【统计理论与方法】

随机森林方法研究综述

方匡南[a,b], 吴见彬[a], 朱建平[a,b], 谢邦昌[a,b]
(厦门大学 a.经济学院计划统计系; b.数据挖掘研究中心, 福建 厦门 361005)

摘要: 随机森林(RF)是一种统计学习理论, 它是利用 bootsrap 重抽样方法从原始样本中抽取多个样本, 对每个 bootsrap 样本进行决策树建模, 然后组合多棵决策树的预测, 通过投票得出最终预测结果。它具有很高的预测准确率, 对异常值和噪声具有很好的容忍度, 且不容易出现过拟合, 在医学、生物信息、管理学等领域有着广泛的应用。为此, 介绍了随机森林原理及其有关性质, 讨论其最新的发展情况以及一些重要的应用领域。

关键词: 随机森林; 分位数回归森林; 生存回归森林; 应用
中图分类号: O212; F222.3 文献标志码: A 文章编号: 1007-3116(2011)03-0032-07

一、引 言

由于传统的分类模型往往精度不高, 且容易出现过拟合问题。因此, 很多学者通过聚集多个模型来提高预测精度, 这些方法称为组合(ensemble)或分类器组合(classifier combination)方法。首先利用训练数据构建一组基分类模型(base classifier), 然后通过对每个基分类模型的预测值进行投票(因变量为分类变量时)或取平均值(因变量为连续数值变量时)来决定最终预测值。为了生成这些组合模型, 通常需要生成随机向量来控制组合中每个决策树的生长。bagging 是早期组合树方法之一, 又称自助聚集(bootstrap aggregating), 是一种从训练集中随机抽取部分样本(不一定有放回抽样)来生成决策树的方法[1]。另外一种方法是随机分割选取, 该方法在每个结点从 k 个最优分割中随机选取一种分割[2]。Ho 关于随机子空间(Random subspace)方法做了很多研究, 该方法通过对特征变量随机选取子集来生成每棵决策树[3]。Amit 和 Geman 定义了很多几何属性以及从这些随机选择属性中寻找每个结点的最优分割[4]。该方法对 Breiman 2001 年提出的随机森林(RF)起了很大的启发作用[5]。

以上这些方法的一个共同特征是, 为第 k 棵决策树生成随机向量 Θ_k, 且 Θ_k 独立同分布于前面的随机向量 $\Theta_1, \ldots, \Theta_{k-1}$。利用训练集和随机向量 Θ_k 生成一棵决策树, 得到分类模型 $h(X, \Theta_k)$, 其中 X 为输入变量(自变量)。比如, 在 bagging 方法中, 随机向量 Θ 可以理解为通过随机扔 N 把飞镖在 N 个箱子上扔中的结果生成, 其中 N 是训练集中的样本记录数。在生成许多决策树后, 通过投票方法或取平均值作为最后结果, 我们称这个为随机森林方法。

随机森林(RF)是一种统计学习理论, 它是利用 bootsrap 重抽样方法从原始样本中抽取多个样本, 对每个 bootsrap 样本进行决策树建模, 然后组合多棵决策树的预测, 通过投票得出最终预测结果。大量的理论和实证研究都证明了 RF 具有很高的预测准确率, 对异常值和噪声具有很好的容忍度, 且不容易出现过拟合。可以说, RF 是一种自然的非线性建模工具, 是目前数据挖掘、生物信息学的最热门的前沿研究领域之一。目前中国对 RF 的研究还是非常少, 因此, 系统地总结整理 RF 最新的理论和应用研究情况很有意义。

收稿日期: 2010-08-10
基金项目: 中央高校基本科研业务费专项资金《基于数据挖掘的数据质量管理研究》(20102210 40); 国家统计局重点项目《金融风险中的统计方法》(2009LZ045)
作者简介: 方匡南, 男, 浙江台州人, 经济学博士, 助理教授, 研究方向: 数据挖掘、金融计量;
吴见彬, 女, 福建宁德人, 硕士生, 研究方向: 数据挖掘。

