**决策树**

自博16 陈斯杰 2016310721

**一、方法与性能**

本次作业中，我设计了三种决策树来对语料库进行分类，具体方法如下：

1. 基于基尼系数，粗暴预剪枝+基尼系数阈值预剪枝
2. 基于基尼系数，基尼系数阈值预剪枝
3. 基于信息增益，增益阈值剪枝

由于决策树经常发生过拟合的，以下的准确率仅仅指测试集上的准确率。

三种决策树的训练数据和分类效果为：

1. 方法选择了30%的数据作为训练集，剩余作为测试集，准确率0.725595238095，耗时1021.873852 secs
2. 方法选择了15%的数据作为训练集，剩余作为测试集，准确率0.708169934641，耗时814.333577 secs
3. 方法选择了8%的数据作为训练集，剩余作为测试集，准确率0.601977657005，耗时672.518412 secs

为了进一步提高分类准确性，我使用了bagging方法。我们使用80%的数据作为训练集，20%的数据作为测试集。在80%的数据中，我们进行10次重采样，每次重采样的覆盖率为20%。然后使用10次重采样的数据训练10个基于基尼系数+基尼系数阈值预剪枝的决策树，根据它们的预测结果投票决定最终分类结果。

10个重采样的分类器单独预测的结果如下：

耗时903.491542 secs

准确率=0.704514

耗时888.620005 secs

准确率=0.710069

耗时888.620005 secs

准确率=0.710069

耗时915.040261 secs

准确率=0.717361

耗时1297.426727 secs

准确率=0.707639

耗时1289.610949 secs

准确率=0.694097

耗时1252.979583 secs

准确率=0.719097

耗时1258.606536 secs

准确率=0.711458

耗时1310.979216 secs

准确率=0.706944

耗时1219.688553 secs

准确率=0.712153

10个分类器投票后，根据众数决定最终预测结果，达到的准确率为78.5417%

**二、运行说明**

0) Sogou\_data文件夹放在同一目录下

1) 先运行sampling.py；

2) 待到1）结束后，可以同时运行bagging1.py ~ bagging10.py；

3) 最后运行reducing.py

**二、附件**

SingleEntropy8perTrain.py、

SingleTree15perTrain.py、

SingleTree30perTrain.py为前文提到的三个单独的决策树分类的代码。

Bagging方法的代码包含：

GiniTree.py（决策树的实现）

sampling.py（训练集和测试集的划分、重采样）

reducing.py（融合多个决策树结果）