DOI: 10.5846/stxb201306031292

张雷,王琳琳,张旭东,刘世荣,孙鹏森,王同立.随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例.生态学报,2014,34(3):650-659.

Zhang L, Wang L L, Zhang X D, Liu S R, Sun P S, Wang T L.The basic principle of random forest and its applications in ecology: a case study of *Pinus yunnanensis*.Acta Ecologica Sinica,2014,34(3):650-659.

随机森林算法基本思想及其在生态学中的应用
——以云南松分布模拟为例

张 雷[1], 王琳琳[2], 张旭东[1], 刘世荣[3,*], 孙鹏森[3], 王同立[4]

(1. 中国林业科学研究院林业研究所,国家林业局林木培育重点实验室,北京 100091; 2. 北京林业大学林学院,北京 100083;
3. 中国林业科学研究院森林生态环境与保护研究所,国家林业局森林生态环境重点实验室,北京 100091;
4. Department of Forest Sciences, University of British Columbia, 3041-2424 Main Mall, Vancouver B.C. Canada V6T 1Z4)

摘要: 通常来讲,生态学者对于解释生态关系、描述格局和过程、进行空间或时间预测比较感兴趣。这些工作可以通过模拟输出值(响应)与一些特征值(即解释变量)的关系来实现。然而,生态数据模拟遇到了挑战,这是因为响应变量和预测变量可能是连续变量或离散变量。需要解释的生态关系通常是非线性的,并且解释变量之间具有复杂的相互作用关系。响应变量和解释变量存在缺失值并不是不常有的现象,奇异值也经常出现在生态数据中。此外,生态学者通常希望生态模型即要易于建立又易要于解释。通常是利用多种统计方法来分析处理各种各样情景中出现的独特的生态问题,这些模型包括(多元)逻辑回归、线性模型、生存模型、方差分析等等。随机森林是一个可以处理所有这些问题的有效方法。随机森林可以用来做分类、聚类、回归和生存分析、评估变量的重要性、检测数据中的奇异值、对缺失数据进行插补等。鉴于随机森林本身在算法上的优势,将就随机森林在生态学中的应用进行总结,对建模过程进行概述,并以云南松分布模拟研究为例,对其主要功能特点进行案例展示。通过对随机森林的一般术语、概念和建模思想进行介绍,有利于读者掌握本方法的应用本质,可以预见随机森林在生态学研究中将得到更多的应用和发展。

关键词: 随机森林; 分类回归树; 变量重要性; 多维数据; 物种分布模拟

The basic principle of random forest and its applications in ecology: a case study of *Pinus yunnanensis*

ZHANG Lei[1], Wang Linlin[2], ZHANG Xudong[1], LIU Shirong[3,*], SUN Pengsen[3], WANG Tongli[4]

1 *Key Laboratory of Forest Silviculture of the State Forestry Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing* 100091, *China*

2 *College of Forestry, Beijing Forestry University, Beijing* 100083, *China*

3 *Key Laboratory of Forest Ecology and Environment of State Forestry Administration, Institute of Forest Ecology, Environment and Protection, Chinese Academy of Forestry, Beijing* 100091, *China*

4 *Department of Forest Sciences, University of British Columbia*, 3041-2424 *Main Mall, Vancouver B.C. Canada V6T 1Z4*

Abstract: Ecological data are often complex. The explanatory and the response variables may be categorical variables or numerical variables. The ecological relationships that need to be defined are often nonlinear and involve high-order interactions between explanatory variables. Missing values for both response and predictor variables are very common, and outliers almost always exist. Random forest (RF), a novel machine learning technique, is ideally suited for the analysis of

基金项目: 国家自然科学基金资助项目(41301056,31290223);中央公益性院所基本科研业务专项资助项目(RIF2012-04);林业公益性行业科研专项资助项目(201104006,200804001);国家"十二五"科技支撑项目课题资助项目(2011BAD38B04)

收稿日期: 2013-06-03; **修订日期:** 2013-09-22

* **通讯作者** Corresponding author.E-mail: liusr@caf.ac.cn

complex ecological data. RF predictors are a ensemble-learning approach based on regression or classification trees. Instead of building one classification tree (classifier) , the RF algorithm builds multiple classifiers using randomly selected subsets of the observations and random subsets of the predictor variables. The predictions from the ensemble of trees are then averaged in the case of regression trees, or tallied using a voting system for classification trees. RF is efficient to support flexible modelling strategies. RF is capable of detecting and making use of more complex relationships among the variables. RF is unexcelled in accuracy among current algorithms and does not overfit. It also generates an internal unbiased estimate of the generalization error as the forest building progresses. Potential applications of RF to ecology include: classification and regression analysis, survival analysis, variable importance estimate and data proximities. Proximities can be used for clustering, detecting outliers, multi-dimensional scaling, and unsupervised classification. RF can interpolate missing value and maintain high accuracy even when a large proportion of the data are missing. RF can handle thousands of input variables without variable exclusion. It runs efficiently on large data bases. RF can also handle a spectrum of response types, including categorical, numeric, ratings, and survival data. Another advantage of the RF is that it requires only two user-defined parameters (The number of trees and the number of randomly selected predictive variables used to split the nodes) to be defined. These two parameters should be optimized in order to improve predictive accuracy. In recent years, RF has been widely used by ecologists to model complex ecological relationships because they are easy to implement and easy to interpret. To understand and use the RF, further information about how they are computed is useful. Here, we summarized the basic principle of RF and showed how RF handle complex data by modelling the geographical distribution of Yunan Pine (*Pinus yunnanensis*) in China. RF is a robust and widely used technique in the field of species distribution modelling (SDM) , since it meets the basic needs of SDM: simulating species distribution and identifying the main drivers of species distribution. In this work, RF showed a high predictive performance in simulating the distribution of Yunan Pine, which was consistent with the multi-dimensional scaling plot that showed it was possible to separate the presences from the absences. We also estimated the relative importance of predictor variables and produced the partial dependence plots for selected predictor variables for random forest predictions of the presences of Yunan Pine. The main aim of the article is to familiarize the reader with the general concepts, terminology and basic principle behind RF. We believe RF will get more applications and development in ecology.

