Assignment 3 报告

夏一凡

兰州大学信息科学与工程学院

2019 计算机科学与技术 2 班

摘要：通过学习随机森林和集成学习算法，调用 sklearn 库中的相关算法进行应用，并使用 UCI 官网中的鸢尾花(Iris)和乳腺肿瘤(breast-cancer-wisconsin)数据集进行分类和回归实验。使用混淆矩阵、accuracy、recall、precision、f1-score 等评价指标对分类器性能进行评价，并对比两种分类器在同一数据集下的性能。使用均误差方(MSE)、平均绝对误差(MAE)、根均方误差(RMSE)、拟合优度(R^2)四个回归指标对随机森林回归器和 AdaBoost 回归器的性能进行评价，并探究不同大小的训练集对于回归器性能的影响，最后对两种回归器性能进行比较分析。

关键词：随机森林、AdaBoost、分类评价指标、回归性能分析

1.引言

本文将介绍随机森林和 Adaboost 的算法原理，并基于原理利用 python 调用 sklearn 库中的相关算法进行实现，其中也包括对数据进行数据预处理、探索性数据分析以及变量相关性分析等过程，并基于热力图选择较为合适的属性列作为目标属性。

基于处理完成的数据集，对数据集进行划分，并利用训练集对算法进行模型训练，利用测试集对随机森林和 Adaboost 分类模型的混淆矩阵进行绘制，并对其进行分析和解释；通过 accuracy、recall、precision、f1-score 等评价指标对模型进行评价，并比较算法性能方面的差异。然后使用均误差方(MSE)、平均绝对误差(MAE)、根均方误差(RMSE)、拟合优度(R^2)等回归指标对随机森林回归器和 AdaBoost 回归器的性能进行评价，并探究不同大小的训练集对于回归器性能的影响，最后对算法的分类和回归模型进行总结分析。

2.算法

2.1 随机森林算法

2.1.1 随机森林算法[1]

RF 使用了 CART 决策树作为弱学习器，并在使用决策树的基础上，RF 对决策树的建立做了改进，对于普通的决策树，会在节点上所有的 n 个样本特征中选择一个最优的特征来做决策树的左右子树划分，但是 RF 通过随机选择节点上的一部分样本特征，这个数字小于 n，假设为n_{sub}，然后在这些随机选择的n_{sub}个样本特征中，选择一个最优的特征来做决策树的左右子树划分。这样进一步增强了模型的泛化能力。

如果$n_{sub} = n$，则此时 RF 的 CART 决策树和普通的 CART 决策树没有区别。n_{sub}越小，

则模型约健壮，当然此时对于训练集的拟合程度会变差。也就是说n_{sub}越小，模型的方差会减小，但是偏倚会增大。在实际算法中，一般会通过交叉验证调参获取一个合适的n_{sub}的值。

随机森林算法流程如下：

(1) 首先，输入为样本集$\mathrm{D}=\{(\mathrm{x}_1,y_1),(x_2,y_2),\ldots(x_m,y_m)\}$;

(2) 随机地选择训练的数据集和样本特征进行 T 轮训练，其中 t=1,2,3,……,T;

1. 对训练集进行第 t 次随机采样，共采集 m 次，得到包含 m 个样本的采样集D_t；
2. 用采样集D_t训练第 t 个决策树模型$G_t(\mathrm{x})$，在训练决策树模型的节点的时候，在节点上所有的样本特征中选择一部分样本特征， 在这些随机选择的部分样本特征中选择一个最优的特征来做决策树的左右子树划分。

(3) 输出最终的强学习器f(x)

(4) 如果是分类算法预测，则 T 个弱学习器投出最多票数的类别或者类别之一为最终类别。如果是回归算法，T 个弱学习器得到的回归结果进行算术平均得到的值为最终的模型输出。

2.1.2 随机森林算法总结[2]

（1）随机森林算法优点：

1. 训练可以高度并行化，对于大数据时代的大样本训练速度有优势。
2. 由于可以随机选择决策树节点划分特征，这样在样本特征维度很高的时候，仍然能高效的训练模型。
3. 在训练后，可以给出各个特征对于输出的重要性
4. 由于采用了随机采样，训练出的模型的方差小，泛化能力强。
5. 相对于 Boosting 系列的 Adaboost 和 GBDT，随机森林实现比较简单。
6. 对部分特征缺失不敏感。

（2）随机森林算法缺点：

1. 在某些噪音比较大的样本集上，RF 模型容易陷入过拟合。
2. 取值划分比较多的特征容易对 RF 的决策产生更大的影响，从而影响拟合的模型的效果。

2.2 AdaBoost 算法

2.2.1 AdaBoost 分类问题算法原理[3]

输入为样本集$\mathrm{T}=\{(\mathrm{x}_1,y_1),(x_2,y_2),\ldots(x_n,y_n)\}$，弱学习器算法，弱学习器迭代次数 M。
输出为最终的强学习器$f(x)$。

（1）初始化样本集权重为

$$D(1)=(w_{11},w_{12},\ldots w_{1N});w_{1i}=\frac{1}{N};i=1,2\ldots N$$

其中，$D(m)$表示第 m 个弱学习器的样本点的权值

（2）对 M 个弱学习器，m=1,2,3,...M。

1. 使用具有权值分布$D(m)$的训练数据集进行学习，得到基分类器 $G_m(x)$
2. 计算弱分类器$G_m(x)$在训练数据集上的分类误差率e_m

$$e_m=P(G_m(x_i)\neq y_i)=\sum_{i=1}^{N}w_{mi}\,I(G_m(x_i)\neq y_i)$$

3. 计算弱分类器$G_m(x)$的权重系数α_m：

$$\alpha_m = \frac{1}{2} ln \frac{1-e_m}{e_m}$$

4. 更新训练数据集的样本权值分布

$$w_{m+1,i} = \frac{w_{mi}}{Z_m} \exp\left(-\alpha_m y_i G_m(X_i)\right), i = 1,2,3 \dots, N$$

这里，Z_m是规范化因子，将W_{mi}的值规范到 0-1 之间，使得$\sum_{i=1}^{N} W_{mi} = 1$

$$Z_m = \sum_{i=1}^{N} W_{mi} \exp\left(\alpha_m y_i G_m(X_i)\right)$$

（3）利用加权平均法构建基本分类器的线性组合：$f(x) = \sum_{m=1}^{M} \alpha_m G_m$，并构建最终分类器：

$$G(x) = sign\left(f(x)\right) = sign(\sum_{m=1}^{M} \alpha_m G_m)$$

2.2.2 AdaBoost 回归问题算法原理

AdaBoost 回归算法变种很多，下面以 Adaboost R2 回归算法为例介绍回归算法过程。
输入为样本集$\mathrm{T} = \{(\mathrm{x}_1, y_1), (x_2, y_2), \dots (x_m, y_m)\}$，弱学习器算法，弱学习器迭代次数 K。
输出为最终的强学习器$f(x)$。
（1）初始化样本集权重为

$$D(1) = (w_{11}, w_{12}, \dots w_{1m}); w_{1i} = \frac{1}{m}; i = 1,2 \dots m$$

（2）对于 k=1,2,...K

1. 使用具有权重的样本集来训练数据，得到弱学习器$G_k(x)$
2. 计算训练集上的最大误差

$$E_k = \max|\mathrm{y}_i - G_K(x_i)| \mathrm{i}=1,2...\mathrm{m}$$

3. 计算每个样本的相对误差:

