVI)、归一化绿蓝差异指数(normalized green-blue difference index, NGBDI)、归一化绿红差异指数(normalized green-red difference index, NGRDI)、超绿指数(excess green vegetation index, ExG)、超红指数(excess red vegetative index, ExR)、超绿减红指数(excess green minus excess red index, ExG-ExR)、红色减绿色(red minus green, RmG)、绿色减蓝色(green minus blue, GmB)、绿蓝与红绿差异(green-blue difference red-green, GBdRG)、红绿蓝(red green blue, RGB)等[14]。光谱特征因子的提取方法为:以DSM节点为中心,上、下、左、右4个方向各扩展25像素,即51×51像素块为卷积单元,计算其等权均值作为该节点的光谱特征值。阈值25像素为试验获得的经验值,其过小易受噪声影响,过大则会引起光谱特征过于平滑而失去节点的固有特征。

几何特征因子主要包括DSM节点所在位置的平面坐标与高程值、归一化高度、高度标准差、法向量、曲率、粗糙度以及节点各方向高差等。不同的几何特征反映了DSM表面不同的形态特征。

2　特征提取及相关性分析

2.1　试验数据

试验数据取自湘西某林区,海拔650 m。航空摄影获取的徕卡ADS100 L1级影像数据,影像地面分辨率大于0.2 m,每个像素大小为5 μm,航摄时航线飞行方向为正东西方向,平均飞行高度为2 500 m,航向重叠度为65%,旁向重叠度为35%。航空影像于2017年10月获得,试验区植被覆盖率为90%,平均坡度为40°。通过提取DSM格网节点处植被高度,得到植被高度最大值为33.24 m,平均值为5.07 m。DOM大小为5 201×5 201像素,对DSM进行网格划分,共划分了506×506个网格,网格间隔2 m,单幅DSM共有256 036个节点。试验区航空影像如图2所示。

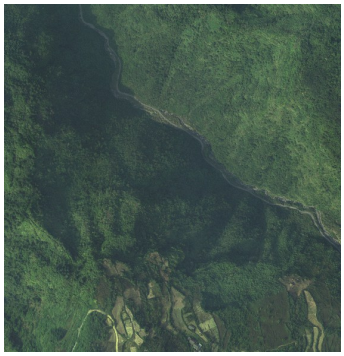

图2　航空影像图

Fig. 2　Aerial image map

2.2　植被高度与特征因子的相关性分析

因植被高度与几何特征因子、光谱特征因子之间不是简单的线性关系,故无法确定其具体的函数关系。最大信息系数(maximal information coefficient, MIC)[20]可用来度量大数据集中变量间的相关程度,据此可发现并识别大数据集中两个变量间所有重要的函数关系;Kendall相关系数[21]是用来度量两个随机变量相关性的统计值的。为了评价植被高度与特征因子间的相关性,根据MIC、Kendall相关系数对植被高度与各特征因子进行相关性指标计算。试验区共有256 036个节点,在使用全部样本数据进行各特征因子相关性计算时,MIC的计算过程耗时较长。为减少计算时长,同时增加不同指标的对比性,可在总样本中随机抽取30%的数据,用以计算植被高度与各特征因子的相关性。考虑到随机抽取一次样本进行试验具有偶然性,故随机抽取10组样本,分别进行相关性试验,计算其相关性指标值的平均值及标准差,结果见表1~2。

为了确保相关性指标值的真实性和准确性,随机选择部分特征因子,使用全部数据进行相关性指标计算,得到的相关性指标值的大小及其排序与使用随机抽取总样本30%数据计算得到的结果没有明显差异。

表1 光谱特征因子与植被高度之间的相关性指标值

Table 1 Correlation index values between spectral feature factor and vegetation height

