基于 Gradient Boosting(梯度提升)算法分类男女性别

摘 要

本实验旨在通过身高和体重两项数据实现对男女性别的准确分类。采用名为 Training_setdata 的数据集，包含 Gender、Height 和 Weight 三项数据。我们使用 Gradient Boosting(梯度提升) 算法建立分类模型，通过调整参数网络（n_estimators、learning_rate、max_depth、min_samples_split、min_samples_leaf）进行网格搜索，找出最佳参数。实验过程中，我们比较了 Gradient Boosting 模型与其他机器学习模型，如 KNN、Logistic 回归、Random Forest、SVC、Bayes-best、Decision Tree 以及深度学习模型（DNN）。

实验结果显示，Gradient Boosting(梯度提升) 算法在精确度上略高于其他模型，但在召回率、F1 分数、混淆矩阵、交叉验证分数以及 ROC-AUC 曲线等方面表现优异。具体而言，模型的召回率达到了 92.97%，F1 分数为 0.9252，表明模型在准确率和召回率之间取得了良好的平衡。ROC-AUC 曲线的面积达到 0.97，显示模型在正例和负例的区分上非常出色。

综合来看，Gradient Boosting(梯度提升) 算法在性别分类问题上表现卓越，为相关领域的性别识别任务提供了一个可行的解决方案。

关键词： Gradient Boosting(梯度提升)算法

一、问题重述

1.1 问题提出

通过身高和体重两项数据实现对男女性别的准确分类。

二、问题分析

2.1 数据集预处理

使用一个名为 Training_setdata 的数据文件，包含 Gender,Height,Weight 三项数据。

2.2 问题的分析

三、模型假设

1. 数据准确无误差。
2. 不考虑极端个例。

四、符号说明

符号	定义
Gender	性别
Height	身高（cm）
Weight	体重（kg）
GradientBoostingClassifier	梯度提升算法模型对象
param_grid	参数网络
n_estimators	弱学习器
learning_rate	弱学习器贡献的参数
max_depth	决策树的最大深度
min_samples_split	拆分内部节点所需的最少样本数
min_samples_leaf	叶节点处需要的最小样本数
accuracy	精确度

五、模型建立与求解

5.1 实验步骤

1. 加载并预处理数据，得到训练集和测试集，绘制数据集对应的散点图(图1)，以便直观观察数据进而选择合适的模型。

2. 创建GradientBoostingClassifier模型对象，并进行网格搜索，找出最佳参数，进行多次实验。

3. 训练和预测模型，计算和打印评价指标，绘制和显示 ROC 曲线。

4.与其他机器学习模型对比得出结论。

5.2.1 数据预处理

对 Gender 使用标注将男性标注为 1，女性标注为 0。使用 pandas 的 read_csv 函数读取 CSV 文件，并从 data 中提取'Height'和'Weight'作为特征，存储在 X 中，从 data 中提取'Gender'作为标签，存储在 y 中。使用 train_test_split 函数将数据划分为训练集和测试集，测试集的大小为 30%。

