应用昆虫学报 *Chinese Journal of Applied Entomology* 2013,50(4):1190 - 1197. DOI:10.7679/j.issn.2095 - 1353.2013.163

随机森林模型在分类与回归分析中的应用*

李欣海**

(中国科学院动物研究所 北京 100101)

摘 要 随机森林(random forest)模型是由 Breiman 和 Cutler 在 2001 年提出的一种基于分类树的算法。它通过对大量分类树的汇总提高了模型的预测精度,是取代神经网络等传统机器学习方法的新的模型。随机森林的运算速度很快,在处理大数据时表现优异。随机森林不需要顾虑一般回归分析面临的多元共线性的问题,不用做变量选择。现有的随机森林软件包给出了所有变量的重要性。另外,随机森林便于计算变量的非线性作用,而且可以体现变量间的交互作用(interaction)。它对离群值也不敏感。本文通过 3 个案例,分别介绍了随机森林在昆虫种类的判别分析、有无数据的分析(取代逻辑斯蒂回归)和回归分析上的应用。案例的数据格式和 R 语言代码可为研究随机森林在分类与回归分析中的应用提供参考。

关键词 随机森林, 分类树, 判别分析, 回归, 机器学习

Using "random forest" for classification and regression

LI Xin-Hai**

(Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China)

Abstract "Random forest" is an algorithm developed by Breiman and Cutler in 2001. It runs by constructing multiple decision trees while training and outputing the class that is the mode of the classes output by individual trees. It has improved performance over single decision trees, and it is much more efficient than traditional machine learning techniques, e. g. artificial neural networks, especially when the dataset is large. Random forest can handle up to thousands of explanatory variables. It can be used to rank the importance of variables when the R package "random. forest" is implemented. It is suitable for demonstrating the nonlinear effect of variables, and it can model complex interactions among variables. Random forest is robust for outliers. In this paper, three examples are used to introduce how to use random forest for a discrimination problem (the dependent variable has multiple categories) for presence-absence data (the deperdent variable has two categories), and for regression(the dependent variable is a continuous variable).

Key words random forest, classification tree, discriminant analysis, regression, machine learning

随机森林(random forest)是一种基于分类树(classification tree)的算法(Breiman,2001)。这个算法需要模拟和迭代,被归类为机器学习中的一种方法。经典的机器学习模型是神经网络(Hopfield,1982),有半个多世纪的历史了。神经网络预测精确,但是计算量很大。20 世纪 80 年代 Breiman 等人(1984)发明了分类和回归树(classification and regression tree,简称 CART)的算法,通过反复二分数据进行分类或回归,计算量大大降低。2001 年 Breiman 和 Cutler 借鉴贝尔实验室的 Ho 所提出的随机决策森林(random decision forests)(Ho,1995,1998)的方法,把分类树组合成随机森林(Breiman,2001a),即在变量(列)的使用和数据(行)的使用上进行随机化,生成很多分类树,再汇总分类树的结果。后来 Breiman 在机器学习杂志上发表了他和 Cutler 设计的随机森林的算法(Breiman,2001a)。这篇文章被大量引用(根据 Google Scholar,该文章至 2013 年被引用 9 000 多次),成为机器学习领域的一个里程碑。

随机森林在运算量没有显著提高的前提下提

* 资助项目:中国科学院战略性先导科技专项(XDA05080701)和环保部公益项目(201209027)。

**通讯作者,E-mail:lixh@ioz.ac.cn

收稿日期:2013-06-09,接受日期:2013-06-26

高了预测精度。随机森林对多元共线性不敏感,结果对缺失数据和非平衡的数据比较稳健,可以很好地预测多达几千个解释变量的作用(Breiman,2001b),被誉为当前最好的算法之一(Iverson *et al.*,2008)。在机器学习的诸多算法中,随机森林因高效而准确而备受关注,在各行各业得到越来越多的应用(Cutler *et al.*,2007;Genuer *et al.*,2010)。

