**数据挖掘 实验报告**

院系＿＿软件学院＿ 专业＿软件工程＿

年级 2015级 课程班级＿

实验时间 2018 年＿4 ＿月 26＿日

实验名称 数据的预处理和可视化

指导老师 刘昆宏

# 第一部分 实验内容

使用决策树和knn，自行通过代码实现而非调用算法函数库，实现bagging算法，并比较结果；

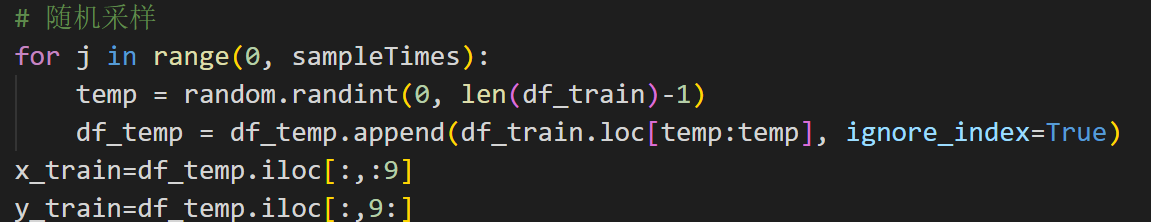
使用决策树实现随机森林算法，分析单纯的bagging和random forest 效果的差异；分析scikitlearn自带的random forest与自己写的算法效果是否存在差异；

请用可视化的图形展示、比较实验结果。

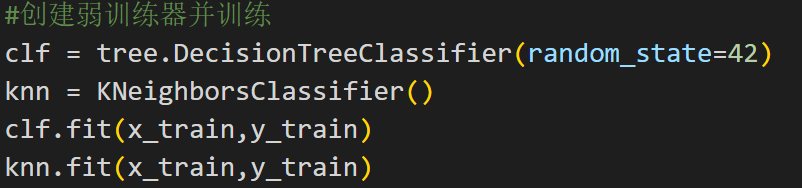
# 第二部分 实验过程描述

Bagging：

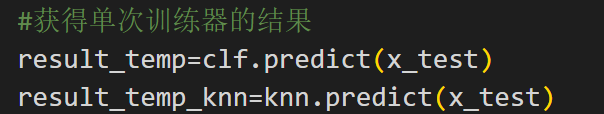
1. 读取数据
2. 构造bagging函数
   1. 随机采样



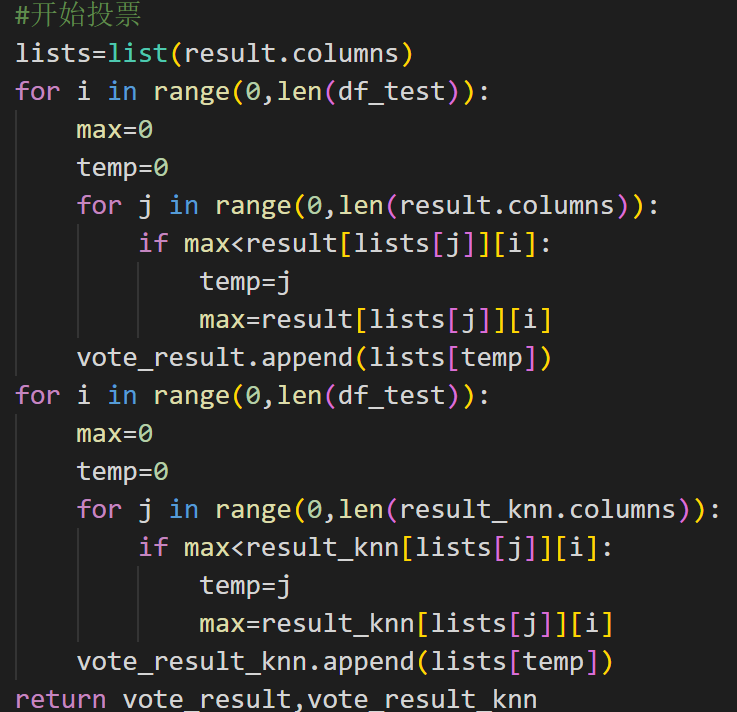
* 1. 创建弱训练器
  2. 用随机采样得到的数据进行训练



* 1. 收集弱训练器的预测结果



* 1. 投票选出bagging算法最后的结果



1. 可视化输出
   1. 比较相同弱分类器数量时，数据集大小不同导致的结果差异
   2. 比较数据集大小相同时，弱分类器数量不同导致的结果差异

RandomForest:

