分析报告

根据数据的复杂程度，选择了第三个Titanic数据集进行分类及聚类。首先对数据进行统计分析，确定属性的变量类型及统计特性，主要观察属性的缺失情况，结果包括Age和Fare两个数值属性，其余为标称属性，各项属性的频数分别等信息如train\_info所示。

1. 缺失值处理

首先，剔除了数据中不重要的属性，例如PassengerId、Name等属性；其余的属性中，Age的缺失值较多，采取单独处理，属性Fare和Embarked有个别缺失值，分别采取中位数填充和众数填充的方法来填补缺失值。

1. 分类

本次分别采用随机森林和Xgboost两种算法对数据进行分类，调用（library）randomForest和xgboost两个程序包。

随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定，具有准确率高、不容易过拟合、抗噪能力强等优点。

程序实现：rf\_model <- randomForest(formula,train)

pred <- predict(rf\_model,test)

用随机森林模型进行训练，得到训练集的分类情况如下：

|  |  |  |
| --- | --- | --- |
| Target\_tf  Pred\_train\_tf | 0 | 1 |
| 0 | 533 | 58 |
| 1 | 16 | 284 |

正确率、灵敏度、特异度分别为91.69%，95.30%，90.19%，将测试集数据得到的预测值提交到kaggle网站上，得到的评分为0.76555。

Xgboost算法，全程eXtreme Gradient Boosting，是在GBDT的基础上对boosting算法进行的改进，内部决策树使用的是回归树，具有速度快、效果好、能处理大规模数据、支持自定义损失函数等优点。

