DOI:10.12171/j.1000−1522.20190356

雷相东，中国林业科学研究院资源信息研究所研究员、博士、森林经理与林业统计研究室主任、博士生导师、首席专家。中国林学会森林经理分会常务理事、中国系统工程学会林业系统工程专业委员会副主任、中国自然资源学会森林资源委员会副主任、国家林草局森林经营国家创新联盟秘书长。国家林业局“百千万人才工程”省部级人选。主要从事森林生长模型与模拟、多功能森林经营及规划等森林经理方面的研究。近年来在气候敏感的森林生长模型、适应性森林经营模拟和及立地质量评价等森林经理学科基础和前沿领域，取得了一系列成果。主持和参加国家科技支撑计划、国家自然科学基金项目、863 项目、林业行业公益性科研专项等 20 余项。曾任亚太森林恢复与可持续管理组织 APFNet“多功能林业示范基地建设试点示范项目”首席专家。在国内外学术期刊上发表论文 100 余篇，其中 SCI 收录 20 余篇，参编专著 10 部。获梁希林业科学技术奖一等奖 1 项、二等奖 3 项。

机器学习算法在森林生长收获预估中的应用

雷相东
(中国林业科学研究院资源信息研究所，北京 100091)

摘要:森林生长收获预估是森林经理学的一个重要方向，采用模型技术进行森林生长收获估计是森林经营决策的重要前提。传统的统计模型如线性及非线性回归模型、混合效应模型、分位数回归、度量误差模型等统计方法已被广泛应用于研究林木生长，但这些统计方法在应用时常常需满足一定的统计假设前提，诸如数据独立、正态分布和等方差等。由于森林生长数据的连续观测和层次性，上述假设通常难以满足。近年来随着人工智能技术的发展，机器学习算法为森林生长收获预估提供了一种新的手段，它具有对输入数据的分布形式没有假设前提、能够揭示数据中的隐含结构、预测结果好等优点，但在森林生长收获预估中的应用仍十分有限。文章对分类和回归树、多元自适应样条、bagging 回归、增强回归树、随机森林、人工神经网络、支持向量机、*K* 最近邻等方法在森林生长收获预估中的应用、软件及调参等进行了综述，讨论了机器学习方法的优势和挑战，认为机器学习方法在森林生长收获预估方面有很大的潜力，必将得到广泛应用，并和传统统计模型相结合成为生长收获模型发展的一种趋势。

关键词：森林生长收获预估；回归；分类；机器学习算法
中图分类号：S758.5　　**文献标志码**：A　　**文章编号**：1000−1522(2019)12−0023−14
引文格式：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 Xiangdong. Applications of machine learning algorithms in forest growth and yield prediction［J］. Journal of Beijing Forestry University, 2019, 41(12): 23−36.

Applications of machine learning algorithms in forest growth and yield prediction

Lei Xiangdong
(Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China)

Abstract: Forest growth and yield prediction is an important field of forest management science, and modelling forest growth and yield is key to forest management decision-making. The traditional statistical

收稿日期: 2019−09−11　**修回日期:** 2019−11−27
基金项目: 国家自然科学基金项目(31870623)。
作者简介: 雷相东，研究员，博士生导师。主要研究方向：森林生长模型与模拟。Email：xdlei@ifrit.ac.cn　地址：100091 北京市海淀区香山路东小府 2 号中国林业科学研究院资源信息研究所。
本刊网址: http://j.bjfu.edu.cn；http://journal.bjfu.edu.cn

growth models such as linear and nonlinear regression model, mixed-effect model, quantile regression, variable-in-error model are often applied under certain statistical assumptions, such as the data are independent, normally distributed and homoscedastic. The above requirements are usually difficult to be met for forest data with repeated observation and hierarchy. With the development of AI techniques, machine learning provides a new way for forest growth modeling, with the advantages of no requirements on data distribution, extracting deep knowledge from the data, and high accuracy. The applications in forest growth and yield are still less than other domains. We reviewed the main machine learning algorithms including classification and regression tree (CART), multivariate adaptive regression splines (MARS), bagging regression, boosted regression tree (BRT), random forest (RF), artificial neural networks (ANN), *k*-nearest neighbors (*k*-NN), and support vector machine (SVM), parameter tuning, software, advantages and challenge. We conclude that machine learning would be widely applied with great potential and its combination with traditional statistical methods would become a trend in forest growth and yield prediction.

Key words: forest growth and yield prediction; regression; classification; machine learning algorithm

森林生长收获模型是森林经营决策的重要工具，在森林资源数据更新、生长收获预估和森林经营规划中都有着重要的作用[1]。按建模分辨率可分为单木模型、径阶模型、林分模型和景观模型；按模型建模方法可分为基于生理过程的过程生长模型、基于观测数据的统计模型或经验模型和结合二者优点的混合模型[2−4]。与过程模型相比，经验模型具有精度高、方便森林经营应用等优点而得到广泛应用。它通常用一个或一组数学函数（回归方程），来描述森林生长与林分状态如年龄、立地、密度等的关系。线性及非线性回归模型、混合效应模型、分位数回归、度量误差模型等统计方法已被广泛应用于研究林木生长[1]。但这些统计方法在应用时常常需满足一定的统计假设前提，诸如数据独立、正态分布和等方差等。由于森林生长数据的连续观测和层次性，上述假设通常难以满足。此外，森林生长是一个复杂的非线性过程，它受到遗传、气候、立地、竞争、干扰等多个因子及其交互作用的影响，因此经验模型还面临着模型选型、变量筛选和参数收敛的挑战。随着人工智能技术的发展，机器学习算法为森林生长收获预估提供了一种新的手段，它具有对输入数据的分布形式没有假设前提、能够很好地处理因变量和自变量之间复杂的关系、深度挖掘数据中有价值的信息、能够揭示数据中的隐含结构、获得更好的预测模型等优点[5−6]，已经广泛应用于遥感和生态领域[7−8]。与遥感和生态领域相比，人工神经网络是较早、较成熟的用于森林生长收获预估的算法[9−15]，但其他机器学习算法的应用并不够广泛和深入[16−17]。本文介绍了主要的机器学习算法，对其在森林生长收获预估中的应用进行分析的基础上，指出了存在的问题和发展趋势。

1 主要机器学习算法及应用

机器学习已成为人工智能发展的重要驱动力，广泛用于数据挖掘等多个领域。机器学习的焦点是学习和构造能够从数据中建立预测或描述模型的学习算法，其目标是使学得的模型能很好地适用于新样本[18−19]。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法，既可用于回归，也可以用于分类问题。对于机器学习算法，有不同的分类。本文按递归划分方法、集成学习算法和黑箱方法进行分析。主要包括分类和回归树法、多元自适应回归样条、随机森林、Bagging、Boosting、支持向量机、神经网络和 *k*-最近邻算法等。

1.1 递归划分方法

递归划分方法主要包括分类和回归树法（classification and regression tree，CART）、多元自适应回归样条法（multivariate adaptive regression splines，MARS）。

1.1.1 分类和回归树（CART）

CART，隶属决策树（decision tree），是简单的决策模型，通过按一定规则持续拆分数据，每次将数据划分为两个相对一致的子集，直至达到目标，从而形成树状结构（图 1）。它是一个递归的过程，可用于分类和回归。它采用一种启发式算法，找到最优的拆分变量和最优的切分点，即选择拆分变量和它的取值将输入空间划分为两部分，然后重复这个操作。通过比较不同划分的误差来找到最优的拆分点和拆分变量，即用真实值和划分区域的预测值的最小二乘来衡量。其中一种拆分变量的选择规则是该节点的总平方和（total sum of squares，SST）和子节点的 SST 之差Δ_{SST} 最大[20]。

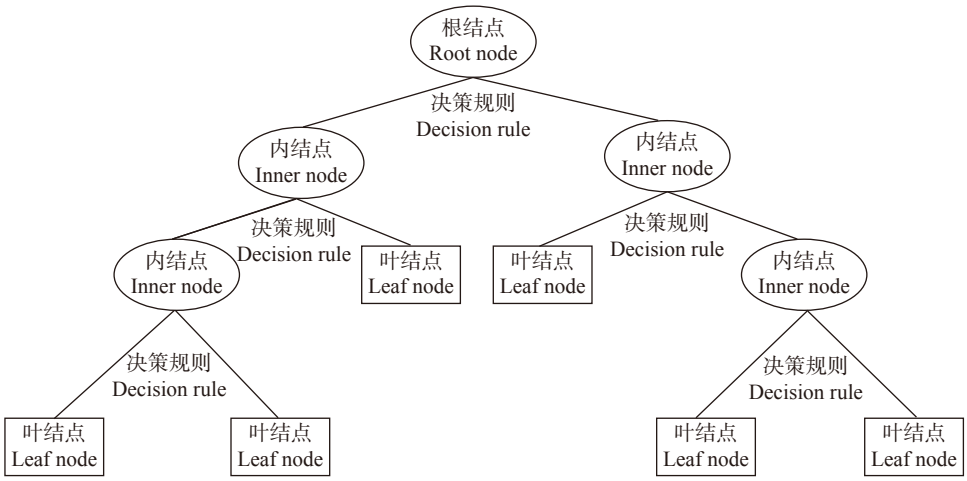

图 1　决策树分枝示意

Fig. 1　Sketch map of decision tree branching

$$\Delta_{SST}=\sum_{i\in 父结点}(y_i-\bar{y})^2-\left\{\sum_{i\in 子结点1}(y_i-\bar{y}_1)^2+\sum_{i\in 子结点2}(y_i-\bar{y}_2)^2\right\} \tag{1}$$