随机森林的算法最初以 FORTRUN 语言编码(Liaw,2012)。现在可以通过 R 语言或 SAS 等工具实现。R 语言是一种用于统计分析和绘图的语言和操作环境(R Development Core Team,2013)。它是自由、免费、源代码开放的软件,近年来已经成为国际学术领域应用最广的统计工具。在国内,R 语言也在迅速普及。本文基于 R 语言介绍随机森林的应用。R 语言中有两个软件包可以运行随机森林,分别是 randomForest(Liaw,2012)和 party。本文介绍 randomForest 的用法。

本文面向没有或只有初步 R 语言基础的生态学工作者,以 3 个案例,通过运行案例中给出的 R 语言代码,读者可以运行随机森林的算法,进行分类或回归分析,得到变量的重要性、模型的误差等指标,并可以进行预测。Breiman 发表随机森林后,有若干文章深入探讨其算法(Biau,2012),变量的比较(Archer and Kirnes,2008;Groemping,2009)和变量间的交互作用(Winham *et al.*,2012)等。本文旨在介绍随机森林的应用方法,不涉及其本身的算法,也不涉及同其他平行方法的比较。

1　随机森林的原理

同其他模型一样,随机森林可以解释若干自变量($X1$、$X2$、…、Xk)对因变量 Y 的作用。如果因变量 Y 有 n 个观测值,有 k 个自变量与之相关;在构建分类树的时候,随机森林会随机地在原数据中重新选择 n 个观测值,其中有的观测值被选择多次,有的没有被选到,这是 Bootstrap 重新抽样的方法。同时,随机森林随机地从 k 个自变量选择部分变量进行分类树节点的确定。这样,每次构建的分类树都可能不一样。一般情况下,随机森林随机地生成几百个至几千个分类树,然后选择重复程度最高的树作为最终结果(Breiman,2001a)。

2　随机森林的应用

随机森林可以用于分类和回归。当因变量 Y 是分类变量时,是分类;当因变量 Y 是连续变量时,是回归。自变量 X 可以是多个连续变量和多个分类变量的混合。在下面 3 个案例中,判别分析和对有无数据的分析是分类问题,对连续变量 Y 的解释是回归问题。

2.1　在判别分析中的应用

别判分析(discriminant analysis)是在因变量 Y 的几个分类水平明确的条件下,根据若干自变量判别每个观测值的类型归属问题的一种多变量统计分析方法。判别与分类在统计学概念上有所交叉,在本文中不强调两者的区别。案例 1 中有 3 种昆虫(A、B 和 C)形态接近,不过可以通过 4 个长度指标(L1、L2、L3 和 L4)进行种类的识别。具体数据如表 1。

表 1　3 种昆虫及其用于分类的 4 个量度指标

Table 1　The four length indices for classifying three insect species

物种 Species	量度 Length			
	L1	L2	L3	L4
A	16	27	31	33
A	15	23	30	30
A	16	27	27	26
A	18	20	25	23
A	15	15	31	32
A	15	32	32	15
A	12	15	16	31
B	8	23	23	11
B	7	24	25	12
B	6	25	23	10
B	8	45	24	15
B	9	28	15	12
B	5	32	31	11
C	22	23	12	42
C	25	25	14	60
C	34	25	16	52
C	30	23	21	54
C	25	20	11	55
C	30	23	21	54
C	25	20	11	55

通过运行下列 R 语言代码,可以得到随机森

林的结果 RF1。R 语言中的“#”表示注释,其后面的语句不被执行。当随机森林用于分类时,其结果 RF1 包含混淆矩阵(confusion matrix)(表 2),显示判别分析的错误率。

in stall. packages(“randomForest”) #安装随机森林程序包(每台计算机只需安装一次)

library(randomForest) #调用随机森林程序包(每次运行都要调用)

insect < - read. csv(" d:/data/insects. csv", header = TRUE) #从硬盘读入数据到对象 insect

RF1 < - randomForest (insect [, c (‘ L1 ’, ‘ L2 ’, ‘ L3 ’, ‘ L4 ’)], insect [, ‘ species ’], importance = TRUE, ntree = 10000) #运行随机森林模型

