集成学习实验报告

161920319 张一帆

0x00 实验目的

通过实现两种集成学习的方法**AdaBoost**和**RandomForest**来时间集成学习的方法和比较两种算法.

0x01 实验过程

本次实验选取的数据集是**UCI**数据集**Adult**, 并进行了一定的处理.

AdaBoost

本次实验实现的**AdaBoost**基分类器采用了**sklearn**中集成好的决策树方法, 其中参数为**max_depth=2**来保证每个基分类器都是一个弱分类器, 采用**AUC**作为评价分类器性能的评价指标, 通过调用**sklearn**算法包对**AUC**指标进行计算.

本次实现的**AdaBoost**算法并没有将所有的分类器存储起来, 而是在训练过程中将对应分类器的个数和AUC值存储起来, 每10个分类器记录一次, 并采取**5-折交叉验证**来验证**AdaBoost**在训练数据集上的精度.

RandomForest

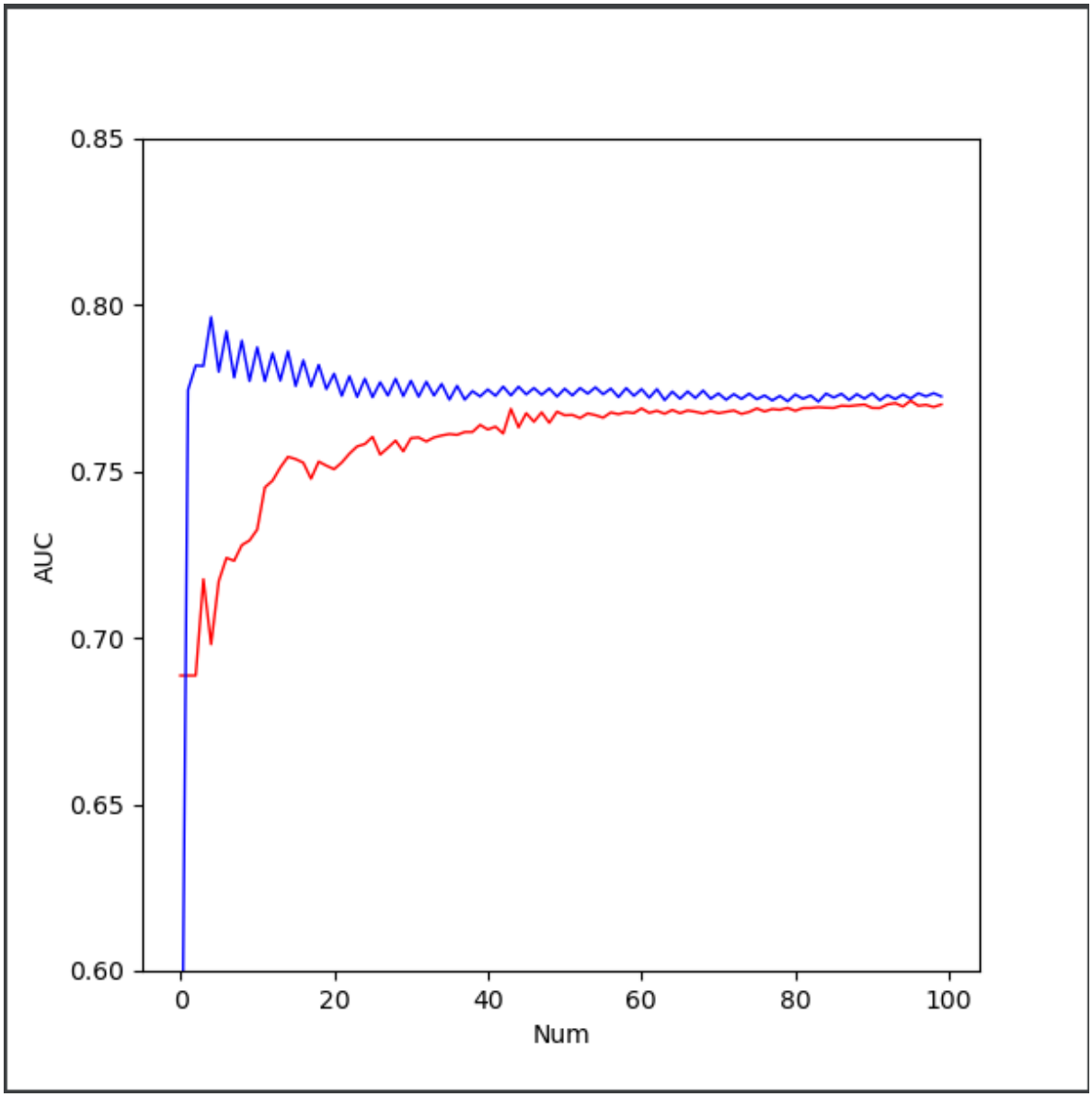
本次实现的随机森林算法基学习器同样采用了**sklearn**中的决策树算法, 因其自带随机属性故随机森林算法中的决策树随机部分可以减少很大的工作量, 而关于数据集的选取, 本次实验采取了有放回的选取方式, 假设一共有m条数据, 则一共选取m次构成数据集来训练决策树, 最后采用投票法来判断每个训练样本的类别, 算法伪代码如下:

```
Input:
    训练集 D,
    基学习器算法F
    训练轮数T
过程:
5-折数据集, 选取其中4份作为数据集, 1份作为验证集
for t = 1, 2, ..., T do
   运行RandomForest, 有放回的选取m个样本作为训练集, 根据5-折交叉验证产生的测试集来计算AUC的值, 并且每10次记录AUC的值
end for
重复5次步骤1, 并取平均
画出AUC随基分类器个数的表格

```

同样的, 本次实现的**RandomForest**算法同样采用**AUC**来评价分类器性能, 并且通过**5-折交叉验证**来验证在训练集上的精度.

实验结果



蓝色线为**RandomForest**, 红色线为**AdaBoost**, 可以看出**AdaBoost**的**best_T**为**13** 此时测试集的**AUC**为
0.7271719023163822, **RandomForest**的**best_T**为**4**, **best_AUC**为**0.783058**