- 如果是线性误差，则$e_{ki} = \frac{|y_i - G_k(x_i)|}{E_k}$
- 如果是平方误差，则$e_{ki} = \frac{(y_i - G_k(x_i))^2}{(E_k)^2}$
- 如果是指数误差，则$e_{ki} = 1 - \exp\left(\frac{-|y_i - G_k(x_i)|}{E_k}\right)$

4. 计算回归误差率

$$e_k = \sum_{i=1}^{m} w_{ki} e_{ki}$$

5. 计算弱学习器的系数

$$\alpha_k = \frac{e_k}{1 - e_k}$$

6. 更新样本集的权重分布为

$$w_{k+1,i} = \frac{w_{ki}}{Z_k} \alpha_k^{1-e_{ki}}$$

这里是Z_k规范化因子

（3）构建最终强学习器为：

$$f(x) = G_{k^*}(x)$$

其中，$G_k*(x)$是所有 $ln\frac{1}{\alpha_k}, k = 1,2,\ldots.K$的中位数值对应序号$k^*$对应的弱学习器。

2.2.3 Adaboost 算法的正则化[4]

为了防止 Adaboost 过拟合，我们通常也会加入正则化项，这个正则化项我们通常称为步长。定义为v,对于前面的弱学习器的迭代

$$\mathrm{f}_k(\mathrm{x}) = f_{k-1}(\mathrm{x}) + \alpha_k G_k(\mathrm{x})$$

如果我们加上了正则化项，则有

$$\mathrm{f}_k(\mathrm{x}) = f_{k-1}(\mathrm{x}) + v\alpha_k G_k(\mathrm{x})$$

v的取值范围为$0 \le v \le 1$。对于同样的训练集学习效果，较小的v意味着我们需要更多的弱学习器的迭代次数。通常我们用步长和迭代最大次数一起来决定算法的拟合效果。

2.2.4 Adaboost 算法总结[5]

Adaboost 的主要优点有：

1）Adaboost 作为分类器时，分类精度很高

2）在 Adaboost 的框架下，可以使用各种回归分类模型来构建弱学习器，非常灵活。

3）作为简单的二元分类器时，构造简单，结果可理解。

4）不容易发生过拟合

Adaboost 的主要缺点有：

1） 对异常样本敏感，异常样本在迭代中可能会获得较高的权重，影响最终的强学习器的预测准确性。

3.实验设计和结果分析

3.1 数据集介绍

（1）鸢尾花数据集[6]：是从 UCI dataset repository 中下载的数据集 IRIS，该数据集可用于分类问题，该数据集包括 150 个样本，鸢尾花有三个亚属，分别是山鸢尾（Iris-setosa）、变色鸢尾（Iris-versicolor）和维吉尼亚鸢尾（Iris-virginica）。该数据集一共包含 4 个特征变量，1 个类别变量。特征变量存储了其萼片和花瓣的长宽，共 4 个属性，鸢尾植物分三类，属性列详细信息如表 3.1-1 所示。数据预处理阶段仅对类别标签进行 labelencoder 处理形成最终的类别标签：0，1，2。数据的分布情况如图 3.1-1 所示。

No	属性	数据类型	字段描述
1	sepal_len	Float	花萼长度
2	sepal_width	Float	花萼宽度

No	属性	数据类型	字段描述
3	petal_len	Float	花瓣长度
4	petal_width	Float	花瓣宽度
5	class	String	花卉种类

表 3.1-1 鸢尾花数据集属性列详细信息

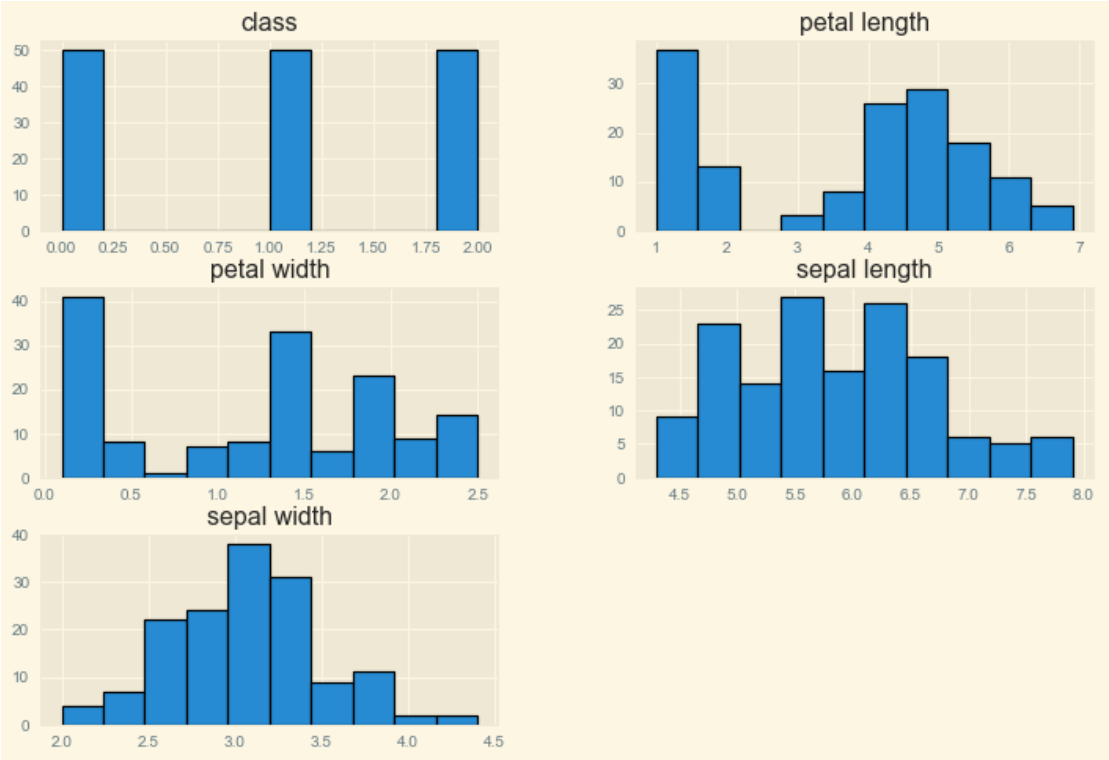

图 3.1-1 鸢尾花数据分布情况

（2）威斯康星乳腺癌数据[7]：是从 UCI dataset repository 中下载数据集 Breast Cancer Wisconsin (Diagnostic)，该数据集可用于分类问题，该数据集包括 699 个样本，每个样本含有 10 个变量，其中缺失值的数量为 16（在“Bare Nuclei”列中），标签的分布情况为: Benign: 458 (65.5%);Malignant: 241 (34.5%),数据集的详细信息如表 3.1-2 所示。并对数据进行数据预处理和探索性数据分析，将分类指标 2 和 4 映射成 0 和 1，并对“Bare Nuclei”列的异常值“？”所在的行进行删除处理，得到最终的数据集 cancer，并通过可视化来查看数据的分布情况，如图 3.1-2 所示。

No	属性	数据类型	字段描述
1	Sample code number	Int	示例代码号
2	Clump Thickness	Int	团厚

No	属性	数据类型	字段描述
3	Cell Size	Int	细胞大小
4	Uniformity of Cell Shap	Int	细胞形状
5	Marginal Adhesion	Int	边缘附着力
6	Single Epithelial Cell Size	Int	单个上皮细胞大小
7	Bare Nuclei	Int	裸核
8	Bland Chromatin	Int	淡色染色质
9	Normal Nucleoli	Int	正常核仁
10	Mitoses	Int	有丝分裂
11	Class	Int	诊断类别