二、随机森林原理与性质

(一) 原理

随机森林分类(RFC) 是由很多决策树 分类模型$\{h(X, \Theta_k), k = 1, \cdots\}$ 组成的组合分类模型, 且参数集$\{\Theta_k\}$ 是独立同分布的随机向量, 在给定自变量 X 下, 每个决策树分类模型都由一票投票权来选择最优的分类结果。RFC 的基本思想: 首先, 利用 bootstrap 抽样从原始训练集抽取 k 个样本, 且每个样本的样本容量都与原始训练集一样; 其次, 对 k 个样本分别建立 k 个决策树模型, 得到 k 种分类结果; 最后, 根据 k 种分类结果对每个记录进行投票表决决定其最终分类, 详见图 1。

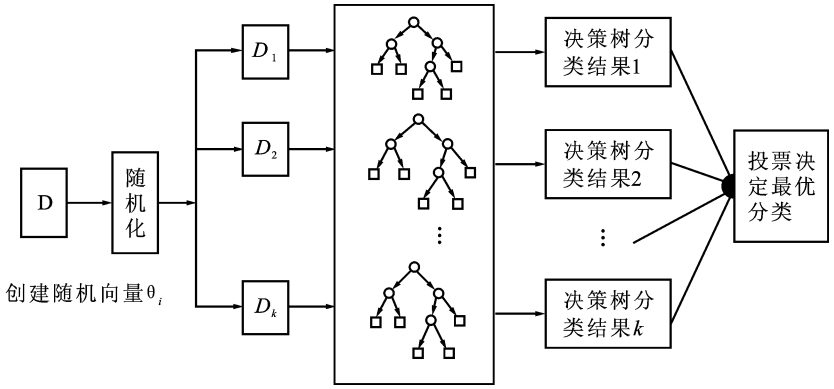

图 1 RF 示意图

RF 通过构造不同的训练集增加分类模型间的差异, 从而提高组合分类模型的外推预测能力。通过 k 轮训练, 得到一个分类模型序列$\{h_1(X), h_2(X), \cdots, h_k(X)\}$, 再用它们构成一个多分类模型系统, 该系统的最终分类结果采用简单多数投票法。最终的分类决策:

$$H(x) = \arg\max_Y \sum_{i=1}^{k} I(h_i(x) = Y) \qquad (1)$$

其中, $H(x)$ 表示组合分类模型, h_i 是单个决策树分类模型, Y 表示输出变量(或称目标变量), $I(\cdot)$ 为示性函数。式(1) 说明了使用多数投票决策的方式来确定最终的分类。

(二) 收敛性

给定一组分类模型$\{h_1(X), h_2(X), \cdots, h_k(X)\}$, 每个分类模型的训练集都是从原始数据集 (X, Y) 随机抽样所得, 由此可以得到其余量函数 (margin function):

$$\mathrm{mg}(X, Y) = av_k I(h_k(X) = Y) - \max_{j \neq k} av_k I(h_k(X) = j)$$

余量函数用来测度平均正确分类数超过平均错误分类数的程度。余量值越大, 分类预测就越可靠。外推误差(泛化误差) 可写成:

$$\mathrm{PE}^* = P_{X,Y}(\mathrm{mg}(X, Y) < 0)$$

当决策树分类模型足够多, $h_k(X) = h(X, \Theta_k)$ 服从于强大数定律。

可以证明, 随着决策树分类模型的增加, 所有序列 $\Theta_1 \cdots$PE* 几乎处处收敛于

$$P_{X,Y}(P_\Theta(h(X, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X, \Theta) = j) < 0)$$