式中：y_i 表示属于父结点、子结点 1 或子结点 2 集合的第 i 个值，$\bar{y}$、$\bar{y}_1$和$\bar{y}_2$分别表示父结点、子结点 1 和子结点 2 集合的平均值。

CART 在森林生长收获中的一个最好直接的应用就是枯死和存活分析，因为树木存活和死亡本身就是一个二分类问题。Dobbertin 和 Biging[21] 基于美国加利福尼亚州 569 块针叶混交林中 4 722 株西黄松（*Pinus ponderosa*）和 6 810 株白冷杉（*Abies concolor*）数据，开展了基于 CART 的枯死研究，并和 Logistic 回归模型进行比较，结果表明 CART 对西黄松枯损的预估精度在 28% ~ 36% 之间，对白冷杉枯损的预估精度在 11% ~ 17% 之间，CART 模型和 Logistic 模型预估枯损的能力相当；但 CART 和 Logistic 回归得到的自变量有所差异，CART 能发现 Logistic 回归不能包括的重要自变量。Fan 等[22] 以美国密苏里州成熟栎类混交林为对象，基于 455 块 35 000 株单木调查数据，采用 CART 方法进行了单木存活分析，其输入变量包括树种、胸径、树冠等级和大于对象木的断面积，发现树种和树冠等级是影响枯死的两个最重要的变量；与传统的 Logistic 回归方法相比，其不需要分布假设，而且能通过树给出自变量的重要层次，对于非竞争引起的枯死（输入变量主要是分类变量）研究尤其有用；其最大的优势在于能将存活能力显著不同的树进行分组，并能通过树状图很直观地进行解释，便于在实践中应用。

除枯死外，CART 还用于其他生长预测。如 Adamec 和 Drápela[23] 比较了回归树、非线性混合效应模型和最小二乘法在树高曲线中的表现，发现 CART 的预测结果不如其他方法，但认为是一种用户友好的方法，可以在允许误差较大时使用；Aertsen 等[24] 在地位指数研究中也发现类似的结果。Räty 和 Kangas[25] 还将该方法用于芬兰森林蓄积模型的局部建模单元的划分。

总的来说，CART 很容易反映非线性关系和变量的交互作用，不受异常值和共线性的影响[26]，能处理大量数据，尤其是具有层次结构特征的数据[27]，用树直观反映变量的重要性。但该方法需要大量的样本才能保证类间的区分和最终节点的样本数，单棵树模型的结果不稳定、不能反映数据空间结构、大型回归树比回归模型更难解释。

1.1.2　多元自适应回归样条（MARS）

MARS 最初由 Friedman[28] 提出，通过使用基函数将自变量和因变量的关系分段进行线性回归拟合。其建模过程受 CART 的启发，与 CART 不同的是，通过递归划分和样条函数的结合，MARS 创建的是具有连续导数的连续模型，而不是树节点处的不连续分支[28−29]。节点和自变量的选择都是通过基函数和详尽搜索来选择的。通过大量的节点的连接可以描述任何不规则的曲线。其模型形式可以表示为：

$$\hat{f}(x)=\beta+\sum_{n=1}^{N}\beta_n B_n(x) \tag{2}$$

式中：$\hat{f}(x)$为因变量；x 为自变量；β 表示截距项；$B_n(x)$表示基函数，它可以是单个链接函数，或两个或多个铰链函数的乘积；β_n表示第 n 个基函数的系数；N 表示基函数或节点的总数。

Chojnacky 和 Heath[30] 基于缅因州 2 491 块固定样地数据，采用 CART、MARS 方法得到影响倒木的生物量的林分因子，发现林分断面积、生物量、枯立木数量和伐桩数量都是预测倒木生物量的重要变量。Hart 和 Laroque[31] 基于加拿大阿尔伯塔 Jasper 国家公园的树木年轮数据，采用 MARS 研究气候与

生长间的关系，通过对结点的分析来确定气候-生长关系的阈值。Moisen 和 Frescino[32] 还将其用于大面积森林属性的估计。Ou 等[33] 关于单木直径生长预测的研究中，发现 MARS 的效果要远低于其他几种机器学习方法。

该方法可以被视为对线性模型一种扩展，因多元自适应回归样条的非线性和基函数的选择，使得它不仅适用于处理高维问题，而且能够自动捕捉变量之间的非线性和交互作用[28,34−35]。

1.2 集成学习方法

集成学习通过选择一种结合策略将若干基学习器集合成一个强学习器来实现预测。按照基学习器之间是否存在依赖关系可以分为两类：基学习器之间不存在强依赖关系，代表算法是 bagging 和随机森林（random forest）系列算法；存在强依赖关系，代表算法是 boosting 系列算法。

1.2.1 bagging 回归和随机森林

bagging 是 bootstrap aggregating 的缩写，又称袋装技术。bagging 回归是针对回归树算法中单棵树结果的不稳定性和误差提出的。它通过对原始数据重抽样（bootstrap）法产生多个训练数据集，建立多棵树，用其平均值来作为最终结果从而减少方差。在 Bagging 的每轮随机采样中，训练集中大约有 36.8% 的数据没有被采中，这部分数据称之为袋外数据（out-of-bag data）。这些数据没有参与训练集模型的拟合，因此可以用来检测模型的泛化能力[20]。

随机森林（random forest, RF）是 Bagging 回归的进化版，是一类以决策树为基学习器的集成学习方法，由 Breiman[36] 于 2001 年提出并得到广泛应用。该方法通过重抽样（bootstrap）手段构建一系列基学习器，并将这些基学习器的预测结果组合起来输出。在每个基学习器的构建过程中，每个树结点在选择拆分变量时，首先随机地从全部 P 个变量中选取 $k(1 \leqslant k \leqslant P)$个，然后从中寻找一个最优划分变量，因此称为随机森林[18,29]。值得指出的是，组合了众多分类与回归树的 RF 模型失去了单棵 CART 的那种简单的树形结构，且无法直观地给出预测变量与响应变量之间的关系，造成模型的可解释性降低。因此，一些学者[36−38] 给出了预测变量的相对重要性以及偏依赖图（partial dependence plots）两种方法来提高 RF 模型的可解释性。相对重要性可通过结点纯度提升法和 OOB 置换法来确定[39]。结点纯度提升法利用变量在结点处分裂产生的结点纯度的提升来度量，OOB 置换法利用变量被置换前后均方误差的改变量来度量。

虽然 RF 在遥感等领域应用非常普遍，但在森林生长收获预估方面应用并不多见。Weiskitte 等[40] 用 RF 研究气候驱动的地位指数和 GPP 预测，并模拟未来气候变化对地位指数和 GPP 的影响，RF 模型能解释 75% 的地位指数变异。Bond-Lamberty 等[41] 用 RF 算法研究了气候和干扰对加拿大北方针叶林直径的生长的影响，发现模型大约能解释 23%～44% 的直径生长变异。Kilham 等[42] 将 RF 算法用于德国东南部森林采伐木选择和采伐蓄积预测，发现 RF 算法要优于广义线性混合模型的结果；它能自动定义变量间的关系和交互作用。欧强新等[43] 关于单木直径生长预测的研究中，发现随机森林模型具有一定的统计可靠性，产生的变量重要性和偏依赖图具有合理的林学意义。其他的应用包括树木削度的估计[44]。

RF 的主要优点有：训练可以高度并行化；可以给出各个变量对于输出的重要性；模型泛化能力强；对部分特征缺失不敏感[20]。但在在某些噪音比较大的样本集上，RF 容易陷入过拟合；取值划分比较多的特征容易对 RF 的决策产生更大的影响，从而影响模型拟合的效果。

1.2.2 Boosting 回归树

增强回归树（boosted regression tree, BRT）也是一类基于集成学习的方法。该方法是对 CART 的提升，通过整合 CART 和提升（boosting）两种算法，使计算结果的稳定性和精度得到提高[45]。与 RF 不同，BRT 模型在构建的过程中，其基学习器的学习是顺序进行的，是专门用来产生互补的模型；各模型不是同等重要，而是根据其表现进行加权，即性能好的模型对最终预测结果有更大的影响。而 RF 则是各学习器的简单组合[46]。具体地，对于既定的损失函数（如回归中的均方误差）和基预测器，BRT 通过对训练数据集重抽样以得到一系列新的训练数据集，依据每个新训练数据集来生成一个相应的基预测器：首先，BRT 通过最小化损失函数来构建一个基学习器，第二个基学习器依据第一个基学习器的梯度（如残差）来构建，依次类推，直到达到用户指定的迭代次数为止[47]。BRT 也能够给出预测变量的相对重要性以及偏依赖图，其计算方法与 RF 相同。与 BRT 类似的还有梯度提升回归树（gradient boosting regression tree, GBRT）有时也称为 GBDT。与 BRT 不同的是，它的采样不是放回抽样；每一次计算都是为了减少上一次的残差，为了减少这些残差，可以在残差减少的梯度（gradient）方向上建立一个新模型，最后将每阶段模型加权相加得到最后的结果。与 Bagging 和 RF 相比，BRT 同时降低了方差和误差[48]。