RF1 #显示模型结果,包括误差率和混淆矩阵(表 2)

其中 insect 是一个包含 5 个变量 20 个记录的数据表。insect[,c(‘L1’,‘L2’,‘L3’,‘L4’)]表示昆虫的量度,是一个 4 乘以 20 的矩阵;insect[,‘species’]表示昆虫的物种类别,是 20 个物种名组成的一个向量。表 2 显示模型对 A 的判别错误率为 28.6% ,对 B 和 C 的判别错误率为 0。

表 2 随机森林(用于分类时)的混淆矩阵显示昆虫分类误差

Table 2 Random forest outputs a confusion matrix showing the classification error

	A	B	C	分类误差 Class error
A	5	2	0	0.286
B	0	6	0	0
C	0	0	7	0

注:每行表示实际的类别,每列表示随机森林判定的类别。The row indicates real classification; the column indicates predicted classification.

随机森林的结果内含判别函数,可以用下列代码根据新的量度判断昆虫的物种类别。

new. data < - data. frame(L1 = 20, L2 = 50, L3 = 30, L4 = 20) #一个新的昆虫的量度

predict(RF1, new. data, type = “prob”) #判别该量度的昆虫归类为 A、B 和 C 的概率

predict(RF1, new. data, type = “response”) #判别该量度的昆虫的类别

该案例中,该量度判别为 A、B 和 C 的概率分别为 82.4% 、9.4% 和 8.2% 。随机森林将其判别为 A。

2.2 对有无数据的分析

对于有或无、生或死、发生或不发生等二分变量的分析,一般用逻辑斯蒂回归(logistic regression)的方法。逻辑斯蒂回归实质上是对因变量 *Y* 作两个分类水平的判别。逻辑斯蒂回归对自变量的多元共线性非常敏感,要求自变量之间相互独立。随机森林则完全不需要这个前提条件。Breiman 在 2001 年发表了具有革命意义的文章,批判了当前主流的统计学方法,指出经典模型如逻辑斯蒂回归经常给出不可靠的结论,而随机森林准确而可靠。

案例 2 以朱鹮为例,说明该方法的具体应用。朱鹮的巢址选择受环境变量的影响(Li *et al.*, 2006,2009;翟天庆和李欣海,2012)。假设朱鹮选择一个地方营巢的概率取决于下列自变量:土地利用类型(森林、草地、灌丛或农田等)、海拔、坡度、温度、降水、人类干扰指数等。该问题的因变量为朱鹮 1981 年至 2008 年间的 532 个巢(Y = 1),以及在朱鹮巢区的系统选择的(等间距)2 538 个点(Y = 0)(图 3:A);自变量为这 3 070 个地点对应的 8 个环境变量。应用随机森林对朱鹮巢址选择进行分析的 R 语言代码如下:

ib is < - read. csv(́d:/data/ibis. csv ́, header = TRUE) #从硬盘读入数据

ibis $ use < - as. factor(ibis $ use) # 定义巢址选择与否(0 或 1)为分类变量。这是因变量 Y。

ibis $ landcover < - as. factor (ibis $ landcover) #定义土地利用类型为分类变量

RF2 < - randomForest(ibis[, c(‘ elevation ’, ‘ footprint ’, ‘ GDP ’, ‘ landcover ’, ‘ pop ’, ‘ slope’, ‘prec_ann’, ‘t_ann’)], ibis[,‘use’], importance = TRUE, ntree = 1000) #运行随机森林