这说明了为什么 RFC 方法不会随着决策树的增加而产生过度拟合的问题, 但要注意的是可能会产生一定限度内的泛化误差[5]。

(三) 泛化误差、强度和相关系数的 OOB 估计

RF 是决策树的组合, 用 bagging 方法产生不同的训练集, 也就是从原始训练集里利用 bootstrap 抽样生成新的训练集, 对每个新的训练集, 利用随机特征选取方法生成决策树, 且决策树在生长过程中不进行剪枝。用 bagging 方法生成训练集, 原始训练集 D 中每个样本未被抽取的概率为$(1 - 1/N)^N$, 这里 N 为原始训练集 D 中样本的个数。当 N 足够大时, $(1 - 1/N)^N$ 将收敛于 $1/e \approx 0.368$, 这表明原始样本集 D 中接近 37% 的样本不会出现在 bootstrap 样本中, 这些数据称为袋外(Out-Of-Bag, OOB) 数据, 使用这些数据来估计模型的性能称为 OOB 估计。之所以使用 bagging 方法, 是因为, 一方面 RFC 使用随机特征时可以提高精度, 另一方面还可以使用 OOB 数据估计组合树的泛化误差(PE*) 以及强度(s) 和相关系数(ρ)。对于每一棵决策树, 我们都可以得到一个 OOB 误差估计, 将森林中所有决策树的 OOB 误差估计取平均, 即可得到 RF 的泛化误差估计。Breiman 通过实验证明, OOB 误差是无偏估计。用交叉验证(CV) 估计组合分类器的泛化误差时, 可能导致很大的计算量, 从而降低算法的运行效率, 而采用 OOB 数据估计组合分类器的泛化误

差时,可以在构建各决策树的同时计算出 OOB 误差率,最终只需增加少量的计算就可以得到。相对于交叉验证,OOB 估计是高效的,且其结果近似于交叉验证的结果。Tibshirani、Wolpert 和 Macready 提出了许多装袋预测的泛化误差方法,并建议使用 OOB 估计作为泛化误差估计的方法[6-7]。Breiman 研究了装袋分类模型(bagged classifier)的误差估计,利用实际例子证明了使用 OOB 估计和使用相同样本容量的测试集的精度一样,因此他认为使用 OOB 估计的话,就没有必要再使用测试集[8]。使用 OOB 数据还可以估计强度和相关系数。它提供了一个内部估计,从而有助于理解分类精度以及如何提高精度。

(四)随机特征选取

随机特征(输入变量)选取,指 RF 为了提高预测精度,引入随机性,减小相关系数而保持强度不变,每棵决策树都使用一个从某固定概率分布产生的随机向量,可使用多种方法将随机向量合并到树的生长过程。目前主要方法有随机选择输入变量(Forest-RI)和随机组合输入变量(Forest-RC)。很多文献都证明通过随机特征选取的方法相对于其他方法具有更低的泛化误差。比如,Dieterrich 认为随机分割选择比 bagging 方法更好,Breiman 认为在输出变量中引入随机噪声数据方法也优于 bagging 方法[2,9]。

1. 随机输入变量选取。RF 最简单的随机特征选取是在每一个结点随机选取一小组(比如 F 个)输入变量进行分割,这样决策树的结点分割是根据这 F 个选定的特征,而不是考察所有的特征来决定。然后,利用 CART 方法完全生长树,不进行修剪,有助于减少树的偏倚。一旦决策树构建完毕,就使用多数表决的方法来组合预测,把这样的过程称为随机选择输入变量(Forest random inputs,简称 Forest-RI)。在 RF 构建过程中选择的输入变量个数 F 是固定的。为了增加随机性,可以使用 bagging 方法为 Forest-RI 产生 bootstrap 样本。RF 的强度和相关性都依赖于 F 的大小,如果 F 足够小,树的相关性趋向于减弱;另一方面,分类模型的强度随着输入变量数 F 的增加而提高。由于在每一个结点仅仅需要考察输入变量的一个子集,这种方法显著减少算法的运行时间。

2. 基于随机变量线性组合的随机森林。假如只有很少的输入变量,比如 M 值不大,用 Forest-RI 法从 M 中随机选择 F 个作为随机特征,这样可能提高模型的强度,但同时也扩大相关系数。另外一种方法是用许多输入变量的线性组合来定义更多的随机特征来分割树,比如由 L 个变量线性组合作为一个输入特征。在一个给定的结点,L 个变量是随机选取的,以它们的系数作为权重相加,每一个系数都是在 $[-1, 1]$ 之间的均匀分布随机数。生成 F 个线性组合,并从中选取最优的分割。这个过程称为 Forest-RC(Forest random combinations)。