光谱特征因子	MIC		Kendall系数	
	均值	标准差	均值	标准差
RGB	0.141 7	0.004 35	−0.126 3	0.001 65
VDVI	0.039 8	0.001 05	0.050 0	0.002 30
NGBDI	0.041 5	0.000 99	−0.073 7	0.001 77
NGRDI	0.144 7	0.001 66	0.211 4	0.003 89
ExG	0.039 9	0.001 00	0.050 8	0.002 15
ExR	0.099 5	0.001 43	−0.173 6	0.001 24
RmG	0.130 4	0.001 38	−0.193 5	0.001 64
GmB	0.039 5	0.000 74	−0.062 8	0.001 78
GBdRG	0.039 1	0.000 83	−0.030 1	0.002 17
ExG-ExR	0.140 1	0.001 93	0.204 7	0.001 59
Lab颜色空间L值	0.064 1	0.001 11	−0.119 2	0.001 66
Lab颜色空间a值	0.033 3	0.000 76	−0.030 1	0.002 17
Lab颜色空间b值	0.049 9	0.000 89	−0.106 0	0.001 76

表2 几何特征因子与植被高度之间的相关性指标值

Table 2 Correlation index values between geometric feature factor and vegetation height

几何特征因子	MIC		Kendall系数	
	均值	标准差	均值	标准差
坐标X	0.051 2	0.000 52	−0.004 8	0.001 55
坐标Y	0.176 7	0.002 49	0.264 3	0.002 87
高程H	0.166 1	0.001 34	0.214 6	0.001 67
邻域高度标准差	0.088 3	0.001 21	0.229 7	0.002 15
东向高差	0.103 3	0.002 24	0.224 0	0.002 12
南向高差	0.098 9	0.001 23	−0.200 3	0.001 51
西向高差	0.125 1	0.002 05	−0.258 9	0.002 38
北向高差	0.087 8	0.001 35	0.172 2	0.001 48
东南向高差	0.047 4	0.001 06	0.018 4	0.001 62
西南向高差	0.182 9	0.002 90	−0.322 3	0.003 40
西北向高差	0.047 2	0.000 78	−0.062 3	0.001 84
东北向高差	0.149 5	0.002 47	0.274 2	0.002 24
法线向量X分量	0.114 8	0.001 70	0.242 8	0.002 24
法线向量Y分量	0.091 3	0.001 15	−0.192 9	0.001 57
法线向量Z分量	0.081 6	0.000 98	0.222 5	0.002 21
表面粗糙度	0.089 0	0.001 26	0.232 3	0.002 21
高斯曲率	0.025 9	0.000 70	0.001 2	0.000 87
平均曲率	0.073 7	0.001 43	0.204 0	0.009 29
最大曲率	0.072 1	0.001 40	0.196 6	0.002 12
最小曲率	0.060 3	0.001 07	0.165 1	0.000 97

由表1~2可知:根据指标值的相对大小,不同特征因子在不同指标下相关性大小的趋势基本一致;同一特征因子在不同的评价指标下,相关性指标值有较大差别,这表明植被高度与特征因子之间存在复杂的函数关系。

3 植被高度建模及参数优化

3.1 植被高度模型参数优化

为了提高植被高度模型的精度,选择对模型精度影响较大的3个参数进行优化,这3个参数分别为节点分裂时参与判断的最大特征值、各个回归估计量的最大深度、弱学习器的最大迭代次数,其他参数均为默认值。针对某一参数进行优化时,需要固定其他参数及试验数据。由于原始数据量较大,在进行3个参数优化时,均随机抽取原始数据的30%作为训练集,10%作为测试集,并将其作为固定数据进行试验。

为探究植被高度模型各个回归估计量的最大深度最优参数值,设最大迭代次数为100,最大特征值为33,其他参数均为默认值,从每棵独立树的深度为1开始试验,结果如图3所示。由图3可以看出,均方根误差E_{rms}随着深度的增加先减小后增大,在深度为12时,E_{rms}最小,此时模型精度最高;当深度大于12时,E_{rms}随着深度的增加而增加,深度测试到18时中止测试。

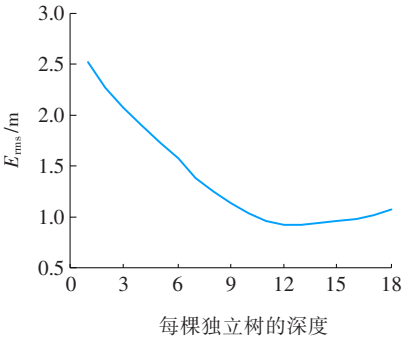

图3 深度与E_{rms}的关系

Fig. 3 Relationship between the depth and the E_{rms}

探究GBDT模型的最大特征值时,试验参数设置如下:最大迭代次数为100,最大深度为12,其他参数均取默认值。由于原始数据共有33种特征因子,故特征值取值范围为1~33,试验结果如图4所

