|  |  |
| --- | --- |
| 班级：数大20-2  学号：2020305267  2020305297  姓名：彭睿、王子亮 | 实验课程：机器学习  实验名称：集成学习的对比分析实验  实验时间：2023-4-7 |
| 实验内容 | 使用bagging、GBDT和RF算法，分别对iris数据、breast\_cancer以及Boston房价案例进行分析，撰写课程报告，详细分析这三类算法的预测准确性，并与基准算法cart算法进行比较，在同一个图中绘制性能指标曲线。 |
| 实验代码 | 已上传”最优参数.html”、”默认参数.html”、”调参测试.html”，代码、结果以及必要的注释包含在内。 |
| 运行结果及分析 | 为了更好的理解本次实验的结果，在撰写实验报告前先提出以下七个问题：   1. **对三个数据集的理解（数据的类型：离散、连续）** 2. **三个不同的数据集是使用Classifier还是regressor（预测方法的选择）** 3. **用什么来判断分类或回归方法的准确性（准确率accuracy函数还是r^2函数，他们的意义是什么）** 4. **对数据集进行一个初步的训练得出准确性（使用默认的训练参数）** 5. **调整每个训练器的参数，如何调整使得每个训练器的训练结果最优，提高模型的性能表现** 6. **实验中的过拟合问题（交叉检验集与叶子节点数量、分类树深度、剪枝方法、最小样本量）** 7. **得出结论，四个不同的算法（CART、BAGGING、GDBT、RF)算法对三个数据集来说，运用哪一种算法进行训练得出的训练结果准确率最高。**   **数据集的理解**  Iris数据集包含了三种不同的鸢尾花（山鸢尾、变色鸢尾、维吉尼亚鸢尾）的50个样本，每种鸢尾花的四个特征（萼片长度、萼片宽度、花瓣长度、花瓣宽度）均有测量值。  Breast\_cancer数据集包含了诊断乳腺肿瘤的结果（良性、恶性）以及一些医学属性，如肿瘤大小、均匀度等，共计30个属性。  Boston房价数据集是一个回归问题，包含了14个变量，其中13个是用于描述房屋和周围环境的因素（如房间数量、犯罪率、房屋年龄等），最后一个是房屋价格的中位数。  **使用Classifier还是Regressor**  Iris数据集和Breast\_cancer数据集是分类问题，使用的是Classifier；Boston房价数据集是回归问题，使用的是Regressor。  **准确性度量指标**  对于分类问题，我们使用准确率（Accuracy）作为性能度量指标，其意义是分类器正确分类的样本数与总样本数之比；对于回归问题，我们使用R^2（R-Square）作为性能度量指标，其意义是模型拟合数据的程度，其取值范围在0到1之间，越接近1表示模型的预测能力越强。  初步训练得出准确性  **默认参数结论**  我们使用默认的训练参数对三个数据集进行了初步的训练，得到如下的准确性：（结果保留两位小数）  Iris数据集：  CART算法：准确率为1；  Bagging算法：准确率为1；  GBDT算法：准确率为1；  RF算法：准确率为1。  Breast\_cancer数据集：  CART算法：准确率为0.91；  Bagging算法：准确率为0.95；  GBDT算法：准确率为0.97；  RF算法：准确率为0.96。  Boston房价数据集：  CART算法：R^2为0.59；  Bagging算法：R^2为0.75；  GBDT算法：R^2为0.74；  RF算法：R^2为0.79。  IMG_256  **以数据集为X轴的准确率散点折线图**  **调整训练器参数**  为了得到更好的性能表现，我们对每个训练器的参数进行了调整。具体调整的方法如下：  对于CART算法，我们使用熵的增量来界定节点分裂的准则，对（splitter）修改为random，在选择属性是采用随机的策略，避免过度拟合。）min\_samples\_split）设置为21避免过度拟合  对于Bagging算法，我们增加了基分类器的数量（n\_estimators），从而提高了模型的稳定性和准确性。  对于GBDT算法，我们调整了学习率（learning\_rate）和基分类器的数量（n\_estimators），从而提高了模型的泛化能力和准确性。  