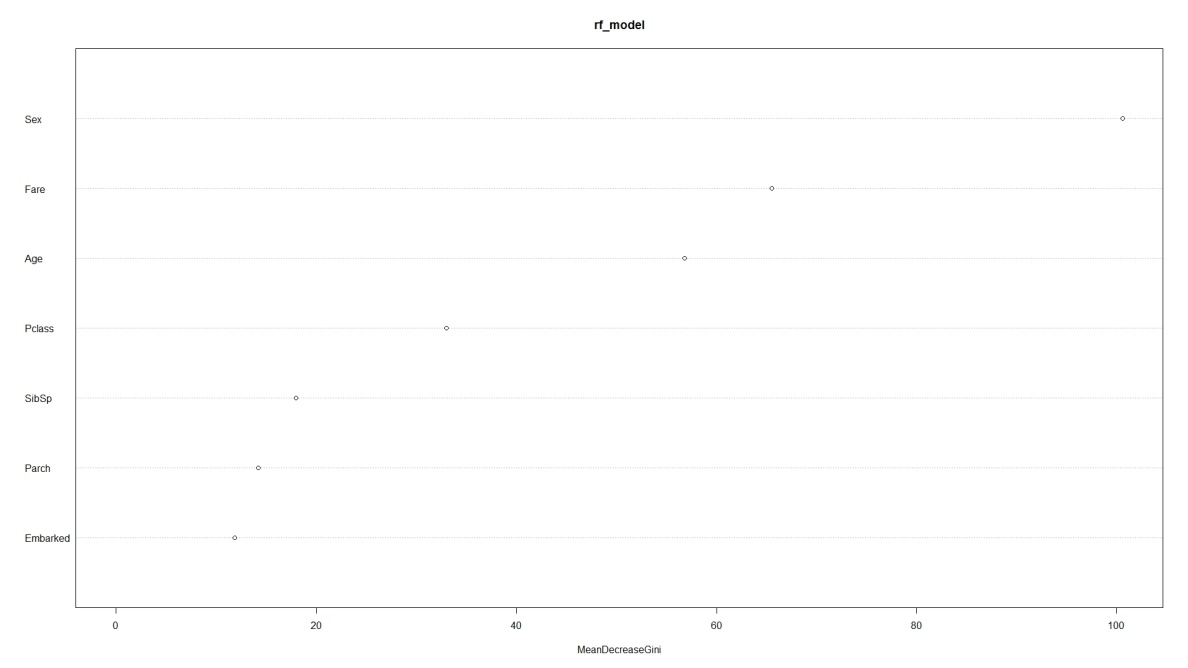
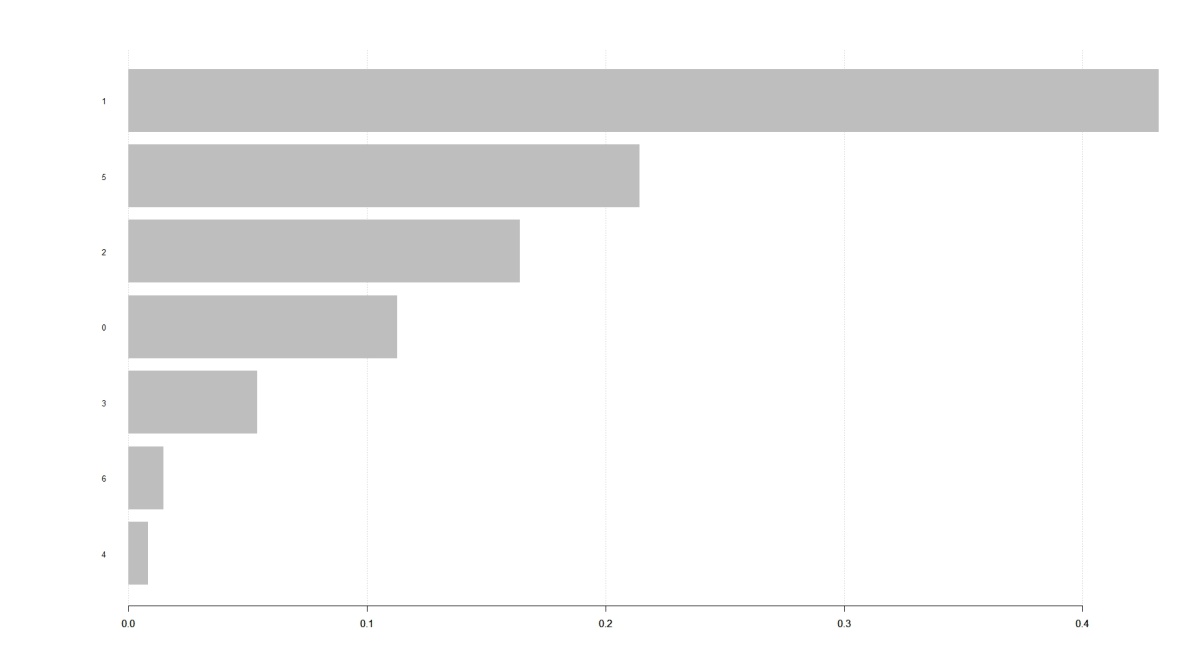
程序实现：xgb\_model <- xgboost(data = train\_data,label = target, eta=0.1,max\_depth = 5,nrounds = 100)

用Xgboost算法得到训练集的分类情况如下所示：

|  |  |  |
| --- | --- | --- |
| Target\_tf  Pred\_train\_tf | 0 | 1 |
| 0 | 534 | 37 |
| 1 | 15 | 305 |

正确率、灵敏度、特异度分别为94.16%，95.31%，93.52%，将测试集数据得到的预测值提交到kaggle网站上，得到的评分为0.7751。

综合来看，Xgboost算法的结果略高于rondamForest的结果，但同时结合训练集的结果看，都存在一定的过拟合问题。另外，采用两种算法对数据集进行分类，分别得到属性重要性排序如下图所示：



Xgboost的结果对应为'Sex','Fare','Age','Pclass','SibSp','Embarked','Parch'，随机森林对应结果为'Sex','Fare','Age','Pclass','SibSp','Parch','Embarked'，可见两算法的结果大体一致。