表 3.1-2 乳腺癌数据集属性列详细信息

图 3.1-2 乳腺癌数据集数据分布情况

3.2 实验设计

3.2.1 分类任务设计

该任务采用乳腺肿瘤数据集和鸢尾花数据集。乳腺肿瘤数据集由于数据不平衡，因此在实验过程需按比例分割数据。在对数据集的类别标签进行便签编码后，按照训练集：测试集=7:3 进行训练集和测试集的划分，然后调用 sklearn 库中的随机森林分类器和 AdaBoost 分类器，用训练集训练模型，然后用测试集进行分类预测，输出预测结果的混淆矩阵、accuracy、recall、precision、f1 等评价指标，并对两个模型在相同数据集上的分类性能进行评价分析。

3.2.2 回归任务设计

该任务采用乳腺肿瘤数据集和鸢尾花数据集。在忽略两个数据集中的类别属性后，通过绘制数据集的热力图分析属性之间的相关性，如图 3.2-1 和图 3.2-2 所示，以乳腺肿瘤数据集为例，通过观察热力图我们可以发现，属性列 Uniformity of Cell Size 与其他属性的相关性

较强，因此选择第二列属性作为回归目标属性，且第一列和第 9 列数据与回归目标属性的相关性较弱，因此对两列不予考虑，得到最终属性列。然后分别对数据集按照训练集：测试集=0.2，0.3….0.8 进行训练集、测试集的划分，分别训练随机森林和 AdaBoost 回归器，考察不同大小的训练集对于回归器性能影响，并分别用 MSE、MAE、RMSE、R^2 四个指标评价两种回归器的性能。

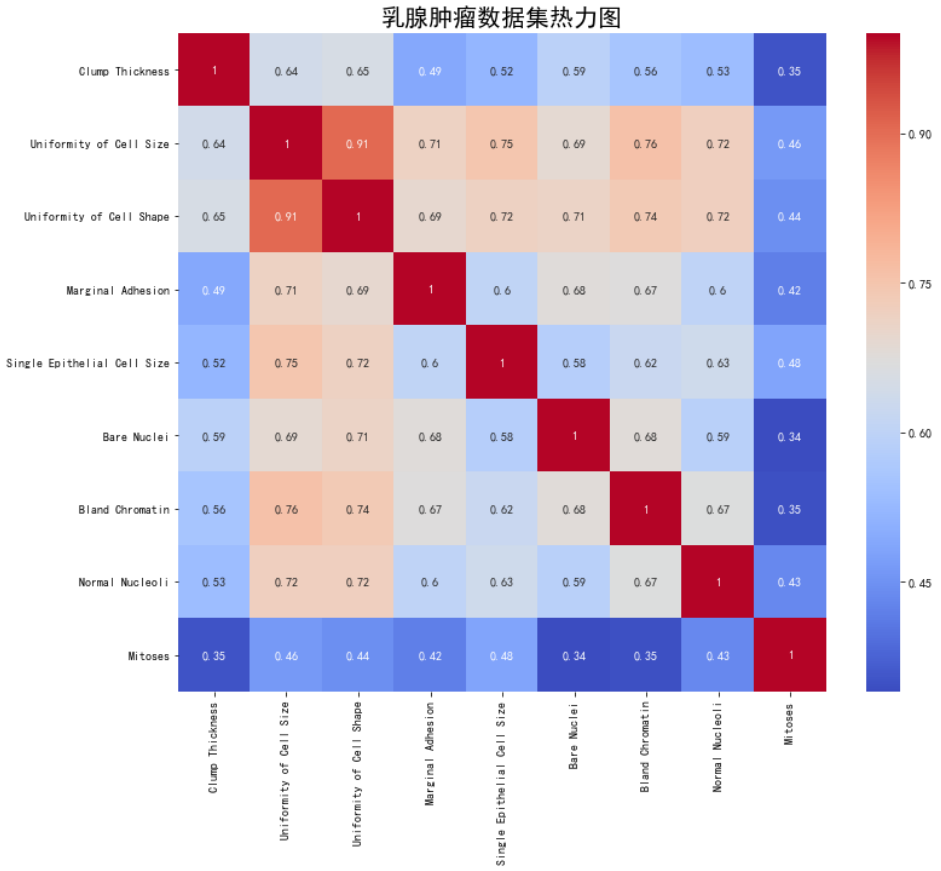

图 3.2-1 乳腺癌数据集热力图

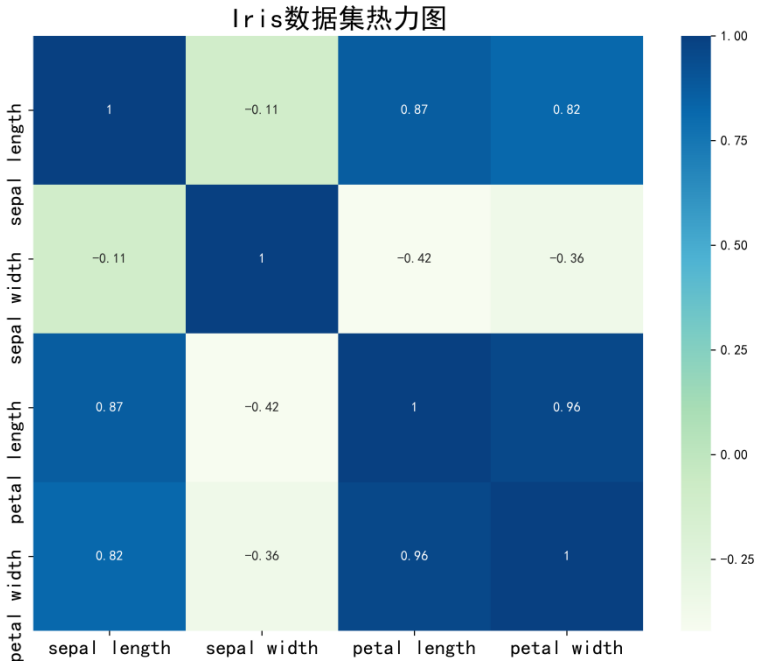

图 3.2-1 鸢尾花数据集热力图

3.3 实验结果分析

3.3.1 分类实验结果分析

3.3.1.1 乳腺肿瘤数据集结果分析

在乳腺肿瘤数据集下分别使用随机森林和 AdaBoost 分类器进行分类预测后得到的混淆矩阵（图 3-3-1、图 3-3-3）和评价指标（图 3-3-2、图 3-3-4）如下，以随机森林分类器为例 ，观察绘制的混淆矩阵，可以发现：在主对角线上的元素均为预测正确的类别，且通过颜色可以看出该模型的分类准确率较高，我们可以知道预测值为 良性 的样本中有 124 个为良性，3 个为 恶性；基于此我们可以计算一些混淆矩阵的模型参数，以准确率为例：准确率= TP +FN/ (TP + FP+TN+FN)=（124+70）/(124+70+6+3) = 95.57%,与我们在评价指标中所看到的一致，通过混淆矩阵能够帮助我们分析每个类别的误分类情况，从而帮助我们分析调整。[8]

