

基于机器学习算法的樟子松立木材积预测

孙铭辰, 姜立春*

(东北林业大学林学院, 森林生态系统可持续经营教育部重点实验室, 黑龙江 哈尔滨 150040)

摘要:【目的】通过非线性和多种机器学习算法构建并对比不同的立木材积模型,为樟子松(*Pinus sylvestris* var. *mongolica*)立木材积的精准预测提供理论依据。【方法】以大兴安岭图强林业局 184 株樟子松伐倒木数据为基础,建立非线性二元材积模型(NLR),并通过十折交叉检验和袋外数据(OOB)误差检验的方法得到 3 种最优机器学习算法,包括:反向神经网络(BP)、 ε -支持向量回归(ε -SVR)和随机森林(RF)。对比分析不同模型间的差异,得到最优立木材积模型。【结果】机器学习算法在立木材积的拟合和预测中均优于传统二元材积模型,具体拟合结果排序为 RF>BP> ε -SVR>NLR。其中 RF 的决定系数(R^2)比传统模型的提高了 2.00%,均方根误差(RMSE)、相对均方根误差(RMSE%)、平均绝对误差(MAE)分别降低了 22.90%、22.93%、36.34%,且与真实值相比平均相对误差(MRB)的绝对值更低,证明了 RF 在立木材积预测中的优越性。【结论】机器学习算法作为一种新兴的建模方法可以有效地提高立木材积的预测精度,为森林资源的精准调查和经营管理提供新的解决方案。

关键词:樟子松;二元材积模型;BP 神经网络; ε -支持向量回归(ε -SVR);随机森林(RF)

中图分类号:S791.253

文献标志码:A

开放科学(资源服务)标识码(OSID):

文章编号:1000-2006(2023)01-0031-07



Standing volume prediction of *Pinus sylvestris* var. *mongolica* based on machine learning algorithm

SUN Mingchen, JIANG Lichun*

(Key Laboratory of Sustainable Forest Ecosystem Management, Ministry of Education,
School of Forestry, Northeast Forestry University, Harbin 150040, China)

Abstract: 【Objective】Using various, nonlinear machine learning algorithms, different volume models were constructed and compared to provide a theoretical basis for the accurate prediction of the volume of *Pinus sylvestris* var. *mongolica*. 【Method】A total of 184 felled *Pinus sylvestris* var. *mongolica* trees in the Tuqiang Forestry Bureau of the Greater Khingan Mountains were used to establish a nonlinear binary volume model (NLR). Three optimal machine learning algorithms were obtained using the K -fold cross test and OOB error test, including back propagation neural network (BP), ε -support vector regression (ε -SVR), and random forest (RF). An optimal volume model was obtained by comparing and analyzing the differences between the different models. 【Result】The results showed that the machine learning algorithm was superior to the traditional binary volume model in the fitting and prediction of standing volume, and the specific order was RF > BP > ε -SVR > NLR. Compared with the traditional model, the R^2 of RF increased by 2.00%; the RMSE, RMSE% and MAE decreased by 22.95%, 22.93% and 36.34%, respectively; and the absolute value of MRB was lower than the real value, which proved the superiority of RF in volume prediction. 【Conclusion】Machine learning algorithms can effectively improve the accuracy at which standing volume can be predicted, providing a new solution for the accurate investigation and management of forest resources.

Keywords: *Pinus sylvestris* var. *mongolica*; binary volume model; BP neural network; ε -support vector regression (ε -SVR); random forest (RF)

收稿日期 Received: 2021-04-12

修回日期 Accepted: 2021-08-04

基金项目:国家自然科学基金项目(31570624);黑龙江省应用技术与开发计划(GA19C006);黑龙江省头雁创新团队计划,国家林业和草原科学数据中心黑龙江子平台(2005DKA32200-OH)。

第一作者:孙铭辰(178059314@qq.com)。*通信作者:姜立春(jlichun@nefu.edu.cn),教授。

引文格式:孙铭辰,姜立春. 基于机器学习算法的樟子松立木材积预测[J]. 南京林业大学学报(自然科学版), 2023, 47(1): 31-37.
SUN M C, JIANG L C. Standing volume prediction of *Pinus sylvestris* var. *mongolica* based on machine learning algorithm[J].
Journal of Nanjing Forestry University (Natural Sciences Edition), 2023, 47(1): 31-37. DOI: 10.12302/j.issn.1000-2006.202104014.

