随机森林

《递归划分》第六章报告

任宣霏

数学学院

2022 年 10 月 23 日

目录

基于森林的分类和预测方法是许多科学和工程领域最常用的非参数统计方法之一，特别是机器学习和高通量基因组数据的分析。本章首先讨论随机森林与确定性森林的构造，然后讨论解决实际问题具体需要多大的森林。

为什么要构造随机森林?

树方法的局限性:

- 树结构很容易不稳定：即使数据扰动小。这是所有逐步模型选择过程的通病。
- 解决“大 p 小 n”问题：具有大量变量和有限观测值的现象。
- 生成更多的树为统计推断提供更多经验的解决方案。

森林的优势:

- 稳定，不易因数据干扰出现预测误差。
- 每棵树不一定好，组合起来效果好!
- 很多树木为我们更丰富地利用信息（变量）提供了机会，从而使我们更深入地了解数据。

注记

森林中的每棵树没必要也不应该修剪到“最小”尺寸，根据 Breiman 讨论，将“好”的模型放到“委员会”(Committee)中，效果会适得其反。

森林构建

假设我们有 n 个观测值和 p 个预测变量。

1. (bootstrap) 在原始样本中又放回地抽取 n 个样本;
2. 在每个结点处，从 p 个变量中随机抽取 q 个，进行分裂;
3. 根据之前的递归划分方法，生成一棵树;
4. 重复上述步骤生成更多树。

基于森林的分类，是通过所有树的多数票划分。也就是说，用每一棵树做一次判断，如果大多数树判断为阳性，则森林预测为阳性。

如果跳过上述步骤 2，则称为袋装法（bagging, i.e. (bootstraping and aggregating)

森林构建

有关森林的一些讨论：

- 森林里需要多少棵树？下一节将会讨论最小森林的问题，我们希望随机森林能达到最好的预测效果的同时树之间的相关性最弱，这样可以让随机森林的大小保持在最低水平。
- 如果不修剪里面的树，是否会过拟合？这也是上一节讨论过的，森林里的每一棵树不必是“最优”的树。Breiman 指出，强大数定律保证了不存在过拟合问题，随着森林增大，预测误差收敛。泛化误差的上界跟森林中树的预测效果和树之间的相关性直接相关。
- q 个变量的选择？常用的有 $\log(p)$ 或 $\sqrt{p}$. 在实践中，应该对“平等对待所有预测因子”这一想法保持警惕，不同预测因子不应该全部被平等对待。具体后面会讨论。

最小森林

森林在解决问题的同时带来了挑战：解释性。关键的想法是缩小森林规模同时实现两个目标：

1. 保持类似（甚至更好）的预测精度；
2. 将森林中树木数量减少的可控水平。

我们从预测精度的角度考虑，来缩小森林规模。核心想法是：如果从森林中移除某树对整体预测准确率影响最小，则删除该树。如果用 p_F 表示森林 F 的预测精度，$p_{F_{-T}}$ 为森林 F 中删除某棵树 T 后的预测精度，那么定义准确率之间的差异：

$$\Delta_{-T} = p_F - p_{F_{-T}}. \tag{1}$$

具有最小 Δ_{-T} 的树 T^p 是不重要的，因此可以将其移除。

$$T^p = \underset{T \in F}{\arg\min}(\Delta_{-T}) \tag{2}$$

最小森林

从规模为 N_f 的原始森林中逐个删除树，可以得到一系列有 i 棵树的子森林。$i = 1, \cdots, N_f - 1$.
令 $h(i)$ 为具有 i 棵树的子森林的预测效果。如果 $h(i)$ 只有一个对应的实现，则最优规模：

$$i_{opt} = \underset{i=1,\cdots,N_f-1}{\arg\max} \ (h(i)). \tag{3}$$

如果某个 $h(i)$ 有 M 个对应的实现，则计算其平均值 $\bar{h}(i)$ 和标准差 $\hat{\sigma}(i)$, 再用 1-SE 准则产生一个更稳健、更简洁的模型。
具体为：先找到

$$i_m = \underset{i=i\cdots M}{\arg\max}(\bar{h}(i)), \tag{4}$$

然后选择误差在 $\bar{h}(i)$ 的一个标准差 $\hat{\sigma}(i)$ 范围内的最小子森林，作为最佳森林大小 i_{opt}.