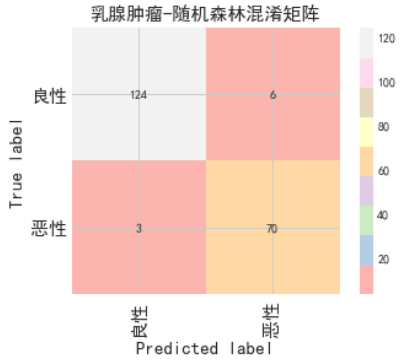

图 3-3-1：乳腺肿瘤随机森林分类器下混淆矩阵

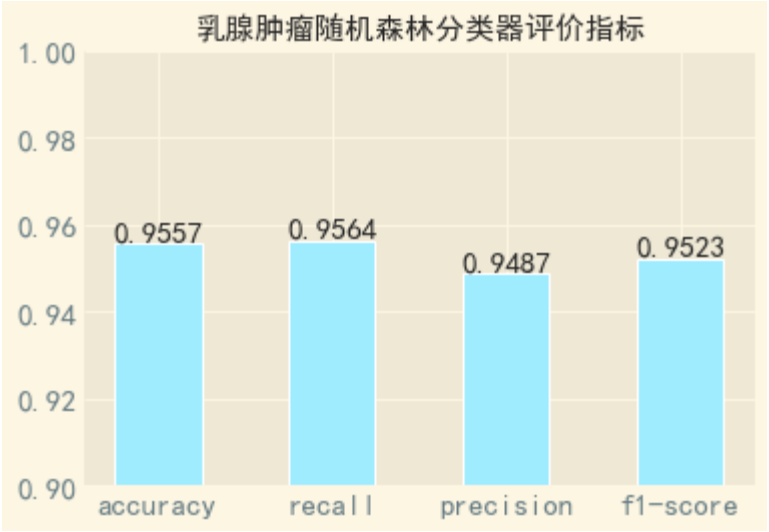

图 3-3-2：乳腺肿瘤随机森林分类器下评价指标

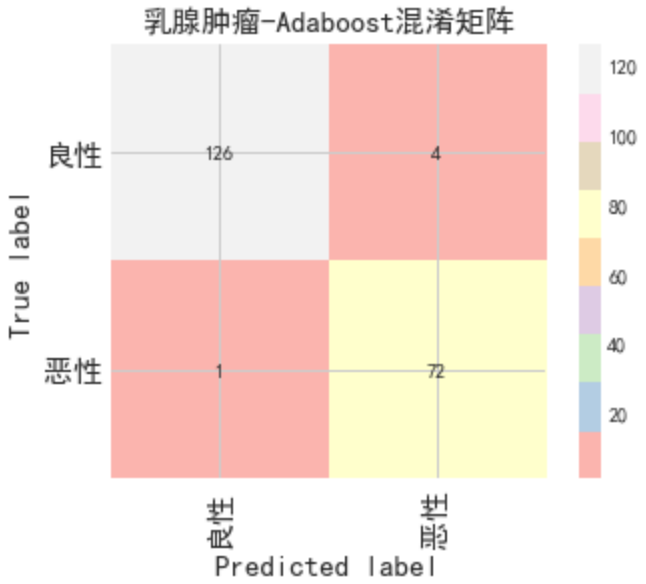

图 3-3-3：乳腺肿瘤 AdaBoost 分类器下混淆矩阵

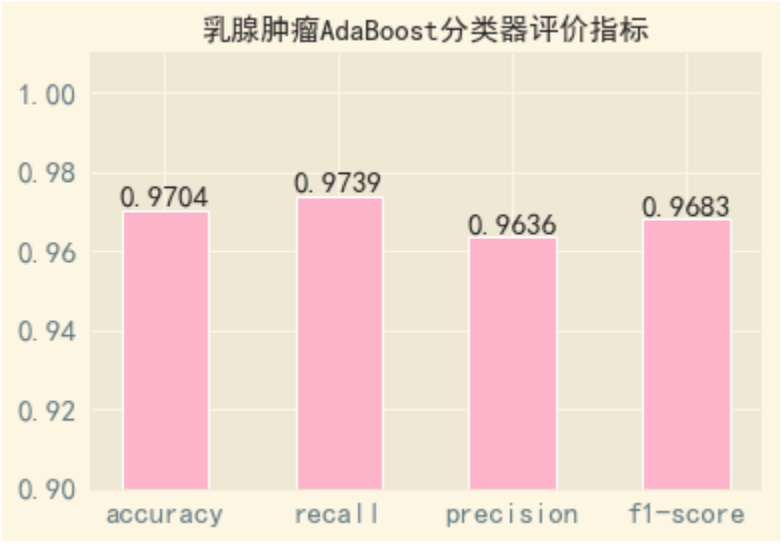

图 3-3-4：乳腺肿瘤 AdaBoost 分类器下评价指标

从上面的四幅图可以看出，乳腺肿瘤数据集在两种分类器下的分类效果都比较好，准确率都高于 95%，且通过混淆矩阵可以看出，假阴性情况较少，有恶性肿瘤的情况基本都能被确诊出。

3.3.1.2 鸢尾花数据集结果分析

在鸢尾花数据集下分别使用随机森林和 AdaBoost 分类器进行分类预测后得到的混淆矩阵（图 3-3-5、图 3-3-7）和评价指标（图 3-3-6、图 3-3-8）如下，通过观察可以发现，两种算法对于鸢尾花数据集的分类效果均较好，但是均对变色鸢尾类别标签的一个样本预测为维吉尼亚鸢尾，除此之外，随机森林模型还误分了一个样本，基于该次数据集的划分，Adaboost 的分类效果略优于随机森林模型。