varImpPlot(RF2) #图示自变量对的巢址选择的重要性

从图 1 可以看到不同指标指示的变量重要性会略有差距,但是差距不会很大。

随机森林可以给出每个自变量对因变量的作用。下列 R 代码给出海拔对巢址选择的影响,结

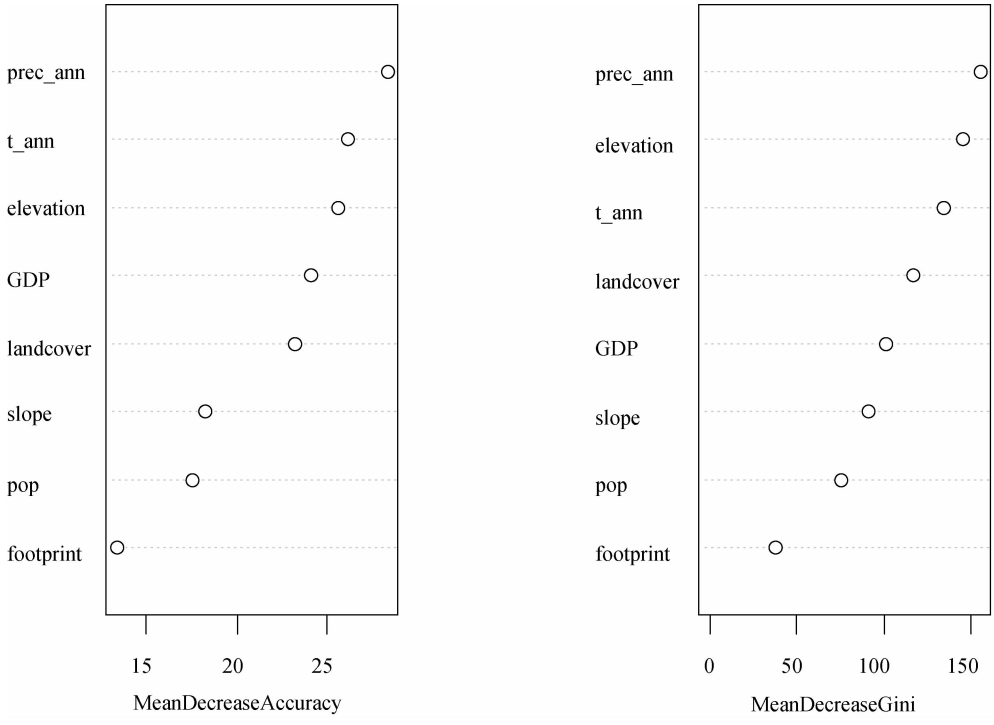

图 1　随机森林对影响朱鹮巢址选择的自变量的重要性进行排序*

Fig. 1　Ranking variable importance that associated with nest site selection of the crested ibis by random forest*

* MeanDecreaseAccuracy 衡量把一个变量的取值变为随机数,随机森林预测准确性的降低程度。该值越大表示该变量的重要性越大(Liaw,2012)。MeanDecreaseGini 通过基尼(Gini)指数计算每个变量对分类树每个节点上观测值的异质性的影响,从而比较变量的重要性。该值越大表示该变量的重要性越大。prec_ann 是年总降水量; t_ann 是年平均温度;elevation 是海拔;GDP 是国内生产总值;landcover 是土地利用类型;slope 是坡度;pop 是人口密度;footprint 是人类干扰指数。

果在图 2 中,表示中等程度的海拔最适宜营巢。

```
prtialPlot(RF2, ibis, elevation, "0", main = '', xlab = 'Elevation (m)', ylab = "Variable effect")
```

