随机森林降维

一、简介

在数据挖掘过程中，维度过高使很多学习器无法工作或效率降低，所以降维是数据预处理过程的一项必要任务。从原始维度中选择一些子集的降维方式称为特征选择，从一组特征中选出一些最有效的特征，使构造出来的模型更好。特征选择的优点：避免过度拟合，改进预测性能；使学习器运行更快，效能更高；剔除不相关的特征使模型更为简单，容易解释。

此次课题我们使用两种方法进行降维，首先使用sbf函数进行过滤，然后利用rfe函数进行封装；另外一种是将过滤后的变量按照重要性排序，依次代入神经网络、逻辑回归与支持向量机模型，通过查看模型精确度确认最终将使用的变量。

二、过滤法

利用sbf（Selection By Filtering）函数，本次使用的是随机森林函数与交叉验证抽样方法，具体如下：

data.filter <- sbf(newdata,mdrrClass,

sbfControl = sbfControl(functions=rfSBF,

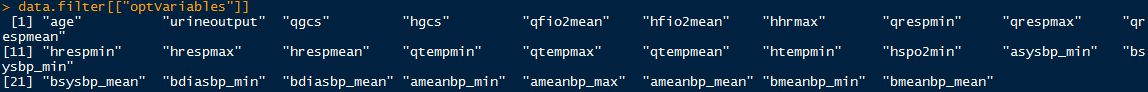
method='cv'))

newdata是输入的矩阵，行为病例，列为变量；mdrrClass是输出的列向量，代表病例的存活与死亡；sbfControl函数用来设置sbf函数的控制参数（其中functions：用于设置模型拟合、预测和特征选择的一系列函数，可以是lmSBF(线性回归),rfSBF(随机森林),treebagSBF(袋装决策树),ldaSBF(线性判别分析法),nbSBF(朴素贝叶斯)和caretSBF(自定义函数)。method：指定抽样方法，可以是boot(BootStrap抽样),cv(交叉验证抽样)）。

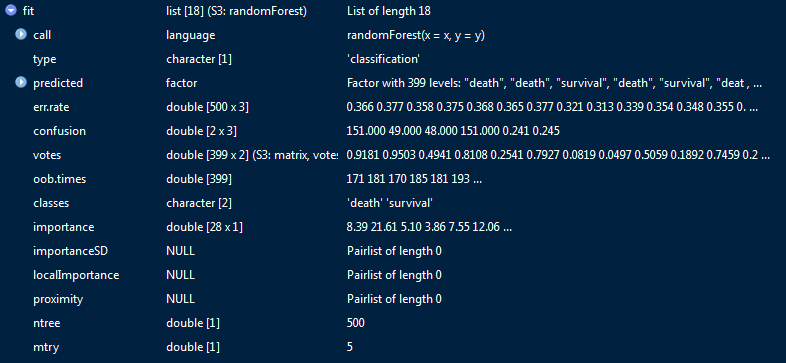
得到的输出结果在optVariables中，挑选出28个特征变量：



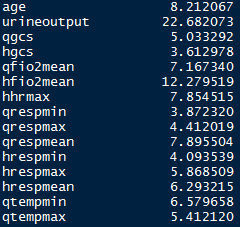
其中展开optVariables可以查看该方法筛选出的具体变量如下：



以及28个特征变量每个变量的重要性分数排序importance：



展开importance变量可查看28个变量对应的重要性评分，分值与重要程度成正比，具体如下：



三、封装法

封装法就是特征选择过程与训练过程整合在一起，以模型的预测能力作为衡量特征子集的选择标准。封装法可以选择出高质量的子集，但速度会比较慢。本课题所用封装法的输入是上一步过滤法得到的变量，对这28个变量进行进一步的筛选，即下面表格中的RF过滤封装降维法。

利用rfe（Recursive Feature Elimination）函数，具体如下：

profile <- rfe(x,mdrrClass,size=

c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30),

rfeControl = rfeControl(functions=rfFuncs,method='cv'))

x是上一步过滤法得到的输入变量;mdrrClass是输出的列向量，代表病例的存活与死亡；sizes是通过一个整数向量，指定需要保留查看的特征（最后筛选出的变量个数与该设定无关）; rfeControl是指定rfe函数（降维函数）的控制参数（本课题封装法使用随机森林函数“rfFuncs”和十折交叉验证抽样方法“cv”）

得到的输出结果，在optVariables中可查看筛选出的21个变量：



筛选过程随着变量变化的预测结果如下图所示，可见本课题的随机森林封装降维法最终选出了21个特征变量，此时准确率最高。

