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， PANSS中每一分类的分数作为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)。

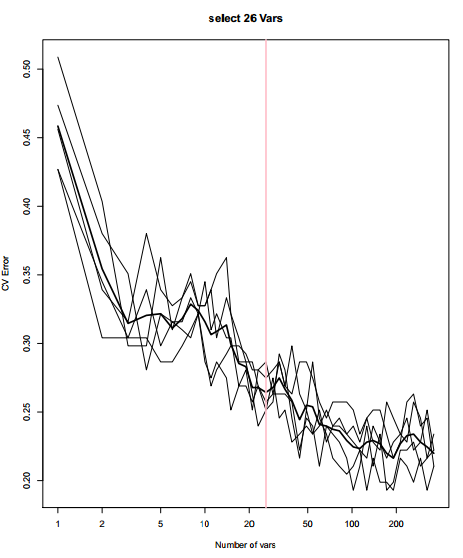


Figure 1 变量-误差曲线图

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)。

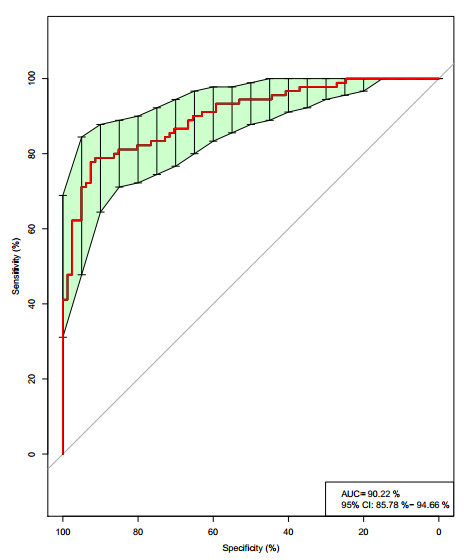


Figure 2 ROC图

本次分类中，得到的AUC为90.22%。