**集成方法学习报告**

**闫庆菲**

**数据科学与大数据技术班**

**摘要：**集成方法是通过将多个模型的结果进行综合，获得一个更准确的结果，这样通常可以获得比单个学习模型更好的预测性能。本次学习过程主要学习了随机森林和Adaboost。

**1.引言：**

本次学习过程主要包括对bagging方法的经典模型随机森林模型和boosting方法的经典模型Adaboost模型在不同数据集上的比较使用，以加深对两种方法及其经典模型的理解。

**2.算法：**

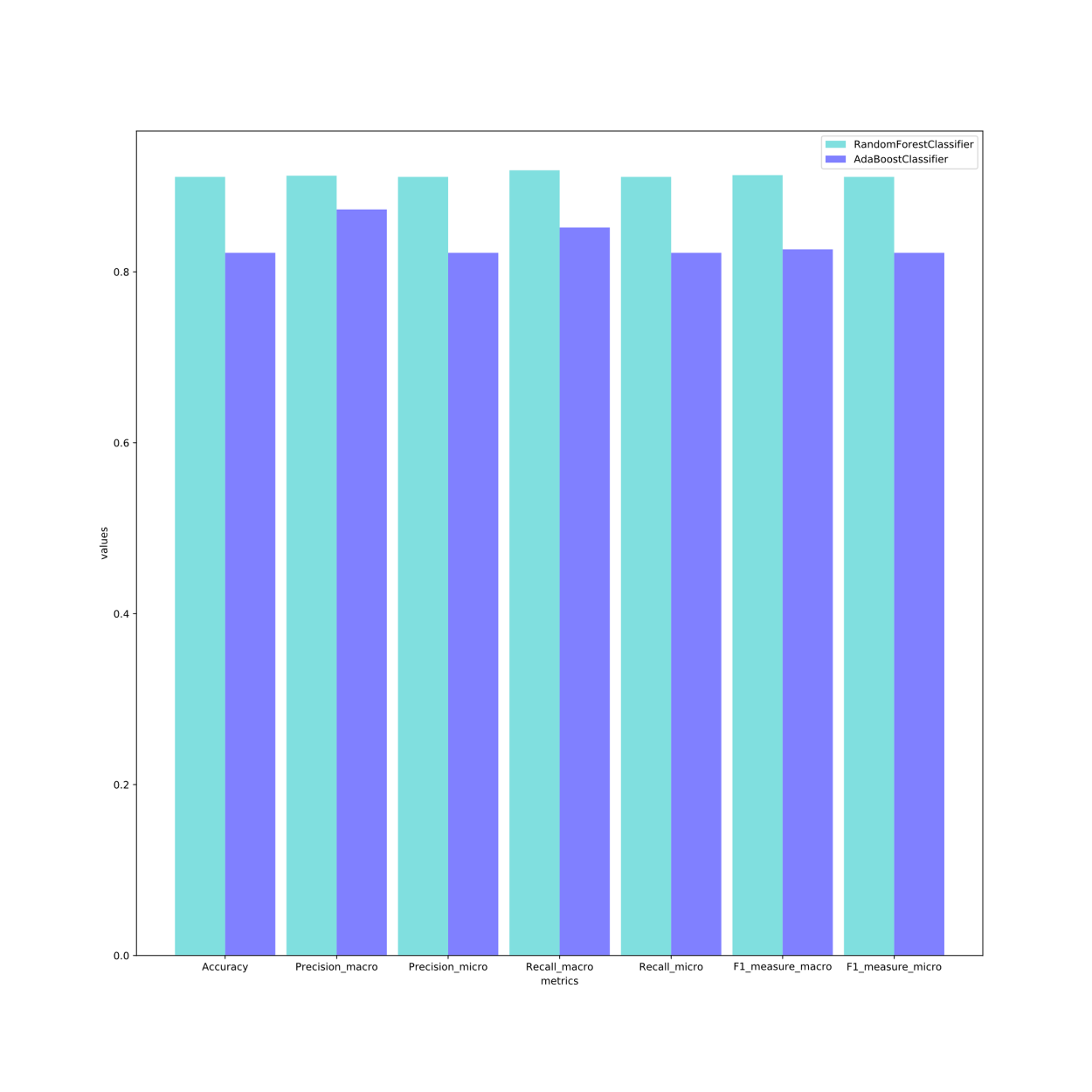
2.1.随机森林：

随机森林模型是基于集成方法Bagging的经典模型。Bagging要求首先从训练数据集中用自主抽样的方法获取多份抽样数据，然后利用抽样的数据为训练集并行训练多个基模型，然后平均所有基模型的预测结果得到集成模型。如果是分类模型，就采用多数投票的方法得到最终的结果，如果是回归模型就将所有模型的结果取平均作为最终的结果。随机森林是最具代表性的Bagging算法，其训练的基模型为决策树。

