机器学习作业实验报告—集成学习

161910126 赵安

1、实验目的

本次实验中将结合两种经典的集成学习思想：***Boosting***和***Bagging***，对集成学习方法进行实践。

本次实验选取***UCI数据集Adult***（二分类数据集）。

由于**Adult**是一个类别不平衡数据集，本次实验选用***AUC***作为评价分类器的评价指标，通过调用***sklearn***算法包中***metrics.roc_auc_score***函数对**AUC**进行计算。

2、实验过程

以下代码均采用***python3.8***编写，基分类器调用***sklearn***中的基本决策树实现，***max_depth***均设置为**2**.

（1）在***BoostMain.py***中对***Adaboost***算法进行了实现，分别设置不同的学习器数目，计算**AUC**。

（2）在***RandomForestMain.py***中对**随机森林**算法进行了实现，别设置不同的学习器数目，计算**AUC**。

随机森林的算法伪代码如下：

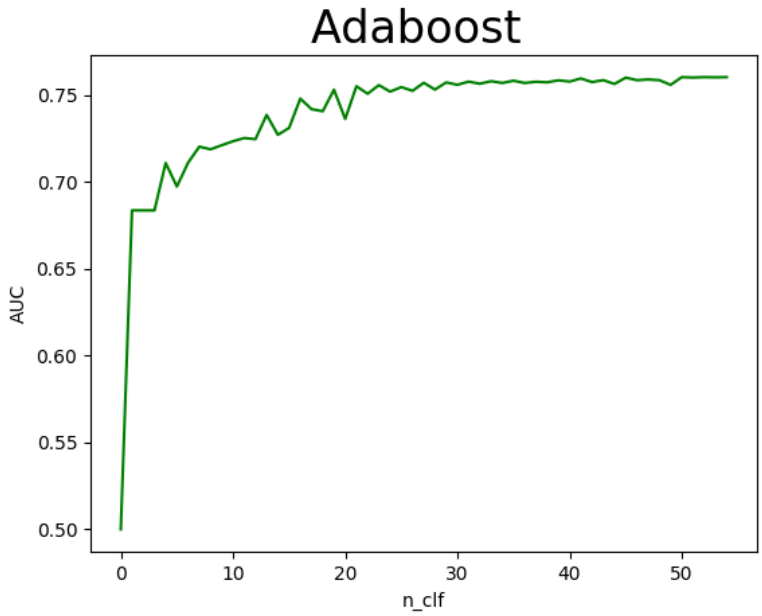
```
def Sampling_function(x_data, y_data, k):
    Generates an unordered array (length = number of attributes)
    Randomly select k required attributes
    randomly sampling the attributes of the training set
    sanpling the training set
    return dataset_sampled

def random_forest(x_train, y_train, x_test_my,y_test_my,number of Weak learner
,attributes sampled):
    initialize the prediction results
    for i in range(number of Weak learner):
        x_train_sampled, y_train_sampled = Sampling_function(x_train, y_train,
k)
        vote on the forecast results
        Calculate AUC
    return AUC
```

（3）对上述***Adaboost***算法和随机森林算法的验证过程做了些许修改，增加在训练集上使用5折交叉验证进行**AUC**验证评价。在***RandomForestMain.py***文件中做出折线图，该折线图以基分类器数目为横轴，**AUC**指标为纵轴，图中两条线对应***Adaboost***和随机森林。