并对训练集文件绘制散点图以便直观展示数据如图 1。

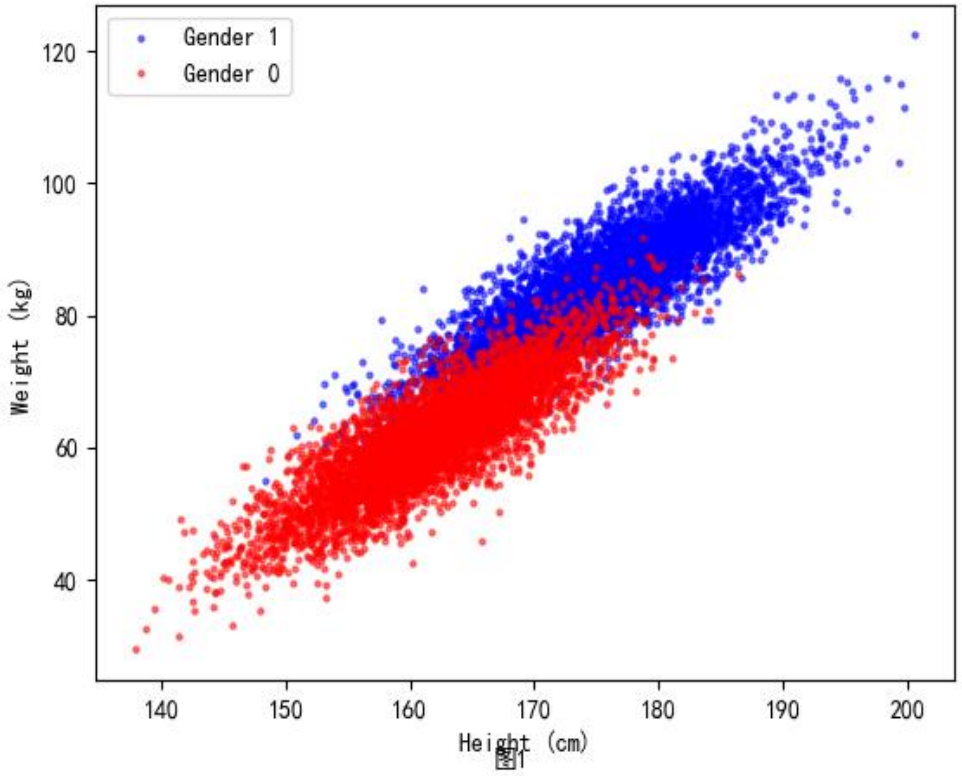


图1

5.2.2 性别与身高，体重建立二分类模型

5.2.2.1 Gradient Boosting(梯度提升)算法的建立

回归创建 GradientBoostingClassifier 模型对象，并定义一个字典 param_grid 存放 n_estimators,learning_rate,max_depth,min_samples_split, min_samples_leaf 五项参数，使其成为 Gradient Boosting(梯度提升)算法的参数网络，并创建一个 GridSearchCV 对象，用于在参数网格上进行网格搜索，找出最佳参数将其存储在存储在 gb_best 中。使用 gb_best.predict(X_test)方法以使用最佳模型对测试集进行预测，并将结果存储在 y_pred_best 中。

5.2.2.2 评价指标

使用 sklearn.metrics 模块中的 accuracy_score 方法计算模型 accuracy; recall_score 方法计算模型召回率；f1_score 方法计算 F1 分数；confusion_matrix 方法生成混淆矩阵；roc_curve 和 auc 方法绘制 ROC 曲线和 AUC 值，最后使用 sklearn.model_selection 模块的 cross_val_score 方法进行交差验证检查是否欠拟合或过拟合。

5.2.2.3 Gradient Boosting(梯度提升)算法的评估

准确率（Accuracy）：这是模型预测正确的样本占总样本的比例。模型的准确率为 0.923，这意味着模型预测正确的概率为 92.3%。

召回率（Recall）：这是模型正确预测的正例占所有实际正例的比例。模型的召回率为 0.9297，这意味着模型能够找出 92.97%的实际正例。

F1 分数（F1 Score）：这是准确率和召回率的调和平均值，用于同时考虑准确率和召回率。模型的 F1 分数为 0.9252，说明模型在准确率和召回率之间达到了良好的平衡。

混淆矩阵（Confusion Matrix）：用于描述模型的性能。模型的混淆矩阵显示，模型正确预测了 1340 个负例和 1429 个正例，同时错误预测了 123 个负例和 108 个正例，准确度很高。

所绘制的 ROC 曲线和 AUC 值如图 2，ROC 曲线的面积 CAUC=0.97，意味着你的模型在将正例和负例区分开来方面做得非常好，几乎接近完美。表明模型的性能非常优秀。

交叉验证分数（Cross-validation scores）：这是在不同的数据子集上评估模型性能的一种方法。模型的交叉验证分数在 0.907 到 0.924 之间，这说明模型在不同的数据子集上的性能相对稳定，没有出现过拟合或欠拟合的情况。

