Statistics and Application 统计学与应用, 2023, 12(2), 255-260
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.122026

Hans汉斯

随机森林与传统经典方法在回归与分类问题中的比较

董娅婷

云南师范大学数学学院，云南 昆明

收稿日期：2023年3月6日；录用日期：2023年3月26日；发布日期：2023年4月14日

摘　要

随机森林最早是由Breiman提出的，是机器学习的算法之一。本文以一个回归，一个分类的数据为基础，利用10折交叉验证的方法比较传统经典回归和分类方法与随机森林的预测效果。对于回归数据，分别用逐步回归、岭回归、偏最小二乘回归、线性回归和随机森林做预测对比，10折交叉验证结果显示随机森林的预测效果比传统回归方法的预测效果好。对于分类数据，分别用混合线性判别分析、线性判别分析、logistic回归和随机森林进行分类对比，10折交叉验证结果显示随机森林的分类效果比传统分类方法的预测效果好。

关键词

随机森林，经典回归方法，经典分类方法，交叉验证，机器学习

Comparison of Random Forest and Traditional Classical Method in Regression and Classification Problems

Yating Dong

School of Mathematics, Yunnan Normal University, Kunming Yunnan

Received: Mar. 6th, 2023; accepted: Mar. 26th, 2023; published: Apr. 14th, 2023

Abstract

Random Forest was first proposed by Breiman as one of the algorithms for machine learning. Based on one regression and one categorical data, this paper uses the 10-fold cross-validation

文章引用: 董娅婷. 随机森林与传统经典方法在回归与分类问题中的比较[J]. 统计学与应用, 2023, 12(2): 255-260.
DOI: 10.12677/sa.2023.122026

method to compare the prediction effect of traditional classical regression and classification methods with random forests. For the regression data, stepwise regression, ridge regression, partial least squares regression, linear regression and random forest were used for prediction comparison, and the 10-fold cross-validation results showed that the prediction effect of random forest was better than that of traditional regression method. For the categorical data, mixed linear discriminant analysis, linear discriminant analysis, logistic regression and random forest were used for classification comparison, and the results of 10-fold cross-validation showed that the classification effect of random forest was better than that of the traditional classification method.

Keywords

Random Forest, Classical Regression Methods, Classical Classification Methods, Cross-Validation, Machine Learning

1. 引言

1.1. 随机森林

随机森林是一种基于分类树的算法，是机器学习的算法之一，最早是由 Breiman [1]提出的。

随机森林是从原始数据中用自助法放回地抽样多次，得到一定数量的自助法样本，对所有样本建立一个决策树，对于各个节点，从每个节点的所有竞争的解释变量中随机选取几个作为竞争拆分的变量，对于回归，默认是选取三分之一的解释变量来进行竞争拆分；对于分类，默认是解释变量数目的平方根来竞争拆分[2]。随机森林的每棵树都不剪枝，让其生长，所有决策树的结果取平均值就是回归最终的预测结果，所有决策树的分类结果最多的类别就是分类最终的结果[2]。

1.2. 传统经典回归方法

本文使用逐步回归、岭回归、偏最小二乘回归和线性回归四种传统的经典回归对混凝土的抗压强度做预测。

逐步回归方法是在解释变量很多时，选取最重要的变量进行建模，得到既简单，预测误差又小的模型。可以采用向前、后退的方法进行逐步回归分析。

解释变量的数据矩阵为 $X=\left\{x_{ij}\right\}_{n\times p}$，残差平方和为 $\sum_{i=1}^{n}\left(y_i-\alpha-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2$，岭回归[2]的系数满足：

$$\left(\hat{\alpha}^{(ridge)},\hat{\beta}^{(ridge)}\right)=\underset{(\alpha,\beta)}{\arg\min}\sum_{i=1}^{n}\left[\left(y_i-\alpha-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right]。$$

偏最小二乘回归[3]就是在响应变量和解释变量中先分别找到一个因子，这两个因子在任何可能性的成分中最相关，接着在选定的一对因子的正交空间中在选择一对最相关的因子，这样下去直到选定的因子有足够的代表性即可。