随机森林可以通过下列代码预测任何地点朱鹮营巢的概率(图 3)

```
pred <- predict(RF2, ibis, type = "prob") #计算原数据 ibis 中 3 070 个地点被朱鹮选择营巢的概率
# 绘制图 3(A)
plot(ibis $ x, ibis $ y, type = "n", xlab = '经度 Longitude', ylab = '纬度 Latitude') #绘制坐标轴
for (i in 1:length(ibis $ x)){ #循环语句,从 1 到 3 070
    if(ibis $ use[i] ! = 1) points(ibis $ x[i], ibis $ y[i], col = "grey80", cex = .8, pch = 19) #非营巢点为灰色
    if(ibis $ use[i] = = 1) points(ibis $ x[i], ibis $ y[i], col = "black", cex = .8, pch = 19) #营巢点为黑色
}
#绘制图 3B,颜色深的营巢概率高
plot(ibis $ x, ibis $ y, type = "n", xlab = '经度 Longitude', ylab = '纬度 Latitude') #绘制坐标轴
for (i in 1:length(ibis $ x)){ #循环语句,从 1 到 3 070
    # 根据每个点朱鹮营巢的概率显示该点的颜色深度
    pints(ibis $ x[i], ibis $ y[i], col = gray.colors(10)[round(pred[i] * 10)], cex = .8, pch = 19)
}
```