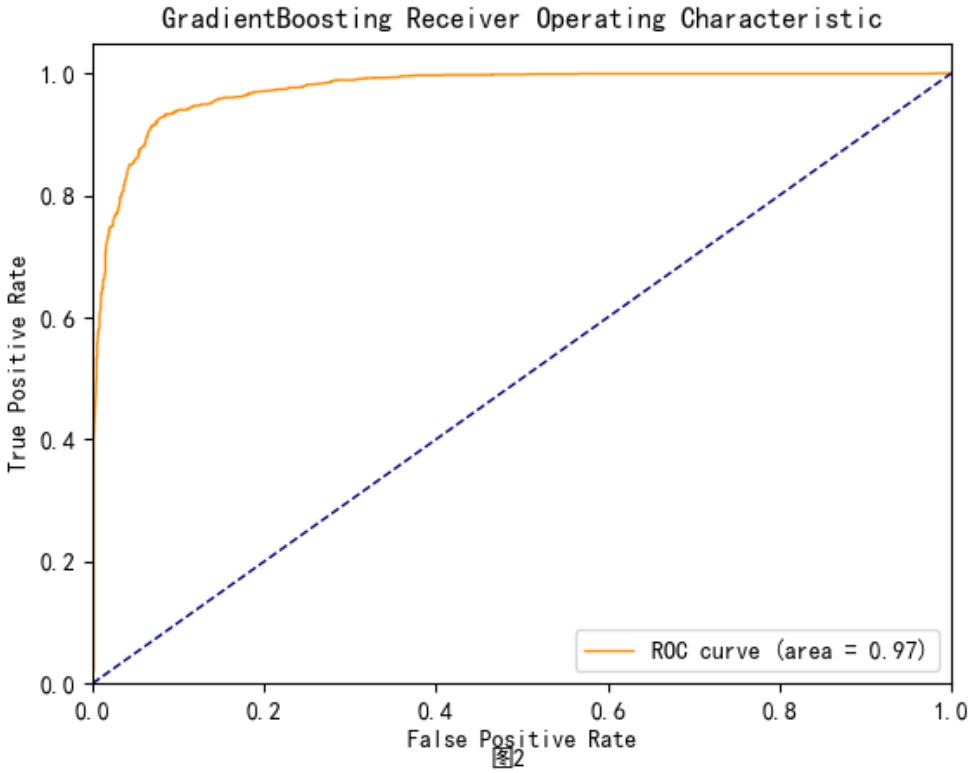


图2

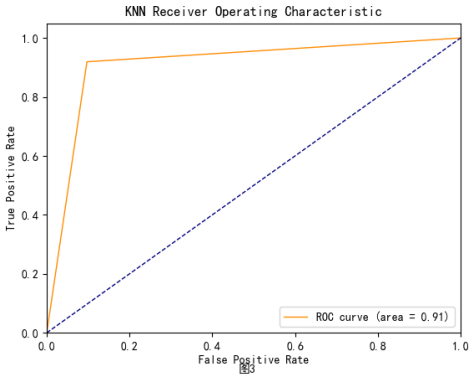
5.2.2.4 模型对比

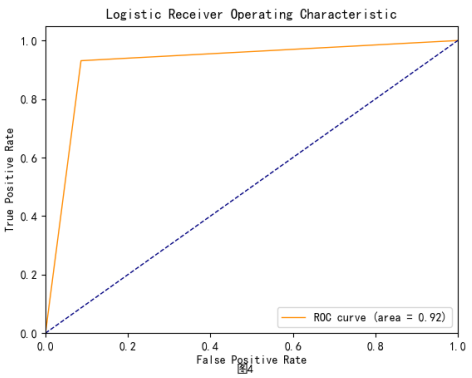
本测试对所有模型使用了相同的训练集 Training_setdata.csv，对 Gender 使用标注将男性标注为 1，女性标注为 0。使用 train_test_split 函数将数据划分为训练集和测试集，测试集的大小为 30%。对比结果如表 1。不同模型的 ROC-AUC 曲线图如图 3 至图 9 所示。