1. 聚类

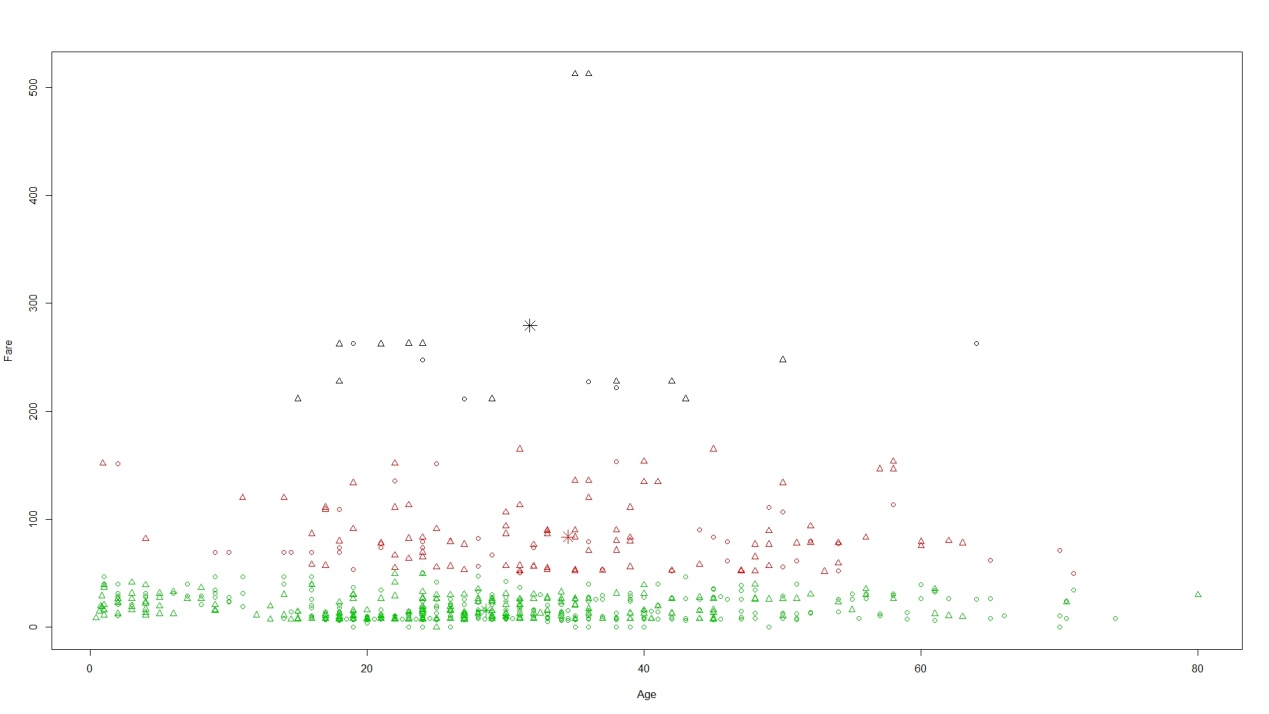
采用K-means和Fuzzy Analysis Clustering两种方法进行聚类，调用（library）cluster程序包实现。

K-means算法是硬聚类算法，是典型的基于原型的目标函数聚类方法的代表，它是数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。K-means算法以欧式距离作为相似度测度，它是求对应某一初始聚类中心向量V最优分类，使得评价指标J最小。算法采用误差平方和准则函数作为聚类准则函数。

程序实现：km <- kmeans(data.matrix(train),centers = 3,nstart = 10)

本次分别试验了聚类个数为3、4、5的结果，其中centers=3时效果最好，结果如下所示：

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster  Target | 1 | 2 | 3 |
| 0 | 6 | 46 | 497 |
| 1 | 14 | 95 | 233 |