3. 随机特征数的确定。在实际的研究中,RF 的随机特征数 F 应该取多少比较合适?不同的随机特征数的选取对模型的强度和相关系数、泛化误差等有何影响?Breiman 研究了随机特征数与强度和相关系数的关系以及随机特征数与泛化误差的关系,发现对于样本量较小的数据集(比如小于 1 000),随着随机特征个数的增加,强度基本保持不变,但是相关系数会相应增加;测试集误差和 OOB 误差比较接近,都随着随机特征数的增加而增加,但 OOB 误差更加稳健。但对于大样本量(比如大于 4 000),结果与小样本量不同,强度和相关系数都随着随机特征数的增加而增加,而泛化误差率都随随机特征数的增加而略微减少。Breiman 认为相关系数越低且强度越高的 RF 模型越好[5]。

三、随机森林的扩展

(一)随机生存森林

随机生存森林(RSF)由 Ishwaran 提出的,是 Breiman 随机森林的衍生[10]。RSF 利用 bootsrap 重抽样方法从原始样本中抽取多个样本集,并对每个样本集建立生存分析树,最后将这些树的预测结果进行综合;与经典随机森林类似,RFS 在每个结点处,只随机抽取 m 个变量建模,而不是将全部自变量都作为分割点的选择范围。

记 $(\boldsymbol{X}, T, \delta), (\boldsymbol{X}_1, T_1, \delta_1), \cdots, (\boldsymbol{X}_n, T_n, \delta_n)$ 为随机元,其中 $\boldsymbol{X}$ 为 d 维离散空间中的特征向量,$T = \min(T^0, C)$ 为所观测的生存时间,$\delta = I(T^0 \leqslant C)$ 是取值为 $\{0, 1\}$ 的二元截尾变量,其中 T^0 为真实事件时间,独立于截尾时间 C。若 $\delta = 0$,则样本 i 在时间 T_i 右截尾;反之,若 $\delta = 1$,则样本 i 在时间 T_i 内未失败。假设 $\boldsymbol{X}$ 独立于 δ,则 $(\boldsymbol{X}, T, \delta)$ 的联合分布为 P,$\boldsymbol{X}$ 的边际分布记为 μ,对于 χ 中的所有子集 A 有,$\mu(A) = P(\boldsymbol{X} \in A)$,并假设对所有的 $\mu(A) > 0$,当 $A \neq \phi$。集合 $\{(\boldsymbol{X}_i, T_i, \delta_i)\}_{1 \leqslant i \leqslant n}$ 被称为训练集,用来建立森林。RSF 算法的步骤:(1)从训练集中抽取出 bootstrap 样本集,对每个样本集都建立一个二元递归生存树。(2)在每棵生存树生长时,每个结点随机选择 p 个候选变

量进行分裂。选择使子结点生存值差异最大的分裂。(3) 让生存树尽可能的生长,直到每个终结点的样本数不小于 $d_0 > 0$。(4) 计算每棵树的生存函数,森林的组合值就是平均生存函数。计算生存函数时采用 KM 估计法,$N(T)$ 为生存树的终结点集。

此外,Ishwaran 等还证明了随机生存森林的一致性,并认为随机生存森林(RSF)往往显著优于其他生存分析方法,尤其是对于高维数据[11]。

(二) 分位数回归森林

分位数回归森林是 Nicolai 在 2006 年提出的,是 Breiman 随机森林的衍生[12]。Nicolai 从数学上证明分位数回归森林具有一致性。分位数回归森林可以被看作一个适应性近邻分类和回归过程[13]。对于每一个 $X = x$,都可以得到原始 n 个观察值权重集合 $\omega_i(x), i = 1, 2, \cdots, n$。

记 θ 为随机参数向量,决定决策树的生长(比如在每个结点对哪些变量分割)。对应的决策树记为 $T(\theta)$。记 B 为 X 的域,也就是 $X: \Omega \to B \subseteq R^p$,其中 $p \in N_+$ 是自变量的个数(维度)。决策树的每一个叶结点 $l = 1, \cdots, L$ 都对应一个 B 的矩形空间。记每一个叶结点 $l = 1, \cdots, L$ 的矩形空间为 $R_l \subseteq B$。对于每一个 $x \in B$,当且仅当一个叶结点 l 满足 $x \in R_l$(也就是 x 沿着决策树的生长被归类的对应叶结点),记决策树 $T(\theta)$ 这样的叶结点为 $l(x, \theta)$。