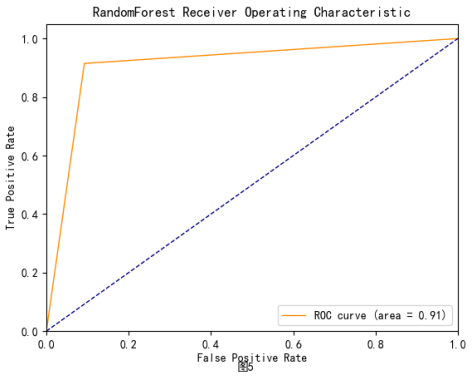
表 1: 不同模型对比

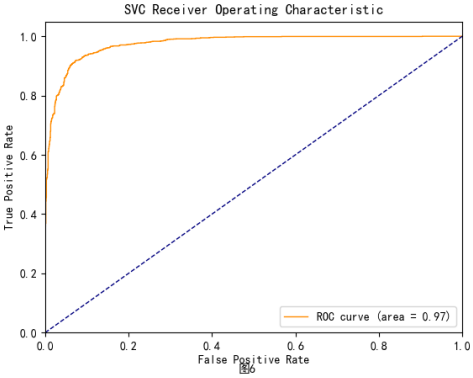
Model	Accuracy	Recall	F1	Cross-validation	Confusion Matrix	AUC
boosting	0.923	0.9297	0.925	[[1340 123] [108 1429]]	[0.9115 0.924 0.907 0.921 0.9235]	0.97
KNN	0.911	0.919	0.913	[[1321 142] [124 1413]]	[0.9 0.904 0.8995 0.8985 0.902]	0.91
Logistic	0.922	0.931	0.924	[[1337 126] [106 1431]]	[0.917 0.925 0.912 0.919 0.9245]	0.92
Random Forest	0.911	0.910	0.913	[[1334 129] [137 1400]]	[0.901 0.913 0.902 0.903 0.909]	0.91
SVC	0.919	0.924	0.921	[[1337 126] [116 1421]]	[0.9125 0.926 0.91 0.914 0.9255]	0.97
Bayes-best	0.916	0.938	0.919	[[1305 158] [94 1443]]	[0.9115 0.9225 0.909 0.9145 0.915]	0.92
Decision Tree	0.876	0.876	0.878	[[1281 182] [190 1347]]	[0.8645 0.88 0.864 0.88 0.8725]	0.88
DNN	0.9757	0.94	0.94	[[4737 263] [639 4361]]	/	0.91

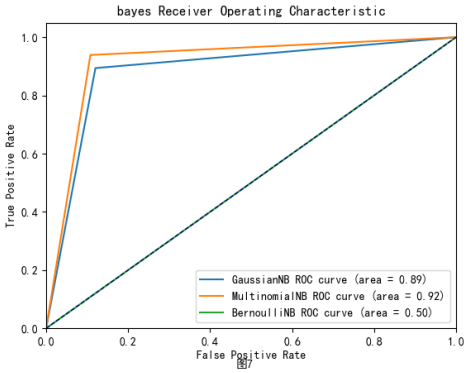
注：使用红色标注的模型为本实验所使用模型，使用蓝色标注的模型为深度学习模型。朴素贝叶斯算法选择表现最佳的。

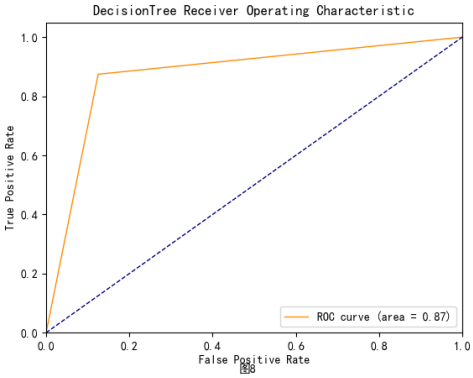

KNN Receiver Operating Characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.91)


Logistic Receiver Operating Characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.92)


RandomForest Receiver Operating Characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.91)


SVC Receiver Operating Characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.97)


bayes Receiver Operating Characteristic
True Positive Rate
False Positive Rate
GaussianNB ROC curve (area = 0.89)
MultinomialNB ROC curve (area = 0.92)
BernoulliNB ROC curve (area = 0.50)


DecisionTree Receiver Operating Characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.87)