1. 读取数据
2. 利用Bagging构造随机森林
3. 可视化输出

# 第三部分 结论与分析

1. Bagging算法基本原理：

Bagging使用随机放回采样，在选择合适的弱分类器后，通过随机样本的训练，使得多个弱分类器对不同的特征值有着比较针对性的适应性，在多个分类后，通过投票的形式得到Bagging算法最后的结果。

1. RandomForest算法基本原理：

和Bagging的原理类似，但是在训练弱分类器——决策树时，优化决策树的分类，根据信息增益进行分类，最后将生成的决策树组成随机森林，投票输出算法结果

1. 可视化数据分析

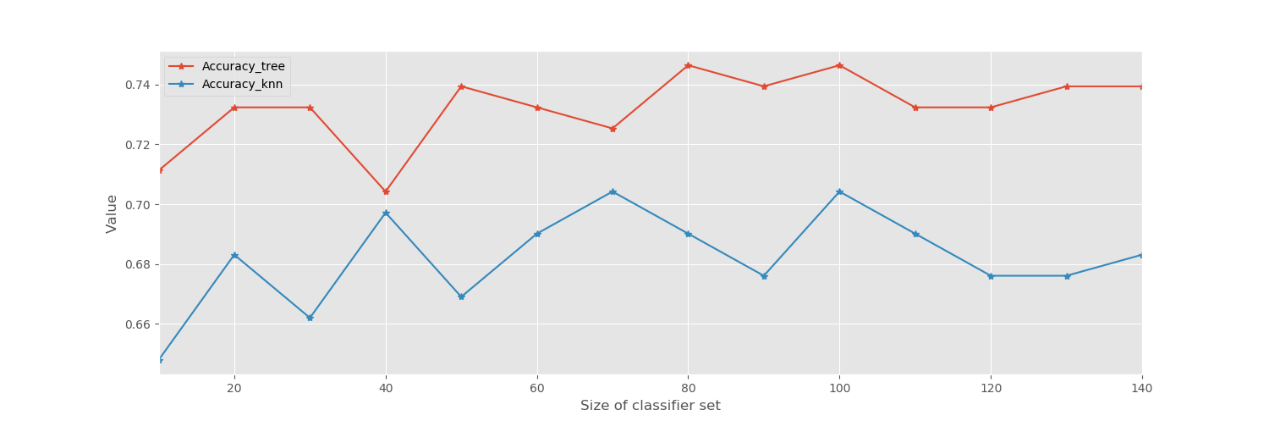
（在训练弱分类器时使用的时完全相同的随机样本，即同一份随机样本训练KNN和决策树）