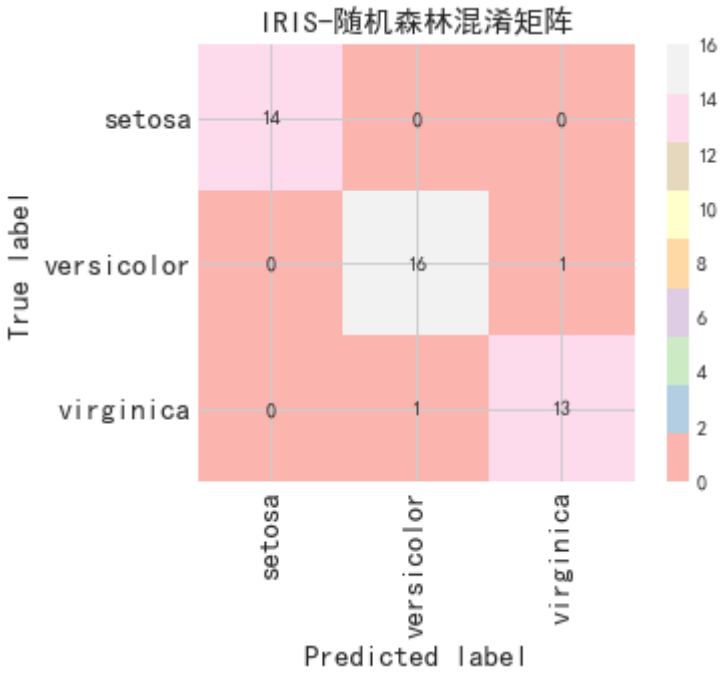

图 3-3-5：鸢尾花随机森林分类器下混淆矩阵

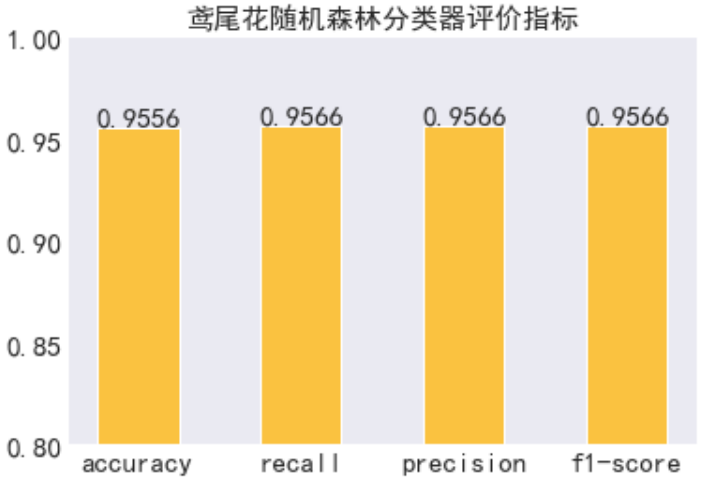

图 3-3-6：鸢尾花随机森林分类器下评价指标

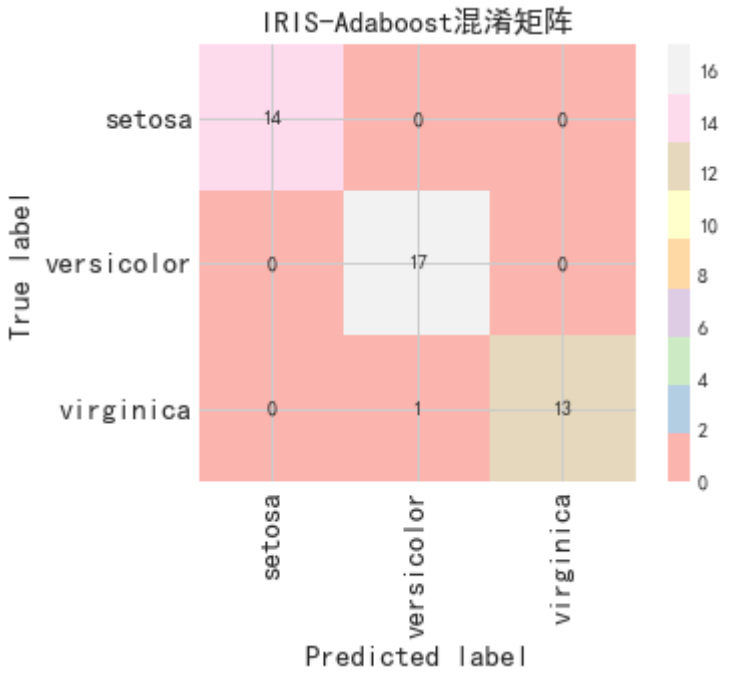

图 3-3-7：鸢尾花 AdaBoost 分类器下混淆矩阵

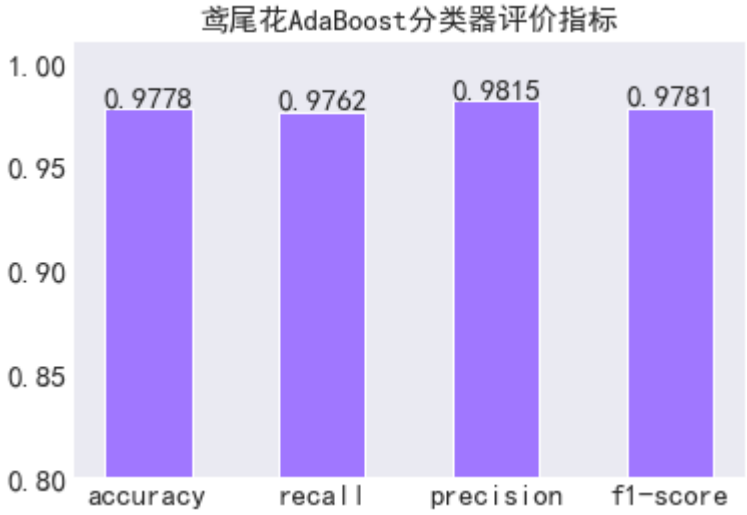

图 3-3-8：鸢尾花 AdaBoost 分类器下评价指标

3.3.1.3 算法分类性能比较

乳腺肿瘤和鸢尾花数据集在两种分类器下的评价指标对比如图 3-3-9 和图 3-3-10 所示。从评价指标图中可以看出，在相同的数据集下，使用相同训练集和测试集，AdaBoost 分类器的分类效果普遍优于随机森林。这与 Adaboost 在迭代过程不断提高分错样本的权重，使得错误率足够小时停止迭代。因此在分类效果上 Adaboost 明显优于随机森林。