重要性得分

一种解决森林“难以解释”问题的方案是量化森林中的信息。例如，识别森林中的“重要”的预测因子。
如果可以识别出森林中比较“重要”的变量，便可以作为一种变量选择的方法。后续对于这些“重要”因子，可以使用其他更加简单的方法，如分类树。
下面会介绍几种变量的重要性度量。

注记

对于重要性，受预测变量的数量 p、q 和森林中树的数量以及相关性因素的影响，重要性得分幅度可能有较大波动。在实际的数据分析中，排名的先后比幅度的大小更加重要。

重要性得分

1. 基尼重要性：将每次基于变量 k 分类结点的基尼指数减少量求和。问题是：偏好具有多种类别的预测变量。
2. 深度重要性：考虑基于变量 k 分裂时的深度，在浅层分裂被认为是更重要的。令 L 为分裂时的深度，S 为变量的 χ^2 检验统计量。再将所有分裂对应的 $2^{-L}S$ 相加。
3. 置换重要性：首先用森林对于样本分类。对于变量 k，我们可以统计森林中每棵树用变量 k 分类正确的投票数。然后将样本中变量 k 的水平随机置换得到新样本（新样本的变量 k 值与 y 无关，用 k 分类无意义）。比较两者的正确投票数，差值作为变量 k 的置换重要性。
4. 最大条件重要性：森林中所有使用变量 k 的分裂，所产生的的最大 χ^2 统计量。

不确定性预测变量的随机森林

一般来说，我们的分析是基于准确无误的观测到预测因子进行的，或者我们假设如此，然而情况并非总是如此。
比如说，流行病学研究几乎都会涉及“种族”，但一些个体确实是半白半黑或其他比例等，也就是说我们得到的某变量水平并不确定。
为了继续使用随机森林方法，我们需要从具有不确定性预测变量的数据集中生成一个确定预测因子的随机森林。具体构造方法如下：
不妨假设 x_1 是唯一具有不确定性的分类变量，它具有 K 个可能的水平。对于第 i 个个体，$x_{i1} = k$ 的概率记作 $p_{ik}(\Sigma_{k=1}^{K} p_{ik} = 1)$.
我们用 x_{i1} 具有随机性的原始数据集合 $\{x_{i1}, \cdots, x_{ip}, y_i\}_{i=1}^{n}$ 来生成确定性集合 $\{z_{i1}, \cdots, x_{ip}, y_i\}_{i=1}^{n}$，其中 z_{i1} 以概率 $(p_{i1}, \cdots, p_{iK})$ 在 $1, \cdots, K$ 中随机抽取。

加权特征选择的随机森林

- 构建随机森林的第二步是选择一个预测变量的子集来分裂一个结点。假如抽到每个变量的概率相同，那么一些因素会被埋没。
- 例如，在 SWAS（有关基因的研究）中，预测变量为数百万基因型和很少的环境变量。标准的随机森林过程对于识别重要的潜在环境变量是无效的，因为单就数量而言，会被基因型数量所淹没。
- 一种简单的替代方法是使用每个预测变量进行单变量检验，例如计算每个 χ^2 统计量。然后用 χ^2 值的单调函数定义一个抽样概率，而不用等概率，即“更重要”的变量更容易被抽到。
- 不过，对于预测误差的大小和发现重要预测因子的能力来看，相似大小的加权随机森林表现得更出色。

确定性森林

观察结果：当特征数量相对于样本数量较大时，我们会发现结构相似的树具有相似的分类性能。
这一观察结果驱动 Zhang 等人提出了具有相似结构和功能的树的森林。这个森林可以提供比任何单棵树更精确和生物可解释性更强的分类规则，并可重现，所以称为确定性森林。
一种形成确定性森林的简单方法：

1. 选择预先指定的数字，比如 20，用于根节点的顶部分裂。
2. 选择预先指定的数字，比如 3，用于根节点的两个子节点的顶部分裂。

这种顶部分裂一共可以产生 $20 \times 3 \times 3 = 180$ 棵结构和性能相似的树，且没有随机性。