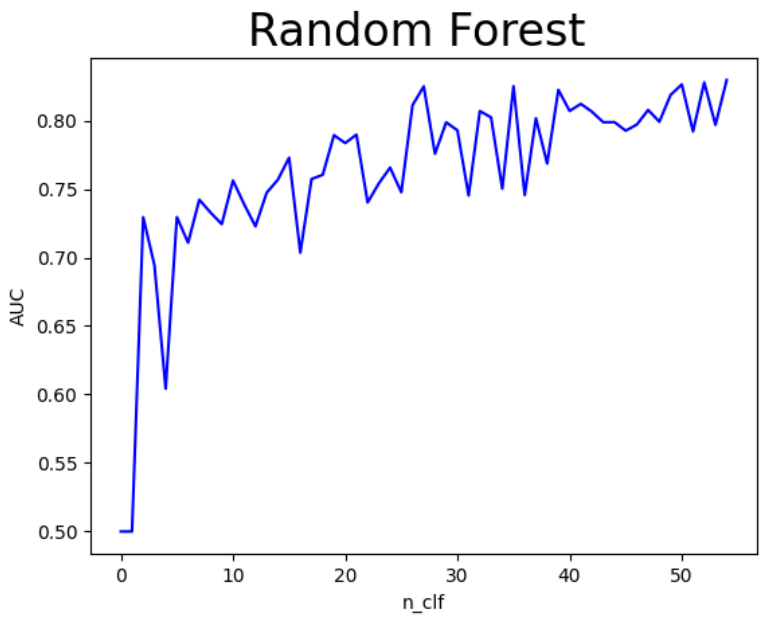
（4）根据参数调查结果，对***Adaboost***和随机森林选取最好的基分类器的数目，在训练数据集上进行训练，得到在测试集上的**AUC**指标。

3、实验结果

使用**Adaboost**算法得到如下结果：

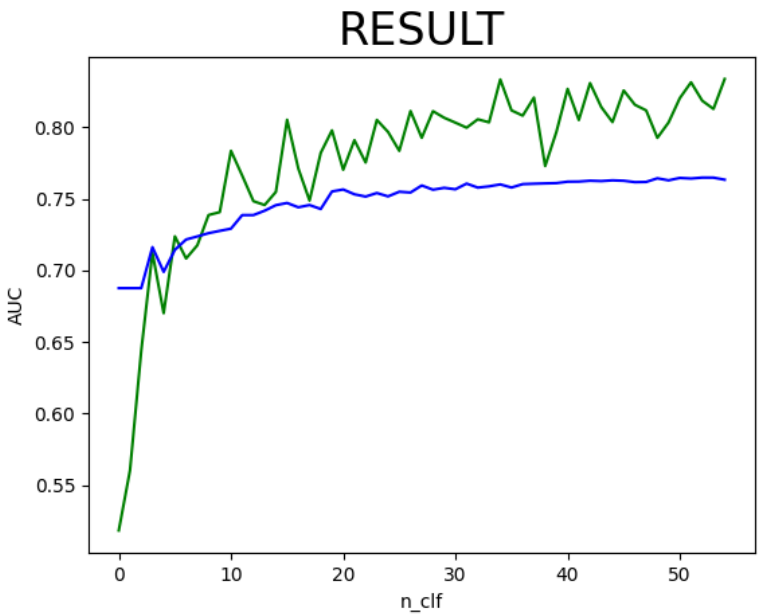


使用**随机森林**算法得到如下结果：



随机森林结果波动较大，应该是某个地方出错了。。。

对**Adaboost**算法与**随机森林**算法使用5折交叉验证，得到**AUC**评价：



根据上述实验进行参数调查，对于**Adaboost**而言，选取基分类器数量为 30，得到测试集上的**AUC**指标为：0.7577994338108869

```
BoostMain
C:\Users\EKKO\AppData\Local\Programs\Python\Python38\python.exe "D:/Machine Learning/XIGUA/Homework_集成学习/BoostMain.py"
AUC: 0.7577994338108869

进程已结束，退出代码为 0
```

对于随机森林而言，选取基分类器数量为——，得到测试集上的**AUC**指标为：0.7598456121598303

```
C:\Users\EKKO\AppData\Local\Programs\Python\Python38\python.exe "D:/Machine Learning/XIGUA/Homework_集成学习/RandomForestMain.py"
AUC: 0.7598456121598303

进程已结束，退出代码为 0
```