Key Words: random forest; classification and regression tree; variable importance; multi-dimensional scaling; specie distribution modelling

经典统计模型中的回归和分类(判别) 可以写成公式形式,但是另外一些回归和分类方法体现在算法之中,其具体形式是计算机程序,这些方法广泛应用于数据挖掘之中。广义来讲算法模型包含经典模型,只是由于算法模型与经典模型发展的过程和思维方式不同而已。算法模型主要发展于最近几十年,它得益于计算机技术的不断进步。在处理海量数据集上,在对付被称为维数诅咒的巨大变量数目时,在无法假定总体分布的情况下,在面对众多竞争模型方面,算法模型较经典建模有很多不可比拟的优越性[1]。随机森林就是算法模型中的一种,是一种比较新的机器学习技术。随机森林是由 Leo Breiman 和 Cutler Adele 在 2001 年开发完成的一种数据挖掘方法,它是一种现代分类与回归技术,同时也是一种组合式的自学习技术[2]。组合学习的思路是在对新的实例进行分类的时候,把若干个单个分类器集成起来,通过对多个分类器的分类结果进行某种组合来决定最终的分类,以取得比单个分类器更好的性能。如果把单个分类器比作一个决策者的话,组合学习的方法就相当于多个决策者共同进行一项决策。

生态数据通常是多维的,变量之间关系复杂且呈非线性,并且测量变量之间有许多缺失值。传统的统计方法在分析这些数据时遭到了挑战,尤其是线性统计方法,比如广义线性模型,不足以揭示更复杂的过程透露出的格局和关系[3]。对于这些数据,

需要更灵活和稳健的分析方法,这些方法可以处理非线性关系、高阶相关性和缺失值。除了这些优势外,这些方法还必须简单易于理解并给出合理的结果解释。分类和回归树是分析复杂生态数据的理想工具。在一般分类与回归树中,当决策树的输出变量(因变量)是分类变量的时候叫分类树,当决策树的输出变量为连续变量时称为回归树。回归树不用假设经典回归中的诸如独立性、正态性、线性或者光滑性等等,无论自变量是数据变量还是定性变量都同样适用,然而它需要更多的数据来保证结果合理。分类回归树可以防止由于训练样本存在噪声和数据缺失引起的精度降低,但它也有与生俱来存在的缺点,如分类规则复杂、收敛到非全局的局部最优解和过度拟合等。随机森林具有一般分类回归树的所有优点,但又克服了其缺点。随机森林可以用来做分类、聚类、回归和生存分析。随着计算机的发展,随机森林算法的构建与实现已变得十分容易,最近几年在生态学上的应用得到了较大的发展[4]。鉴于随机森林本身在算法上的优势,本文将就随机森林在生态学中的应用进行总结概况,通过介绍其基本数学思想,以及应用案例分析,以期更有利于此方法在生态学中的广泛的应用。

1 随机森林概述

对于一个数学方法的掌握关键在于掌握问题的实质及方法的使用范围,以便在生态学研究中对方法进行正确应用并合理解释结果。

1.1 基本思想

假设读者知道单个分类树的构建,单个分类回归树的基本数学思想可以参考张雷等[5]等的相关研究。随机森林通过自助法随机选择(不是全部)向量生长成分类“树”,每个树都会完整生长而不会修剪。并且在生成树的时候,每个节点的变量都仅仅在随机选出的少数几个变量中产生。即在变量(列)的使用和数据(行)的使用上进行随机化。通过这种随机方式生成的大量的树被用于分类和回归分析,因此被称为“随机森林”。森林中每一棵树依赖于一个随机向量,森林中的所有向量都是独立同分布的。最终的决策树是通过对潜在的随机向量树进行“投票”表决生成的,即随机森林选择具有最多投票的分类。如果目的是回归,则由这些树的结果的平均得到因变量的预测值。随机森林在运算量没有显著提高的前提下提高了预测精度。分类树汇总后比任何单个个体更精确的必要充分条件是,汇总的分类树是多样化的并且比随机性个体的预测精度表现更好[6]。随机森林是统计学上对于此方法的简单称呼,正确的称呼应为随机森林分类器。

1.2 建模过程

随机森林模型具有两个非常重要的自定义参数:分类树的数量(k)和分割节点的随机变量的数量(m)。这些参数必须进行优化,以使数据处理过程中常规错误出现的次数最小。随机森林模型可以通过有放回抽样以及不同树演化过程中随机改变预测变量组合来增加分类树的多样性。每一个分类树可以通过原始数据集(X)中的一个自助法取样子集(X_i)进行生长,并且利用随机选择的 m 个预测变量中的最佳预测变量进行节点分割[7]。这种分类方法与标准的分类树方法(如:分类回归树)有一点不同,即并没有以整个预测变量中最好的分割变量进行节点分割。