对于RF算法，我们调整了基分类器的数量（n\_estimators）、最大特征数（max\_features）和最大深度（max\_depth），从而提高了模型的准确性和鲁棒性。  **我们是42组，所以所有的random\_state参数都为42**  经过参数调整，我们得到了如下的准确性：（结果保留两位小数）  Iris数据集：  CART算法：准确率为1.00；  Bagging算法：准确率为1.00；  GBDT算法：准确率为1.00；  RF算法：准确率为1.00。  Breast\_cancer数据集：  CART算法：准确率为0.97；  Bagging算法：准确率为0.98；  GBDT算法：准确率为0.97；  RF算法：准确率为0.99。  Boston房价数据集：  CART算法：R^2为0.65；  Bagging算法：R^2为0.78；  GBDT算法：R^2为0.77；  RF算法：R^2为0.85。  IMG_256  **以数据集为X轴的准确率散点折线图**  **实验中的过拟合问题**  在本次实验中，我们使用了几种不同的机器学习算法来训练数据集，其中包括了Bagging、GBDT和RF算法。在使用这些算法时，一个重要的问题是过拟合。过拟合指的是机器学习模型在训练数据上表现非常好，但在测试数据上表现较差的情况。这通常是由于模型过于复杂或训练数据过少而导致的。过拟合会导致模型在真实数据上的泛化能力不足，因此在机器学习中需要尽可能避免过拟合。在本实验中，我们可以通过**调整模型参数**来减少过拟合的风险。例如，在决策树算法中，可以通过**限制树的深度、叶子节点数量或剪枝来降低模型复杂度**。在集成学习算法中，可以通过**调整基分类器的数量**或权重来提高模型的稳定性。此外，我们还可以通过交叉验证等技术来评估模型的泛化能力，并选择最佳模型。在实际应用中，还可以增加训练数据的数量和多样性来降低过拟合的风险。  **调参前后对比**  在 iris 数据集上，四个算法的准确率都已经达到了 1.0，因此调参没有带来准确率的提升。  在 breast\_cancer 数据集上，BAGGING 和 RF 两个算法的准确率分别提升了 3.5% 和 2.5%，GDBT 算法的准确率保持不变，CART 算法的准确率提升了 6%。  在 boston 数据集上，BAGGING 和 GDBT 两个算法的准确率分别提升了 2.9% 和 1.3%，RF 算法的准确率下降了 0.6%，CART 算法的准确率提升了 4.6%。  综上所述，调参后不同算法的准确率提升程度不同，具体情况要根据数据集和具体参数来定。在这些数据集上，调参带来的准确率提升幅度并不是很大，但是在一些特定的应用场景中，调参可能会带来显著的提升。 |
| 实验总结 | 本次实验使用了bagging、GBDT和RF算法对三个不同数据集（iris、breast\_cancer、Boston房价）进行了分析，并与基准算法cart算法进行了比较。实验结果表明，在默认参数下，四个算法在iris数据集上表现相当好，其准确率都为1.0，而在breast\_cancer数据集和Boston房价数据集上，GBDT和RF算法表现更好，其准确率比基准算法cart算法要高。  在进行调参后，四个算法在iris数据集上的表现没有变化，其准确率仍为1.0。而在breast\_cancer数据集和Boston房价数据集上，GBDT和RF算法的准确率均有所提高，特别是在breast\_cancer数据集上，RF算法的准确率达到了0.9912，而在Boston房价数据集上，GBDT和RF算法的准确率也有所提高，但是与基准算法cart算法相比，其准确率比基准算法cart算法要高。  综合来看，本次实验表明，对于不同的数据集，选择合适的算法是非常重要的。在iris数据集上，四个算法的表现基本相同，而在breast\_cancer数据集和Boston房价数据集上，GBDT和RF算法表现更好，这说明对于复杂的数据集，集成学习算法可以更好地提高准确率。此外，在调参过程中，适当地调整算法的参数可以进一步提高算法的准确率，但是需要注意过度调参可能会导致过拟合的问题。 |