立木材积是森林资源调查的重要指标,也是计算森林蓄积量和生物量的主要依据^[1-4]。因此,在森林经营管理中评价经济效益和生态效益时,立木材积的研究一直以来都备受关注。传统立木材积的计算通常利用已知的一元、多元立木材积表,或通过拟圆锥法、质心法、重要性采样法等进行估测^[5-6]。随着人们对森林资源调查的不断深入,针对不同地区不同树种相继建立了不同类型的材积模型^[7-8]。但森林生长是一个复杂连续且具有随机性的非线性生长过程,通过拟合立木材积模型虽然能填补材积预测的空白,但受制于模型和变量的选择以及对模型先验知识的累积,且预测精度受区域变化影响较大,增加了森林资源调查和经营管理的难度。

机器学习算法理论始于20世纪中叶,相比传统模型,机器学习算法可以在没有先验知识的前提下对数据进行拟合,分析数据中不同变量之间复杂、动态的内部结构^[9],而且机器学习算法的适用性更广,应用更加方便,能很好地克服数据中可能存在的缺失点、噪音、多重共线性和异方差等现象^[10],在生物遗传、信息技术和金融工程等领域已有广泛应用。近年来随着统计软件技术的发展,机器学习算法在林业上也得到了一定的应用。Guan等^[11]通过胸径及其年增长量建立4种人工神经网络模型,成功地对红松的生存率进行了预测;Maria^[12]通过人工神经网络模型估算树皮材积,发现相比于非线性模型的均方根误差(RMSE)降低了6.02%;Maria等^[13]通过对比分析不同的非线性模型和 ε -支持向量机回归模型对4种黑凯木树皮材积的预测,得出 ε -支持向量机回归模型的Furnival指数(FI)值分别比3种非线性模型均有所降低,且与真实值更接近。Colin等^[14]结合LiDAR数据建立多个模型对森林生物量进行估算,结果表明支持向量机回归为最优模型。目前,已有部分机器学习算法应用于立木材积中^[15-19],但鲜有对不同类型算法同时进行对比和分析的研究。

本研究以大兴安岭樟子松(*Pinus sylvestris* var. *mongolica*)为研究对象,利用Matlab 2019b软件建立3种目前应用较为广泛的机器学习算法:反向神经网络模型(back propagation, BP)、 ε -支持向量机回归模型(ε -support vector regression, ε -SVR)和随机森林模型(random forest, RF),并与传统二元材积模型作对比,评价最优模型,以期为提高樟子松的立木材积预测精度和科学经营提供理论依据。

1 材料与方法

1.1 数据来源

供试樟子松来源于大兴安岭图强林业局(122°18'28"~123°28'10"E, 52°15'35"~53°33'42"N)。将树木伐倒后测量其带皮胸径、树高,并对15个相对树高的带皮直径,利用区分求积法计算樟子松带皮立木材积。通过散点图排除异常点后得到184株样木,以5 cm为一个径级分为10个径阶,按7:3分径阶随机抽样。最终得到训练样本129株、测试样本55株。样木调查统计量如表1所示。

建模前对样本数据进行min-max标准归一化处理,使其统一介于[0,1],得到的泛化结果通过反归一化还原,并与真实值对比和评价。

表1 樟子松调查因子统计量

Table 1 Survey factor statistics of *Pinus sylvestris* var. *mongolica*

样本 sample	统计量 statistics	胸径/cm DBH	树高/m tree height	材积/m ³ volume
训练样本 train sample	最大值 max	50.70	25.70	2.38
	最小值 min	6.30	5.80	0.01
	平均值 average	27.95	18.28	0.71
	标准差 SD	12.12	4.33	0.60
测试样本 test sample	最大值 max	50.50	25.60	2.37
	最小值 min	7.60	6.30	0.02
	平均值 average	28.87	18.85	0.75
	标准差 SD	12.16	4.26	0.61

1.2 模型建立方法

1.2.1 二元材积模型

传统立木材积方程包括一元、多元线性和非线性模型,采用林业上应用较为广泛的非线性二元材积模型(non-linear regression models, NLR)作为立木材积计算公式^[20-21]。

$$V = a_0 D^{a_1} H^{a_2} \quad (1)$$

式中: V 为材积; D 为胸径; H 为树高; a_0 、 a_1 、 a_2 为方程的参数。

1.2.2 反向神经网络模型

反向神经网络模型(back propagation, BP)是由输入层、隐含层和输出层组成的反向传播网络模型,不同层通过神经元相互连接,但相同层神经元互不相连。通过激活函数将上一层的输入转化为下一层的输出,如输出结果超出期望误差,则通过误差逆向传播算法修正各神经元函数的权值和阈值,并不断重复上述过程以达到降低误差的目的^[22]。经研究表明,3层BP神经网络即可拟合任

何非线性曲线。

以胸径 D 、树高 H 作为输入节点,材积 V 作为输出节点,建立结构为 $2:S:1$ 的 3 层神经网络模型。其中隐含层激活函数选用 tansig , 输出层激活函数选用 purelin 函数。利用经验公式 $S = \sqrt{a+b} + c$ (式中, S 为隐含层神经元个数, a 、 b 分别为输入层和输出层节点数, c 为 1 到 10 之间的整数) 可得隐含层神经元个数为 3~12。通过适当更改隐含层训练算法和神经元个数可以有效提高网络拟合的速度和精度^[23], 并用提前结束训练的方法防止模型过拟合。