包含了 k 个分类树的随机森林模型的建模过程如下:(1)当 i 从 1 变到 k 时:建立一个包含原始数据集 X 中三分之二数据量的自助法子集 X_i;以上述子集 X_i 为基础,在每个节点上随机选择 m 个预测变量,并在这些随机变量中选择一个最好的进行节点分割分类,建立一个不需修剪的深度最大的分类树。(2)通过 k 个分类树的反馈信息预测新数据。如果是分类,利用 k 个分类树组合中的多数选票,如果是回归计算平均值。由于随机森林只采用了三分之二的数据进行建模,因此不会产生模型过度拟合。

在建立一个随机森林模型的过程中,k 和 m 是非常重要的两个自定义参数,并用于树的生长。模型对预测变量的形式没有严格的要求,既可以是定量变量也可以是定性的描述;不需要把分类变量转化成设计变量或者虚变量。

在随机森林中没有必要进行交叉验证或采用独立数据建立误差无偏估计,因为随机森林在建模过程中实现了对常规误差进行无偏估计:(1)当 i 从 1 变到 k 时:以原始数据不同的自助法样本子集 X_i 建立分类树。每一子集数据中应包含原始数据集三分之二的数据元素,其它三分之一数据则被称为范围外数据。在建立第 i 个分类树的过程中,这些范围外

数据并不发挥作用。在第 i 个分类树建成后,对这些范围外数据再进行分类。(2) 在整个模型运行结束后,原始数据集中的每一个变量平均都会在三分之一 k 个分类树建立过程中成为一次范围外变量。或者可以说,在三分之一 k 个分类树的建立过程中,数据集中的每个数据都进行了一次分类处理。所有分类过程中所有范围外数据元素被错误分类的比例称为"范围外"错误。

这种范围外错误是常规错误的一个无偏估计。Breiman[2,3] 证明随机森林模型会针对常规错误形成一个极限值。范围外误差具有相当精度的前提条件是分类树必须足够多。随着分类树数量(k)的增加,这种错误会逐渐收敛。因此,在设定分类树数量时,分类树数量必须足够多,以使其满足常规错误可以逐渐收敛的要求。可以通过两个描述独立分类树计算精确度和分类树多样性的指标来推导常规误差的上限[2,3]。它们是森林中每个分类树的强度和森林中任意两个分类树间的相关性。随机森林的预测精度与单个分类树的强度和分类树之间的相关性有关[2]。具有较低错误的分类树是一个强分类器。这里的分类强度和相关性并不是自定义变量。减少分割节点的随机变量的个数(m)则会降低分类强度和相关性。降低树的分类强度则会增加整个森林的误差。而降低分类树间的相关性则可以减少森林的误差。增加 m 值则可以同时增大分类树的相关关系和分类强度,因此 m 值必须最佳化以获得最小的森林误差。同时也需要指出,随着预测变量的增加,用于最佳预测的树的个数也增加[7]。确定树个数的最好方法是比较分析利用随机森林的预测结果与利用部分森林的预测结果之间的差异,当部分森林的预测结果与所有森林的预测结果一样好的时候,说明树的选择个数最佳。

1.3　主要功能

1.3.1　变量重要性

变量重要性是一个非常难以定义的概念,因为变量之间可能具有相互作用。许多分类和回归统计模型通过统计显著性和(AIC)等指标来选择预测变量从而间接评估变量的重要性。而随机森林所采取的方法是完成不同的,对于每个分类树,范围外观测值有一个误分类比例。为了评价一个特定的预测变量的重要性,随机森林对于范围外数据这个变量的值做随机的序列改变,而所有其它变量保持不变的情况下,随后修改的范围外数据通过分类树得到新的预测值。通过分析范围外错误序列改变时范围外误差的增加情况来估计某一预测变量的重要程度。这个范围外误差的增加程度与预测变量的重要性具有一定的比例关系。修改的范围外数据和原始的范围外数据的误分类比例的差值,除以标准差,就是变量的重要性。当预测变量具有较强的解释能力,且存在多重共线性的时候,逐步选择程序和基于标准的变量选择程序将会留下 1—2 个典型的变量,剔除其余的变量,而随机森林可以把变量的重要性扩展到所有的变量中,并能识别这些重要的变量。随机森林避免了剔除重要的变量,这些变量可能在生态学上有具有重要意义,但是它们也可能与其它变量有相关性。

1.3.2　数据点相似性

数据集中任何两点之间的相似度或者接近度被定义为两个数据在同一个分类树末节点上出现次数的比例。通过建立一个 $N\times N$ 维相似矩阵(N 为数据点的个数),矩阵中的每一个元素都代表了随机森林每个树中两个相应的数据点落到同一个末节点上的比例。根据经验可以知道,相似的数据点在分支末端聚集的机率要明显高于不相似数据点的聚集机率。计算相似度的时候计算量比较大[9]。相似性可以用于缺失值填充、检测奇异值、聚类和对数据进行降维并可视化。