增强回归树广泛用于树木死亡[49−51]、进界[52]、林

分断面积[53]、生物量和碳储量估计[54−57]、生物量生长[58]和地位指数预测[59−60]等方面。Sproull 等[50]基于波兰北部地区 65 块挪威云杉(*Picea abies*)样地数据，利用 BRT 分析了地形因子和林分变量因子对甲壳虫虫害引起的林木死亡率的影响，结果表明，在测试集上，模型仅能够解释 12% ~ 24% 的死亡率变异，而在训练集上能够解释 56% ~ 62% 死亡率变异。Lin 等[55]采用 BRT 构建了天然常绿阔叶混交林地上碳储量与林分密度、地形、植物功能多样性、群落加权平均功能属性之间的预估模型，结果表明 BRT 模型解释了 73% 的地上碳储量变异。Ren 等[58]利用 BRT 模型分析了林分因子、地形因子以及土壤因子对生物量平均生长量的影响，结果表明模型可解释了 52.9% 的生物量生长量变异，且 3 类因子对生物量生长量的影响依次递减。Razakamanarivo 等[54]基于马达加斯加岛中部高地 41 块红桉(*Eucalyptus robusta*)人工林样地，发现一元线性模型、多元线性模型和 BRT 模型分别能够解释 33%、60% 和 74% 地上碳储量变异，以及 52%、48% 和 85% 地下碳储量变异。Fricker 等[61]用 BRT 检验了影响最大树高的环境因子，输入变量包括气候因子、地形因子、土壤、林分因子，发现 BRT 模型能够模拟生态过程中的非线性关系，可以很好地服务于森林经营管理。

1.3　黑箱方法

包括人工神经网络(artificial neural networks, ANN)、支持向量机(support vector machine, SVM)以及 *k*-最近邻(*k*-nearest neighbors, *k*-NN)，由于将输入转换为输出是通过一个模糊的“箱子”来处理的，称为黑箱(black box)过程。

1.3.1　人工神经网络 ANN

人工神经网络系统出现于 20 世纪 40 年代，它是通过模仿人类大脑神经网络处理和记忆信息，属于最早提出的机器学习方法之一[20,62]。ANN 由输入层、隐藏层和输出层 3 部分组成(图 2)，其因变量可以是一个或多个，隐藏层同样也可以是一个或多个。它的原理是把上层节点的值加权到下层节点，经过非线性变化，最终到输出层节点，然后根据误差的大小反馈到前面的层，修改权值，再重新加权平均，反复这样训练，直到误差在可接受的范围。ANN 的处理过程主要通过激活函数来实现，它赋予神经网络非线性的特性，将输入转换为输出。常用的激活函数有线性函数、单位阶跃函数、sigmoid 函数、双曲正切函数、softmax 函数等[63−64]。ANN 适用于分类和预测问题，能模拟更复杂的模式，对数据的基本关系不需要做出假设，但计算量大，训练缓慢，容易过度拟合。由于是“黑箱”，它主要用于预测。其学习能力来源于它的拓扑结构，或相互连接的神经元的结构，通过层的数目、信息传播的方向(前向、后向或递归)、每一层的结点数 3 个特征形成不同的形式。除经典的 BP 神经网络外(backpropagation neural networks)，随着深度学习的发展，出现了深度神经网络 DNN(deep neural networks)、卷积神经网络 CNN(convolutional neural networks)、循环神经网络 RNN(recurrent neural networks)等算法。但这些方法在森林生长收获预估中很少应用。

ANN 是较早用于森林生长收获预估的机器学习方法。在单木枯死和存活预测方面，Hasenauer 等[65]比较了 Logistic 回归模型和 ANN 模型对挪威

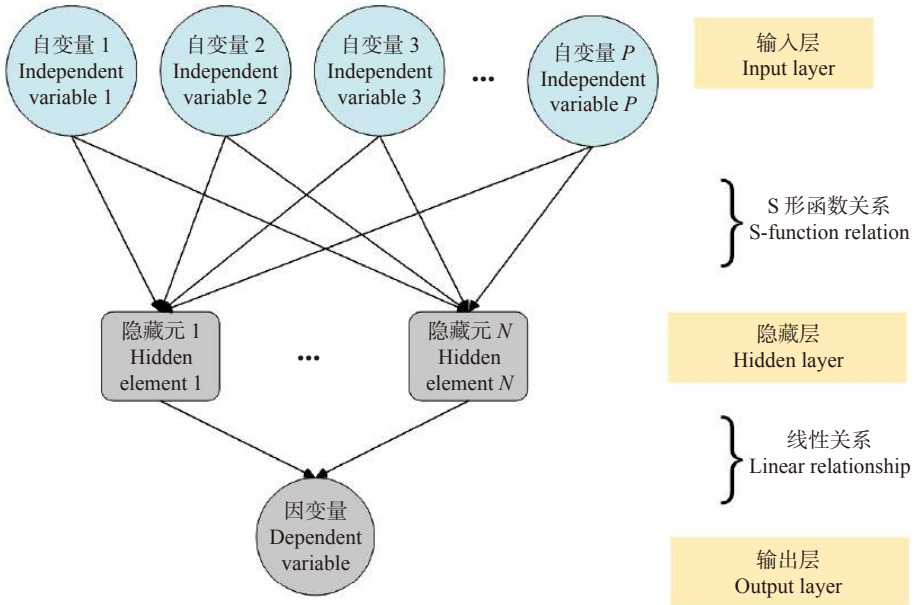

图 2　ANN 示意图

Fig. 2　Schematic diagram of ANN

云杉单木枯死的预测，结果表明 ANN 模型的模型表现要优于 Logistic 回归模型。Castro 等[66] 基于巴西北部 65 块桉树（*Eucalyptus*）人工林样地，将树高、胸径、立地指数、年龄以及与距离无关的竞争指数等作为输入，利用 ANN 模型估算了各径阶林木的年枯死率，ANN 模型对各径阶林木每年枯死率预测的确定系数介于 0.579 ~ 0.799 之间。Reis 等[67] 以亚马逊热带雨林为对象，建立了基于 ANN 的单木存活和枯死模型，发现存活估计精度约在 90%，远高于枯死估计；类似的结果如 da Rocha 等[68] 对巴西森林枯死、Bayat 等[69] 对伊朗森林枯死和存活的估计。在单木胸径、断面积和树高生长预测方面，Soares 等[70] 用多层感知机 ANN 预测了桉树不同高度部位的直径和立木材积，取得了满意的结果；Ashraf 等[71] 用加拿大新斯科舍省 1 000 块固定样地数据，用 ANN 建立了单木断面积生长模型，其确定系数在 0.38 ~ 0.60，ANN 优于传统回归方法；Vieira 等[72] 基于巴西 398 块桉树样地，将胸径、树高、优势高、基因材料作为输入，用 ANN 估计胸径和高生长，结果表明 ANN 优于回归方法。马翔宇等[73] 以东北林业大学帽儿山实验林场白桦（*Betula platyphlla*）人工林为研究对象，用林分内单木相对直径、林分密度指数、地位指数和林分年龄作为输入变量，以单木胸径生长量作为输出变量，构建了 ANN 模型，总体拟合精度为 96.86%。在林分平均高和优势高估计方面，沈剑波等[74] 基于吉林省的 168 块长白落叶松（*Larix olgensis*）人工林样地的林龄、林木竞争因子和立地因子数据，利用 ANN 模型预测了林分平均高，结果表明 ANN 模型解释了 84.55% 的平均高变异；其他的研究也显示了较高的预估精度[11,15,75−76]。Reis 等[77] 基于 36 个多期观测样地，用 ANN 预测单木胸径生长，确定系数达 0.98。其他的应用包括生物量和碳估计[78−81]、立地指数[82]、林分蓄积量和其他因子[83−85]、单木削度和材积[44,86]、树高曲线[87]、直径分布参数估计[88]。

1.3.2 支持向量机 SVM

该方法于 1992 年由 Boser 等提出，它使用一种称为超平面的边界将数据划分为具有近似值的组。通过分析输入变量和输出变量的数量关系，对新观测的输出变量进行预测。用于分类时，用超平面分类包括线性可分的数据、线性不可分的数据及非线性空间 3 种情况。对于非线性问题，需要通过核函数来解决。常用的核函数如线性核、多项式核、S 形核、径向基函数核等[20]。与统计中的回归分析一样，它也遵循离差平方和最小的原则进行超平面参数估计，但为了降低过拟合风险，采用 ε 不敏感损失函数[89]，即当观测值与预测值的绝对偏差不大于事先给定的 ε 值时，该误差不计入损失函数。SVM 在解决小样本、非线性、高维的回归和分类问题上有很大优势。但该方法属于一种“黑箱”模型，难以解释；对缺失数据敏感；计算复杂度高，且处理大规模数据的效率很低；当数据中自变量有太多定性变量或定性变量的水平太多时往往无法运作。