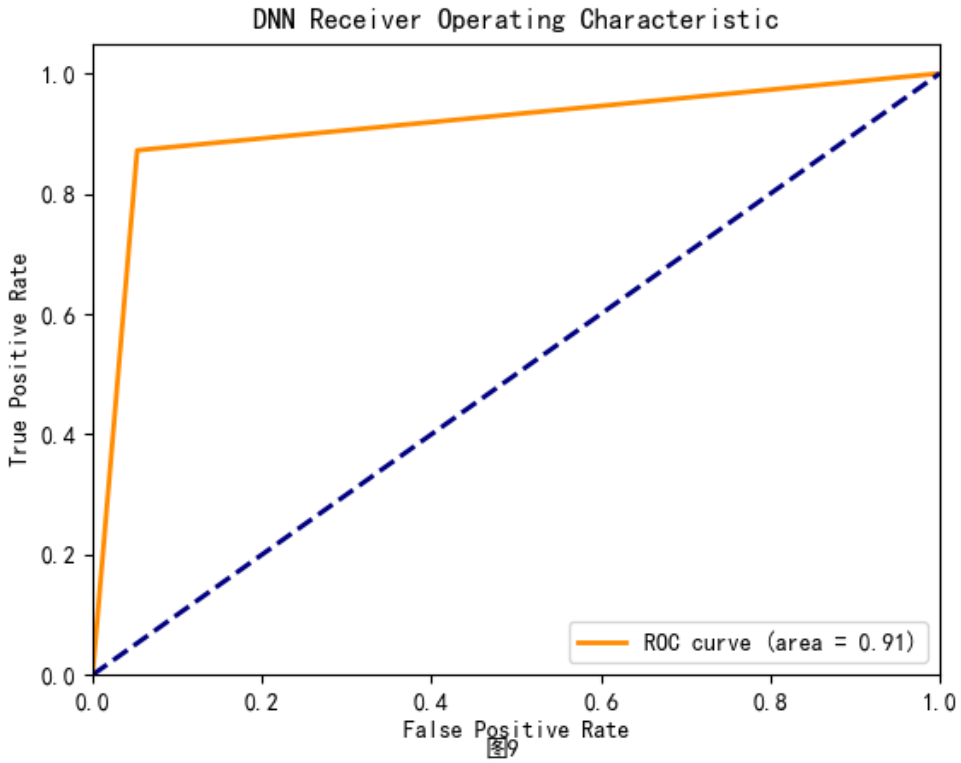


图9

由实验数据表明 Gradient Boosting(梯度提升)算法精确度略高于其他所实验的机器学习模型，低于 DNN（深度神经网络）。但召回率，F1 分数，混淆矩阵，交叉验证分数以及 ROC-AUC 曲线均优于其他模型，表明模型性能非常优秀。

5.2.2.5 实验结论与体会

通过本次实验,我掌握了 Gradient Boosting(梯度提升)算法的原理和应用，熟悉了 Gradient Boosting(梯度提升)算法的优化问题和求解方法，使用 Gradient Boosting(梯度提升)算法对人的性别进行了分类，使用参数网络找出生成最优测试模型，评估了 Gradient Boosting(梯度提升)算法的分类性能，使用 Accuracy，Recall，F1，Cross-validation，Confusion Matrix 等指标，并绘制了 ROC-AUC 曲线。通过本次实验，我发现 Gradient Boosting(梯度提升)算法是一种强大的分类模型，它可以处理线性不可分与非线性不可分的数据，但需要根据数据集的特点进行调整和优化。通过本次实验，我也发现 Gradient Boosting(梯度提升)算法预排序过程的空间复杂度过高，不仅需要存储特征值，还需要存储特征对应样本的梯度统计值的索引，相当于消耗了两倍的内存训练时间,导致训练会非常长,因此需要对数据进行降维,以提高 Gradient Boosting(梯度提升)算法的效率和准确性。

六、参考文献

[1]任宇博.基于 Boosting 的设计模式识别方法研究[D].大连海事大学,2022.DOI:10.26989/d.cnki.gdlhu.2022.000144

[1]章敏.基于平方 Hinge 损失的梯度 Boosting 算法研究[D].江西师范大学,2021.DOI:10.27178/d.cnki.gjxsu.2021.001385

[1]陈强.基于局部相似性的多类 Boosting 分类方法研究[D].河南师范大学,2021.DOI:10.27118/d.cnki.ghesu.2021.000433

七、附录

源代码仓库地址：https://github.com/YUZHEthefool/Gender-Classification-Based-on-Gradient-Boosting-Algorithm.git
你可以保存至本地并复现此实验

author：thefool