(1) 奇异值检测

奇异值通常定义为需要从数据总体中移除的个体。奇异值即是与其它个体相似性较小的个体。因此,类型 j 中的奇异值就是与类型 j 中所有其它个体相似性较小的个体。

(2) 数据缺失插补

随机森林有两个方式来代替缺失值,第 1 个方式较快速。如果第 m 个变量不是分类变量,随机森林计算类型 j 中这个变量所有值的中值,然后利用这个中值来代替类型 j 中第 m 个变量的所有缺失值。如果第 m 个变量是分类变量,所选择的代替值是类型 j 中非缺失值中频率最高的那个。

第 2 个替代方法计算量较大,但是处理大量的缺失值的时候比第一个方法表现要好。它仅仅插补训练数据中的缺失值。它首先对缺失值进行粗略的

不精确的修补,然后运行森林树计算相似性。如果 $x(m,n)$ 是一个缺失的数量值,通过第 n 个个体与非缺失个体相似性作为权重,求取第 m 个变量中非缺失值的平均值,以这个平均值代替缺失值。如果缺失值是分类变量,利用非缺失值中的出现频率最高的那个类型,频率由相似性作为权重。利用新的填充值进行内部森林迭代构建,发现新的填充值,再次迭代。一般的经验是 4—6 次迭代即可。

Cutler 等[9]指出随机森林具有较强的缺失值修补能力,原始值与修补值的平均值大约是一样的,修补值比真实值更集中且标准差比真实值小。这种收敛是基于回归修补程序的典型特征。当数据集中大量的数据被修补的时候,Breiman 和 Cutler[10]警告说,在随后利用这些修补值来分类的时候,范围外估计的分类正确率可能会被高估。

(3) 非监督自学习

相似性可以用来作为传统的聚类运算法则的输入量来检测多元数据的分组,但并不是所有的多元结构都采用聚类的形式。随机森林利用非监督学习来检测一般的多元结构,并且不用对数据内的聚类特征做假设。一般的步骤是: 原始数据记为分类一。相同的数据中每个变量的数值进行独立置换组成类型二。因此类型二具有独立随机变量分布特征,类型二每个变量都和原始数据中相应的变量一样都具有相同的单变量分布特征。类型二因此破坏了原始数据的非独立结构。这个两类型问题可以通过随机森林来模拟。可以使随机森林所有选项都应用到原始非标签数据中。如果这两类的范围外误分类误差,例如 40%或更多,这意味着变量对于随机森林来讲是高度的独立性。如果非独立性没有起到巨大的作用,进行分类的可能性比较低。如果误分类误差低,那么非独立性起到了重要的作用。把它描述为一个两分类的问题具有很大益处,可以高效修补缺失值,可以检测奇异值,可以测量变量重要性,可以执行定标(标度)(如果原始数据具有标签,非监督分类通常保留原始标度的结构),最大的益处应该是聚类的可能性[10]。

(4) 多维尺度分析

多维尺度分析是分析数据的统计方法之一。多维尺度分析用于反映多个研究事物间相似(不相似)程度,通过适当的降维方法,将这种相似(不相似)程度在低维度空间中用点与点之间的距离表示出来,这有可能帮助识别那些影响事物间相似性的潜在因素。随机森林是目前常用的处理高维数据的机器学习法之一。该方法无需事先指定参数的分布特征,并且可以评价每个预测变量对结局的预测能力; 同时利用内部交叉验证评价其预测错误率并能够保证有较高的准确性。因其表现突出,不断地被应用于高维数据的研究分析中。随机森林产生数据点的相似矩阵,这个相似矩阵内所有个体的值在 0—1 之间,这个值是数据点之间的距离。双变量十进制的多维尺度分析图是距离矩阵前两个主成分散点分布图。

1.3.3 偏相关图

偏相关图可以把少量变量对"黑箱"分类回归模型(包括随机森林、推进树、支持向量机、人工神经网络)预测值的影响进行作图表达。一般来讲,分类或回归函数会依赖于许多预测变量。分类或者回归函数对某一个变量(X_j)的偏依赖性可以表达为函数对其余变量的期望。在实际中,通常固定变量 X_j,并且对其它变量的所有组合的预测函数进行平均。这个过程需要对训练数据集中变量 X_j 的每个值根据所有数据进行预测。而在随机森林执行偏依赖图分析中没有应用训练数据集中变量 X_j 的值,而是应用训练数据集中变量 X_j 变程范围内等距离的分段数据,分析时需要指定有多少个分段。这个特征对于变量 X_j 中数值较大非常有用。对两个变量的偏依赖图定义为函数对这两个变量之外的其余变量的条件期望,对两个变量的偏依赖图是三维透视图。

1.3.4 生存分析

当响应变量是生存时间或失效时间,有或者没有核查的时候,随机森林可以为每个明确的预测变量组合计算完全的非参数生存曲线。生存分析是随机森林的一个扩展,通常称为随机生存森林,适用于右截尾的生存资料[11]。随机生存森林极大的仿制了随机森林,继承随机森林的优点,同时又在传统的生存分析方法里有新的突破。其中有两点需要强调: 一是有 3 个参数需要设置(分类树的数量、分割节点的随机变量的数量和分割规则); 二是高度的数据驱动型模型,不需要模型假设。最后一个特征对于生存分析特别有利。在传统的生存分析方法中,常用到的方法都依赖很强的限制性假设,如比例风

险率函数的假定。利用这些方法总是需要考虑预测变量和风险之间的相关性是否能合理的模拟,预测变量间的非线性影响或高阶次相互作用能否包含其中。然而这些问题在随机森林内可以自动处理。