选用目前较为常用的几种训练算法:①梯度下降算法,包括变学习率动量梯度下降算法(GDX)、变学习率梯度下降算法(GDA)、弹性梯度下降算法(RP);②共轭梯度算法,包括 Powell-Beale 共轭梯度算法(CGB)、Fletcher-Reeves 共轭梯度法(CGF)、Polak-Ribiere 共轭梯度法(CGP)、Scaled 共轭梯度算法(SCG);③拟牛顿及其他优化算法,包括拟牛顿算法(BFG)、一步正割算法(OSS)、Levenberg-Marquardt 算法(LM)。

1.2.3 ε -支持向量回归模型

ε -支持向量回归模型(ε -support vector regression, ε -SVR)是一种基于结构风险最小化原理的超平面模型。通过核函数将低维空间中难以区分的向量通过非线性变换映射至高维空间中使其线性可分,借助惩罚因子(C)和核参数(γ)降低数据与超平面之间的残差,以达到提高模型精度的目的^[17]。因此改变核函数类型以及 C 、 γ 参数可以有效降低误差,得到最优 ε -SVR 模型。常用的核函数方法包括线性核函数(line)、多项式核函数(polynomial)、径向基核函数(RBF)等。其中 C 、 γ 参数通常选用网格搜索法,即设定两个参数的范围,按一定步长进行组合建模。但这种方法费时费力且精度不高,因此本研究选用遗传算法(genetic algorithm, GA)进行参数寻优。

GA 是通过模拟物种进化过程研发的一种全局搜索优化算法。通过生成一个初始群体并对数据进行编码、选择、杂交、变异,不断生成新的组合,计算每个组合的适应度,通过“优胜劣汰”的方法不断筛选最优个体,并以适应度达到最大时的结果作为最优参数,以十折交叉检验对参数的解释能力进行评估。综上所述,GA 是一种通用性很强的参数寻优方法,在机器学习也有着广泛的应用^[24]。

1.2.4 随机森林回归模型

随机森林回归模型(random forest, RF)是一种

基于 bagging 回归的集成学习算法。通过对样本数据进行随机且有放回重复采样得到多个样本组合,并通过节点分裂和随机特征变量的随机抽取形成多个决策树构成“森林”,对每个决策树得到的结果进行加权平均,作为样本的回归结果^[25]。由于每次采样中总有大约 1/3 的样本未被选用,它们被称为袋外数据(out-of-bag, OOB)。因此可利用这部分数据计算袋外错误率(out-of-bag error)代替交叉检验作为验证模型泛化能力的标准,使得随机森林可以有效地避免过拟合现象^[26]。

通过控制变量的方法对决策树个数(ntree)、最小叶子大小(minleaf)、随机抽取变量个数(mtry)3 个参数进行寻优,以得到最优的 RF 模型。其他参数均选用软件默认设置。

1.3 模型评价与检验指标

选用决定系数(R^2)、平均绝对误差[MAE, 式中记为 $\sigma(\text{MAE})$]、均方根误差[RMSE, 式中记为 $\sigma(\text{RMSE})$]、相对均方根误差[RMSE%, 式中记为 $\sigma(\text{RMSE}\%)$]评价模型拟合和预测的能力,并通过平均相对误差[MRB, 式中记为 $\sigma(\text{MRB})$]检验模型预测偏差的大小。为评价模型的无偏估计能力,选用 z 检验判断模型泛化结果与真实值是否存在显著差异^[19]。

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

$$\sigma(\text{RMSE}) = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}; \quad (3)$$

$$\sigma(\text{RMSE}\%) = \sigma(\text{RMSE}) \times 100\%; \quad (4)$$

$$\sigma(\text{MAE}) = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}; \quad (5)$$

$$\sigma(\text{MRB}) = \frac{100}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{y_i} \right). \quad (6)$$

式中: y_i 为实际值; $\bar{y}$ 为实测样本均值; $\hat{y}_i$ 为预测值; n 为样本数, i 为样本数据。

2 结果与分析

2.1 NLR 模型的拟合结果

通过最小二乘法得到 3 个参数分别为 $a_0 = 3.7060 \times 10^{-5}$, $a_1 = 1.5126$, $a_2 = 1.5662$ 。最终二元材积表达式为 $V = 3.7060 \times 10^{-5} D^{1.5126} H^{1.5662}$ 。

2.2 BP 模型的拟合结果

通过胸径 D 、树高 H 和材积 V 建立 3 层 BP 模型,采用十折交叉检验的方法对训练样本分别进行算法优选和神经元参数寻优。为保证不同算法在统一标准下对比分析,模型的其他参数统一设置如下:隐含层神经元个数为 8、最大训练次数为