SVM 在遥感中得到广泛应用，在森林生长收获预估方面的应用很少。林卓等[85] 利用 ANN 和 SVM 估算杉木（*Cunninghamia lanceolata*）人工林样地蓄积，发现两者分别能够解释 93.5% 和 93.6% 的蓄积变异。Che 等[90] 用 SVM 来预测杉木的林分断面积，发现要优于传统方法。

1.3.3 *k*-最近邻（*k*-NN）

k-NN 的基本思想是在已有数据中找到与新的输入变量 X_0 相近的观测（X_0 的近邻），将这些近邻输出值的平均值作为 X_0 的预测值。即将样本中包含的 n 个观测数据看成 p 维（p 个输入变量）特征空间中的点，并根据 X_0 的 k 个邻体$(y_1,y_2,\cdots,y_k)$依函数 $F(y_1,y_2,\cdots,y_k)$计算输出$\hat{y}_0$：

$$\hat{y}_0 = \frac{\sum_{X_i \in N_k(X_0)} y_i}{k} \tag{3}$$

式中：$N_k(X_0)$是 X_0 的 k 个近邻的集合。

因此 *k*-NN 算法的关键，一是表征与 X_0 的邻近关系的距离指标，二是 k 值的确定，三是加权的核函数[20]。

与 SVM 类似，KNN 方法在基于遥感的林分参数估计方面有大量应用，但在林分生长收获预估方面较少。仅有的如 Maltamo 和 Kangas[91] 将 KNN 用于林分断面积分布预测。

2 软件及参数调节

R 语言作为功能强大、免费且源代码开放的数据分析工具，在建模领域得到越来越多的应用。针对不同的机器学习算法，R 都提供了相应的包来执行机器学习建模任务[92]。此外，在利用机器学习算法进行林分生长预估时，大多数机器学习算法都需要对超参数（hyper-parameters）进行设置，超参数配置不同，学习得到的模型的性能往往有显著的差别，或者说对应不同的模型。因此，在进行机器学习模型评估和选择时，除了要选择适用的机器学习算法外，还需要对算法超参数进设置，即“参数调优”或简称“调参”。表 1 给出了第 1 节中提到的主要机器学习算法的 R 程序包和参数调节（parameter tuning）方法[20,93−96]。

3 模型评价与比较

3.1 模型评价

在机器学习中，将学习器在训练集上的误差称

为“训练误差”（training error）；在新样本上的误差称为“泛化误差”（generalization error）。泛化误差小的学习器为理想的模型。同时测试集与训练集最好互斥，即测试样本未在训练集中使用。交叉验证（cross

表 1　R 软件中机器学习包及调参方法

Tab. 1　R packages and parameter tuning of machine learning algorithms

机器学习算法 ML algorithm	R 程序包 R package	可调超参数 Adjustable hyper-parameter	调参方法 Method of parameter tuning	默认值（回归） Default value (regression)
CART	rpart	复杂度参数 Complexity parameter (cp)	cp取介于0 ~ 1的实数。可基于交叉验证（如10折交叉验证），建立调参网格对cp进行调优（如cp可设置为0.000 1，0.001，0.01，0.1等不同值） The value of cp is between 0 and 1. The optimal one could be obtained by grid search (with different cp values of 0.000 1, 0.001, 0.01, 0.1 for example）	cp = 0.01
MARS	earth	①自变量间交互的阶数 The degree of interaction of input variables (degree) ②模型中最大的项数 Maximum number of terms (including intercept) in the model (nprune)	degree为大于等于1的整数，Hastie等[93]建议为交互项degree设置一个上限（如degree ⩽ 3）。2 ⩽ nprune ⩽ nk，nk的计算公式为：nk = min(200, max(20, 2 × ncol(*x*)))，式中，min()与max()分别表示取最小值和取最大值，ncol(*x*)表示自变量*x*的总数。可基于交叉验证，建立调参网格对degree和prune进行调优 The value of degree is the integer ⩾ 1. Hastie et al.[93] suggested degree should be set an upper limit (degree ⩽ 3, for example). 2 ⩽ nprune ⩽ nk, nk = min(200, max(20, 2 × ncol(*x*))), where ncol(*x*)is the number of input variables. The values of degree and prune could be attained by grid search based on cross validation	degree = 1 nprune无默认值 No default provided for nprune
Bagging回归树 Bagging regression tree	ipred	决策树的数量 The number of decision trees (nbagg)	nbagg取值为大于等于1的整数，该参数取值仍需依具体数据而定，为保证预估结果的可靠性且不影响计算效率，可将nbagg设置为大于25的值（如50等） The value of nbagg is the integer ⩾ 1. It could be set as the integer ⩾25 (50 for example) for prediction reliability and computation efficiency	nbagg = 25
RF	randomForest	①随机森林中决策树的数目 The number of decision trees (ntree) ②树节点随机抽选的变量个数 The number of input variables randomly sampled as candidates at each node (mtry)	ntree和mtry取值均为大于1的整数，当ntree在500以后整体误差便趋于稳定，但仍需依据具体数据而定，为保证预估结果的可靠性且不会影响计算效率，ntree可以取大于500的值（如1 000等）；1 ⩽ mtry ⩽ *P*, *P*为全部自变量数目，可基于交叉验证，建立调参网格对mtry进行调优 The values of ntree and mtry are the integers ⩾ 1. The bias could be stable when ntree is larger than 500. It could be set as the integer ⩾ 500 (1 000 for example) for prediction reliability and computation efficiency. mtry is less than the number of all input variables *P*, and the optimal value could be attained by grid search based on cross validation	ntree = 500； mtry为全部自变量数目的三分之一（取整） The value of mtry is the one-third of all input variables (integer)
BRT	gbm	①损失函数的形式 The name of the distribution (distribution) ②决策树的数目（或称迭代次数） Integer specifying the total number of trees to fit (n.trees) ③学习速率（或称收缩参数） The learning rate or step-size reduction (shrinkage) ④再抽样比率 The fraction of the training set observations randomly selected to propose the next tree in the expansion (bag.fraction) ⑤变量交互的深度 The maximum depth of variable interactions (interaction.depth)	对于回归问题distribution设置为gaussian；bag.fraction (0 < bag.fraction ⩽ 1)，Friedman[96]推荐将bag.fraction设置在0.5左右；shrinkage (0 < shrinkage ⩽ 1)，n.trees为大于1的整数，shrinkage 影响超参数n.trees的取值，Ridgeway[95]建议将shrinkage 的取值范围设置在0.01至0.001之间，同时n.trees取值介于3 000至10 000之间；interaction.depth取值为大于等于1的整数，为了平衡计算开销和模型性能，该超参数可尝试若干值进行调优（如1，3，5，7，9）。可基于交叉验证，建立调参网格对上述超参数进行调优 gaussian is assumed for regression; bag.fraction is suggested to be set as 0.5 by Friedman[96], shrinkage is suggested to be between 0.01 and 0.001 and n.trees between 3 000 and 10 000 by Ridgeway[95]. Shrinkage affects the value of n.trees. The value of interaction.depth is a integer ⩾ 1, and several specific values could be tested (1, 3, 5, 7, 9 for example) for the balance of model performance and computation efficiency. All optimal values of these hyper-parameters could be attained by grid search based on cross validation	Distribution = gaussian； n.trees = 100； shrinkage = 0.1； bag.fraction = 0.5； interaction.dept = 1
SVM	kernlab	①核函数 Kernel function (kernel) ②代价参数 Cost of constraints violation (*C*)	对于回归问题，线性核函数以及径向基核函数是两类常用的kernel；*C*为大于0的实数（如0.01，0.1，1，10，100，…）。可基于交叉验证，建立调参网格对上述超参数进行调优 Linear or radical basis functions are commonly used for regressions. The optimal values of *C* (0.01, 0.1, 1, 10, 100, … for example) could be attained by grid search based on cross validation	kernel为径向基核函数； Radical basis function; *C* = 1

表 1(续)

Tab.1(Continued)