对于一个新的数据 $X = x$,单棵决策树 $T(\theta)$ 的预测可以通过叶结点 $l(x, \theta)$ 的观测值取平均获得。假如一个观测值 X_i 属于叶结点 $l(x, \theta)$ 且不为 0,令权重向量 $\omega_i(x, \theta) = \frac{1\{X_i \in R_l(x, \theta)\}}{\#\{j: X_j \in R_l(x, \theta)\}}$,权重 $\omega_i(x, \theta)$ 之和等于 1。

在给定自变量 $X = x$ 下,单棵决策树的预测通过因变量的观测值 $Y_i (i = 1, \cdots, n)$ 的加权平均得到。单棵决策树的预测值为 $\mu(x) = \sum_{i=1}^{n} \omega_i(x, \theta) Y_i$。

使用 RF 方法,条件均值 $E(Y \mid X = x)$ 近似于 k 棵决策树预测值的平均,每个决策树随机向量 $\theta_t (t = 1, \cdots, k)$ 都是独立同分布的。令 $\omega_i(x)$ 为 k 棵决策树 $\omega_i(x, \theta_t)$ 的均值,$\omega_i(x) = k^{-1} \sum_{t=1}^{k} \omega_i(x, \theta_i)$,然后 RF 的预测可记为 $\mu(x) = \sum_{i=1}^{n} \omega_i(x) Y_i$。

所以在给定 $X = x$ 下,Y 的条件均值的估计等于所有因变量观测值的加权和。权重随自变量 $X = x$ 的变化而变化,且当给定 $X = X_i (i \in \{1, \cdots, n\})$ 下 Y 的条件分布与 $X = x$ 下 Y 的条件分布越相似,权重越大。

四、随机森林的应用

近 10 来年,随机森林在国外得到了迅速发展,在医学、管理学、经济学等众多领域得到了广泛的应用。Nicolai 基于分位数回归和 Breiman 的随机森林提出了分位数回归森林方法,该方法可提供因变量的全部条件分布的信息,不仅仅是条件均值[12]。Sexton 和 Laake 研究和比较了 bagging 与 RF 的估计结果,发现对于大的数据集,基于 bootstrap 的 RF 估计效果更好[14]。Brence 和 Brown 在改进稳健 RF 回归算法的基础上提出了 booming 算法[15]。

与传统的分类算法相比,RF 具有高准确性等优点,所以近 10 年来,RF 的理论和方法在许多领域都有了比较迅速的发展: 在生物信息学方面,Parkhurst 等使用 RF 研究了沙滩细菌密度与其它变量的影响关系,Smith 等用 RF 对细菌追踪数据进行了研究,并和判别分析方法进行了比较,Alonso 等利用生物标记寄生虫进行鱼类种群判别[16-18]; 在生态学方面,Gislason 等利用 RF 方法对土地的覆盖面积进行了研究,并发现 RF 与其它组合算法相比训练更快,Jan 等基于 RF 和 Logistic 模型建立了方法生态水文分布模型,比较发现 RF 的预测误差小于 Logistic 模型[19-20]; 在医学上,Lee 等利用 RF 技术对助肺 CT 图像进行肺结节的自动检测,同时还在 RF 中加入了(CAC)[21]; 在遗传学上,Diaz - Uriate 等利用 RF 方法进行基因识别,Chen 和 Liu 进行了蛋白质相互关系的研究[22-23]; 在遥感地理学上,Pal、Ham 等、Gislason 等分别都利用 RF 分类器进行了遥感研究[24-26]。此外 Xu 等还进行语言的自动学习,Auret 等还将 RF 应用到时间序列中,对时间序列的变化点进行检测[27-28]。

RF 在经济管理领域也有一定的应用,特别是在客户流失度预测方面: Bart 把 RF 应用于客户关系管理中,发现 RF 的效果都要优于普通线性回归和 Logistic 模型[29]; Xie 等在 RF 中融合了抽样技术和成本惩罚,并以银行客户数据为例进行了客户流失预测[30]; Coussement 等比较了 SVM、Logistic 模型和 RF 客户流失预测能力,当且仅当 SVM 加入最优参数自动选择器后优于 Logistic 模型,但 RF 始终优于 SVM[31]; Burez 等还将加权随机森林应用到客户流失预测中,通过与标准 RF 的比较发现,加权随机森林有更好的预测效果,此外还特别强调了

