随机森林方法介绍

介绍

因为随机森林的稳健性比较好， 不像其他方法一样受样本数目的影响比较大。所以小样本分类本报告中使用随机森林方法。

数据

仍然使用之前分析中的 mOTU profile 以及疾病状态（是否患病）。

步骤

全部的 mOTU profile 作为 X， 是否患病（二分类）作为 Y。通过设置种子使得结果可重复。跟其他模型一样，主要分为两大部分：第一，模型的选择。在本次报告中，模型的选择也就是对变量的选择（从 X 中提取亚集作为新的 X）。第二，用选择好的模型进行预测。也就是用新的 X 预测 Y。

第一部分 模型的选择

模型的选择分为三步。第一步，一般的 k 折交叉验证。因为我们的样本数目数目较少，所以采用 5 折交叉验证。第二步，重复进行 5 折交叉验证 5 次或者 10 次。第三步，变量数目的选择。通过两个方面去选择变量：预测值不等于真实值整体的均值的波动性尽量小，变量的数目适中（10 个到 30 之间都可以）。下面就对这三步进行详细的介绍。
为了方便的表述，先介绍一些用的向量的名称。
n.var：向量中每一个数字表示变量 X 的数目。从大到小排列， 并且最大为全部变量 X 的数目，最小为 1. 例如 n.var = (360, 50, 25, 10, 4, 1) 则表示变量数可以为 360，或者 50，或者 25，或者 10 或者 4 或者 1.
K：n.var 中数值的数目。同样按照上例，则 K 为 6.
cv.pred：记录预测值的列表。一共包含了 K 个向量， 每个向量中有 90 个（样本数目）值。
error: mean(trainY != testY), 因为是二分类，所以用预测值不等于真实值的整体的均值，来衡量分类效果。
为了方便去选择合适的模型，一般我们会画变量数目和 SSE 的折线图。曲线的数目即为重复 CV 的次数， 每条曲线中的点即为当次重复中 n.var 对应的 error.

1. CV

5-fold CV 中， 原始的 X，Y 会被分 5 次，每次分成 5 份。其中 4 份为训练集，1 份为测试集。比如我们把第一次的 CV 中的数据集记为 DataSet1 (DS1). 并且选择 X 的数目时基于 n.var 向

量。对于 n.var 中的第一个元素，比如 360，则使用 DS1 作为其训练集和测试集，得到每一个变量的重要程度（importance），并按照重要程度从高到低对 DS1 中的变量 X 排序。对于 n.var 中第二个元素，比如 50， 从排序后的 DS1 中顺序选 50 个 X，而 Y 不变去做随机森林。以此类推，并将每次测试集的预测值写入 cv.pred 中。最终 1 次 CV 得到的 cv.pred 为 K 个向量，每个向量有 m 个元素。m 为测试集中样本的数目。则重复上面的过程 5 次，最终得到的 cv.pred 中有 K 个向量，每个向量有全部样本的预测值。
对 cv.pred 中的预测值求 error, 得到 K 个 error. 对应到折线图中的一条折线。

2. 重复 CV

使用 R 中的 replicate 函数，重复以上的 CV 过程 5 次。得到 5 组 error, 每组 K 个。对应到折线图中的 5 条折线。对每个 error 取均值，得到第 6 条折线 (Figure 1)。