2.2.Adaboost:  
 Adaboost是Boosting算法的经典模型。Boosting是另一种组合多个基模型从而提升基模型效果的集成方法，不过其采用的是串行训练多个基模型的策略，其基本思想为：从训练数据集中训练得到一个基模型，下一个基模型基于前一个基模型的表现，重点关注前一个基模型分错的数据样本，尽量将前一个基模型分错的数据样本分对，这一操作可以通过提高分错数据的权重、降低分对数据的权重来实现。

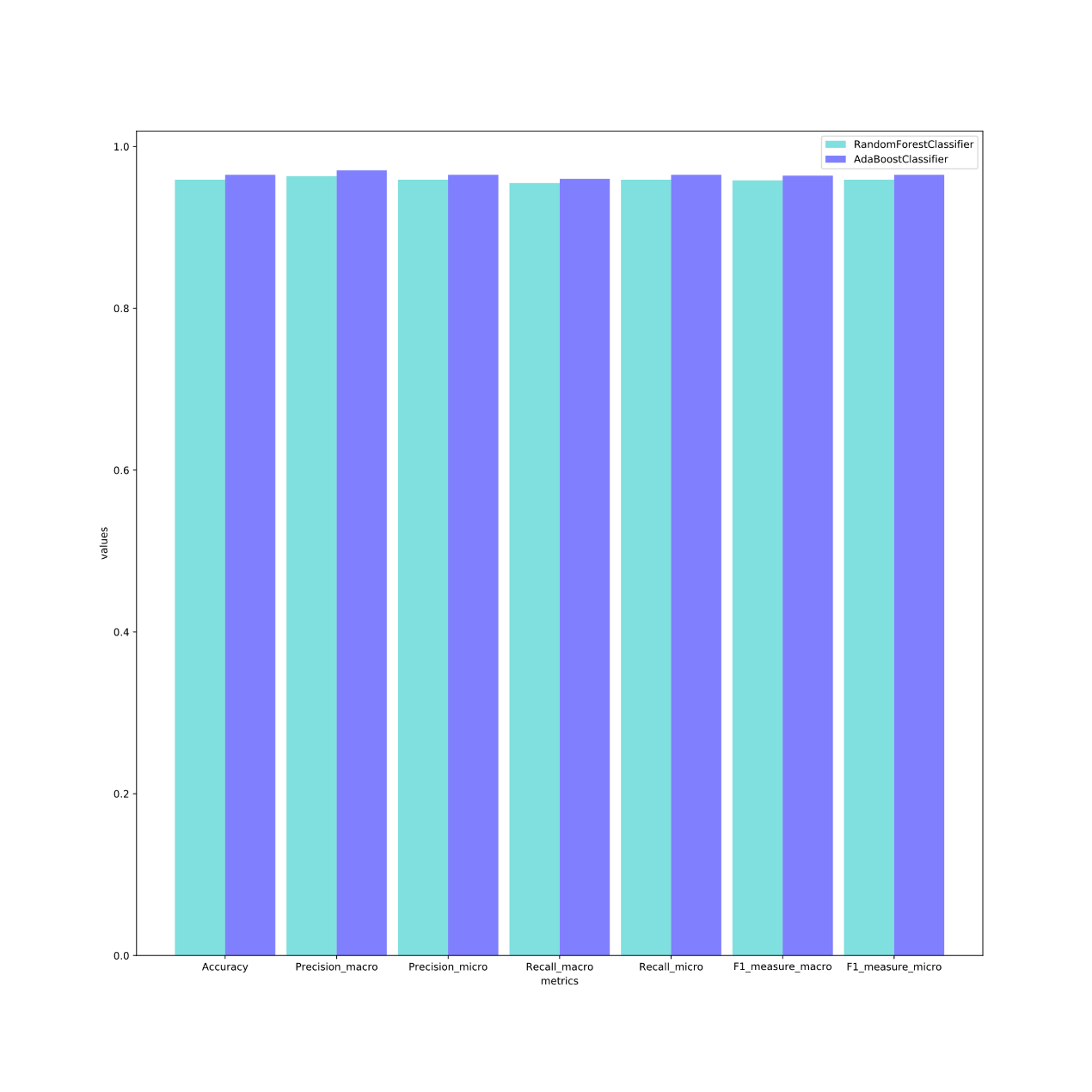
**3.试验及结果分析：**

3.1 分类任务：

随机森林和Adaboost分类模型都分别使用了鸢尾花数据集【1】和乳腺癌数据集【2】进行了训练和

**图3.1-1**

测试。鸢尾花数据集保存在文件D1.data中，其中包含4列连续型属性以及最后一列的类别标签；乳腺癌数据集保存在D2.data文件中，其中包含了32个属性，第一列属性为人员编号，对于模型训练无意义，所以抛弃不用，第二列离散型数据作为类别标签，其他所有的属性均为连续型属性，作为数据样本。训练后经过测试集测试，分别得到了两个模型在不同数据集上测试的性能指标，我用这些指标绘制了图3.1-1和3.1-2：