线性回归是建立解释变量和响应变量之间线性关系的一种统计分析方法[4]。利用线性回归分析预测是当今数据建模领域中最简单、应用最广泛的模型应用。

1.3. 传统经典分类方法

本文使用混合线性判别分析、线性判别分析、logistic 回归这三种传统的经典分类方法对乳腺癌进行分类。

首先，假设分类的响应变量一共有 K 类(K 个水平)，则一个个体属于第 K 类的先验概率 π_k 由频率 $\hat{\pi}_k = n_k / N$ (n_k：第 K 类的样本个数；N：总样本个数)来估计。

如果有 P 个解释变量，则解释变量 X 对应的响应变量 Y 属于第 K 类的后验概率用 $\hat{G}(x) = \arg\max_k \left[f_k(x)\pi_k \right]$ ($f_k(x)$：属于第 K 类的观测值向量的分布函数)来估计。

对于混合线性判别分析[2]，后验概率为：

$$\hat{G}(x) = \arg\max_k \left[\pi_k \sum_{c=1}^{C_k} \omega_{kc} \phi(x \mid \mu_k, \Sigma) \right].$$

对于线性判别分析[2]，后验概率为：

$$\hat{G}(x) = \arg\max_k \left[x^{\mathrm{T}} \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^{\mathrm{T}} \Sigma^{-1} \mu_k + \log(\pi_k) \right].$$

另外，logistic 回归[5]是假设响应变量 $y \sim B(n, p)$，则 $\mu = E(y) = p$，采用 Logit 连接函数 $g(\mu)$，$g(\mu) = \mathrm{logit}(p) = \ln \frac{p}{1-p} = X\beta$，这个广义线性模型就是 logistic 模型。

2. 对回归数据混凝土抗压强度的分析

2.1. 数据说明

混凝土抗压强度数据来自网址：https://archive.ics.uci.edu/ml/machine-learning-databases/concrete/。

数据共有 1030 个观测值，共 9 个变量，除了 Age 其他都是混凝土的原料，其中 Compressive.strength 作为响应变量，其余作为解释变量。变量代表的含义如表 1。

Table 1. The meaning of variable names in the compressive strength data of concrete
表 1. 混泥土抗压强度数据变量名代表的含义

变量名	含义
Cement	水泥
Blast.Furnace.Slag	高炉矿渣
Fly.Ash	粉煤灰
Water	水
Superplasticizer	超塑化剂
Coarse.Aggregate	粗骨料
Fine.Aggregate	细骨料
Age	时间
Compressive.strength	抗压强度

2.2. 随机森林和经典回归方法对混凝土抗压强度数据的预测对比

分别使用逐步回归(step)、岭回归(ridge)、偏最小二乘回归(pls)、线性回归(lm)以及随机森林(RF)这五种方法对混凝土的抗压强度做预测，都使用 10 折交叉验证的方法对预测结果进行比较。10 折交叉验证就是把数据随机的分成 10 份，随机选取 1 份当作测试集，余下的 9 份当作训练集，训练集用来训练模型，对模型进行参数估计，测试集用来预测，然后计算得到平均标准化均方误差(NMSE) [2]。NMSE 越小，说明预测效果相对越好。

对五种方法进行 10 折交叉验证得到的平均 NMSE 如图 1 所示。

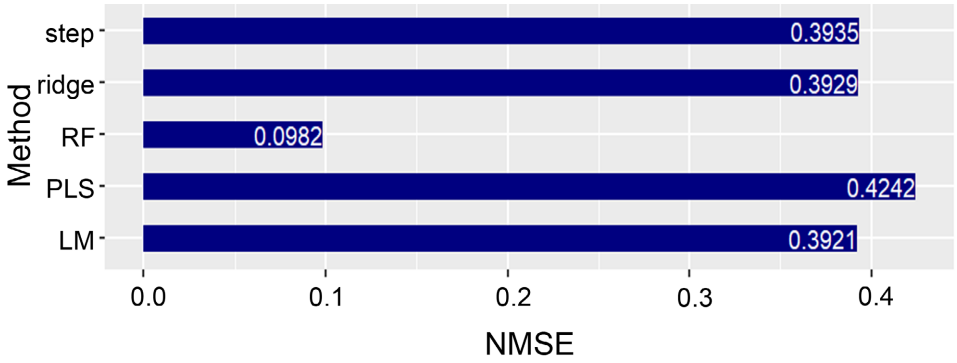

Figure 1. 10-Fold cross-verification of average NMSE by five methods for compressive strength prediction
图 1. 五种方法对抗压强度预测的 10 折交叉验证平均 NMSE