示。由图 4 可以看出，特征值取值为 1~5 时，E_{rms} 波动较大，且 E_{rms} 整体较大；特征值取值为 5~33 时，E_{rms} 呈现整体下降的趋势且仍有波动；在特征值为 28 时，E_{rms} 最小，模型精度最高。

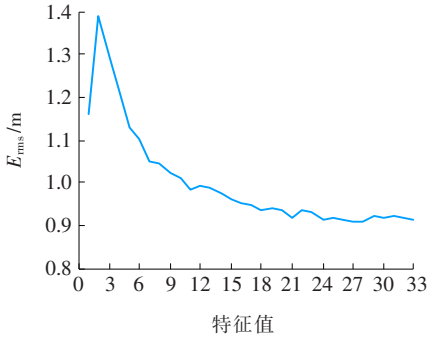

图 4　特征值与 E_{rms} 的关系

Fig. 4　Relationship between the feature number and the E_{rms}

探究 GBDT 模型的最优迭代次数时，设置模型参数如下：最大深度为 12，最大特征值为 28，其他参数取默认值。由于试验数据量较大，随着迭代次数的增加，每次消耗的时间逐渐增加，因此，结合迭代次数与 E_{rms} 整体趋势的具体情况，迭代次数在 10~100 时，设置迭代步长为 10；迭代次数在 100~500 时，设置迭代步长为 25。试验结果如图 5 所示。由图 5 可以看出，迭代次数在 100 之内时，E_{rms} 随着迭代次数的增加而大幅度减小；迭代次数在 100~500 时，随着迭代次数的增加，E_{rms} 出现了一些波动，但整体上 E_{rms} 随着迭代次数的增加而减小。迭代次数为 200~500 时，模型精度的变化幅度很小，为确保模型运行效率及模型精度，防止过拟合，选择迭代次数 200 作为模型最优迭代次数。

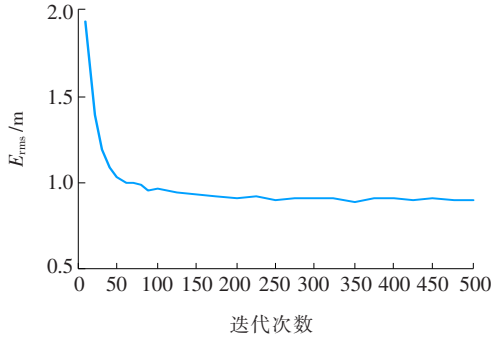

图 5　迭代次数与 E_{rms} 的关系

Fig. 5　Relationship between the iteration time and the E_{rms}

通过以上参数优化试验，得出 GBDT 模型的最优参数如下：最大迭代次数为 200，最大特征值为 28，最大深度为 12，其余参数取默认值。为了验证在不同样本下模型的稳定性，在原始总样本数据中随机多次抽取 30% 作为训练集、10% 作为测试集进行试验，得到的 E_{rms} 的大小基本与本次试验得到的一致，波动幅度在 0.01 m 左右。