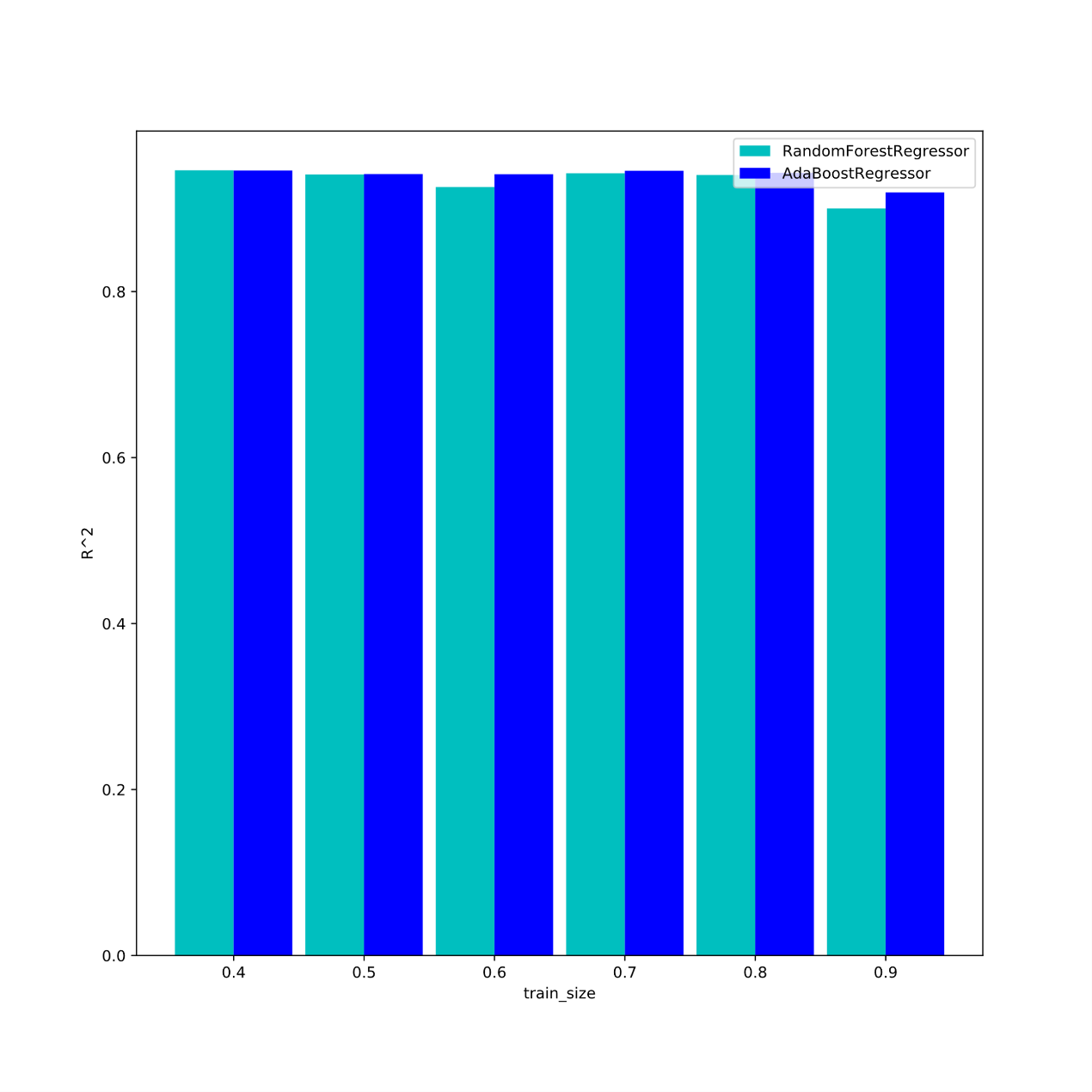


**图3.1-2**

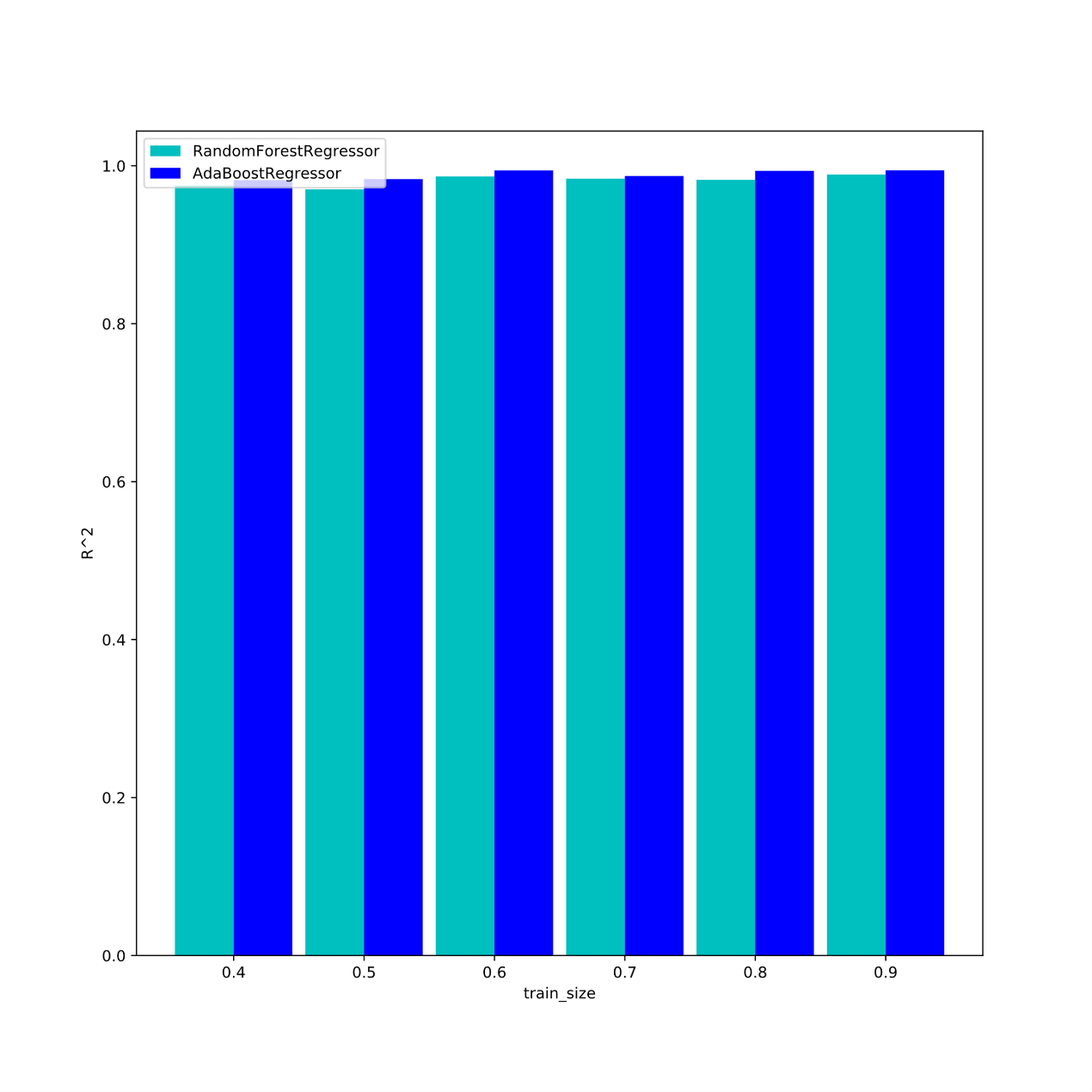
从上图我们可以看出，在鸢尾花数据集上，随机森林分类器的各项指标都要明显好于Adaboost分类器，而在乳腺癌数据集上，Adaboost分类器的各项指标要比随机森林分类器的各项指标略高一些。

3.2 回归任务：

随机森林回归器和Adaboost回归器也用了鸢尾花数据集和乳腺癌数据集，不过为了对比同一回归器在不同大小的训练集训练之后的性能差异，我采用不同的比例划分了测试集和训练集，分别输出了它们的性能值，然后绘制在了图3.2-1和图 3.2-2中：



**图3.2-1**



**图3.2-2**

从上面两张图中我们可以看出，训练集的比重对模型的决定系数影响并不是十分明显，尤其是图3.2-2中，而从图 3.2-1中我们可以看出，训练集占比最大时，决定系数并不是最大，所以在鸢尾花数据集中，并不是训练集越大，模型的效果就越好。

**4.结论：**

**本次工作对于集成方法中的Bagging的经典算法随机森林和Boosting的经典算法Adaboost做了调用，并在不同的数据集上、以及同一数据集不同比例的测试集和训练集上进行训练和测试，对比其性能差异。通过这次工作，加深了我对这两种集成方法的理解以及对随机森林和Adaboost这两种算法的认识。**

**参考文献：**

**【1】**[Index of /ml/machine-learning-databases/iris (uci.edu)](http://archive.ics.uci.edu/ml/machine-learning-databases/iris/)

**【2】**[Index of /ml/machine-learning-databases/breast-cancer-wisconsin (uci.edu)](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/)