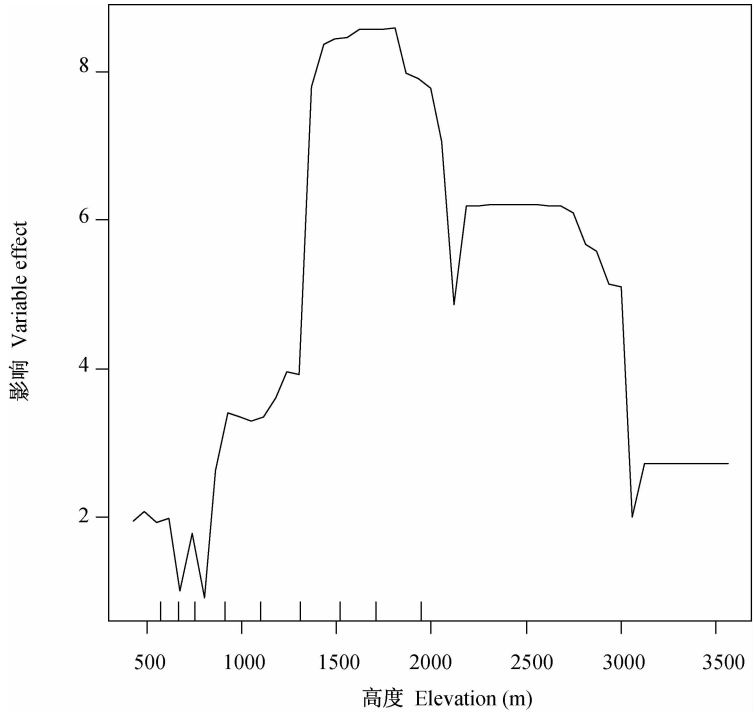

图 2 随机森林算出的海拔对朱鹮巢址选择的影响

Fig. 2 Partial effect of elevation on nest site selection of the crested ibis

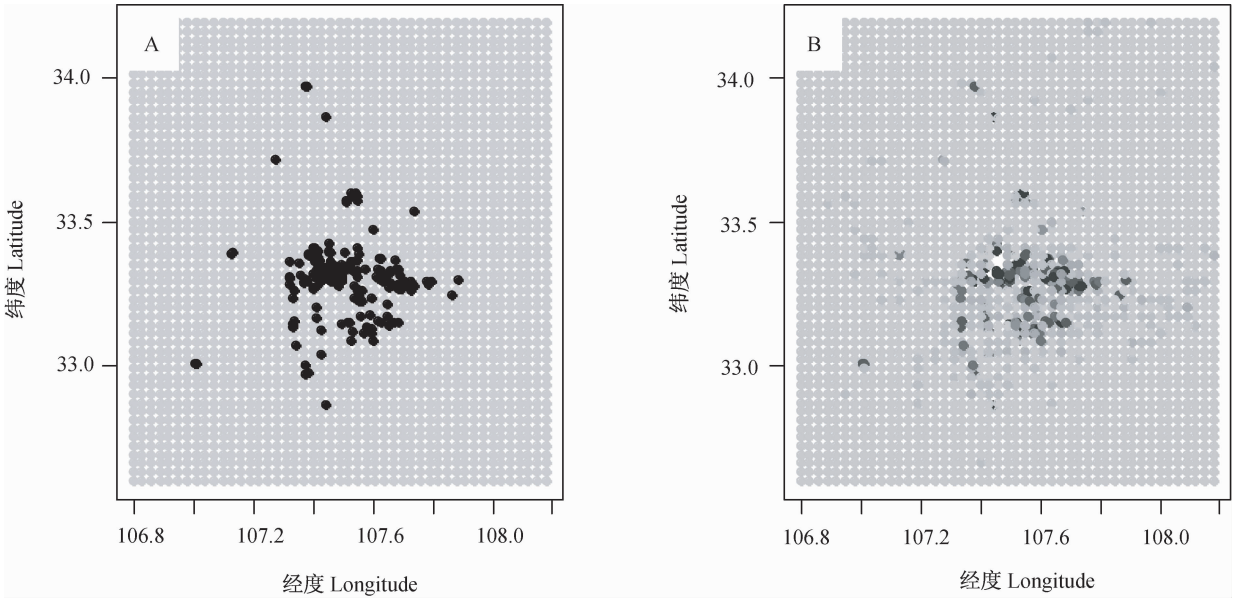

图 3 朱鹮的巢址(黑色)和对照点(灰色)(A)及随机森林算出的每个点朱鹮选择营巢的概率(深色概率高)(B)

Fig. 3 The nest site of the crested ibis (black dots) and the pseudo-absence points (grey dots) (A) and the probability of nest site selection of the crested ibis calculated by random forest (dark color means higher probability) (B)

2.3 回归分析

当因变量 Y 为连续变量时,随机森林通过一组自变量 X 对 Y 进行解释,类似经典的回归分析。

案例 3 依旧以朱鹮为例,介绍随机森林在回归分析上的应用。朱鹮是依赖湿地的鸟类,其生境可以分为一个个相邻的集水区。每个集水区内

朱鹮的巢数同集水区的环境变量相关。用环境变量(包括连续变量和分类变量两个类型)解释集水区内朱鹮的巢数,可以被看作为一个回归的问题。下列代码读取数据并显示数据前 6 行:

```
sheds < - read. csv ( ´d:/data/watersheds4. csv´, header = T) #读取数据
head( sheds) #显示数据 sheds 的前 6 行,如表 3 所示。NA 表示缺失值。
```

表 3　朱鹮栖息地每个集水区内朱鹮的巢数以及环境变量

Table 3　The number of nests and environmental variables for every watershed in the habitat of the crested ibis

巢数 Nests	海拔 Elevation	人类干扰 Footprint	温度 Temperature	稻田 Rice_paddy	水体 Water_body	湿地 Wetland	海拔的变异 Elev_SD
1	597. 83	44. 54	14. 02	0. 14	0. 52	0. 07	197. 54
0	588. 74	32. 41	14. 09	0. 15	0. 08	0. 01	148. 32
0	513. 84	NA	14. 66	0	0. 16	0	28. 84
5	609. 33	30. 2	14. 29	1. 17	1. 03	1. 21	184. 58
0	NA	35. 88	13. 32	0. 18	0. 17	0. 03	NA
2	651. 08	47. 62	14. 41	1. 11	0. 34	0. 38	121. 37

对于缺失数据,R 语言的 randomForest 软件包通过 na. roughfix 函数用中位数(对于连续变量)或众数(对于分类变量)来进行替换。

```
Dat. fill < - na. roughfix( sheds) #用中位数或众数替代缺失值
RF3 < - randomForest( Nests ~ Elevation + Footprint + Temperature + Rice_paddy + Water_body + Wetland + Elev_SD, data = Dat. fill, ntree = 5000, importance = TRUE, na. action = na. roughfix, mtry = 3) #运行随机森林
RF3 #模型结果,显示残差的平方,以及解释变异(环境变量 X 对巢数 Y 的解释)的百分率
plot( RF3) #误差的分布图(图 4)
```

mtry 指定分类树每个节点用来二分数据的自变量的个数。如果 mtry 没有被指定,随机森林用缺省值。对于分类(判别)分析(Y 是分类变量),缺省值是自变量总数的平方根;如果是回归分析(Y 是连续变量),缺省值是自变量总数的 1/3。