AUC 指标的优越性[32]; Coussement 等认为提高客户流失预测精度,一方面是充实客户的资料库,另一方面是选取一个好的方法,比较了 RF、SVM、Logistic 模型后发现 RF 是预测效果最好的[33]。RF 除了预测客户的流失度外,还被应用到客户忠诚度预测中。Buckinx 等提出在客户交易数据库中加入客户忠诚度的预测值,同时比较了多元线性回归、RF 和 ANN 的预测效果[34]。

在信用风险评价领域,Fantazzini 等将随机生存森林应用在 SME 信用风险管理中,Hiroyuki 等利用 RF 研究了电力市场的信用风险评估,均取得了很好的效果[35-36]。在国内,林成德等利用 RF 建立企业信用评估的指标体系[37]。在本人所查文献范围内,RF 在金融中的应用还较少,主要有:方匡南等把 RF 应用到基金超额收益方向预测和交易策略中,刘微等使用 RF 进行基金重仓股预测[38-39]。

此外,Keely 等利用 RF 研究社会收入再分配问题[40]。Lessmann 等还利用 RF 进行赛马胜率预测,认为 RF 的预测结果优于传统预测方法,能够带来丰厚的商业利润[41]。

可见 RF 的应用已经比较广泛,对此,Verikas 等在大量实验的基础上探讨 RF 的适用范围,发现 RF 的变量重要性排序结果并非完全可信[42]。

五、小结与讨论

综上所述,RF 是一种有效的预测工具,是一个组合分类器算法,是树型分类器的组合,它集成了 bagging 和随机选择特征分裂等方法的特点,具有以下特征:(1)RF 的精度和 AdaBoost 相当,甚至更好,但运算速度远远快于 AdaBoost,且不容易过拟合。(2)由 bagging 方法产生的 OOB 数据,可用来进行 OOB 估计。OOB 估计可以用来估计单个变量的重要性,也可用来估计模型的泛化误差。(3)能同时处理连续型变量和分类变量。(4)bagging 和随机选择特征分裂的结合,使该算法能较好地容忍异常值和噪声。(5)RF 还可以提供内部误差估计、强度、相关系数以及变量重要性等有用信息。

近年来,RF 在理论和方法上都越来越成熟,并被广泛应用到多种学科中,特别是生物生态领域。研究结果表明,RF 比其它算法确有较大的优势。在经济金融方面的应用目前还较少,有兴趣的学者可以做进一步的研究。中国目前在随机森林领域的研究非常少,本文的目的在于引起更多国内学者关注随机森林的理论、方法和应用研究。

参考文献:

[1] Breiman L. Bagging Preditors[J]. Machine Learning, 1996,24(2).

[2] Dietterich T. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization[J]. Machine Learning, 2000,40(2).

[3] Ho T K. The Random Subspace Method for Constructing Decision Forests[J]. Trans. on Pattern Analysis and Machine Intelligence, 1998,20(8).

[4] Amit Y, Geman D. Shape Quantization and Recognition with Randomized Trees[J]. Neural Computation, 1997,9(7).

[5] Breiman L. Random Forests[J]. Machine Learning, 2001,45(1).

[6] Tibshirani R. Bias, Variance, and Prediction Error for Classification Rules[C]. Technical Report, Statistics Department, University of Toronto, 1996.

[7] Wolpert D H, Macready W G. An Efficient Method to Estimate Bagging's Generalization Error[J]. Machine Learning, 1999,35(1).

[8] Breiman L. Out-of-bag Estimation[EB/OL]. [2010-06-30]. http//stat.berkeley.edu/pub/users/breiman/OOB estimation.ps.

[9] Breiman L. Randomizing Outputs to Increase Prediction Accuracy[J]. Machine Learning, 2000,40(3).

[10] Ishwaran H, Kogalur U B, Blackstone E H, Lauer M S. Random Survival Forests[J]. The Annals of Applied Statistics, 2008,2(3).