2×10^3 、训练目标误差为 1×10^{-5} 、学习速率为 0.05、动量为 0.9,其余参数均采用软件默认数值,并采取提前终止的办法防止模型过拟合。

各算法结果如表 2 所示,可以看出在 3 种梯度下降算法中 RP 算法的 4 个评价指标均明显优于其他两种。通过比较可以看出 4 种共轭梯度算法的精度总体差距不大,其中 CGF 算法的 R^2 略高于其他三者,但在 RMSE、RMSE%、MAE 的表现不如 CGB 算法,拟合误差相对较大。在其他 3 种优化算法中 LM 算法拟合优度和误差检验均为最佳,相比于其他算法 R^2 提高 1.13%~5.26%,RMSE 降低 13.12%~39.81%,RMSE%降低 17.41%~42.46%,MAE 降低 14.97%~45.94%。因此将 LM 算法作为 BP 模型的训练算法。

表 2 不同算法下 BP 神经网络拟合精度
Table 2 Fitting accuracy of BP neural network with different algorithms

类型 type	算法 algorithm	R^2	RMSE	RMSE%	MAE
梯度下降算法 gradient descent algorithms	GDX	0.933 3	0.139 8	19.703 1	0.108 1
	GDA	0.920 4	0.159 5	23.140 5	0.120 8
	RP	0.954 0	0.116 8	16.255 9	0.086 1
共轭梯度算法 conjugate gradient algorithms	CGB	0.954 8	0.110 5	16.128 4	0.076 8
	CGF	0.958 0	0.116 1	16.575 5	0.083 5
	CGP	0.950 2	0.126 6	18.055 4	0.089 0
拟牛顿及其他优化算法 quasi Newton and other optimization algorithms	SCG	0.947 7	0.119 1	16.918 8	0.090 8
	BFG	0.946 6	0.116 6	16.525 3	0.087 5
	OSS	0.932 7	0.132 3	19.482 9	0.098 3
	LM	0.960 8	0.096 0	13.319 9	0.065 3

注:RMSE.均方根误差 root-mean-square error;RMSE%.相对均方根误差,% relative root-mean-square error;MAE.平均绝对误差 mean absolute error。下同。The same below.

根据经验公式计算可得隐含层神经元个数在 3~12 之间,通过试错法依次建模比较,在隐含层神经元个数为 7 时模型 R^2 最高且 RMSE 最低(图 1)。综上所述,最优 BP 模型为训练算法-输入变

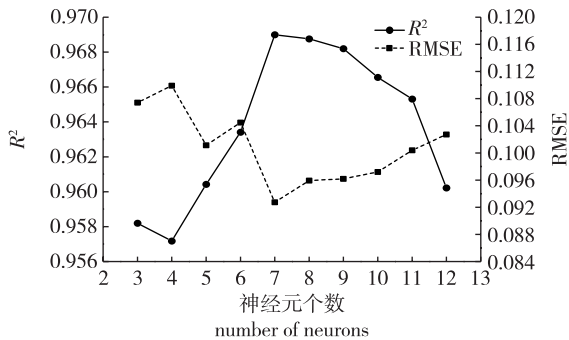


图 1 不同神经元的 BP 模型精度对比
Fig.1 Comparison of BP model accuracy of different neurons

量-神经元个数-输出变量(LM-DH-7-V)。

2.3 ε -SVR 模型的拟合结果

ε -SVR 通过核函数将数据映射至高维空间中,寻找样本期望风险最小的最优超平面实现线性回归。采用遗传算法对 3 种常见的核函数[线性核函数(line)、多项式核函数(polynomial)、径向基核函数(RBF)]进行参数寻优,具体参数设置如下:最大进化数为 200、种群进化数量为 20、变异值为 0.9、不敏感损失参数 $p=0.01$ 、惩罚因子 C 搜索范围(0,30)、核参数 gamma 搜索范围(0,10),并通过十折交叉检验得到 3 组最优的模型参数。

各模型结果如表 3 所示,可以看出精度最高的核函数是 RBF,其次是 polynomial 和 line 核函数。RBF 相比于多项式函数和线性核函数 R^2 分别提高 1.80%和 11.11%,RMSE 降低 21.46%和 52.82%,RMSE%降低 21.44%和 52.83%,MAE 降低 31.69%和 59.02%,由此可见 RBF 不仅可以提高拟合能力,还能很好地降低误差,故选用 RBF 作为 ε -SVR 模型的核函数并建立模型。

表 3 不同核函数下的模型参数和拟合精度
Table 3 Model parameters and fitting accuracy under different kernel functions

核函数 kernel function	C	核参数 gamma	R^2	RMSE	RMSE%	MAE
线性核函数 line	0.117 0	1.697 4	0.874 9	0.212 6	30.120 5	0.132 0
多项式函数 polynomial	0.122 7	9.854 5	0.954 9	0.127 7	18.083 9	0.079 2
径向基核函数 RBF	12.861 1	3.100 1	0.972 1	0.100 3	14.206 7	0.054 1