3　讨论

本文以 3 个案例介绍了随机森林的具体应用。随机森林结构比较复杂,然而它却极端易用,需要的假设条件(如变量的独立性、正态性等)比逻辑斯蒂回归等模型要少得多。它也不需要检查变量的交互作用和非线性作用是否显著。在大多数情况下模型参数的缺省设置可以给出最优或接近最优的结果。使用者可以调节 mtry 的取值来检查模型的缺省值是否给出误差最小的结果。使用者也可以指定所用的分类树的数量。在计算负荷可以接受的情况下分类树的数量越大越好。图 4 可以帮助使用者判断最小的分类树的数量,以便节省计算时间。

目前,人们已经对多种机器学习的模型进行了比较(Kampichler *et al.*, 2010; Li and Wang, 2013),随机森林经常独占鳌头(Kampichler *et al.*, 2010;Li *et al.*, 2012)。随机森林通过产生大量的分类树,建立若干自变量 X 和一个因变量 Y 的关系。随机森林的一个优点是:它的学习过程很快。在处理很大的数据时,它依旧非常高效。随机森林可以处理大量的多达几千个的自变量(Breiman, 2001b)。现有的随机森林算法评估所有变量的重要性,而不需要顾虑一般回归问题面临的多元共线性的问题。它包含估计缺失值的算法,如果有一部分的资料遗失,仍可以维持一定的准确度。随机森林中分类树的算法自然地包括了变量的交互作用(interaction)(Cutler *et al.*, 2007),即一个自变量 $X1$ 的变化导致另一个自变量 $X2$ 对 Y 的作用发生改变。交互作用在其他模型中(如逻辑斯蒂回归)因其复杂性经常被忽略。随机森林对离群值不敏感,在随机干扰较多的情

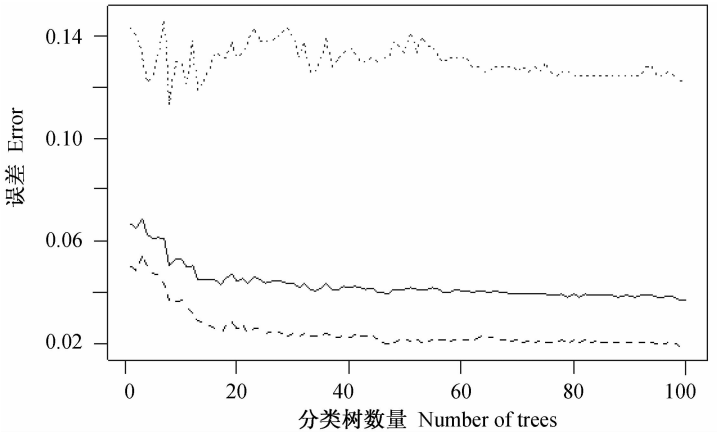

图 4 随机森林的预测误差及其 95% 的置信区间同所用的分类树数量的关系

Fig. 4 The association between prediction error (and its 95% confidence interval) and the number of trees used in random forest

况下表现稳健。随机森林不易产生对数据的过度拟合(overfit)(Breiman,2001b),然而这点尚有争议(Elith and Graham,2009)。

随机森林通过袋外误差(out-of-bag error)估计模型的误差。对于分类问题,误差是分类的错误率;对于回归问题,误差是残差的方差。随机森林的每棵分类树,都是对原始记录进行有放回的重抽样后生成的。每次重抽样大约 1/3 的记录没有被抽取(Liaw,2012)。没有被抽取的自然形成一个对照数据集。所以随机森林不需要另外预留部分数据做交叉验证,其本身的算法类似交叉验证,而且袋外误差是对预测误差的无偏估计(Breiman,2001a)。