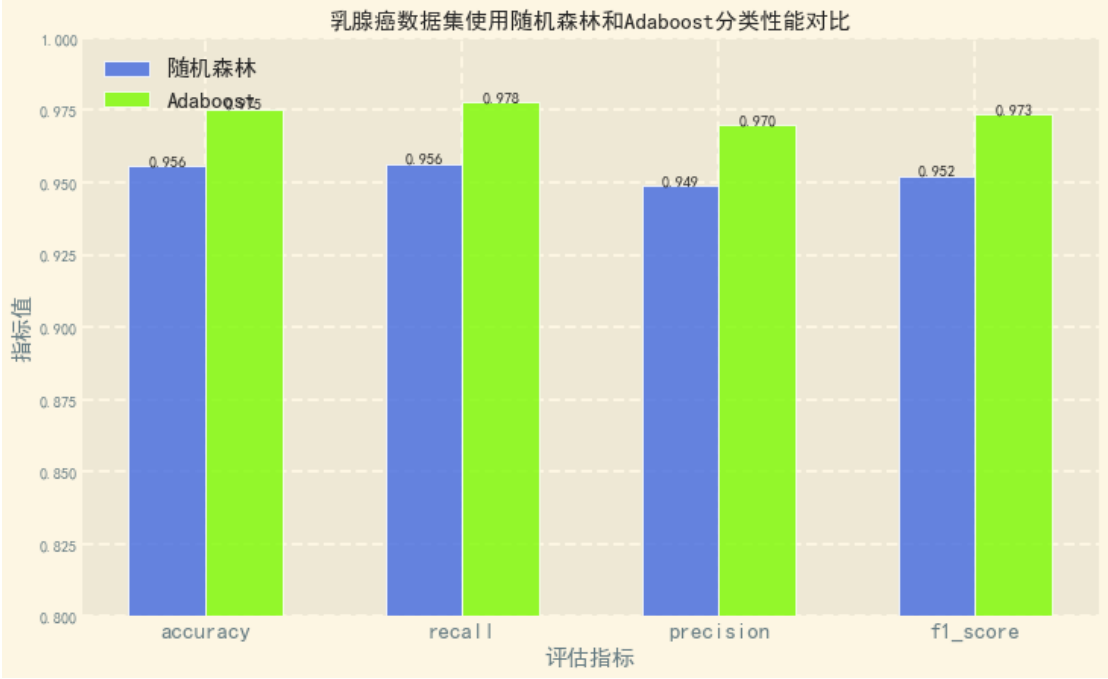

图 3-3-9：乳腺肿瘤使用两种分类器的性能对比

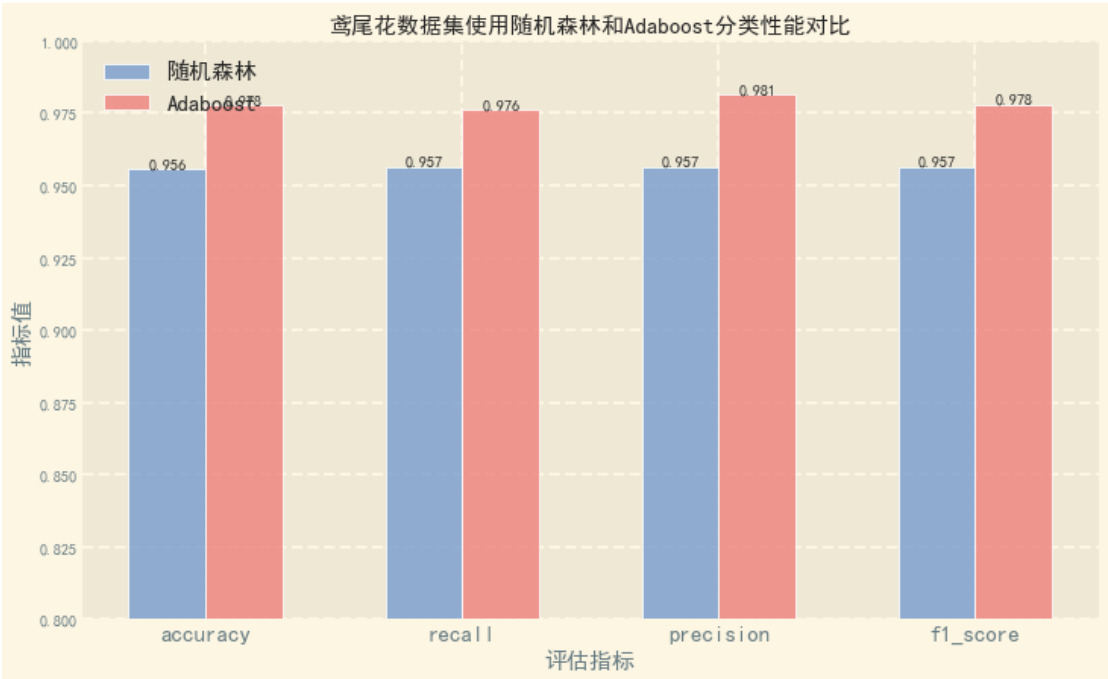

图 3-3-10：鸢尾花使用两种分类器的性能对比

3.3.2 回归实验结果分析

3.3.2.1 乳腺肿瘤数据集分析

在乳腺肿瘤数据集下分别使用随机森林和 AdaBoost 回归器在不同大小训练集下进行回归预测后得到的评价指标（图 3-3-11、图 3-3-12）如下，从图中可以看出，两种回归器对于

乳腺肿瘤中细胞大小的回归效果比较好，拟合优度达到 0.85。此外，随着训练集大小的增大，回归器性能一般出现先提高后降低的趋势，这是由于随着训练集的增大，回归器训练更加充分，但是训练集过大，容易出现过拟合，使得预测结果可能出现较大偏差。对于随机森林回归器，最佳训练集大小为数据集的 60%；对于 AdaBoost 回归器，最佳训练集大小分别出现在数据集大小的 40%时，此时可能出现欠拟合的现象。