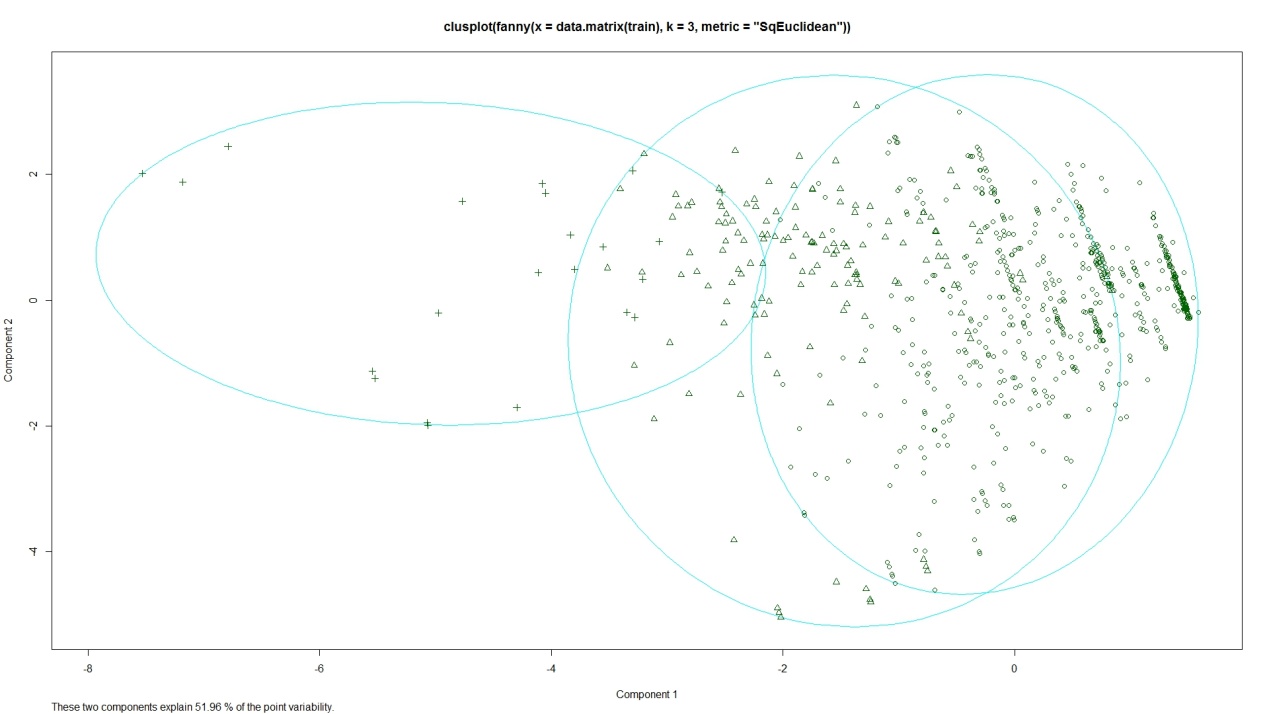
以较为具有代表性的属性Age和Fare作图如上所示，可以看出黑色、红色、绿色分别代表类别1、2、3。通过横坐标的Age属性观察，各个类别在各年龄段之间没有明显的区分；观察纵坐标Fare，从聚类结果可以看出Fare的高低与类别之间具有良好的相关性。

模糊聚类分析是一种采用模糊数学语言对事物按一定的要求进行描述和分类的数学方法。模糊聚类分析一般是指根据研究对象本身的属性来构造模糊矩阵，并在此基础上根据一定的隶属度来确定聚类关系，即用模糊数学的方法把样本之间的模糊关系定量的确定，从而客观且准确地进行聚类。

程序实现：fc <- fanny(data.matrix(train),3,metric = 'SqEuclidean')

各个聚类类别的存活情况如下所示：

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster  Target | 1 | 2 | 3 |
| 0 | 490 | 53 | 6 |
| 1 | 231 | 95 | 16 |



结果可视化如上图所示，我们可以很直观地看出聚类之间有交叉，结果不如K-means方法更加清晰的让我们看到，但综合各个类别的存活情况来看，两种方法的聚类结果还是大体一致的。但图像为fanny函数自动生成，即横纵坐标分别属于哪个属性还有待探究。