* 1. Bagging算法中KNN和决策树的比较：



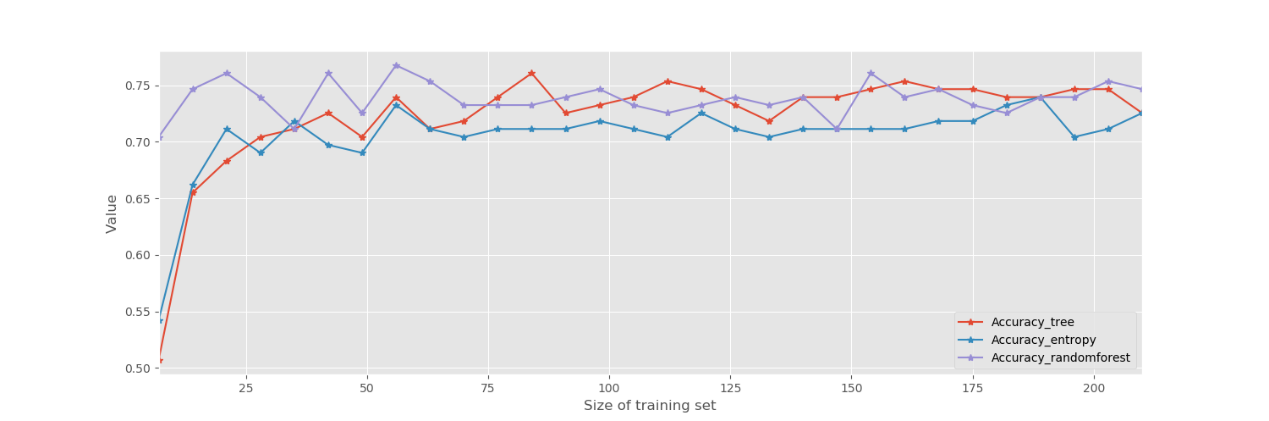
图为比较相同弱分类器数量时，数据集大小不同导致的结果差异

在随机采样的样本数量较少时，决策树明显优于KNN，在数量足够时，二者并无明显差异

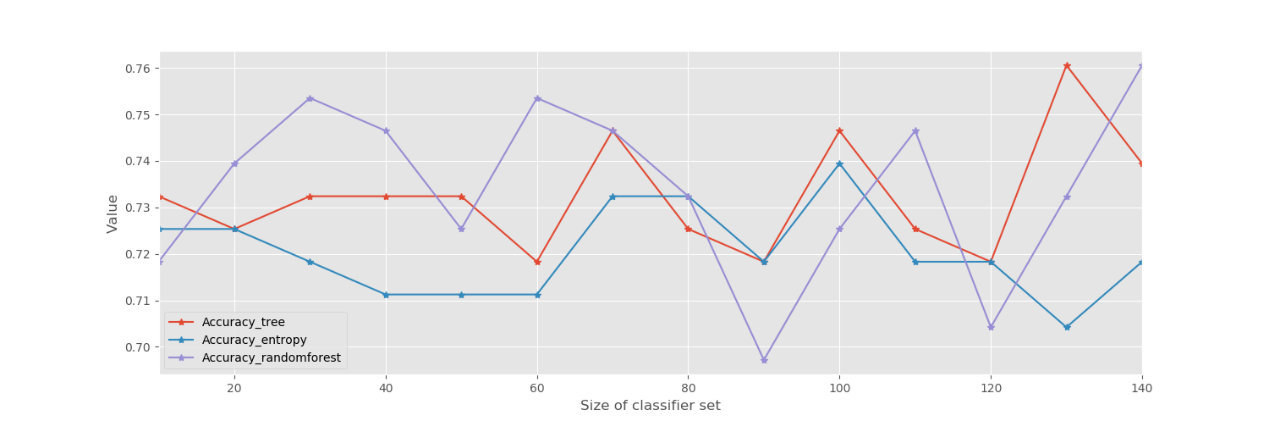


在上图中，发现在随机样本数量在140时，选决策树作为弱分类器的Bagging算法精确度最高，因此将样本数量固定为140，递增弱分类器的数量计算精确度，说明Bagging算法使用决策树作为弱分类器在样本数量不多时的优势较为明显，效率也相对较高