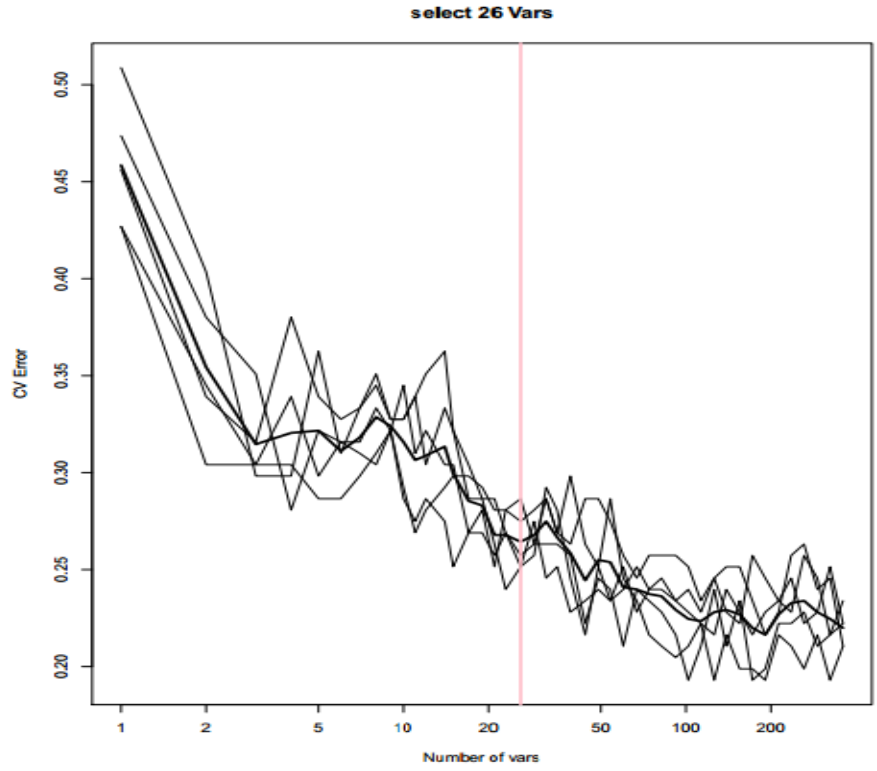

Figure 1 变量-误差曲线图
横坐标为变量的数目，纵坐标为 error. 每条折线表示每次重复得到的 error。中间最黑的一条表示均值。粉红色的竖线即为本次选择的样本数目。

3. 变量选择

通过计算 SSE 均值加标准差，得到的值小并且变量数目合适，则选择该变量数目 p。又通过按照重要程度对菌进行的排序，选择最重要的 p 个菌作为最终的变量。

第二部分 预测

用第一部分中选好的 p 个菌为模型，因为我们样本比较少，所以拿全部的样本做预测。得到的预测值画 ROC，并求得对应的 AUC (Figure 2)。

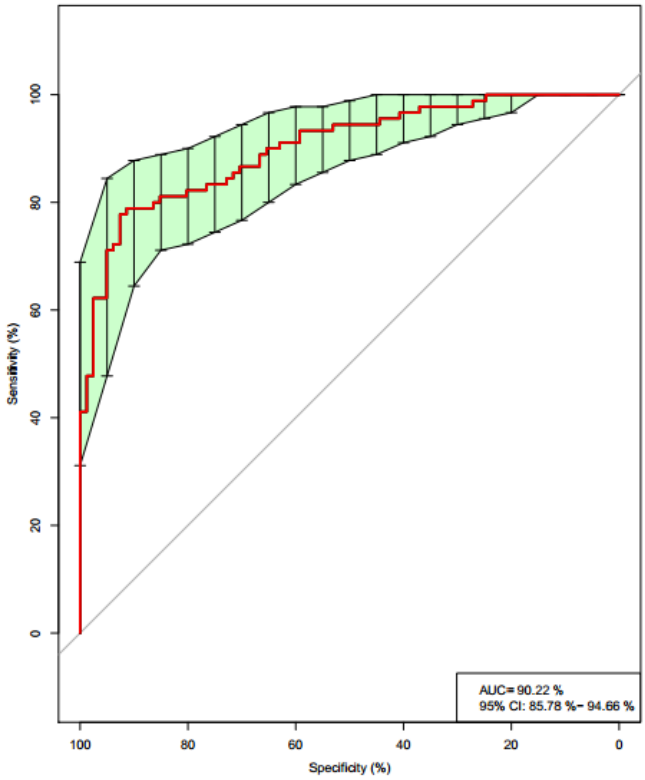

Figure 2 ROC 图

本次分类中，得到的 AUC 为 90.22%。