随机森林的缺点是它的算法倾向于观测值较多的类别(如果昆虫 B 的记录较多,而且昆虫 A、B 和 C 间的差距不大,预测值会倾向于 B)。另外,随机森林中水平较多的分类属性的自变量(如土地利用类型 > 20 个类别)比水平较少的分类属性的自变量(气候区类型 < 10 个类别)对模型的影响大(Deng *et al.*,2011)。总之,随机森林功能强大而又简单易用,相信它会对各行各业的数据分析产生积极的推动作用。

参考文献(References)

Archer KJ, Kirnes RV, 2008. Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.*, 52(4):2249 – 2260.

Biau G, 2012. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13:1063 – 1095.

Breiman L, 2001a. Random forests. *Mach. Learn.*, 45:5 – 32.

Breiman L, 2001b. Statistical modeling: The two cultures. *Stat. Sci.*, 16:199 – 215.

Breiman L, Friedman JH, Olshen RA, Stone CJ, 1984. Classification and Regression Trees. Chapman and Hall. 1 – 359.

Cutler DR, Edwards TC, Jr., Beard KH, Cutler A, Hess KT, 2007. Random forests for classification in ecology. *Ecology*, 88(11):2783 – 2792.

Deng H, Runger G, Tuv E, 2011. Bias of importance measures for multi-valued attributes and solutions// Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN).

Elith J, Graham CH, 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1):66 – 77.

Genuer R, Poggi JM, Tuleau-Malot C, 2010. Variable selection using random forests. *Pattern Recogn. Lett.*, 31(14):2225 – 2236.

Groemping U, 2009. Variable importance assessment in regression: linear regression versus random forest. *Am. Stat.*, 63(4):308 – 319.

Ho TK, 1995. Random decision forest// Proceedings of the 3rd International Conference on Document Analysis and Recognition. 278 – 282.

Ho TK, 1998. The random subspace method for constructing decision forests//IEEE Transactions on Pattern Analysis and

Machine Intelligence. 832 – 844.

Hopfield JJ, 1982. Neural networks and physical systems with emergent collective computational abilities. *PNAS*, 79(8): 2554 – 2558.

Iverson LR, Prasad AM, Matthews SN, Peters M, 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *Forest Ecol. Manage.*, 254: 390 – 406.

Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S, 2010. Classification in conservation biology: A comparison of five machine-learning methods. *Ecol. Inform.*, 5(6):441 – 450.

Li XH, Li DM, Ma ZJ, Schneider DC, 2006. Nest site use by crested ibis: dependence of a multifactor model on spatial scale. *Landscape Ecol.*, 21:1207 – 1216.

Li XH, Tian HD, Li DM, 2009. Why the crested ibis declined in the middle twentieth century. *Biodiversity and Conservation*, 18(8):2165 – 2172.

Li XH, Tian HD, Li RQ, Song ZM, Zhang FC, Xu M, Li DM, 2012. Vulnerability of 208 endemic or endangered species in China to the effects of climate change. *Reg. Environ. Chang.*, DOI: 10.1007/s10113-10012-10344-z.

Li XH, Wang Y, 2013. Applying various algorithms for species distribution modeling. *Integrat. Zool.*, 8(2):124 – 135.

Liaw A, 2012. Package "randomForest". http://stat-www.berkele.edu./users/breiman/Random Forests.

R Development Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Winham S, Wang X, de Andrade M, Freimuth R, Colby C, Huebner M, Biernacka J, 2012. Interaction detection with random forests in high-dimensional data. *Genet. Epidemiol.*, 36:142.

翟天庆,李欣海, 2012. 用组合模型综合比较的方法分析气候变化对朱鹮潜在生境的影响. 生态学报, 32(8): 2361 – 2370.