* 1. 随机森林算法的分析：

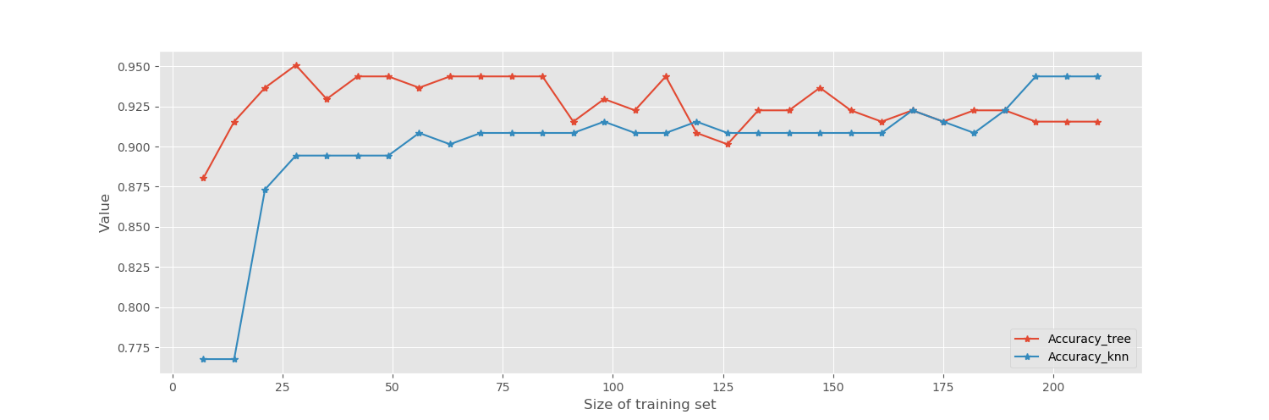


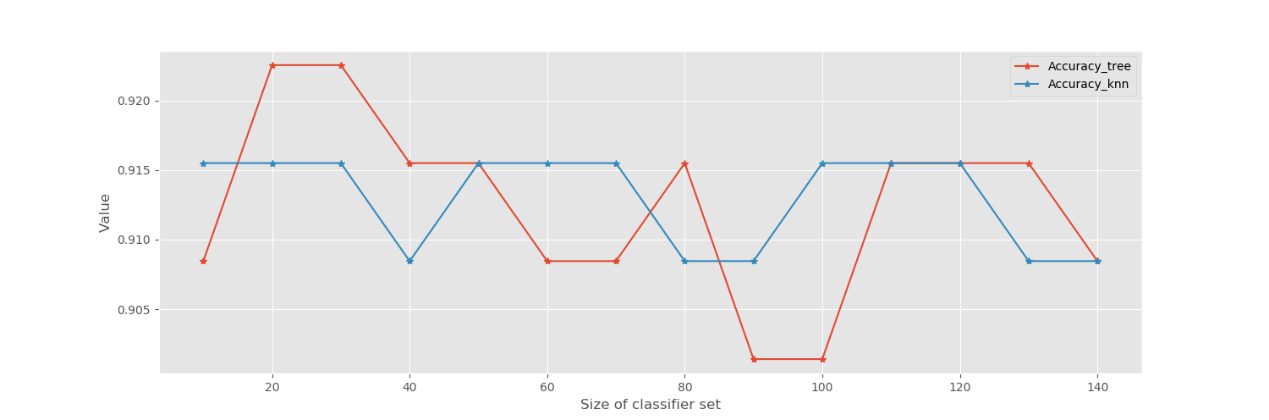
经过查阅资料，在写Bagging算法时使用的默认的决策树已经是根据基尼系数作出的最好分类，因此在实现随机森林时使用的是最大信息增益，二者的差异实在不大，此图为固定了分类器的数量，递增样本量得出的精确度。Scikit-learn的工具包中随机森林的算法与自行实现的bagging算法相比并无明显优势，只是在低样本数量时有较好表现（无实际作用）



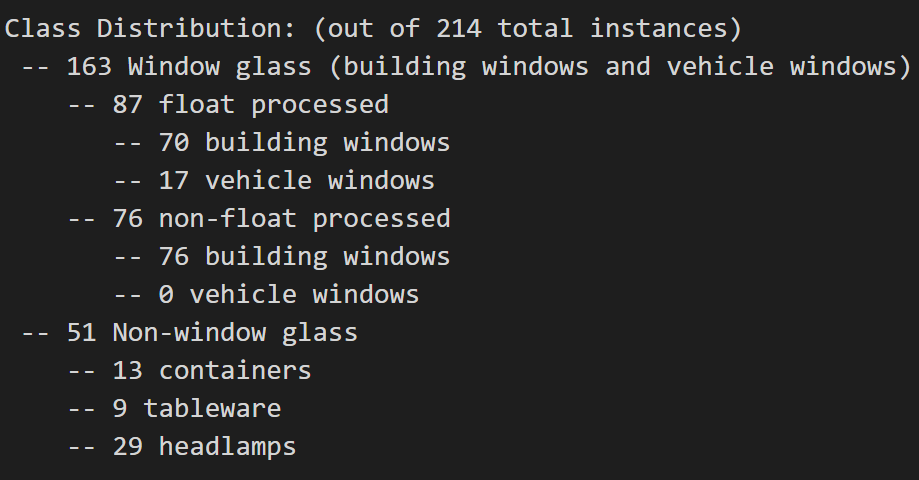
同样的，在固定样本数量为140后，递增弱分类器数量得到了此图。三者的表现无明显差异。

* 1. 将数据标成两类后的Bagging算法：





可以看出精确度有了极大的提升，至少提升了20%，由于随机森林和自己实现的Bagging算法没有太大的差距，因此无需再做比较。数据分为两类后，两个类别之间的属性差异更加明显，因此在分类算法中的精确度极高。

分类依据：

# 第四部分 相关代码

数据预处理：