3.2　试验结果分析

3.2.1　默认参数与优化参数

为分析参数对 GBDT 模型精度的影响，进行如下试验：使用相同的数据，计算 GBDT 模型在默认参数和最优参数下的试验结果，并对预报误差进行对比分析。

在默认参数下，GBDT 算法的试验结果如图 6~7 所示。

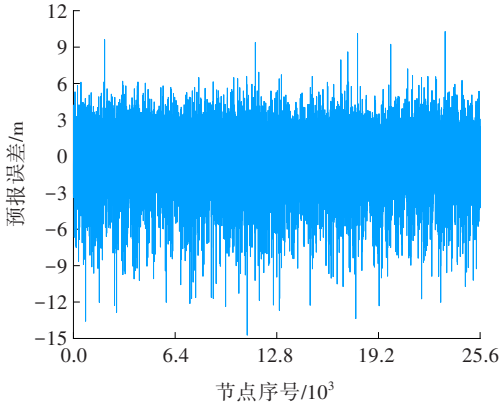

图 6　默认参数下预报误差散点图

Fig. 6　Scatter plot of the forecast error with default parameters

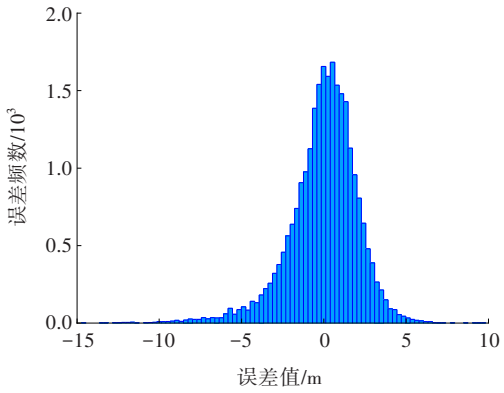

图 7　默认参数下预报误差柱状图

Fig. 7　Bar graph of the prediction error with default parameters

由图 6 可知，默认参数下的模型精度为 2.075 m，判定系数为 0.824。通过统计发现：预报误差在 1.0 m 范围内的节点约占 43.41%；预报误差在 0.5

m范围内的节点约占23.80%;预报误差在0.3 m范围内的节点约占14.24%;预报误差在0.1 m范围内的节点约占4.86%。

在优化参数下,GBDT算法的试验结果如图8~9所示。

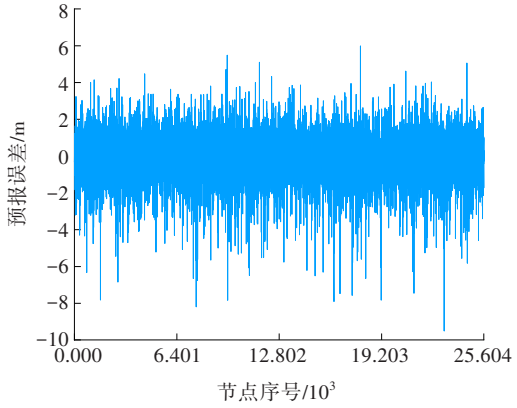

图8 优化参数下预报误差散点图

Fig. 8 Scatter plot of the forecast error with optimization parameters

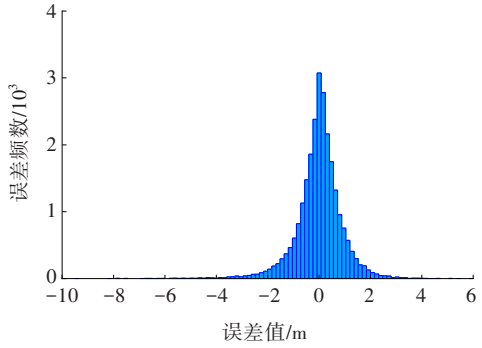

图9 优化参数下预报误差柱状图

Fig. 9 Bar graph of the prediction error with optimization parameters

在优化参数的植被高度建模试验中,模型精度达到0.9 m,判定系数为0.96,运行时间为10 min,模型稳定性较好。通过预报误差的散点图和柱状图可以发现,预报误差正负值较为对称,误差集中在0附近,在"0"处误差频数达到峰值,误差正负值较为对称,呈现出正态分布的特征。通过统计发现:预报误差的均值为-0.012 m,预报误差在1.0 m范围内的节点约占81.82%;预报误差在0.5 m范围内的节点约占56.91%;预报误差在0.3 m范围内的节点约占39.22%;预报误差在0.1 m范围内的节点约占14.38%。通过对比分析默认参数和优化参数下GBDT算法的试验结果发现,采用优化参数,GBDT算法的模型精度更高,模型稳定性更好。

3.2.2 特征因子优化

采用不同的相关性指标对特征因子与植被高度进行相关性分析,在不同的指标下其相关性大小不同,但整体趋势基本一致。本次试验以Kendall相关系数为筛选基础,依次剔除部分相关性较小的特征因子进行植被高度建模,并评价其精度大小,通过优化参数,提高模型精度。本次试验随机抽取5组样本数据,样本数据量均为总样本的30%,每组样本均进行特征因子优化的全部试验,然后计算5次试验结果的均值及其标准差,模型精度与特征因子种类的关系见表3。