1.4　总体优点

随机森林算法的主要优点[10,12]:(1) 当前算法中随机森林的精度是无可比拟的; (2) 可以有效处理大数据集,可以处理没有删减的成千上万的输入变量,即使预测变量数目极大超过观测值数据也同样有效; (3) 在分类过程中给出变量的重要性估计; (4) 在森林建立过程中内部可以产生一个对一般误差的无偏估计,不会产生过度估计; (5) 可以有效处理缺失数据,即使在数据大量缺失的情况也可以维持较高的精度; (6) 可以配平分类总体不平衡数据集的误差; (7) 保存产生的森林并用于未来其它数据预测; (8) 技术原型的计算可以给出变量之间相关性和分类信息; (9) 可以通过计算两两实例之间的相关性用于聚类、奇异点定位或者给出数据集的有意诠释查看; (10) 以上各种功能可以扩展到无标签数据中,进行非监督分类、数据查看和奇异点检测; (11) 提供一个检测变量相互作用的实验方法,对多元共线性不敏感。

2　随机森林在生态学中的应用

2.1　案例分析: 以云南松分布模拟预测为例

物种分布模拟研究是随机森林应用的一个典型领域,基于相关关系的静态模型越来越多的被用于预测、评估环境变化对物种分布的影响[13]。物种分布模型中经常遇到的问题就是海量的数据和大量的预测变量。物种分布模型具有两个目的: 模拟分析物种分布和识别影响物种分布的主导因子。以云南松(*Pinus yunnanensis*)分布模拟研究为例,介绍随机森林本身所具有的特点,将更有利于掌握问题的实质和应用范围。

2.1.1　数据来源

物种分布数据: 云南松地理分布数据从 1∶100 万中国植被图集[14]中提取。本研究在空间分辨率为 8 km 的尺度上开展分析。

预测变量: 本研究采用 17 个预测变量,其中 9 个是气候变量: 年平均温度(MAT, ℃),平均最冷月温度(MCMT,℃),平均最暖月温度(MWMT,℃),气温年较差(MWMT-MCMT)(TD, ℃),大于 5 ℃ 的积温(DD, ℃),平均年降水量(MAP, mm),平均夏季降雨(5—9 月,MSP, mm),年湿热指数((MAT+10)/(MAP/1000))(AHM),夏季湿热指数((MWMT)/(MSP/1000))(SHM)。这 9 个气候变量均是 1961—1990 年连续 30a 数据的平均值。模型所采用的 9 个气候变量是从 ClimateChina[13] 软件中输出。同时也采用了 7 个连续的土壤变量(土壤有机质含量(%),土壤粗砂,细沙,粉砂和粘粒含量(%),土层厚度(cm),pH 值)和 1 个土壤类型变量。土壤数据从 1∶100 万土壤图中提取(本数据以及上文中的 1∶100万植被图都来源于国家自然科学基金委员"中国西部环境与生态科学数据中心"和"地球系统科学数据共享网")

2.1.2　分布模拟预测

利用随机森林建立云南松地理分布预测模型。由于建模过程同时需要物种存在和不存在数据,采用张雷等[13]的方法,把所有没有云南松出现记录的地点与已知地点的环境条件(即 17 个环境变量)进行对比,如果未知分布点中有与已知点环境条件相同的地区(即未知分布点的环境变量完全包含在所有已知地点 17 个环境条件变程范围内),那么这些地区可以认为适合云南松分布,把这些地区从未知点中删除,剩余的未知点可以认为是"完全"不适宜云南松分布的地区,这些地区指定为不存在区,并把其作为模型建模数据。

2.1.3　输出结果

当利用分类变量(存在、不存在)作为响应变量建立模型时,云南松地理分布模拟预测结果是二元分布图(图 1,a);当利用数值变量(0 代表不存在,1 代表存在)作为响应变量建立模型时,预测结果是云南松概率分布图(图 1,b)。随机森林具有较高的预测结果,其中分类分析的预测误差是 0.058; 回归分析的预测误差是 0.044。其预测结果精度较高在在多维度降维图中也有表现,在由前两个主成分维数组成的分布图中,存在值和不存在值较容易分开(图 2)。

变量重要性可以帮助理解那个变量主要影响物种分布。随机森林给出的预测变量重要性估计表明,影响云南松分布的主导因素是热量条件,其次是水分条件,影响因素最小的是土壤条件(图 3)。热

量条件中重要性最大的是气温年较差，水分条件中最大的是平均年降水量，土壤条件中主导因子是有机质含量。变量重要性可以比较那些变量的影响力是相似的。如，年平均温度和积温的影响力是相似的(图 3) 。偏依赖图表明平均最冷月温度与云南省出现概率呈非线性关系，在平均最冷月温度大于 -7℃后，随着平均最冷月温度的增加云南省出现概率增加; 而气温年较差与此相反，随着气温年较差的增加云南松出现概率急剧下降，在气温年较差升高到 18℃时趋于稳定。