2.4 RF 模型的拟合结果

RF 是通过构建多个决策树(ntree)随机选取不同的变量(mtry),并以每个决策树最小叶子节点(minleaf)均值作为模型最后回归的结果,因此参数的大小直接影响模型拟合的精度。经研究发现,模型误差大小随 ntree 增加而逐渐减少至某一定值后,便不再发生变化。因此为确保模型充分训练,预设 ntree 个数为 2 000。采用控制变量的方法测定最优 minleaf 的数量,在默认 mtry 值下对比 minleaf 为 1、3、5、10、20 时的 OOB 误差值。结果如图 2 所示,可以看出随着 minleaf 值逐渐减小,模型的误差也逐渐减小,因此得出最优的 minleaf 值为 1。同时从图 2A 中也可以看出,当 ntree 大于 100 后,OOB 误差基本不再下降,意味着在建立 100 颗决策树后模型已完成了训练。为缩短训练时间,提

高模型运行的效率,故将 `ntree` 的值调整为 100。

由于模型的输入变量只有 2 个,故 `mtry` 的取值范围为 [1, 2],调整参数后对不同 `mtry` 值重新建立模型,由图 2B 所示两条误差曲线随着 `ntree` 的

提高最终重叠在一起,说明改变 `mtry` 值不会对模型误差产生影响。但是在拟合过程中可以看出,当 `mtry` 为 2 时模型收敛速度明显优于 `mtry` 为 1,故 `mtry` 值设置为 2。

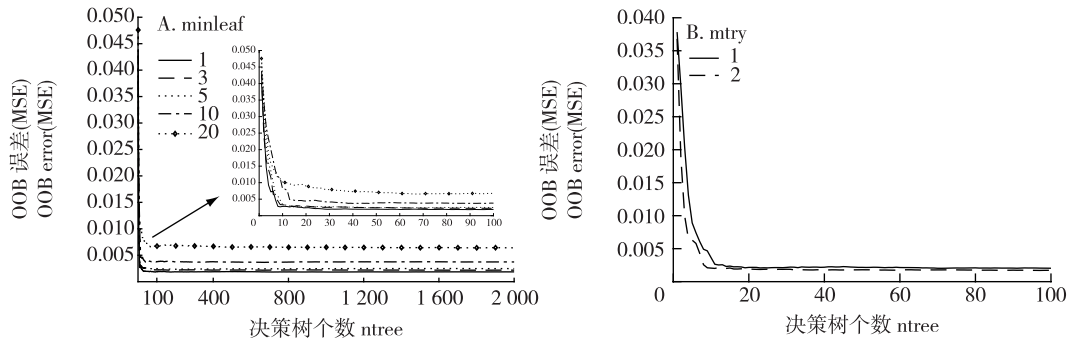


图 2 不同 `minleaf` 和 `mtry` 下随机森林的 OOB 误差曲线 (MSE 为归一化后的模型误差)

Fig.2 The OOB error curve of random forest under different `minleaf` and `mtry`
(MSE is the normalized model error)

2.5 模型评价与检验

选用最优参数对训练样本建模,并用测试样本验证模型的预测能力。结果如表 4 所示,机器学习算法无论在拟合和预测中均明显优于传统的二元材积模型,且机器学习算法中训练样本与测试样本的精度差异不大,证明模型并未发生过拟合现象。通过训练样本得到的拟合评价结果可以看出,4 种模型的 R^2 均高于 0.96,且 RMSE 均低于 0.12,说明模型都具备较好的拟合能力。其中 RF 模型表现最好,相比于 NLR、BP、 ϵ -SVR 模型, R^2 分别提高 3.43%、2.05%、2.45%, RMSE 降低 67.30%、59.40%、62.31%, RMSE% 降低 67.29%、59.38%、62.28%, MAE 降低 71.48%、60.96%、59.52%。通过计算 MRB 可以看出,所有模型的拟合结果均略高于真实值,其中 RF 模型的偏差最小。对 4 种模型的拟合评价作对比,除在 MAE 中 ϵ -SVR 略优于 BP 模型,其余检验精度均满足 RF>BP> ϵ -SVR>NLR。

表 4 模型拟合和预测结果评价

Table 4 Model fitting and prediction result evaluation						
模型 model	数据集 dataset	R^2	RMSE	RMSE%	MAE	MRB
NLR	训练 train	0.963 0	0.115 6	16.380 8	0.076 8	-5.699 5
	测试 test	0.953 1	0.130 7	17.503 5	0.091 9	-4.900 3
BP	训练 train	0.976 0	0.093 1	13.191 5	0.056 1	-2.090 4
	测试 test	0.962 8	0.116 4	15.584 4	0.075 1	-1.636 3
ϵ -SVR	训练 train	0.972 2	0.100 3	14.206 7	0.054 1	-2.594 5
	测试 test	0.960 4	0.120 1	16.077 6	0.075 3	1.283 5
RF	训练 train	0.996 0	0.037 8	5.358 5	0.021 9	-1.550 8
	测试 test	0.972 2	0.100 7	13.489 4	0.058 5	0.169 0