[11] Ishwaran H, Udaya B, Kogalur. Consistency of Random Survival Forests[J]. Statistics and Probability Letters, 2010, 80(13/14).

[12] Nicolai, Meinshausen. Quantile Regression Forests[J]. Journal of Machine Learning Research, 2006,7(6).

[13] Lin Y, Jeon Y. Random Forests and Adaptive Nearest Neighbors[J]. Journal of the American Statistics Assoccation, 2006,101(474).

[14] Sexton J, Laake P. Standard Errors for Bagged and Random Forest Estimators[J]. Computational Statistics & Data Analysis, 2009,53(1).

[15] Brence J R, Brown D E. Improving the Robust Random Forest Regression Algorithm[R]. Systems and Information Engineering Technical Papers, Department of Systems and Information Engineering, University of Virginia, 2006.

[16] Parkhurst D F, Brenner K P, Dufour A P, Wymer L J. Indicator Bacteria at Five Swimming Beaches—Analysis Using Random Forests[J]. Water Research, 2005, 39(7).

[17] Smith A, Sterba-Boatwright B, Mott J. Novel Application of a Statistical Technique, Random Forests, in a Bacterial Source Tracking Study[J]. Water Research, 2010, 44(14).

[18] Perdiguero-Alonso D, Montero F E, A Kostadinova, Raga J A, Barrett J. Random Forests, a Novel Approach for Discrimination of Fish Populations Using Parasites as Biological Tags[J]. International Journal for Parasitology, 2008, 38(12).

[19] Gislason P O, Benediktsson J A, Sveinsson J R. Random Forests for Land Cover Classification[J]. Pattern Recognition Letters, 2006, 27(4).

[20] Jan P, Bernard D B, Niko E C V, Roeland S, Sven D, Piet D B, Willy H. Random Forests as a Tool for Ecohydrological Distribution Modelling[J]. Ecological Modelling, 2007, 207(2/4).

[21] Lee S L A, Kouzania A Z, Hu E J. Random Forest Based Lung Nodule Classification Aided by Clustering[J]. Computerized Medical Imaging and Graphics, 2010, 34(7).

[22] Diaz-Uriate R, Andres S A D. Gene Selection and Classification of Microarray Data Using Random Forest [J]. BMC Bioinformatics, 2006, 7(3).

[23] Chen X W, Liu M. Prediction of Protein-protein Interactions Using Random Decision Forest Framework [J]. Bioinformatics, 2006, 21(24).

[24] Pal M. Random Forest Classifier for Remote Sensing Classification [J]. Remote Sens, 2005, 26(1).

[25] Ham J, Chen Y C, Crawford M P, Ghosh J. Investigation of the Random Forest Framework for Classification of hyperspectral Data[J]. IEEE Trans. Geosci. Remote Sens, 2005, 43(3).

[26] Gislason P O, Benediktsson J A, Sveinsson J R. Random Forests for Land Cover Classification[J]. Pattern Recogn. Lett, 2006, 27(4).

[27] Xu P, Jelinek F. Random Forests and the Data Sparseness Problem in Language Modeling [J]. Computer Speech & Language, 2007, 21(1).

[28] Auret L, Aldrich C. Change Point Detection in Time Series Data with Random Forests[J]. Control Engineering Practice, 2010, 18(8).

[29] Larivi re B, Poel D V D. Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques[J]. Expert Systems with Applications, 2005, 29(2).

[30] Xie Y, Li X, Ngai E W T, Wei Y Y. Customer Churn Prediction Using Improved Balanced Random Forests[J]. Expert Systems with Applications, 2009, 36(3).

[31] Coussement K, Poel D V D. Churn Prediction in Subscription Services: An Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques[J]. Expert Systems with Applications, 2008, 34(1).

[32] Burez J, Poel D V D. Handling Class Imbalance in Customer Churn Prediction[J]. Expert Systems with Applications, 2009, 36(3).

[33] Coussement K, Poel D V D. Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers[J]. Expert Systems with Applications, 2009, 36(3).

[34] Buckinx W, Verstraeten G, Poel D V D. Predicting Customer Loyalty Using the Internal Transactional Database[J]. Expert Systems with Applications, 2007, 32(1).