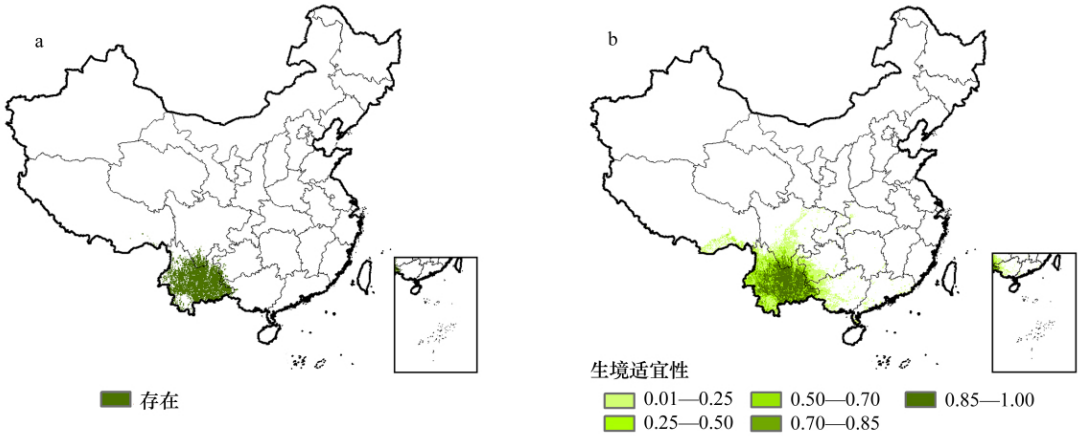

图 1 云南松分布模拟预测图 (a) 二元(存在/不存在) 分布图，(b) 概率分布图 [5]

Fig.1 Geographical distribution of *Pinus yunnanensis* predicted by random forests (a) Binary (presence-absence) distribution map; (b) Probability distribution map [5]

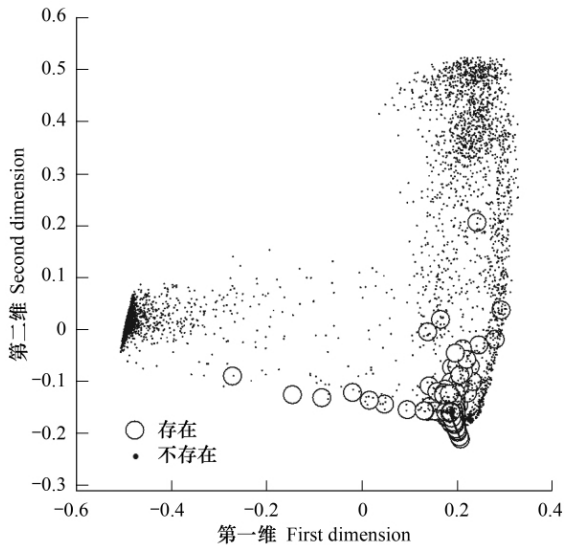

图 2 云南松多维预测变量降维图

Fig. 2 Random forest-based multi-dimensional scaling plot of presence vs. absence for *Pinus yunnanensis*

2.2 在生态学研究中的应用

随机森林在生态学中的应用发展迅速，目前与分类和回归相关的生态学研究中都有应用案例。基于组合方法的随机森林在预测物种分布时比其它模型的预测表现要好。如张雷等 [5,13] 比较分析了基于组合方法的物种分布模型(其中包括随机森林) 与其它常规模型(广义线性模型、广义加法模型和分类回归树) 在模拟预测物种分布时的差异，发现基于组合方法的随机森林预测精度高于其它常规模型。Peters 等 [8] 成功利用随机森林预测了比利时峡谷地区依赖地下水生长的植被物种的分布情况，并采用 4 种模型精度评估方法比较了随机森林与多元逻辑回归模型在预测植被类型分布时的精度，虽然两种方法的预测精度都较高，但随机森林预测精度显著高于后者。同样地，Cutler 等 [9] 采用 10 种模型精度评估方法比较分析了随机森林与线性判别分析、逻辑回归模型和分类回归树在物种分布模拟研究中的预测精度，发现随机森林的预测精度普遍高于其它 3 个方法。

随机森林在景观分类中也有一定的应用，如植被分类、土地利用分类。由于随机森林只利用了一个随机子集进行节点分割，并且树完全生长不需修建，因此它的训练速度较快。Pal [15] 比较分析了随机森林与支持向量机在遥感土地利用分类过程中的分类精度和模型训练时间等方面的差异，发现随机森林预测精度稍高于支持向量机，模型训练时间稍少于后者。Gislason 等 [16] 把随机森林用于多元遥感数据和地理数据分类，并与分类回归树、助推法和自助

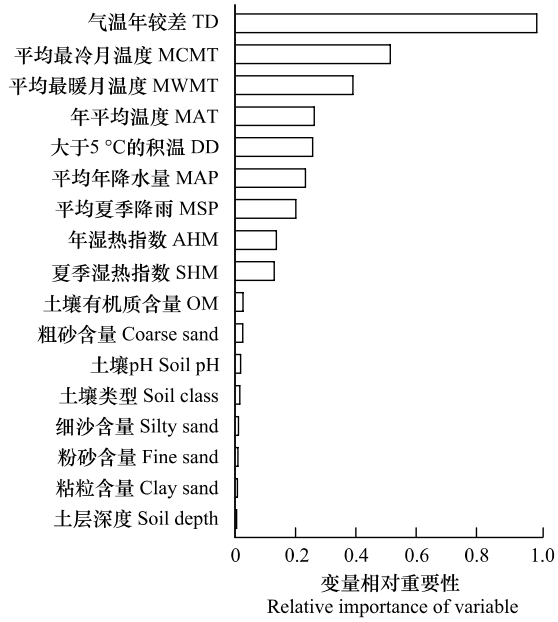

图 3　预测变量相对重要性

Fig.3　Variable importance plot for predictor variables

TD: Temperature difference between mean warmest month temperature and mean coldest month temperature; MCMT: Mean coldest month temperature; MWMT: Mean warmest month temperature; MAT: Mean annual temperature; DD: Growing degree-days; MAP: Mean annual precipitation; MSP: Mean annual summer precipitation; AHM: Annual heat: moisture index; SHM: Summer heat: moisture index; OM: Organic matter