注:MRB.平均相对误差。

根据已建立的模型对测试样本预测,可以看出 4 种模型都有较好的预测能力, R^2 均高于 0.95, RMSE 均低于 0.14,且与拟合结果相同,精度均满足 RF>BP> ϵ -SVR>NLR。相比于其余三者,RF 模型的 R^2 分别提高 2.00%、0.98%、1.23%, RMSE 降低 22.95%、13.49%、16.15%, RMSE% 降低 22.93%、13.44%、16.10%, MAE 降低 36.34%、22.10%、22.31%。结合模型预测结果图(图 3)和 MRB 发现,NLR 与 BP 模型的预测结果略高于真实值, ϵ -SVR 与 RF 的预测结果略低于真实值,通过比较 MRB 绝对值可知二元材积模型的预测偏差最大,RF 模型偏差最小。

为客观评价模型的无偏估计能力,故对模型的拟合和预测结果进行置信度为 95%的 z 检验,以验证泛化结果与真实值的分布是否一致。如表 5 所示,可以看出 4 种模型的 P 值均远大于 0.05,故保留原假设,即结果与真实值不存在显著差异。

表 5 模型的 z 检验结果

Table 5 Z-test results of the model				
模型 model	数据集 dataset	z 值 z value	P	结论 conclusion
NLR	训练 train	0.033 3	0.973 5	不存在明显差异
	测试 test	0.147 8	0.882 5	不存在明显差异
BP	训练 train	0.031 3	0.975 1	不存在明显差异
	测试 test	-0.027 3	0.978 2	不存在明显差异
ϵ -SVR	训练 train	-0.222 7	0.823 8	不存在明显差异
	测试 test	-0.120 1	0.904 4	不存在明显差异
RF	训练 train	0.021 7	0.982 7	不存在明显差异
	测试 test	0.085 1	0.932 2	不存在明显差异

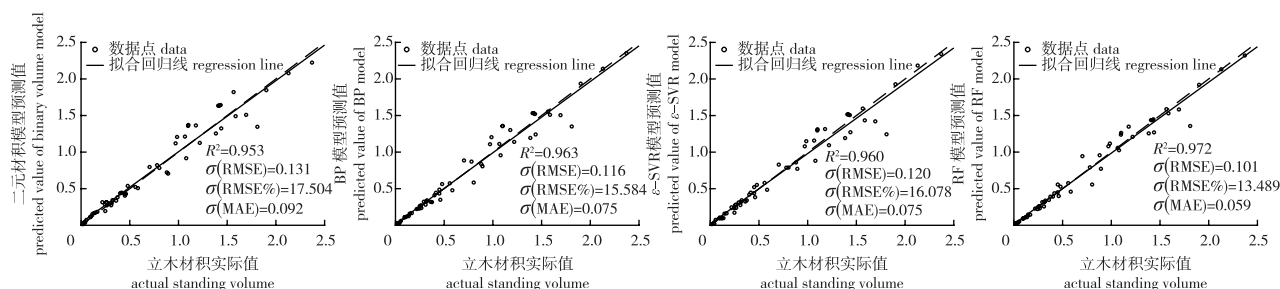


图3 4种模型预测线性回归拟合图

Fig.3 The prediction results of four models

综上所述,RF 为预测大兴安岭樟子松立木材积的最优模型。

3 讨论

利用大兴安岭 184 株樟子松单木材积数据建立传统二元材积模型,并与 3 种常见的机器学习算法进行对比。研究表明,机器学习算法可以更好地预测大兴安岭樟子松材积。通过对测试样本的预测结果对比分析可以看出,预测精度大小排序为 RF>BP> ϵ -SVR>NLR。其中,RF 模型的预测精度最高,相比于传统模型,RF 模型的 R^2 提高 2.00%、RMSE 降低 22.95%、RMSE%降低 22.93%、MAE 降低 36.34%,且与真实值相比偏差更低。通过减少训练样本后重新建模,RF 模型精度仍高于传统模型,由此可以看出 RF 模型在樟子松立木材积预测中的优越性。其余两种机器学习算法中,BP 模型虽略优于 ϵ -SVR 但与 RF 模型仍有较大差距。这是因为 BP 模型所选用的 LM 算法与 ϵ -SVR 更适用于中小样本^[27-28],而建模所选用的样本数量相对较大,此时 RF 模型在大样本中的适应性优势才被凸显出来。但 RF 模型同样存在一定的缺点,在处理噪声较大或小样本中易发生过拟合现象^[9],因此在面对此类问题时应先对数据做预处理,避免对模型精度造成影响。