表3 模型精度与特征因子种类的关系

Table 3 Relationship between the model accuracy and the number of feature factor type

特征因子种类	最优参数	E_{rms}均值/m	E_{rms}标准差	时间/min
使用全部特征因子	迭代次数200,最大特征值为28,深度为12	0.902	0.002 5	10.0
剔除相关性较小的6个特征	迭代次数200,最大特征值为22,深度为12	0.897	0.003 0	7.0
剔除相关性小于0.15的特征	迭代次数200,最大特征值为16,深度为12	0.888	0.003 1	5.0
剔除相关性小于0.20的特征	迭代次数200,最大特征值为10,深度为13	0.840	0.003 1	3.0
剔除相关性小于0.22的特征	迭代次数200,最大特征值为9,深度为13	0.963	0.003 4	2.5

由表3可知,逐步剔除相关性相对较小的特征因子后,模型精度、运行效率逐步增加,模型精度最高达到0.840 m,运行时间提高到3 min;随着剔除的特征因子的相关性越来越大时,模型精度会降低。

4 讨论

通过对比默认参数与优化参数下的试验结果可知:参数优化后的模型预报误差更加收敛,集中

在0附近,符合正态分布的特征。通过与文献[14]中采用支持向量机算法的试验结果进行对比,在同样的数据中,支持向量机算法的精度为0.97 m,运行时间为25 min,GBDT算法的精度略高于支持向量机算法的精度,在运行效率上,也明显优于支持向量机算法。在最优参数下,对最终的植被高度模型进行预测,试验结果如图10~12所示。

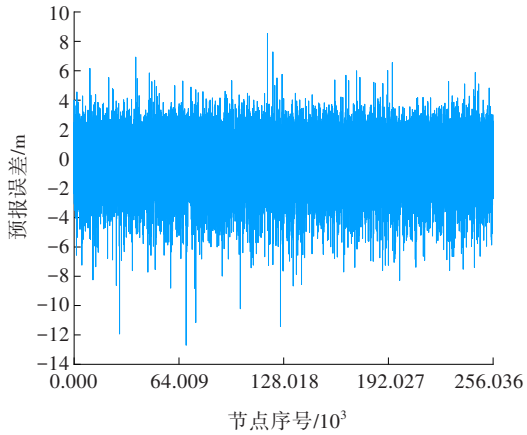

图10　预报误差散点图

Fig. 10　Scatter plot of the forecast error

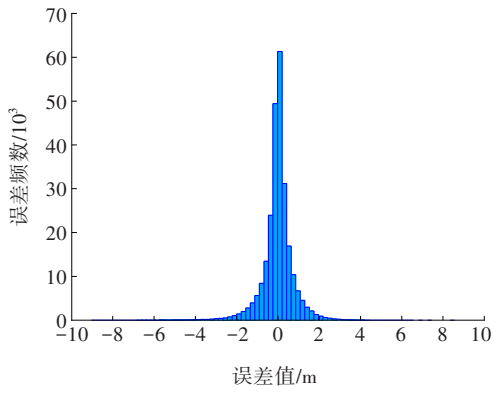

图11　预报误差柱状图

Fig. 11　Bar graph of the prediction error

在对整个试验区进行植被高度预测时,其模型精度为0.758 m,判定系数为0.98,运行时间为10 min。通过与试验区影像图进行对比,反演得到的植被高度等值线符合影像图中植被、地面等地物地貌的分布特征。使用本研究数据,文献[14]的最终模型精度为0.893 m,运行时间为70 min。使用文献[14]中影像1的数据进行最终的森林植被高度建模时,支持向量机算法的精度为0.868 m,GBDT算法的精度为0.645 m。通过与文献[14]采用的支持向量机算法进行对比,在精度和运行效率上,GBDT算法均优于支持向量机算法。