[35] Figini S, Fantazzini D. Random Survival Forests Models for SME Credit Risk Measurement[J]. Methodology and Computing in Applied Probability, 2009, 11(1).

[36] Yasushi U, Hiroyuki M. Credit Risk Evaluation of Power Market Players with Random Forest[J]. Transactions on Power and Energy, 2008, 128(1).

[37] 林成德,彭国兰. 随机森林在企业信用评估指标体系确定中的应用[J]. 厦门大学学报: 自然科学版, 2007, 46(2).

[38] 方匡南,朱建平. 基于随机森林方法的基金超额收益方向预测与交易策略研究[J]. 经济经纬, 2010(2).

[39] 刘微,罗林开,王华珍. 基于随机森林的基金重仓股预测[J]. 福州大学学报: 自然科学版, 2008, 36(1).

[40] Keely L C, Tan C M. Understanding Preferences for Income Redistribution[J]. Journal of Public Economics, 2008, 92(516).

【统计理论与方法】

基于Lyapunov指数渐近分布的混沌特征存在性检验
——以人民币汇率波动序列为例

朱新玲[1], 黎 鹏[2]
(1. 武汉科技大学 管理学院, 湖北 武汉 430081; 2. 中南民族大学 经济学院, 湖北 武汉 430074)

摘要: 当前对于动态系统是否具有混沌特征的判断主要根据最大Lyapunov指数是否大于零进行, 但是要得到混沌现象存在的充分证据, 还需通过相应的假设检验过程, 判断最大Lyapunov指数是否在统计意义上显著大于零。在探讨Lyapunov指数渐近分布的基础上给出了最大Lyapunov指数是否大于零的假设检验过程, 并以人民币汇率波动序列为例进行相应的实证测算。

关键词: Lyapunov指数; 渐近分布; 假设检验
中图分类号: F222.1 文献标志码: A 文章编号: 1007-3116(2011)03-0038-04

一、问题的提出及检验意义

为了判断某动态系统是否具有混沌特征, 其核心工作就是在此系统中找到混沌存在的证据, 现有的研究一般先计算该系统的最大Lyapunov指数, 并根据其是否大于零作为混沌是否存在的判断(1983年, 格里波基证明只要最大Lyapunov指数大于零, 就可以肯定混沌的存在)。但是从统计检验的

收稿日期: 2010-11-12
基金项目: 教育部人文社会科学研究项目《基于非线性和高阶矩视角的人民币汇率波动行为研究》(09YJC790209); 武汉科技大学校基金项目《金融危机背景下人民币汇率波动特征研究》(2009xz41)
作者简介: 朱新玲, 女, 江苏泰兴人, 经济学博士, 讲师, 研究方向: 金融计量;
黎 鹏, 男, 湖北武汉人, 经济学硕士, 讲师, 研究方向: 保险精算。

[41] Lessmann S, Sung M-C, Johnson J E V. Alternative Methods of Predicting Competitive Events: An Application in Horserace Betting Markets[J]. International Journal of Forecasting, 2010, 26(3).
[42] Verikas A, Gelzinis A, Bacauskiene M. Mining Data with Random Forests: A Survey and Results of New Tests[J]. Pattern Recognition, 2011, 44(2).

A Review of Technologies on Random Forests

FANG Kuang-nan[a,b], WU Jian-bin[a], ZHU Jian-ping[a,b], SHIA Bang-chang[a,b]
(a. Department of Statistics, School of Economics; b. Data Mining Center, Xiamen University, Xiamen 361005, China)

Abstract: Random Forests is a statistical learning theory, using bootsrap re-sampling method form sample sets, and then combining the tree predictors by majority voting so that each tree is grown using a new bootstrap training set. It is widely applied in medicine, bioinformatics, economics and other fields, because of its high prediction accuracy, good tolerance of noisy data, and the law of large numbers they do not overfit. In this paper we first introduce the concept of random forest and the latest research, then provide some important aspects of applications in economics, and a summary is given in the final section.

Key words: Random Forests; Quantile Regression Forests; Survival Regression Forests; application
(责任编辑: 杜一哲)