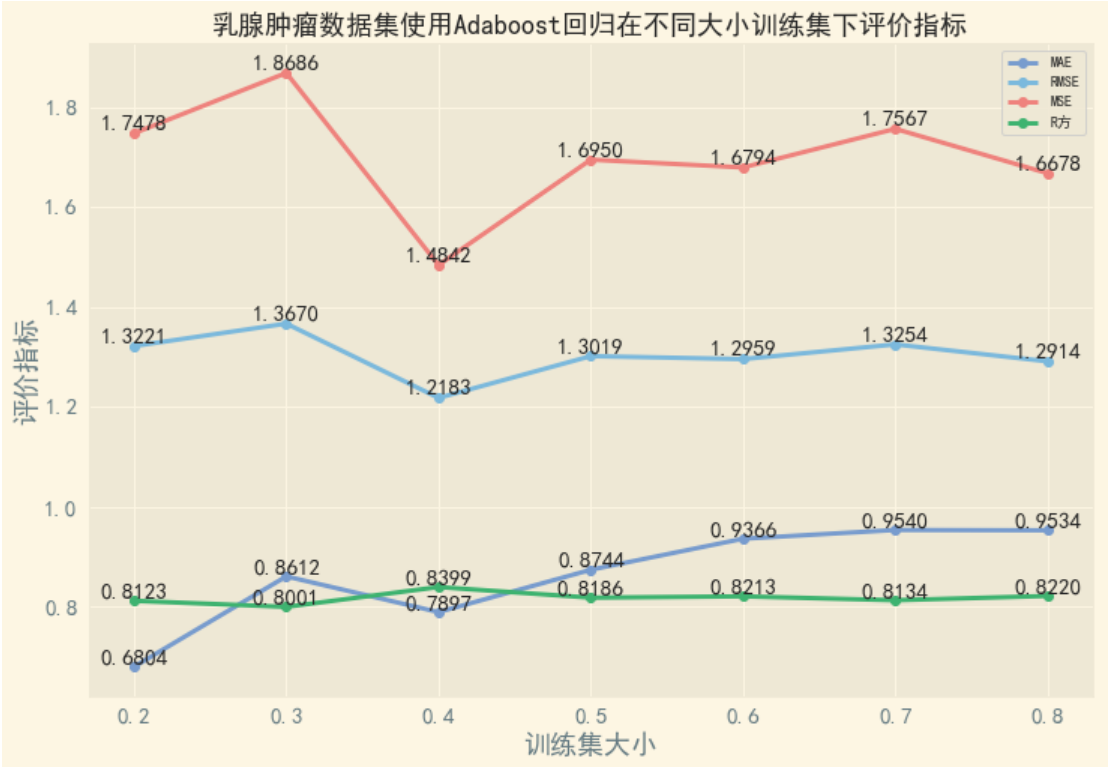

图 3-3-11：乳腺肿瘤 AdaBoost 回归器评价指标

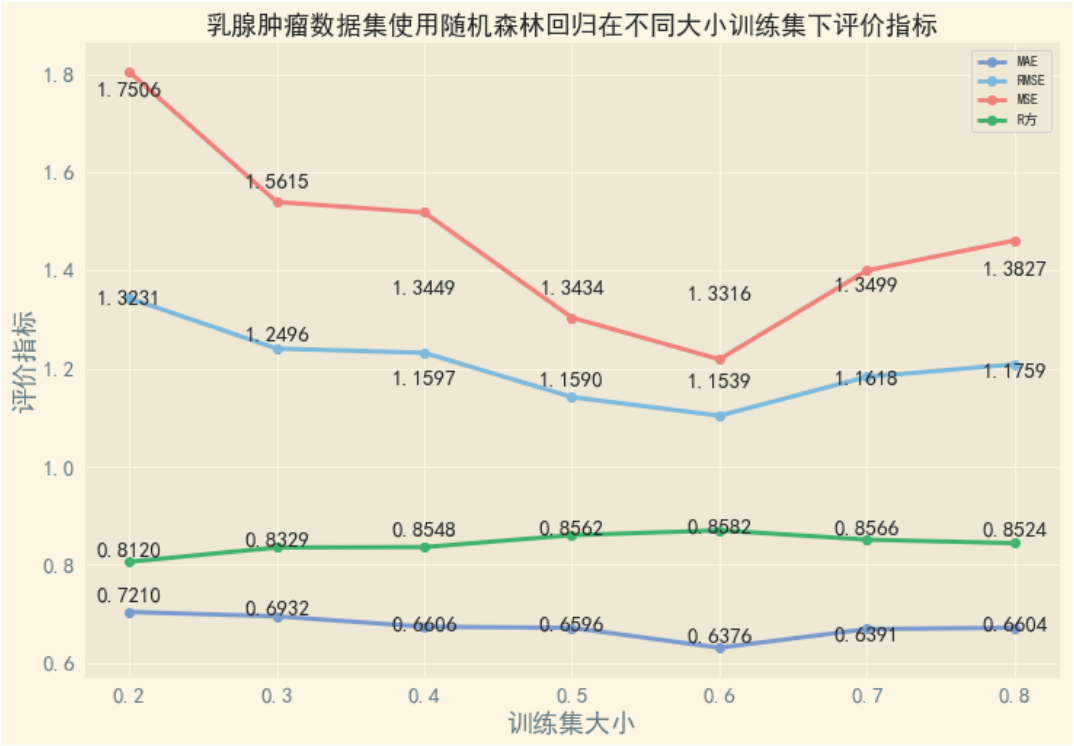

图 3-3-12：乳腺肿瘤随机森林回归器评价指标

3.3.2.1 鸢尾花数据集分析

在鸢尾花数据集下分别使用随机森林和 AdaBoost 回归器在不同大小训练集下进行回归预测后得到的评价指标（图 3-3-13、图 3-3-14）如下，由评价指标图可以看出，对于鸢尾花数据集 petal length 回归效果特别好，拟合优度能达到 0.96，这是由于鸢尾花数据集中 sepal length, petal width 与目标属性的相关性较强，使得预测效果特别好。当训练集增大到数据集的 80%，回归器性能达到最优，但是可能会出现过拟合的现象。

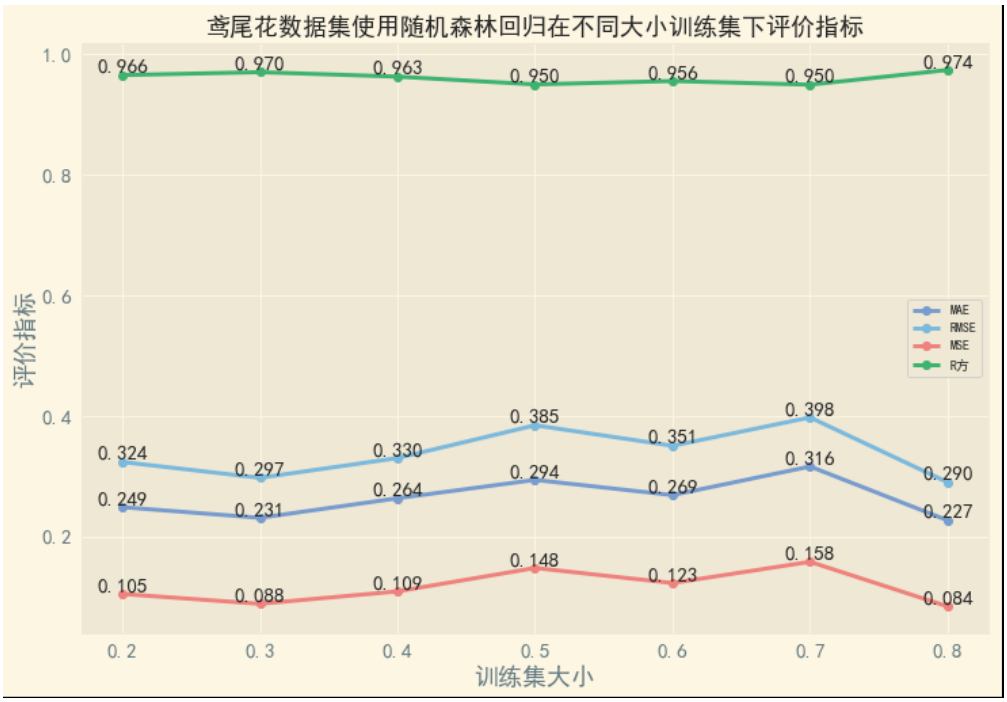

图 3-3-13：鸢尾花数据集随机森林回归器评价指标

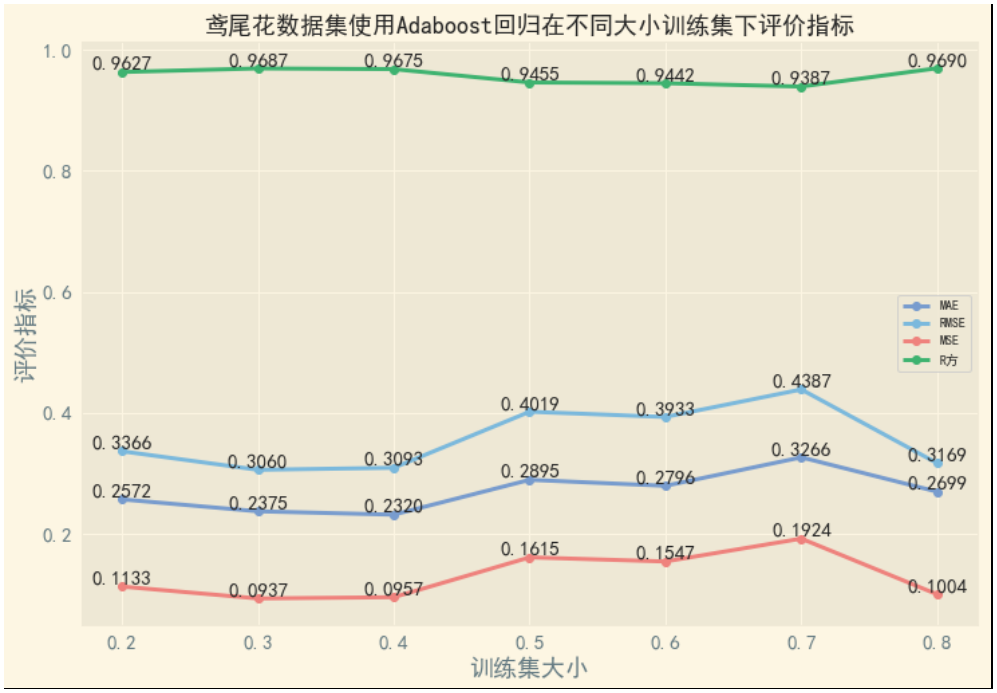

图 3-3-14：鸢尾花数据集 AdaBoost 回归器评价指标

3.3.1.3 算法回归性能比较

乳腺肿瘤和鸢尾花数据集在两种回归器下的评价指标对比如图 3-3-15 和图 3-3-16 所示。从图中可以看出，在乳腺肿瘤数据集下，使用相同的训练集和测试集，随机森林回归器的回归效果优于 AdaBoost 回归器。而在鸢尾花数据集下，使用相同训练集和测试集，两个回归器效果相近，AdaBoost 略优于随机森林回归器。说明，乳腺肿瘤数据集而言，随机森林回归器的性能更高，预测效果更好。