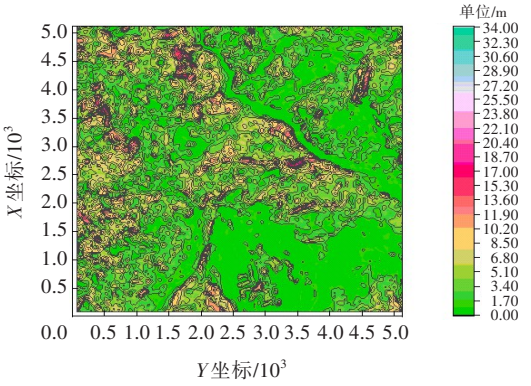

图12　植被高度等值线图

Fig. 12　Contour map of vegetation height

5　结论

1)目前利用航空摄影测量建立森林地区植被高度模型的研究才刚起步,本研究提出利用DOM与DSM,提取光谱特征与几何特征因子共33个。

2)采用MIC、Kendall相关系数对植被高度与特征因子进行相关性分析,并以此作为筛选特征因子的依据,采用GBDT算法,建立植被高度预测模型,通过优化参数,植被高度预测模型的精度达到0.840 m;对整个试验区进行植被高度预测时,模型精度达到0.758 m。

3)通过与支持向量机算法进行对比,GBDT算法在精度和运行效率上均优于支持向量机算法。此方法具有可行性、可靠性,为植被高度模型研究提供了新的思路与方法,但特征因子的提取与优化、航空影像数据与雷达数据融合处理等问题尚需要进一步研究。

〔参考文献〕

[1]　沈鹏,汪长城,朱建军,等.融合升降轨的极化干涉SAR三层模型植被高度反演方法[J].测绘学报,2017,46(11):1868-1879.DOI:10.11947/j.AGCS.2017.20170122.
SHEN Peng, WANG Changcheng, ZHU Jianjun, et al. Vegetation height inversion method with three-layer

model by fusing the ascending and descending PollnSAR data[J]. Acta Geodaetica et Cartographica Sinica, 2017, 46(11): 1868-1879. DOI: 10.11947/j. AGCS. 2017.20170122.

[2] 朱建军,付海强,汪长城.InSAR林下地形测绘方法与研究进展[J].武汉大学学报(信息科学版),2018,43(12):2030-2038.DOI:10.13203/j.whugis20180266.
ZHU Jianjun, FU Haiqiang, WANG Changcheng. Methods and research progress of underlying topography estimation over forest areas by InSAR[J]. Geomatics and Information Science of Wuhan University, 2018, 43(12): 2030-2038. DOI:10.13203/j.whugis20180266.

[3] 智长贵,丁雷,范文义.基于光束法空中三角测量理论测量林分平均高度的方法[J].东北林业大学学报,2009,37(3):29-31.DOI:10.3969/j.issn.1000-5382.2009.03.012.
ZHI Changgui, DING Lei, FAN Wenyi. Measurement method of stand average height based on theory of bundle aerotriangulation[J]. Journal of Northeast Forestry University, 2009, 37(3): 29-31. DOI: 10.3969/j.issn.1000-5382.2009.03.012.

[4] 郭鹏,武法东,戴建国,等.基于机载LiDAR数据的农田区植被高度估测研究[J].干旱区地理,2017,40(6):1241-1247. DOI: 10.13826/j. cnki. cn65-1103/x. 2017.06.014.
GUO Peng, WU Fadong, DAI Jianguo, et al. Estimation of vegetation height in farmland region based on airborne LiDAR data[J]. Arid Land Geography, 2017, 40(6): 1241-1247.DOI:10.13826/j.cnki.cn65-1103/x.2017.06.014.

[5] NIE S, WANG C, ZENG H C, et al. Above-ground biomass estimation using airborne discrete-return and full-waveform LiDAR data in a coniferous forest[J]. Ecological Indicators, 2017, 78:221-228. DOI:10.1016/j.ecolind.2017.02.045.

[6] 段祝庚,曾源,赵旦,等.机载激光雷达森林冠层高度模型凹坑去除方法[J].农业工程学报,2014,30(21):209-217. DOI:10.3969/j.issn.1002-6819.2014.21.025.
DUAN Zhugeng, ZENG Yuan, ZHAO Dan, et al. Method of removing pits of canopy height model from airborne laserradar[J]. Transactions of the Chinese Society of Agricultural Engineering , 2014, 30(21): 209-217. DOI: 10.3969/j.issn.1002-6819.2014.21.025.