机器学习算法 ML algorithm	R 程序包 R package	可调超参数 Adjustable hyper-parameter	调参方法 Method of parameter tuning	默认值(回归) Default value (regression)
ANN	nnet	①隐藏节点个数 The number of hidden node (size) ②权重衰减 Weight decay (decay)	size为大于等于0的整数，一般使用的size确定方法为，size= $\sqrt{P+O}+m$，P表示输入层自变量的个数，O表示输出层因变量的个数，m的取值为0 ~ 10之间的整数；decay为0至0.1的实数。可基于交叉验证(如10折交叉验证)，建立调参网格对上述超参数进行调优size= $\sqrt{P+O}+m$ Size is the integar ⩾ 0 and determined as size= $\sqrt{P+O}+m$, where P is the number of input variables, O is the number of the output variables, m is a integer between 0 and 10. Decay is a real between 0 and 0.1. The optimal values of these parameters could be attained by grid search based on cross validation	decay = 0 size无默认值 no default provided for size
k-NN	caret	近邻点的个数 The number of nearest neighbors (k)	k为大于等于1的整数，通常k的取值在3 ~ 10范围内[94]。可基于交叉验证, 建立调参网格对k进行调优 k is the integer ⩾ 1 and often set as between 3 and 10 according to Lantz[94]. The optimal value could be attained by grid search based on cross validation	k = 5

validation)是评价机器学习模型的一种主要方法，最常用的是 k-折交叉验证(k-fold cross validation)，即把数据分成 k 份，每次把其中的一份作为测试集，其余的 k-1 份作为训练集，共做 k 次，从而得到 k 个误差统计量，以其平均值做为模型的评价指标。k 最常用的取值为 10，称为 10 折交叉验证。由于测试集的选择是随机的，交叉验证结果不唯一，但多次交叉验证的结果差别不会太大。在生长收获预估中，主要归结为分类和回归两类问题。对于回归问题，传统回归中的误差统计量如均方误差、平均误差、相对误差的平均值都可以用来评价模型。对于分类问题，主要采用混淆矩阵，计算平均误判率、查准率和查全率，以及 ROC 曲线来度量[18]。

3.2 模型比较

如前所述，迄今为止，产生了大量的机器学习算法。每种算法都有其优缺点。表 2 给出了主要机器学习算法的优缺点[20,92−93]。

4 问题与展望

总体来看，机器学习已在森林生长模型领域开始应用，表现出了机器学习方法的优势和潜力，并成为生长收获模型发展的一种趋势。但目前该方法在森林生长收获预估中的应用仍然十分有限，除人工神经网络外，其他机器学习算法的发展应用比较缓慢。一方面是由于机器学习算法的可解释性，即不像传统回归那样明确；另一方面是人们习惯了统计学，而对机器学习算法不太熟悉。机器学习算法远远超过本文提到的类型，不同算法之间的比较也需要更多的检验和验证。可以预料，机器学习在森林生长收获中的应用会超出我们的想象，并不断地为人们认识森林的生长提供新的知识。

4.1 机器学习的优势及局限

机器学习方法有 3 个主要优点：(1)不用像参数估计方法中要求的假设前提就可以模拟复杂的非线性关系；(2)能同时利用和评价大量或含噪音的数据；(3)快速实现的能力。大多算法能自动地检验变量间的交互作用，对异常值不敏感等。与传统统计回归相比，机器学习不能产生变量的 p 值，很难确定自由度和单个最优模型，但可通过偏依赖图来说明变量间的关系。不像传统统计更关注模型假设和最优模型，机器学习更强调数据探索、结果的解释和问题本身。对机器学习最大的批评就是黑箱问题。实际上，每种算法内部都有其工作原理，只是对使用它的人来说可能是黑箱。因此，黑箱论并不是一种客观的评价。我们不能因此而排斥它，而是想办法使其越来越透明，从而被更多的人理解和应用。此外，模型应用时应关注过拟合问题，即对训练数据有好的预测结果，但对检验数据则较差(泛化误差)。泛化误差是真正检验模型对于新数据的预测效果。维数灾难(curse of dimensionality)或属性个数也是需要考虑的问题，通常需要降维或进行重要变量的初步筛选[97]。当出现问题数据，如类间不均衡、定性变量过多、异常值或缺失数据过多时，也需要通过重采样等办法进行处理。

4.2 与传统统计方法的比较和结合

如上所述，虽然机器学习算法在一些方面较传统统计方法有一定优势，但是直接比较两种方法的表现是很困难的，因为没有统一的评价标准，结果常常是与个例相关[97−98]。在森林生长收获预估中，存在大量的非线性关系和交互作用，机器学习方法可能更有用，但选择哪种方法需要研究人员针对具体问题和数据而定。一些情况下传统统计方法

表 2　主要机器学习算法优缺点

Tab. 2　Advantages and disadvantages of main machine learning algorithms

机器学习算法 ML algorithm	优点 Advantage	缺点 Disadvantage
CART	与分布无关；受共线性和异常值影响小；简单，容易解释，能自动处理交互作用；能处理连续和分类变量 Distribution independent; less affected by collinearity and outliers; easy explained; deal with interactions, continuous and category variables	单棵树的结构不稳定，容易出现过拟合，大的决策树不易解释 Unstable structure for single tree; over-fitting; difficulty in explaining big decision trees
bagging回归 Bagging regression	与分布无关；调节参数少；受共线性和异常值影响小；较CART泛化能力强；能处理连续和分类变量 Distribution independent; less affected by collinearity and outliers; less tuning parameters; better generalization capacity than CART; deal with interactions, continuous and category variables	不如单棵决策树容易解释，对过多变量会敏感 Not so easy to be explained as single decision tree; sensitive to too many input variables
RF	与分布无关；调节参数少；受共线性和异常值影响小；能产生变量重要性和偏依赖图；与bagging相比，训练出的模型的方差更小，泛化能力强；能处理连续和分类变量 Distribution independent; less affected by collinearity and outliers; less tuning parameters; better generalization capacity than CART; deal with interactions, continuous and category variables	不如单棵决策树容易解释，对过多变量会敏感 Not so easy to be explained as single decision tree; sensitive to too many input variables
BRT	与分布无关；受共线性和异常值影响小，可处理复杂非线性关系；能产生变量偏依赖性和相对重要性图；能处理连续和分类变量 Distribution independent; less affected by collinearity and outliers; deal with complex nonlinear relations, and continuous and category variables; generate variable importance and partial dependence plots	超参数过多，调参复杂 Many hyperparameters, complex parameter tuning
SVM	与分布无关；可以解决高维问题，能处理非线性特征的相互作用，泛化能力强；能处理连续和分类变量 Distribution independent; workable for high-dimensional variables; deal with continuous and category variables; good generalization capacity	黑箱；样本量多时，效率不高；非线性问题较难找到核函数；有太多的定性变量或定性变量水平太多时很难实现，难以解释；易受共线性影响 Black box; low efficiency for large sample size; difficulty in searching kernel function for nonlinear issues and implementing owing to too many category variables or levels; easily affected by collinearity
ANN	与分布无关；可处理非线性数据；能处理连续和分类变量 Distribution independent; workable for nonlinear relations, continuous and category variables	黑箱；存在过拟合风险，自变量过多时预测结果不好；易受共线性影响 Black box; overfitting risk; large prediction errors for many input variables; easily affected by collinearity
KNN	与分布无关；建模简单；可处理复杂非线性关系；能处理连续和分类变量 Distribution independent; easy; workable for nonlinear relations, continuous and category variables	黑箱；*K*值需要调试，样本不平衡时，预测误差较大；易受共线性影响 Black box; testing for *k* values; large prediction errors for imbalanced samples; affected by collinearity
MARS	与分布无关；调节参数少；受共线性和异常值影响小；可处理非线性和变量交互问题；计算效率高；能处理连续和分类变量 Distribution independent; less tuning parameters; unaffected by collinearity and outliers; workable for nonlinear relations and interactions; high computation efficiency	易受局部数据特征的影响 Easily being affected by local data features

更适合。在森林生长收获领域也进行了不同方法的比较[24,33,99−100]，但结果并不一致。如 Temesgen 和 Hoef[100] 基于 3 356 块主要树种为花旗松（*Pseudotsuga menziesii*）、西黄松和加洲铁杉（*Tsuga heterophylla*）样地的气候数据、地形数据以及空间位置数据，利用多元线性回归模型、空间线性回归模型、RF 以及 *k*-NN 模型估算了林分地上生物量，发现空间线性回归模型的预测效果优于其他几种模型；Jevšenak 和 Levanič[101] 研究年轮生长预测时，发现 ANN 方法优于传统的线性模型。传统经验生长模型中的一些问题如生长模型的相容性，如何在机器学习中进行考虑；将传统统计方法和机器学习方法相结合，如利用机器学习算法进行重要变量的筛选，以及通过机器学习算法的集成来减少模型的不确定性和误差都是未来需要解决的问题。

4.3　大尺度的森林生长收获预估

利用机器学习算法从遥感图像上反演森林参数得到了充分的发展，这也为开展大尺度的森林生长收获预估提供了方法。从遥感图像上提取信息基于机器学习方法反演林分生长收获，将成为一种趋势，对于大尺度的森林生长收获相关决策，将会起到重要的作用。如 Görgens 等[102] 对桉树蓄积量的估计和基于 Lidar 的林分因子估计。此外，利用机器学习算法建立林分生长收获变量与环境因子等的关系，也可以实现大尺度精细化的生长收获预估如立地质量和生产力[103−104]，揭示大尺度的森林生长规律[105]。

4.4　深度学习的应用

作为机器学习的一个分支，深度学习由于其较高的灵活性和预估精度近年来已经在诸多领域得到应用，尤其是图像和语音识别以及无人驾驶领域。

林业领域也开始应用，如植物物种识别[106−108]、病虫害识别[109]、森林参数遥感反演[110] 等。森林经营者需要预测经营和气候变化条件下林木的生长，随着从单木到林分到景观不同空间尺度、从天到月到年不同时间尺度的生长以及生长相关数据的积累，及深度机器学习方法的完善如模型的解释性、计算速度和代码共享等，它将会在探寻影响森林生长的生物和非生物因子以及影响因子间的非线性交互作用，实现经营和气候变化条件下森林生长精准预测方面发挥重要的作用。

参 考 文 献

[1] Weiskittel A R, Hann D W, Kershaw Jr J A, et al. Forest growth and yield modeling[M]. Chichester: John Wiley and Sons, 2011.