整合法的分类精度进行了比较,随机森林分类精度显著高于分类回归树,与助推法和自助整合法的分类精度相当,但是随机森林的训练速度要快于后两者,尤其是快于助推法。Chan 等[17] 把随机森林应用于高光谱数据生态区制图,发现随机森林和自适应助推法的预测精度几乎一样优异,并且两者的预测精度都高于神经网络分类器,两者在处理高光谱数据时都同样非常有效,但是随机森林的训练速率更快且更稳定。

随机森林在回归分析研究中的应用较少,但也有一些报道。如 Iverson 等[4] 利用随机森林预测了美国东部 134 个树种的丰富度。Prasad 等[18] 利用随机森林模拟了美国东部 4 个树种重要值(表征基面积和丰富度)的空间分布,采用 4 个模型精度评估方法比较分析了随机森林与分类树、多元自适应样条平滑函数和自助整合法在预测 4 个树种分布时的预测精度,发现随机森林多数情况下预测精度都高于其它 3 个模型。

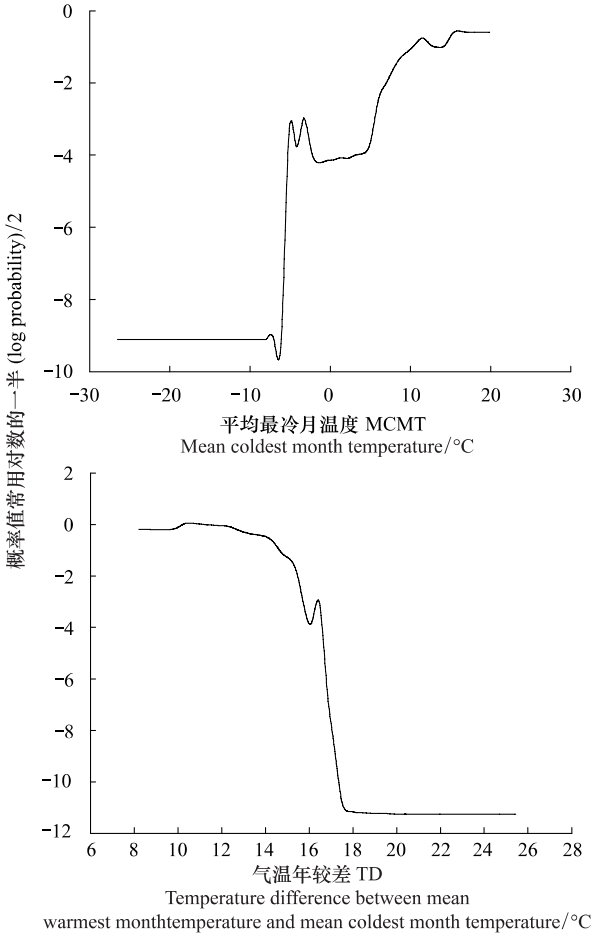

图 4　变量平均最冷月温度(MCMT)和气温年较差(TD)的偏依赖图

Fig.4　Partial dependence plots for selected predictor variables (mean coldest month temperature, MCMT; temperature difference between mean warmest month temperature and MCMT, TD) **for random forest predictions of the presences of *Pinus yunnanensis***

3　小结

综上所述,随机森林是一个非常快捷有效的机器学习方法,是一个组合学习分类器算法,在分类和回归中都有重要的应用。随机森林不是传统的统计推断工具,不适合进行方差分析或者假设检验,也不计算 p 值或回归系数,或置信区间[9]。由于不能检测单个分类树,随机森林被认为是“黑箱”模型。但是它又提供了其它的优点来协助解释,比如变量的重要性,因此这个方法比人工神经网络更易于解释,Prasad 等[18] 认为最好把随机森林定义为“灰箱”模型。对于黑箱/灰箱分类器,可以一次构建两个变量或一个变量的偏依赖图[19]。如果分类函数是由单个变量主导的并且具有低阶次的交互作用,这些图

形将会是分类结果可视化的一个有效工具,但是它们可能对于表示或者解释高阶次的交互作用没有帮助。与传统的分类算法相比,随机森林具有更高的准确性和稳健性等优点,所以近 10 年来,随机森林的理论和方法在许多学科领域都得到了较大的发展。