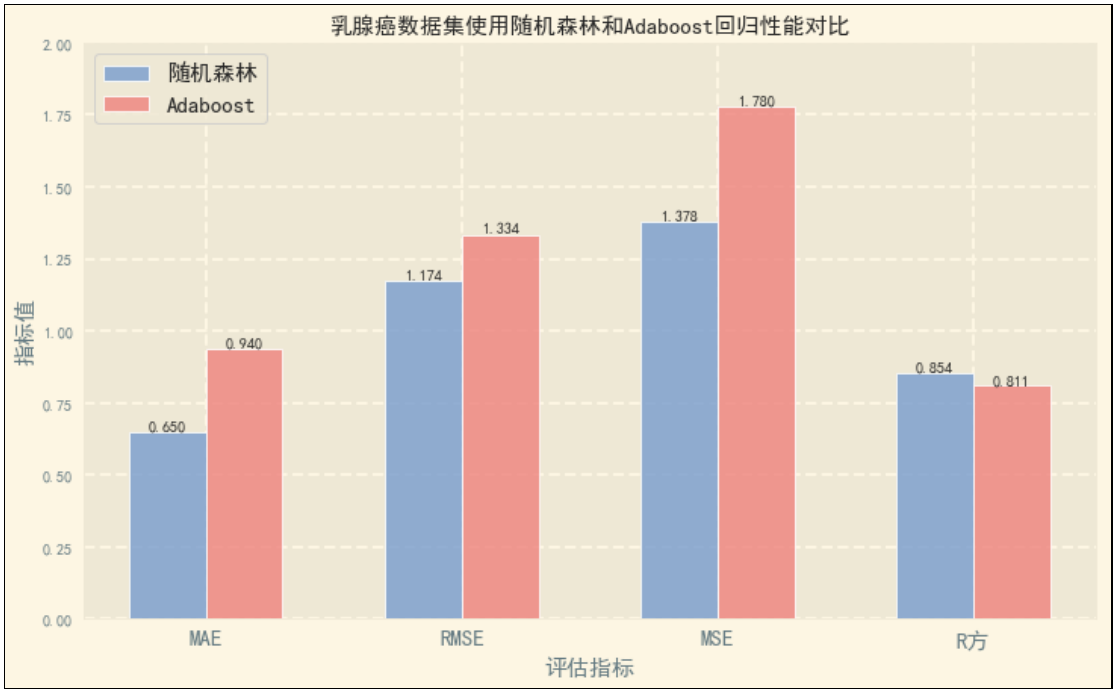

图 3-3-15：乳腺肿瘤使用两种回归器的性能对比

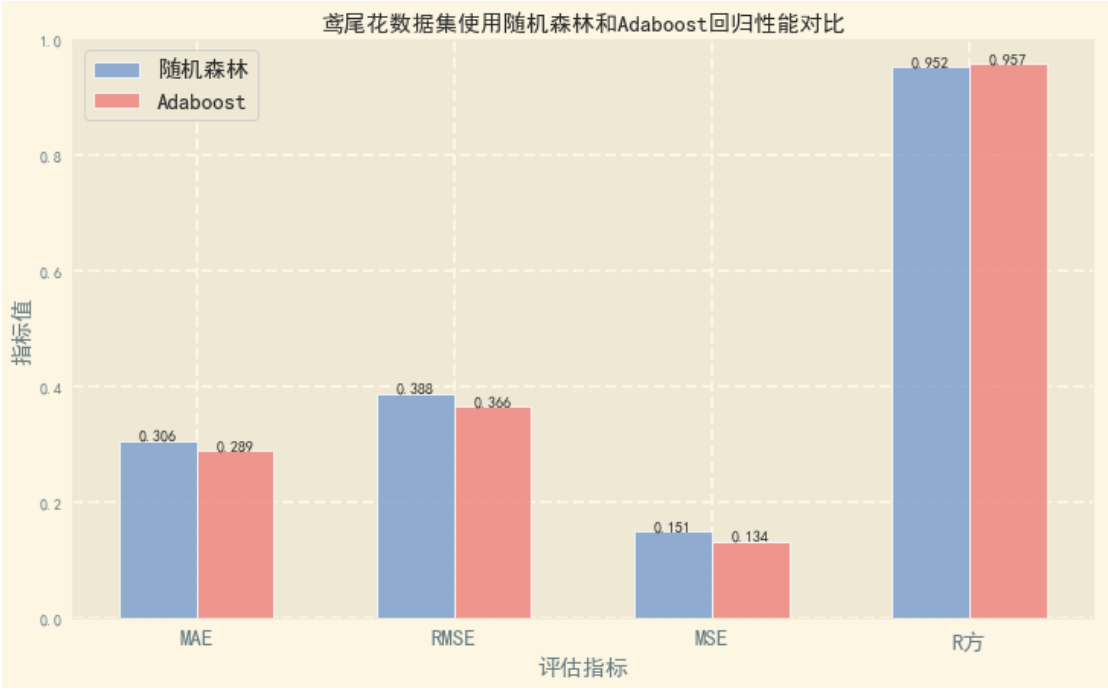

图 3-3-16：鸢尾花数据使用两种回归器的性能对比

4.结论与感悟

4.1 分类任务实验结论分析

在分类实验中，我们绘制了混淆矩阵，并基于混淆矩阵进行模型相关参数的计算，并通过混淆矩阵帮助分析每个类别的误分类情况，并发现假阴性情况较少，有恶性肿瘤的情况基本都能被确诊出。然后在算法比较分析中，可以看出在相同的数据集下，使用相同训练集和测试集，AdaBoost 分类器的分类效果普遍优于随机森林，这与 Adaboost 在迭代过程不断提高分错样本的权重，使得错误率足够小时停止迭代。因此在分类效果上 Adaboost 明显优于随机森林。

4.2 回归任务实验结论分析

在回归分析实验中，我们通过使用均误差方(MSE)、平均绝对误差(MAE)、根均方误差(RMSE)、拟合优度(R^2)四个回归指标对随机森林回归器和 AdaBoost 回归器的性能进行评价，发现随着训练集大小的增大，回归器性能一般出现先提高后降低的趋势，这是由于随着训练集的增大，回归器训练更加充分，模型性能更好，但是当训练集过大，容易出现过拟合现象，使得预测结果可能出现较大偏差。因此在实验过程中应该选择合适的训练集比例。

4.3 探索性数据分析和数据预处理

在本次实验中，探索性数据分析和数据预处理显得十分重要，通过探索性数据分析，发现乳腺癌数据集中的“Bare Nuclei”列具有“？”的异常值，而在数据预处理中，通过 labelencoder 方法对数据集的类别便签进行标签编码，再通过变量相关性分析选择相关性较强的列作为回归目标属性进行模型的训练。[9]

5.参考文献

[1]俞孙泽.对随机森林算法的优化改进的分析[J].中国新通信,2020,22(13):126.

[2] https://www.cnblogs.com/pinard/p/6156009.html .Bagging 与随机森林算法原理小结

[3]Yang Wan Qi,Lu Xin,Li Fu Sheng,Zhao Yan Chun. Study on LOD of Trace Elements by XRF Analysis Using BP & Adaboost and PLS Methods[J]. Advances in Science and Technology,2021,6258.

[4]都承华,龚谊承,张冬阳.基于三种回归器和 VotingRegressor 优化 Adaboost 的血糖集成预测[J].中国卫生统计,2021,38(02):254-256+261.

[5]https://www.cnblogs.com/pinard/p/6133937.html 集成学习之 Adaboost 算法原理小结

[6] 鸢尾花数据集 http://archive.ics.uci.edu/ml/machine-learning-databases/iris

[7] 乳腺癌数据集 http://archive.ics.uci.edu/ml/machine-learningdatabases/breast-cancer-wisconsin/

[8]《机器学习实战：基于 Scikit-Learn 和 TensorFlow》,ISBN-978-7-111-60302-3

[9]张瑞全. 基于偏最小二乘和灰色关联分析的时序预测研究[D].大连理工大学,2017.