[2] 唐守正, 李希菲, 孟昭和. 林分生长模型研究的进展[J]. 林业科学研究, 1993, 6(6): 672−679.
Tang S Z, Li X F, Meng Z H. The development of studies on stand growth models[J]. Forest Research, 1993, 6(6): 672−679.

[3] Peng C H. Growth and yield models for uneven-aged stands: past, present and future[J]. Forest Ecology and Management, 2000, 132(2−3): 259−279.

[4] Huang S L, Ramirez C, McElhaney M, et al. F^3: simulating spatiotemporal forest change from field inventory, remote sensing, growth modeling, and management actions[J]. Forest Ecology and Management, 2018, 415−416: 26−37.

[5] Cutler D R, Edwards T C, Beard K H, et al. Random forests for classification in ecology[J]. Ecology, 2007, 88(11): 2783−2792.

[6] Wu C F, Shen H H, Shen A H, et al. Comparison of machine-learning methods for above-ground biomass estimation based on Landsat imagery[J]. Journal of Applied Remote Sensing, 2016, 10(3): 035010.

[7] Recknagel F. Applications of machine learning to ecological modelling[J]. Ecological Modelling, 2001, 146(1−3): 303−310.

[8] Liu Z L, Peng C H, Xiang W H, et al. Application of artificial neural networks in global climate change and ecological research: an overview[J]. Chinese Science Bulletin, 2010, 55(34): 3853−3863.

[9] Guan B T, Gertner G. Modeling red pine tree survival with an artificial neural network[J]. Forest Science, 1991, 37(5): 1429−1440.

[10] Guan B T, Gertner G. Using a parallel distributed processing system to model individual tree mortality[J]. Forest Science, 1991, 37(3): 871−885.

[11] 李际平, 姚东和. BP 模型在单木树高与胸径生长模拟中的应用[J]. 中南林学院学报, 1996, 16(3): 34−36.
Li J P, Yao D H. Application of BP neural network model to the simulation of breast height diameter and tree-height growth[J]. Journal of Central-South Forestry University, 1996, 16(3): 34−36.

[12] 洪伟, 吴承祯, 何东进. 基于人工神经网络的森林资源管理模型研究[J]. 自然资源学报, 1998, 13(1): 69−72.
Hong W, Wu C Z, He D J. A study on the model of forest resources management based on the artificial neural network[J]. Journal of Natural Resources, 1998, 13(1): 69−72.

[13] 浦瑞良, 宫鹏. 应用神经网络和多元回归技术预测森林产量[J]. 应用生态学报, 1999, 10(2): 129−134.
Pu R L, Gong P. Forest yield prediction with an artificial neural network and multiple regression[J]. Chinese Journal of Applied Ecology, 1999, 10(2): 129−134.

[14] 林辉, 彭长辉. 人工神经网络在森林资源管理中的应用[J]. 世界林业研究, 2002, 15(3): 22−31.
Lin H, Peng C H. Application of artificial neural network in forest resource management[J]. World Forestry Research, 2002, 15(3): 22−31.

[15] 黄家荣, 孟宪宇, 关毓秀. 马尾松人工林单木生长神经网络模型研究[J]. 山地农业生物学报, 2004, 23(5): 386−391.
Huang J R, Meng X Y, Guan Y X. The study on neural network models of individual tree growth in *Pinus massoniana* plantation[J]. Journal of Mountain Agriculture and Biology, 2004, 23(5): 386−391.

[16] Peng C H, Wen X Z. Recent applications of artificial neural networks in forest resource management: an overview[C/OL]// Environmental Decision Support Systems and Artificial Intelligence. AAAI, 1999 [2019−06−16]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.487.7652.

[17] Liu Z L, Peng C H, Work T, et al. Application of machine-learning methods in forest ecology: recent progress and future challenges[J]. Environmental Reviews, 2018, 26(4): 339−350.

[18] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016.
Zhou Z H. Machine leaning[M]. Beijing: Tsinghua University Press, 2016.

[19] Zhou Z H. Machine learning: recent progress in China and beyond[J]. National Science Review, 2018, 5(1): 20.

[20] 吴喜之. 应用回归及分类: 基于 R[M]. 北京: 中国人民大学出版社, 2016.
Wu X Z. Applied regression and classification with R[M]. Beijing: China People's University Press, 2016.

[21] Dobbertin M, Biging G S. Using the non-parametric classifier CART to model forest tree mortality[J]. Forest Science, 1998, 44(4): 507−516.

[22] Fan Z F, Kabrick J M, Shifley S R. Classification and regression tree based survival analysis in oak-dominated forests of Missouri 's Ozark highlands[J]. Canadian Journal of Forest Research, 2006, 36(7): 1740−1748.

[23] Adamec Z, Drápela K. Comparison of parametric and nonparametric methods for modeling height-diameter relationships[J]. iForest-Biogeosciences and Forestry, 2017, 10(1): 1−8.

[24] Aertsen W, Kint V, van Orshoven J, et al. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests[J]. Ecological Modelling, 2010, 221(8): 1119−1130.

[25] Räty M, Kangas A. Localizing general models with classification

and regression trees[J]. Scandinavian Journal of Forest Research, 2008, 23(5): 419−430.

[26] Piramuthu S. Input data for decision trees[J]. Expert Systems with Applications, 2008, 34(2): 1220−1226.

[27] Rejwan C, Collins N C, Brunner L J, et al. Tree regression analysis on the nesting habitat of smallmouth bass[J]. Ecology, 1999, 80(1): 341−348.

[28] Friedman J H. Multivariate adaptive regression splines[J]. Annals of Statistics, 1991, 19(1): 1−67.

[29] Prasad A M, Iverson L R, Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction[J]. Ecosystems, 2006, 9(2): 181−199.

[30] Chojnacky D C, Heath L S. Estimating down deadwood from FIA forest inventory variables in Maine[J]. Environmental Pollution, 2002, 116(Suppl.1): S25−S30.

[31] Hart S J, Laroque C P. Searching for thresholds in climate-radial growth relationships of Engelmann spruce and subalpine fir, Jasper National Park, Alberta, Canada[J]. Dendrochronologia, 2013, 31(1): 9−15.

[32] Moisen G G, Frescino T S. Comparing five modelling techniques for predicting forest characteristics[J]. Ecological Modelling, 2002, 157(2/3): 209−225.

[33] Ou Q X, Lei X D, Shen C C. Individual tree diameter growth models of larch-spruce-fir mixed forests based on machine learning algorithms[J]. Forests, 2019, 10(2): 187.

[34] Lee T S, Chiu C C, Chou Y C, et al. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines[J]. Computational Statistics and Data Analysis, 2006, 50(4): 1113−1130.

[35] Heddam S, Kisi O. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree[J]. Journal of Hydrology, 2018, 559: 499−509.

[36] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5−32.

[37] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics, 2001, 29(5): 1189−1232.

[38] Goldstein A, Kapelner A, Bleich J, et al. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation[J]. Journal of Computational and Graphical Statistics, 2015, 24(1): 44−65.

[39] Strobl C, Boulesteix A L, Kneib T, et al. Conditional variable importance for random forests[J]. BMC bioinformatics, 2008, 9: 307.

[40] Weiskitte A R, Crookston N L, Radtke P J. Linking climate, gross primary productivity, and site index across forests of the western United States[J]. Canadian Journal of Forest Research, 2011, 41(8): 1710−1721.

[41] Bond-Lamberty B, Rocha A V, Calvin K, et al. Disturbance legacies and climate jointly drive tree growth and mortality in an intensively studied boreal forest[J]. Global Change Biology, 2014, 20(1): 216−227.

[42] Kilham P, Hartebrodt C, Kändler R G. Generating tree-level harvest predictions from forest inventories with random forests[J]. Forests, 2019, 10(1): 20.

[43] 欧强新, 雷相东, 沈琛琛, 等. 基于随机森林算法的落叶松-云冷杉混交林单木胸径生长预测[J]. 北京林业大学学报, 2019, 41(9): 9−19.
Ou Q X, Lei X D, Shen C C, et al. Individual tree DBH growth prediction of larch-spruce-fir mixed forests based on random forest algorithm[J]. Journal of Beijing Forestry University, 2019, 41(9): 9−19.

[44] Nunes M H, Görgens E B. Artificial intelligence procedures for tree taper estimation within a complex vegetation mosaic in Brazil[J/OL]. PLoS One, 2016, 11(5): e0154738 [2019−10−02]. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154738.