由图 1 可以看出，经典回归中的逐步回归、岭回归、偏最小二乘回归和线性回归的 NMSE 远远大于随机森林的 NMSE，说明随机森林对混凝土的抗压强度预测效果最好。另外，对混凝土抗压强度的预测，线性回归虽然没有随机森林的预测效果好，但是在经典的回归方法中，线性回归的预测效果最佳，逐步回归、岭回归的预测效果微次于线性回归，偏最小二乘回归的预测效果最不佳。

3. 对分类数据乳腺癌的分析

3.1. 数据说明

乳腺癌数据来自 https://github.com/cystanford/breast_cancer_data/。

数据共有 569 个观测值，共 31 个变量，其中 diagnosis 作为响应变量，其余作为解释变量。变量代表的含义如表 2。

Table 2. Meanings of breast cancer data variable names
表 2. 乳腺癌数据变量名代表的含义

变量名	含义
radius_mean	半径(点中心到边缘的距离)平均值
texture_mean	文理(灰度值的标准差)平均值
perimeter_mean	周长平均值
area_mean	面积平均值
smoothness_mean	平滑程度(半径内的局部变化)平均值
compactness_mean	紧密度(=周长 × 周长/面积 − 1)平均值

Continued

concavity_mean	凹度(轮廓凹部的严重程度)平均值
concave points_mean	凹缝(轮廓的凹部分)平均值
symmetry_mean	对称性平均值
fractal_dimension_mean	分形维数(=海岸线近似 − 1)平均值
radius_se	半径(点中心到边缘的距离)标准差、
texture_se	文理(灰度值的标准差)标准差
perimeter_se	周长标准差
area_se	面积标准差
smoothness_se	平滑程度(半径内的局部变化)标准差
compactness_se	紧密度(=周长 × 周长/面积 − 1.0)标准差
concavity_se	凹度(轮廓凹部的严重程度)标准差
concave points_se	凹缝(轮廓的凹部分)标准差
symmetry_se	对称性标准差
fractal_dimension_se	分形维数(=海岸线近似 − 1)标准差
radius_worst	半径(点中心到边缘的距离)最大值
texture_worst	文理(灰度值的标准差)最大值
perimeter_worst	周长最大值
area_worst	面积最大值
smoothness_worst	平滑程度(半径内的局部变化)最大值
compactness_worst	紧密度(=周长 × 周长/面积 − 1.0)最大值
concavity_worst	凹度(轮廓凹部的严重程度)最大值
concave points_worst	凹缝(轮廓的凹部分)最大值
symmetry_worst	对称性最大值
fractal_dimension_worst	分形维数(=海岸线近似 − 1)最大值
diagnosis	M/B (M：恶性，B：良性)

3.2. 随机森林和经典分类方法对乳腺癌数据的分类对比

分别使用混合线性判别分析(mda)、线性判别分析(lda)、logistic 回归(Logit)以及随机森林(RF)这四种方法对乳腺癌进行分类，也是使用 10 折交叉验证的方法对分类结果进行比较。通过计算 10 折交叉验证的平均误判率来判断分类的效果。误判率越小，说明分类效果相对越好。

对四种方法进行 10 折交叉验证得到的平均误判率如图 2 所示。

由图 2 可以看出，和上述回归结果一样，对乳腺癌的分类，经典分类方法中的混合线性判别分析、线性判别分析和 logistic 回归的误判率远远大于随机森林的误判率，说明随机森林对乳腺癌的分类效果最好。另外，混合线性判别分析对乳腺癌的分类效果虽然没有随机森林的分类效果好，但是在经典的分类方法中，混合线性判别分析的误判率最低，分类效果最佳，logistic 回归的分类效果最不佳。