[7] 贺跃光,刘聪,邢学敏.某钻井水溶开采矿区地表形变D-In SAR监测精度分析[J].长沙理工大学学报(自然科学版),2018,15(3):79-84. DOI:10.3969/j.issn.1672-9331.2018.03.013.
HE Yueguang, LIU Cong, XING Xuemin. Precision analysis of surface deformation D-In SAR monitoring in a drilling water-soluble mining area[J]. Journal of Changsha University of Science & Technology (Natural Science), 2018, 15(3): 79-84. DOI: 10.3969/j. issn. 1672-9331.2018.03.013.

[8] 张兵,朱建军,付海强,等.多基线极化干涉SAR植被高度反演方法[J].测绘工程,2017,26(9):23-27,31.DOI:10.19349/j.cnki.issn1006-7949.2017.09.005.
ZHANG Bing, ZHU Jianjun, FU Haiqiang, et al. Multi-baseline PolInSAR vegetation height inversion method[J].Engineering of Surverying and Mapping,2017, 26(9): 23-27, 31. DOI: 10.19349/j. cnki. issn1006-7949. 2017.09.005.

[9] 解清华,朱建军,汪长城,等.基于S-RVoG模型的PolInSAR森林高度非线性复数最小二乘反演算法[J].测绘学报,2020,49(10):1303-1310. DOI: 10.11947/j. AGCS.2020.20190081.
XIE Qinghua, ZHU Jianjun, WANG Changcheng, et al. A S-RVoG model-based PolInSAR nonlinear complex least squares method for forest height inversion[J]. Acta Geodaetica et Cartographica Sinica, 2020, 49(10): 1303-1310.DOI:10.11947/j.AGCS.2020.20190081.

[10] 罗环敏,陈尔学,程建,等.极化干涉SAR森林高度反演方法研究[J].遥感学报,2010,14(4):806-821.DOI:10.11834/jrs.20100414.
LUO Huanmin, CHEN Erxue, CHEN Jian, et al. Forest height estimation methods using polarimetric SAR interferometry[J].Journal of Remote Sensing, 2010, 14(4): 806-821.DOI:10.11834/jrs.20100414.

[11] 解金卫,索志勇,李真芳,等.基于PolInSAR的植被区高精度数字表面模型反演方法[J].电子与信息学报,2019,41(2):293-301.DOI:10.11999/JEIT180387.
XIE Jinwei, SUO Zhiyong, LI Zhenfang, et al. High-precision digital surface model inversion approach in forest region based on PolInSAR[J]. Journal of Electronics & Information Technology, 2019, 41(2): 293-301.DOI:10.11999/JEIT180387.

[12] 廖中平,陈立,白慧鹏,等.自适应α-shapes平面点云边界提取方法[J].长沙理工大学学报(自然科学版),2019,16(2):15-21.DOI:10.3969/j.issn.1672-9331.2019.02.004.
LIAO Zhongping, CHEN Li, BAI Huipeng, et al. Adaptive Alpha-shapes plane cloud boundary extraction method [J]. Journal of Changsha University of Science & Technology (Natural Science), 2019, 16(2): 15-21. DOI: 10.3969/j.issn.1672-9331.2019.02.004.

[13] 赵利霞,王宏涛,郭增长,等.基于随机森林的倾斜影像

匹配点云分类研究[J].测绘工程,2018,27(12):13-18.DOI:10.19349/j.cnki.issn1006-7949.2018.12.004.
ZHAO Lixia, WANG Hongtao, GUO Zengzhang, et al. A study of classification of point clouds generated by oblique imagery based on random forest[J]. Engineering of Surveying and Mapping, 2018, 27(12): 13-18. DOI: 10.19349/j.cnki.issn1006-7949.2018.12.004.

[14] DENG X S, TANG G, WANG Q Y, et al. A method for forest vegetation height modeling based on aerial digital orthophoto map and digital surface model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60:4404307.DOI:10.1109/TGRS.2021.3093976.

[15] DING C, WANG D G, MA X L, et al. Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees[J]. Sustainability, 2016, 8(11): 1-16. DOI: 10.3390/su8111100.