与机器学习算法相比,传统模型虽然可以通过建立回归方程的方式清晰地表达不同变量间相互影响的关系,但通常需要满足很多检验条件作为假设前提,例如正态性检验、独立性检验、异方差检验等,因此不能同时解释复杂的森林生长因子,且不容易对含有噪声的大尺度数据进行拟合。随着近年来科学技术的发展,机器学习算法的出现很好地解决了传统模型的问题。在其他立木材积的研究中也证明了本研究的结果。Maria 等^[15]和 Bhering 等^[19]发现不同神经网络模型估算立木材积的精度

均高于传统的线性或非线性模型,Mushar 等^[18]在对比不同的机器学习算法与传统模型后发现神经网络模型的精度在分类树种材积比较中精度最高,而在全部类型树种材积比较中支持向量机最高。Wu 等^[17]运用 LiDAR 数据对立木材积进行估算,得出 3 种机器学习模型精度均高于传统模型的结论,其中支持向量机精度高于随机森林,这与本研究结果略有差异。

虽然机器学习算法目前仍存在问题,比如参数寻优没有统一的标准、无法掌握数据在黑箱中运行的过程、易产生过拟合现象^[11,29]等,但关于机器学习算法的研究还有广阔的进步空间。总而言之,机器学习算法作为一种新兴的建模方法,在大兴安岭樟子松立木材积的预测中相比传统的二元材积模型有着明显的优势,在森林经营管理中是一种有效的替代方案。

参考文献(reference):

- [1] MAO P, QIN L J, HAO M Y, et al. An improved approach to estimate above-ground volume and biomass of desert shrub communities based on UAV RGB images[J]. Ecol Indic, 2021, 125: 107494. DOI:10.1016/j.ecolind.2021.107494.
- [2] MUUKKONEN P. Generalized allometric volume and biomass equations for some tree species in Europe[J]. Eur J Forest Res, 2007, 126(2): 157-166. DOI:10.1007/s10342-007-0168-4.
- [3] 廖祖辉. 福建桉树人工林材积表和蓄积量表编制的研究[J]. 福建林业科技, 2005, 32(2): 17-20. LIAO Z H. Studies of the volume table and stocking table compilations of *Eucalyptus* plantations[J]. J Fujian For Sci Technol, 2005, 32(2): 17-20. DOI: 10.13428/j.cnki.fjlk.2005.02.005.
- [4] 吕勇, 刘辉, 王才喜. 杉木林分蓄积量不同测定方法的比较[J]. 中南林业学院学报, 2001, 21(4): 50-53. LV Y, LÜ H, WANG C X. Comparison of the different measurement methods of stand volume[J]. J Central South For Univ, 2001, 21(4): 50-53. DOI:10.3969/j.issn.1673-923X.2001.04.006.
- [5] WIANT H V JR, WOOD G B, WILLIAMS M. Comparison of three modern methods for estimating volume of sample trees using one or two diameter measurements[J]. For Ecol Manag, 1996, 83(1/2): 13-16. DOI:10.1016/0378-1127(96)03708-5.