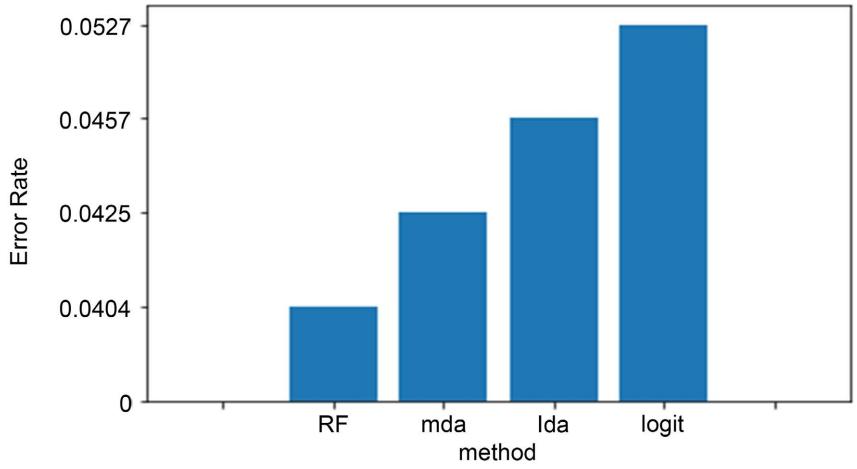

Figure 2. Average false positive rate of 10-fold cross-validation of breast cancer classification by four methods
图 2. 四种方法对乳腺癌分类的 10 折交叉验证平均误判率

随机森林在决定类别时，会评估变量的重要性，做变量选择，可以更好地处理多重共线性的问题。对于缺失值较多的数据，随机森林仍然可以维持较好的准确度，而传统方法需要对缺失数据进行填补，一定程度降低了准确度。

4. 结束语

本文以一个回归，一个分类的数据为基础，比较传统经典回归和分类方法与随机森林的预测效果，两个数据的结果都显示，随机森林的预测效果优于传统经典方法。另外，使用传统经典方法分析此回归数据，结果显示线性回归的预测效果最佳，逐步回归、岭回归的预测效果微次于线性回归，偏最小二乘回归的预测效果最不佳。在此分类数据中，混合线性判别分析的分类效果最佳，logistic 回归的分类效果最不佳。

如今，机器学习成为热门的学习课程之一，人们对多种机器学习的模型进行了比较，通常随机森林的效果最好[6]。随机森林[2]处理高维数据非常高效，也能处理观测值很少的数据，还能处理高阶交互作用和多重共线性问题。随机森林采用了集成算法，它的精度比大多数的单个算法要好，所以准确性高。相信会成为数据分析方法的首要选择。

参考文献

[1] Breiman, L. (2001) Random Forest. *Machine Learning*, **45**, 5-32.

[2] 吴喜之, 张敏. 应用回归及分类: 基于 R 与 Python 的实现[M]. 第二版. 北京: 中国人民大学出版社, 2020. https://doi.org/10.1023/A:1010933404324

[3] 李红梅, 吴喜之, 王涛. 基于纵向数据与多重共线性数据的神经网络与传统方法比较[J]. 统计与决策, 2020, 36(9): 22-25. https://doi.org/10.13546/j.cnki.tjyjc.2020.09.004

[4] Meerasri, J. and Sothornvit, R. (2022) Artificial Neural Networks (ANNs) and Multiple Linear Regression (MLR) for Prediction of Moisture Content for Coated Pineapple Cubes. *Case Studies in Thermal Engineering*, **33**, Article ID: 101942. https://doi.org/10.1016/j.csite.2022.101942

[5] 费宇, 郭民之, 陈贻娟. 多元统计分析: 基于 R [M]. 第二版. 北京: 中国人民大学出版社, 2020.

[6] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报, 2013, 50(4): 1190-1197.