近几年其它来自机器学习领域的分类程序包括助推法、支持向量机和人工神经网络等。所有这些方法和随机森林一样,都是高精度的分类器,可以做分类和回归分析。同样地,和这些程序一样,由随机森林产生的预测变量和响应变量之间的关系并没有简单的表达方式,像诸如公式(逻辑回归)或者象形图(分类树)一样的表示方法,来表示分类函数,这样的缺陷导致生态上的解释是困难的。但随机森林具有两个区别于这些方法的关键特征:一是变量重要性估计,它克服了传统的变量选择方法(如,在一组变量同样好具有高度相关性的组中选择一个或者两个变量)的缺陷。变量重要性主观上可以用来识别生态上重要的影响因子并进行解释,但是不能像变量子集选择方法一样可以自动选择变量子集。如,在入侵物种的研究案例中,随机森林确定的最重要的影响入侵种的变量与以往文献中所指出的一样[9]。二是随机森林可以执行数据排列分析,包括数据相似性分析。数据点之间相似性测量由随机森林自动产生。相似性可以用于缺失值修补,作为传统基于距离和协方差矩阵的多变量模型(例如聚类分析和多维尺度缩放)的输入值,方便分类结果实现作图。

随机森林的潜在应用范围是宽广的。如:(1)进行景观分类,如植被类型或土地类型分类,利用遥感数据对森林类型进行制图,预测大地理范围内森林特征;(2)筛选野生动植物的适宜生境。可以为常见的或稀有物种或入侵种识别特定的栖息地或概率表面;(3)模拟预测环境变化对物种分布的影响。随机森林足够稳健,可以处理物种当前分布区外的环境变化数据;模型具有自动选择机制,可以根据不同的物种选择最佳的输入参数,以使对单个物种内部参数调整花费最小。重要的是,随机森林可以在景观内通过取样位置的预测变量进行外推,理解那些变量是重要的驱动力,并且对变量重要性的认识比其它方法更具可靠性。(4)对具有大量相互作用的复杂生态数据进行数据挖掘。与许多传统分析方法不同,随机森林没有预测变量或者响应变量的分布假设,并且可以处理预测变量数目极大超过观测值数目这种情况[12],因此随机森林为传统的参数和半参数的生态数据分析统计方法提供了其它可选项。因此强烈推荐在预测性的生态模拟研究中应用此方法。本文的目的在于介绍随机森林的基本思想,以引起更多学者关注随机森林在解决生态问题中的价值。

References:

[1] Wu X Z. Statisics: From concepts to data analysis. Higher education Press, Bejing, 2008.

[2] Breiman L. Random forests. Machine learning, 2001, 45(1): 5-32.

[3] De´ath G, Fabricius K E. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 2000, 81(11): 3178-92.

[4] Iverson L R, Prasad A M, Matthews S N, Peters M. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. Forest Ecology and Management, 2008, 254(3): 390-406.

[5] Zhang L, Liu S R, Sun P S, Wang T L. Partitioning and mapping the sources of variations in the ensemble forecasting of species distribution under climate change: a case study of *Pinus tabulaeformis*. Acta Ecologica Sinica, 2011, 31(19): 5749-5761.

[6] Hansen L K, Salamon P. Neural network ensembles. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1990, 12(10): 993-1001.

[7] Liaw A, Wiener M. Classification and Regression by randomForest. R news, 2002, 2(3): 18-22.

[8] Peters J, Baets B D, Verhoest N E C, Samson R, Degroeve S, Becker P D, Huybrechts W. Random forests as a tool for ecohydrological distribution modelling. Ecological Modeling, 2007, 207(2/4): 304-18.

[9] Cutler D R, Jr. T C E, Beard K H, Cutler A, Hess K T, Gibson J, Lawler J J. Random forests for classification in ecology. Ecology, 2007, 88(11): 2783-92.

[10] Breiman L, Cutler A. Random Forests. 2004. http://www.math.usu.edu/—adele/forests/ [Cited 20 May, 2013]

[11] Ishwaran H, Kogalur U B. Random Survival Forests for R. R News, 2007, 7(2): 25-31.

[12] Breiman L. Statistical modeling: The two cultures. Statistical Science, 2001, 16(3): 199-231.

[13] Zhang L, Liu S R, Sun P S, Wang T L. Comparative evaluation of multiple models of the effects of climate change on the potential distribution of *Pinus massoniana*. Chinese Journal of Plant

Ecology, 2011, 35 (11): 1091-1105.

[14] Compiling committee of vegetation maps of 1:1000000 in China. Atlas of vegetation maps of 1:1000000 in China. 2001, Science Press, Beijing.

[15] Pai M. Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 2005, 26(1): 217-22.

[16] Gislason P O, Benediktsson J A, Sveinsson J R. Random Forests for land cover classification. Pattern Recognition Letters, 2005, 27(4): 294-300.

[17] Chan J C W, Paelinckx D. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment, 2008, 112(6): 2999-3011.

[18] Prasad A M, Iverson L R, Liaw A. Newer classification and regression tree techniques: Bagging and Random forests for ecological rrediction. Ecosystems, 2006, 9(2): 181-99.

[19] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer New York. 2001.

参考文献:

[1] 吴喜之. 统计学: 从概念到数据分析. 北京: 高等教育出版社, 2008.

[5] 张雷, 刘世荣, 孙鹏森, 王同立. 气候变化对物种分布影响模拟中的不确定性组分分割与制图——以油松为例. 生态学报, 2011, 31(19): 5749-61.

[13] 张雷, 刘世荣, 孙鹏森, 王同立. 气候变化对马尾松潜在分布影响预估的多模型比较. 植物生态学报, 2011, 35(11): 1091-105.

[14] 中国科学院中国 1:100 万植被图编辑委员会. 1:100 万中国植被图集. 北京: 科学出版社, 2001.