[45] De'ath G. Boosted trees for ecological modeling and prediction[J]. Ecology, 2007, 88(1): 243−251.

[46] Freeman E A, Moisen G G, Coulston J W, et al. Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance[J]. Canadian Journal of Forest Research, 2016, 46(3): 323−339.

[47] Kuhn M, Johnson K. Applied predictive modeling[M]. New York: Springer, 2013.

[48] Elith J, Leathwick J R, Hastie T. A working guide to boosted regression trees[J]. Journal of Animal Ecology, 2008, 77(4): 802−813.

[49] Mezei P, Grodzki W, Blaženec M, et al. Host and site factors affecting tree mortality caused by the spruce bark beetle (*Ips typographus*) in mountainous conditions[J]. Forest Ecology and Management, 2014, 331: 196−207.

[50] Sproull G J, Adamus M, Bukowski M, et al. Tree and stand-level patterns and predictors of Norway spruce mortality caused by bark beetle infestation in the Tatra Mountains[J]. Forest Ecology and Management, 2015, 354: 261−271.

[51] Oguro M, Imahiro S, Saito S, et al. Relative importance of multiple scale factors to oak tree mortality due to Japanese oak wilt disease[J]. Forest Ecology and Management, 2015, 356: 173−183.

[52] Cai W H, Yang J, Liu Z H, et al. Post-fire tree recruitment of a boreal larch forest in Northeast China[J]. Forest Ecology and Management, 2013, 307: 20−29.

[53] De Cauwer V, Fichtler E, Beeckman H, et al. Predicting site productivity of the timber tree *Pterocarpus angolensis*[J]. Southern Forests: a Journal of Forest Science, 2017, 79(3): 259−268.

[54] Razakamanarivo R H, Grinand C, Razafindrakoto M A, et al. Mapping organic carbon stocks in eucalyptus plantations of the central highlands of Madagascar: a multiple regression approach[J]. Geoderma, 2011, 162(3−4): 335−346.

[55] Lin D M, Anderson-Teixeira K J, Lai J S, et al. Traits of dominant tree species predict local scale variation in forest

aboveground and topsoil carbon stocks[J]. Plant and Soil, 2016, 409(1−2): 435−446.

[56] 欧强新, 李海奎, 杨英. 福建地区马尾松生物量转换和扩展因子的影响因素[J]. 生态学报, 2017, 37(17): 5756−5764.
Ou Q X, Li H K, Yang Y. Factors affecting the biomass conversion and expansion factor of Masson pine in Fujian Province[J]. Acta Ecologica Sinica, 2017, 37(17): 5756−5764.

[57] 欧强新, 李海奎, 雷相东, 等. 基于清查数据的福建省马尾松生物量转换和扩展因子估算差异解析: 3 种集成学习决策树模型的比较[J]. 应用生态学报, 2018, 29(6): 2007−2016.
Ou Q X, Li H K, Lei X D, et al. Difference analysis in estimating biomass conversion and expansion factors of masson pine in Fujian Province, China based on national forest inventory data: a comparison of three decision tree models of ensemble learning[J]. Chinese Journal of Applied Ecology, 2018, 29(6): 2007−2016.

[58] Ren Y, Chen S S, Wei X H, et al. Disentangling the factors that contribute to variation in forest biomass increments in the mid-subtropical forests of China[J]. Journal of Forestry Research, 2016, 27(4): 919−930.

[59] Aertsen W, Kint V, De Vos B, et al. Predicting forest site productivity in temperate lowland from forest floor, soil and litterfall characteristics using boosted regression trees[J]. Plant and Soil, 2012, 354(1−2): 157−172.

[60] Mitsopoulos I, Xanthopoulos G. Effect of stand, topographic, and climatic factors on the fuel complex characteristics of Aleppo (*Pinus halepensis* Mill.) and Calabrian (*Pinus brutia* Ten.) pine forests of Greece[J]. Forest Ecology and Management, 2016, 360: 110−121.

[61] Fricker G A, Synes N W, Serra-Diaz J M, et al. More than climate? Predictors of tree canopy height vary with scale in complex terrain, Sierra Nevada, CA (USA)[J]. Forest Ecology and Management, 2019, 434: 142−153.

[62] 王星. 大数据分析: 方法与应用 [M]. 北京: 清华大学出版社, 2013.
Wang X. Big data analysis: methods and applications[M]. Beijing: Tsinghua University Press, 2013.

[63] Ciaburro G, Venkateswaran B. 神经网络: R 语言实现 [M]. 李洪成, 译. 北京: 机械工业出版社, 2018.
Ciaburro G, Venkateswaran B. Neural networks with R[M]. Li H C, trans. Beijing: China Machine Press, 2018.

[64] Ciaburro G, Venkateswaran B. Neural networks with R: smart models using CNN, RNN, deep learning, and artificial intelligence principles[M]. Birmingham: Packt Publishing, 2017.

[65] Hasenauer H, Merkl D, Weingartner M. Estimating tree mortality of Norway spruce stands with neural networks[J]. Advances in Environmental Research, 2001, 5(4): 405−414.

[66] Castro R V O, Boechat Soares C P, Leite H G, et al. Individual growth model for *Eucalyptus* stands in Brazil using artificial neural network[J/OL]. ISRN Forestry, 2013, 2013: Article ID 196832 [2019−05−18]. https://www.hindawi.com/journals/isrn/2013/196832/.

[67] Reis L P, de Souza A L, dos Reis P C M, et al. Estimation of mortality and survival of individual trees after harvesting wood using artificial neural networks in the amazon rain forest[J]. Ecological Engineering, 2018, 112: 140−147.

[68] da Rocha S J S S, Torres C M M E, Jacovine L A G, et al. Artificial neural networks: modeling tree survival and mortality in the Atlantic Forest biome in Brazil[J]. Science of the Total Environment, 2018, 645: 655−661.

[69] Bayat M, Ghorbanpour M, Zare R, et al. Application of artificial neural networks for predicting tree survival and mortality in the Hyrcanian forest of Iran[J]. Computers and Electronics in Agriculture, 2019, 164: 104929.

[70] Soares F A A M N, Flôres E L, Cabacinha C D, et al. Recursive diameter prediction and volume calculation of eucalyptus trees using multilayer perceptron networks[J]. Computers and Electronics in Agriculture, 2011, 78(1): 19−27.

[71] Ashraf M I, Zhao Z Y, Bourque C P A, et al. Integrating biophysical controls in forest growth and yield predictions with artificial intelligence technology[J]. Canadian Journal of Forest Research, 2013, 43(12): 1162−1171.

[72] Vieira G C, de Mendonça A R, da Silva G F, et al. Prognoses of diameter and height of trees of eucalyptus using artificial intelligence[J]. Science of the Total Environment, 2018, 619: 1473−1481.

[73] 马翔宇, 段文英, 崔金刚. 白桦人工林单木生长的人工神经网络模型研究[J]. 森林工程, 2009, 25(3): 30−33, 38.
Ma X Y, Duan W Y, Cui J G. Study on the artificial neural network model of individual tree growth in the *Betula platyphlla* plantation[J]. Forest Engineering, 2009, 25(3): 30−33, 38.

[74] 沈剑波, 雷相东, 李玉堂, 等. 基于 BP 神经网络的长白落叶松人工林林分平均高预测[J]. 南京林业大学学报(自然科学版), 2018, 42(2): 147−154.
Shen J B, Lei X D, Li Y T, et al. Prediction mean height for *Larix olgensis* plantation based on Bayesian-regularization BP neural network[J]. Journal of Nanjing Forestry University (Natural Sciences Edition), 2018, 42(2): 147−154.

[75] 车少辉, 张建国, 段爱国, 等. 杉木人工林胸径生长神经网络建模研究[J]. 西北农林科技大学学报(自然科学版), 2012, 40(3): 84−92.
Che S H, Zhang J G, Duan A G, et al. Modelling tree diameter growth for Chinese fir plantations with neural networks[J]. Journal of Northwest A&F University (Natural Sciences Edition), 2012, 40(3): 84−92.

[76] 龙滔, 覃连欢, 叶绍明. 基于 BP 神经网络连栽桉树人工林生长量预测[J]. 东北林业大学学报, 2012, 40(5): 122−125.
Long T, Qin L H, Ye S M. Prediction for the growth of *Eucalyptus* plantations with continuous-planting rotations based on BP neural network[J]. Journal of Northeast Forestry University, 2012, 40(5): 122−125.

[77] Reis L P, de Souza A L, Mazzei L, et al. Prognosis on the diameter of individual trees on the eastern region of the amazon using artificial neural networks[J]. Forest Ecology and

Management, 2016, 382: 161−167.

[78] Vahedi A A. Artificial neural network application in comparison with modeling allometric equations for predicting above-ground biomass in the Hyrcanian mixed-beech forests of Iran[J]. Biomass and Bioenergy, 2016, 88: 66−76.

[79] Özçelık R, Diamantopoulou M J, Eker M, et al. Artificial neural network models: an alternative approach for reliable aboveground pine tree biomass prediction[J]. Forest Science, 2017, 63(3): 291−302.