- [6] OZCELIK R, WIANT H, BROOKS J. Accuracy using xylometry of log volume estimates for two tree species in Turkey[J]. Scan J Forest Res, 2008, 23 (3): 272 - 277. DOI: 10.1080/02827580801995323.
- [7] 曾伟生. 杉木相容性立木材积表系列模型研建[J]. 林业科学研究, 2014, 27(1): 6-10. ZENG W S. Establishment of compatible tree volume equation systems of Chinese fir[J]. For Res, 2014, 27(1): 6-10. DOI:10.13275/j.cnki.lykxyj.2014.01.002.
- [8] 许晴, 李晓莎, 许中旗, 等. 塞罕坝地区樟子松立木材积表研究[J]. 林业资源管理, 2017(1): 57-62. XU Q, LI X S, XU Z Q, et al. Study on the volume table of scotch pine in Saihanba area[J]. For Resour Manag, 2017(1): 57-62. DOI:10.13466/j.cnki.lyzygl.2017.01.011.
- [9] 雷相东. 机器学习算法在森林生长收获预估中的应用[J]. 北京林业大学学报, 2019, 41 (12): 23 - 36. LEI X D. Applications of machine learning algorithms in forest growth and yield prediction[J]. J Beijing For Univ, 2019, 41(12): 23-36. DOI:10.12171/j.1000-1522.20190356.
- [10] CUTLER D R, EDWARDS T C J, BEARD K H, et al. Random forests for classification in ecology[J]. Ecology, 2007, 88(11): 2783-2792. DOI:10.1890/07-0539.1.
- [11] GUAN B T, GERTNER G. Modeling red pine tree survival with an artificial neural network[J]. For Sci, 1991, 37(5): 1429-1440. DOI:10.1093/forestscience/37.5.1429.
- [12] MARIA J D. Artificial neural networks as an alternative tool in pine bark volume estimation[J]. Comput Electron Agric, 2005, 48(3): 235-244. DOI:10.1016/j.compag.2005.04.002.
- [13] MARIA J D, ÖZÇELİK R, YAVUZ H. Tree-bark volume prediction via machine learning: a case study based on black alder's tree-bark production[J]. Comput Electron Agric, 2018, 151: 431-440. DOI:10.1016/j.compag.2018.06.039.
- [14] COLIN J G, JONGHO I. Forest biomass estimation from airborne LiDAR data using machine learning approaches[J]. Remote Sens Environ, 2012, 125: 80-91. DOI: 10.1016/j.rse.2012.07.006.
- [15] MARIA J D, MILIOS E. Modelling total volume of dominant pine trees in reforestation via multivariate analysis and artificial neural network models[J]. Biosyst Eng, 2010, 105 (3): 306 - 315. DOI: 10.1016/j.biosystemseng.2009.11.010.
- [16] ÖZCELIK R, MARIA J D, BROOKS J R, et al. Estimating tree bole volume using artificial neural network models for four species in Turkey[J]. J Environ Manag, 2010, 91(3): 742-753. DOI: 10.1016/j.jenvman.2009.10.002.
- [17] WU J W, YAO W, CHOI S, et al. A comparative study of predicting DBH and stem volume of individual trees in a temperate forest using airborne waveform LiDAR[J]. IEEE Geosci Remote Sens Lett, 2015, 12(11): 2267-2271. DOI:10.1109/LGRS.2015.2466464.
- [18] MUSHAR S H, AHMAD S S, KASMIN F, et al. Machine learning approach for estimating tree volume[J]. J Phys: Conf Ser, 2020, 1502 (1): 012039. DOI: 10.1088/1742 - 6596/1502/1/012039.
- [19] BHERING L L, CRUZ C D, DE AZEVEDO P L, et al. Application of neural networks to predict volume in eucalyptus[J]. Crop Breed Appl Biotechnol, 2015, 15(3): 125-131. DOI:10.1590/1984-70332015v15n3a23.
- [20] 靳晓东, 姜立春. 基于树干不同形率的樟子松立木材积方程研建[J]. 北京林业大学学报, 2020, 42(3): 78-86. JIN X D, JIANG L C. Equation construction on standing tree volume of *Pinus sylvestris* var. *mongolica* based on different form quotients of trunk[J]. J Beijing For Univ, 2020, 42(3): 78-86. DOI: 10.12171/j.1000-1522.20190047.
- [21] 张明铁. 单株立木材积测定方法的研究[J]. 林业资源管理, 2004(1): 24-26. ZHANG M T. Study on volume measurement of single trees[J]. For Resour Manag, 2004(1): 24-26. DOI:10.13466/j.cnki.lyzygl.2004.01.005.
- [22] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning internal representations by error propagation[J]. Read Cognit Sci, 1988, 323(6088): 399-421. DOI: 10.1016/B978-1-4832-1446-7.50035-2.
- [23] 王轶夫, 孙玉军, 郭孝玉. 基于 BP 神经网络的马尾松立木生物量模型研究[J]. 北京林业大学学报, 2013, 35(2): 17-21. WANG Y F, SUN Y J, GUO X Y. Single-tree biomass modeling of *Pinus massoniana* based on BP neural network[J]. J Beijing For Univ, 2013, 35(2): 17-21. DOI: 10.13332/j.1000-1522.2013.02.018.
- [24] 徐晓明. SVM 参数寻优及其在分类中的应用[D]. 大连: 大连海事大学, 2014. XU X M. SVM parameter optimization and its application in the classification[D]. Dalian: Dalian Maritime University, 2014.
- [25] 曹正凤. 随机森林算法优化研究[D]. 北京: 首都经济贸易大学, 2014. CAO Z F. Study on optimization of random forests algorithm [D]. Beijing: Capital University of Economics and Business, 2014.
- [26] BREIMAN L. Bagging predictors[J]. Mach Learn, 1996, 24(2): 123-140. DOI:10.1007/bf00058655.
- [27] DING S, CHANG X H, WU Q H. A study on approximation performances of improved BP neural networks based on LM algorithms[J]. Appl Mech Mater, 2013, 411/412/413/414: 1935 - 1938. DOI: 10.4028/www.scientific.net/amm.411 - 414.1935.
- [28] 杜军岗, 魏汝祥, 刘宝平. 基于 PSO 优化 LS-SVM 的小样本非线性协整检验与建模研究[J]. 系统工程理论与实践, 2014, 34(9): 2322-2331. DU J G, WEI R X, LIU B P. Non-linear cointegration test and error correction modeling based on LS-SVM optimized by PSO in small sample[J]. Syst Eng-theory Pract, 2014, 34(9): 2322-2331. DOI: CNKI:SUN:XTLL.0.2014-09-015.
- [29] HAJAR A S, AZIZAH A N, HASZLINNA M N. Adoption of machine learning techniques in software effort estimation: an overview[J]. IOP Conf Ser: Mater Sci Eng, 2019, 551(1): 012074. DOI:10.1088/1757-899x/551/1/012074.

(责任编辑 李燕文 孟苗婧)