[16] MENG Q, WENG J. Classification and regression tree approach for predicting drivers merging behavior in short-term work zone merging areas[J]. Journal of Transportation Engineering, 2012, 138(8): 1062-1070. DOI:10.1061/(asce)te.1943-5436.0000412.

[17] 刘非凡,刘忙龙,曹少珺.无线电调频引信谐波提取改进算法[J].兵器装备工程学报,2018,39(5):117-120.DOI:10.11809/bqzbgcxb2018.05.025.
LIU Feifan, LIU Manglong, CAO Shaojun. Improved algorithm for fast harmonic extraction of radio fuze[J]. Journal of Ordnance Equipment Engineering, 2018, 39(5):117-120.DOI:10.11809/bqzbgcxb2018.05.025.

[18] AHMED M, ABDEL-ATY M. Application of stochastic gradient boosting technique to enhance reliability of real-time risk assessment: use of automatic vehicle identification and remote traffic microwave sensor data [J]. Transportation Research Record, 2018, 2386: 26-34. DOI: 10.3141/2386-04.

[19] 翁剑成,付宇,林鹏飞,等.基于梯度推进决策树的日维度交通指数预测模型[J].交通运输系统工程与信息,2019,19(2):80-85,93.DOI:10.16097/j.cnki.1009-6744.2019.02.012.
WENG Jiancheng, FU Yu, LIN Pengfei, et al. GBDT method based on prediction model of daily dimension traffic index[J]. Journal of Transportation Systems Engineering and Information Technology, 2019, 19(2): 80-85,93.DOI:10.16097/j.cnki.1009-6744.2019.02.012.

[20] RESHEF D N, RESHEF Y A, FINUCANEH K, et al. Detecting novel associations in large data sets[J]Science, 2011, 334(6062): 1518-1524. DOI: 10.1126/science.1205438.

[21] 刘乾玉,唐家银.基于Kendall协同系数的产品加速试验失效机理一致性统计检验[J].湖北大学学报(自然科学版),2021,43(6):671-677.DOI:10.3969/j.issn.1000-2375.2021.06.011.
LIU Qianyu, TANG Jiayin. Statistical test of failure mechanism consistency of product in accelerated test based on Kendall synergy coefficient[J]. Journal of Hubei University (Natural Science), 2021, 43(6): 671-677.DOI:10.3969/j.issn.1000-2375.2021.06.011.

Study on vegetation height model based on the gradient boosting decision tree

DENG Xingsheng, WANG Qingyang

(School of Traffic and Transportation Engineering, Changsha University of Science & Technology, Changsha 410114, China)

Abstract: **[Purposes]** The study aims to establish vegetation height model by aerial photogrammetry. **[Methods]** Based on digital orthophoto map and digital surface model, spectral and geometric feature factors were extracted for vegetation height modeling. The correlation between vegetation height and feature factor was analyzed by correlation index, the feature factors were selected. The gradient boosting decision tree algorithm was adopted to establish the vegetation height prediction model, and the accuracy of the model was improved through parameter optimization. **[Findings]** The model accuracy is about 2.000 m under the default parameter. By optimizing the parameters, the model accuracy reaches 0.900 m. Furthermore, the model accuracy enhance to 0.840 m, resulted from excluding some feature factors. By compared with the support vector machine algorithm, the accuracy of the vegetation height model has been increased from 0.893 m to 0.758 m, and the running time has been reduced from 70 minutes to 10 minutes. **[Conclusions]** The accuracy of the vegetation height model can reach to sub meter, when the error of the original modeling data is neglected, and the accuracy of the model remains stable in many experiments.

Key words: vegetation height; gradient boosting decision tree; feature factor; machine learning

Manuscript received: 2021-11-30; **revised**: 2022-06-21; **accepted**: 2022-06-28

Foundation item: Project (2020JJ4601) supported by Natural Science Foundation of Hunan Province

Corresponding author: DENG Xingsheng (1971—) (ORCID: 0000-0003-0158-9227), male, associate professor, research interest: photogrammetry data processing.E-mail: 383500135@qq.com

(责任编辑:刘平;校对:石月珍;英文编辑:陈璐)