[80] Wu C Y, Chen Y F, Peng C H, et al. Modeling and estimating aboveground biomass of *Dacrydium pierrei* in China using machine learning with climate change[J]. Journal of Environmental Management, 2019, 234: 167−179.

[81] 徐奇刚, 雷相东, 国红, 等. 基于多层感知机的长白落叶松人工林林分生物量模型[J]. 北京林业大学学报, 2019, 41(5): 97−107.
Xu Q G, Lei X D, Guo H, et al. Stand biomass model of *Larix olgensis* plantations based on multi-layer perceptron networks[J]. Journal of Beijing Forestry University, 2019, 41(5): 97−107.

[82] Hlásny T, Trombik J, Bošeľa M, et al. Climatic drivers of forest productivity in Central Europe[J]. Agricultural and Forest Meteorology, 2017, 234−235: 258−273.

[83] Lima M B D O, Junior I M L, Oliveira E M, et al. Artificial neural networks in whole-stand level modeling of *Eucalyptus* plants[J]. African Journal of Agricultural Research, 2017, 12(7): 524−534.

[84] Yousefpoor M, Shahraji T R, Eslam B A, et al. The use of artificial neural network to evaluate the effects of human and physiographic factors on forest stock volume[J]. Journal of Applied Sciences and Environmental Management, 2016, 20(4): 1017−1024.

[85] 林卓, 吴承祯, 洪伟, 等. 基于 BP 神经网络和支持向量机的杉木人工林收获模型研究[J]. 北京林业大学学报, 2015, 37(1): 42−54.
Lin Z, Wu C Z, Hong W, et al. Yield model of *Cunninghamia lanceolata* plantation based on back propagation neural network and support vector machine[J]. Journal of Beijing Forestry University, 2015, 37(1): 42−54.

[86] Tavares J I D S, da Rocha J E C, Ebling Â A, et al. Artificial neural networks and linear regression reduce sample intensity to predict the commercial volume of *Eucalyptus* Clones[J]. Forests, 2019, 10(3): 268.

[87] 刘鑫, 王海燕, 雷相东, 等. 基于 BP 神经网络的天然云冷杉针阔混交林标准树高−胸径模型[J]. 林业科学研究, 2017, 30(3): 368−375.
Liu X, Wang H Y, Lei X D, et al. Generalized height-diameter model for natural mixed spruce-fir coniferous and broadleaf forests based on BP neural network[J]. Forest Research, 2017, 30(3): 368−375.

[88] Diamantopoulou M J, Özçelik R, Crecente-Campo F, et al. Estimation of Weibull function parameters for modelling tree diameter distribution using least squares and artificial neural networks methods[J]. Biosystems Engineering, 2015, 133: 33−45.

[89] 薛薇. R 语言数据挖掘方法及应用 [M]. 北京: 电子工业出版社, 2016.
Xue W. Data Mining method with R language and its application[M]. Beijing: Publishing House of Electronics Industry, 2016.

[90] Che S H, Tan X H, Xiang C W, et al. Stand basal area modelling for Chinese fir plantations using an artificial neural network model[J]. Journal of Forestry Research, 2019, 30(5): 1641−1649.

[91] Maltamo M, Kangas A. Methods based on *k*-nearest neighbor regression in the prediction of basal area diameter distribution[J]. Canadian Journal of Forest Research, 1998, 28(8): 1107−1115.

[92] Lantz B. 机器学习与 R 语言 [M]. 李洪成, 许金炜, 李舰, 译. 北京: 机械工业出版社, 2017.
Lantz B. Machine learning with R[M]. Li H C, Xu J W, Li J, trans. Beijing: China Machine Press, 2017.

[93] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction[M]. 2nd ed. New York: Springer, 2009.

[94] Lantz B. Machine Learning with R[M]. Birmingham: Packt Publishing, 2013.

[95] Ridgeway G. Generalized boosted models: a guide to the GBM package[Z/OL]. [2019−10−13]. https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

[96] Friedman J H. Stochastic gradient boosting[J]. Computational Statistics and Data Analysis, 2002, 38(4): 367−378.

[97] Thessen A E. Adoption of machine learning techniques in ecology and earth science[J/OL]. PeerJ PrePrints, 2016, 4: e1720v1 [2019−05−06]. https://peerj.com/preprints/1720.pdf.

[98] Fielding A H. Cluster and classification techniques for the biosciences[M]. London: Cambridge University Press, 2006.

[99] Corona-Núñez R O, Mendoza-Ponce A, López-Martínez R. Model selection changes the spatial heterogeneity and total potential carbon in a tropical dry forest[J]. Forest Ecology and Management, 2017, 405: 69−80.

[100] Temesgen H, Ver Hoef J M. Evaluation of the spatial linear model, random forest and gradient nearest-neighbour methods for imputing potential productivity and biomass of the Pacific Northwest forests[J]. Forestry, 2015, 88(1): 131−142.

[101] Jevšenak J, Levanič T. Should artificial neural networks replace linear models in tree ring based climate reconstructions?[J]. Dendrochronologia, 2016, 40: 102−109.

[102] Görgens E B, Montaghi A, Rodriguez L C E. A performance comparison of machine learning methods to estimate the fast-growing forest plantation yield based on laser scanning metrics[J]. Computers and Electronics in Agriculture, 2015, 116: 221−227.

[103] Wang Y H, Raulier F, Ung C H. Evaluation of spatial predictions

of site index obtained by parametric and nonparametric methods: a case study of lodgepole pine productivity[J]. Forest Ecology and Management, 2005, 214(1/3): 201−211.

[104] 高若楠, 谢阳生, 雷相东, 等. 基于随机森林模型的天然林立地生产力预测研究[J]. 中南林业科技大学学报, 2019, 39(4): 39−46.
Gao R N, Xie Y S, Lei X D, et al. Study on prediction of natural forest productivity based on random forest model[J]. Journal of Central South University of Forestry and Technology, 2019, 39(4): 39−46.

[105] Zhang H, Wang K L, Zeng Z X, et al. Large-scale patterns in forest growth rates are mainly driven by climatic variables and stand characteristics[J]. Forest Ecology and Management, 2019, 435: 120−127.

[106] Guan H Y, Yu Y T, Ji Z, et al. Deep learning-based tree classification using mobile LiDAR data[J]. Remote Sensing Letters, 2015, 6(11): 864−873.

[107] Sun Y, Liu Y, Wang G, et al. Deep learning for plant identification in natural environment[J/OL]. Computational intelligence and neuroscience, 2017, 2017: Article ID 7361042 [2019−05−18]. https://www.hindawi.com/journals/cin/2017/7361042/.

[108] Pearline S A, Kumar V S, Harini S. A study on plant recognition using conventional image processing and deep learning approaches[J]. Journal of Intelligent and Fuzzy Systems, 2019, 36(3): 1997−2004.

[109] Wang G, Sun Y, Wang J X. Automatic image-based plant disease severity estimation using deep learning[J/OL]. Computational intelligence and neuroscience, 2017, 2017: Article ID 2917536 [2019−05−16]. https://www.hindawi.com/journals/cin/2017/2917536/.

[110] Asner G P, Brodrick P G, Philipson C, et al. Mapped aboveground carbon stocks to advance forest conservation and recovery in Malaysian Borneo[J]. Biological Conservation, 2018, 217: 289−310.

（责任编辑 冯秀兰
责任编委 赵秀海）

北京林业大学提出黄河流域生态保护和高质量发展 6 项建议

12 月 6 日，北京林业大学成立发展战略咨询委员会，46 位相关学科领域的两院院士应邀担任发展战略咨询委员会委员。第一次咨询会议就学校在黄河流域生态保护和高质量发展方面着力推动和解决的重大科学技术问题进行咨询。咨询会上，北京林业大学提出了黄河流域生态保护和高质量发展 6 大方面重大科学技术问题的立项建议。

一是针对黄河流域面临的主要生态问题，完善山水林田湖草生态空间格局，构建适于黄河流域生态安全与经济发展的生态保护体系。二是加强黄河流域典型森林群落的演替规律和稳定性研究，提出典型森林资源构建、修复、改造和提效关键技术，做好黄河流域森林资源保育。三是针对黄河流域局部地区生态系统退化、水源涵养功能下降、水土流失与荒漠化严重、水体污染严重等问题，开展黄河流域生态修复与治理研究。四是聚焦黄河流域人居环境高质量发展目标，开展城镇空间格局构建和乡村生态景观营造技术研究，提升黄河流域人居景观系统质量。五是围绕流域生态系统和经济社会发展的相互作用，开展黄河流域高质量发展的治理体系支撑研究，为流域高质量发展重大问题提供咨询服务和决策参考。六是针对黄河流域生态保护、修复和治理现状，进行流域复合生态系统服务价值评估，提升黄河流域生态系统服务评估与决策能力。

委员们一致认为，北京林业大学在黄河流域的科学研究与社会服务方面有着光荣的传统和深厚的积淀，在生态保护、水土保持和生物多样性等领域取得了丰硕的成果。本次提出的立项建议充分体现了北京林业大学发挥学科特色优势，服务重大国家战略的责任担当。

（摘自北京林业大学绿色新闻网）