基于SVM的AdaBoost增强手写数字分类报告

1. 项目概述

本项目实现了基于SVM的AdaBoost算法用于MNIST手写数字分类任务，主要内容包括：

1. 比较线性核函数的SVM和RBF核函数的SVM的性能
2. 从零实现AdaBoost算法，对比决策树桩和线性SVM作为基分类器的性能

2. 数据集与预处理

本实验使用MNIST手写数字数据集，包含28×28像素的手写数字图像。

预处理步骤：

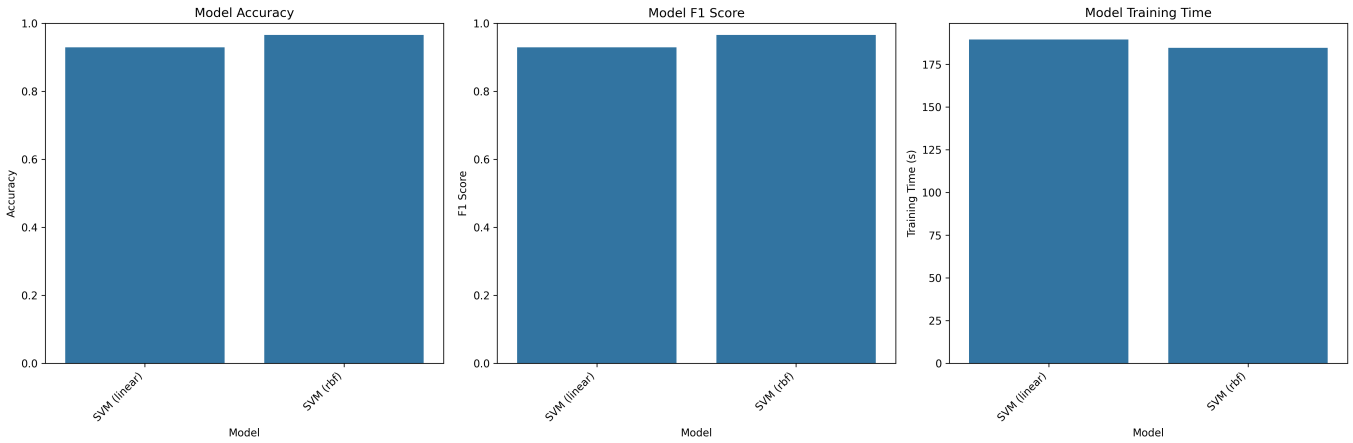
- 将标签转换为整数
- 使用StandardScaler对特征进行标准化
- 将数据集分为训练集和测试集（6/1）（按照MNIST数据集原有的比例）

注：为了便于测试，代码中先使用了sklearn.datasets.fetch_openml加载MNIST数据集，并将其存储在本地。运行时将直接调用本地数据集。

3. SVM模型实现与比较

3.1 线性核与RBF核SVM对比

模型	准确率	F1分数	训练时间(秒)
线性SVM	0.9293	0.9291	199
RBF SVM	0.9660	0.9660	187



3.2 分析

可以看出，RBF核SVM在准确率和F1分数上均优于线性核SVM，且训练时间相近。其原因在于RBF核能够更好地捕捉数据的非线性特征，而线性核SVM适用于线性可分的数据。

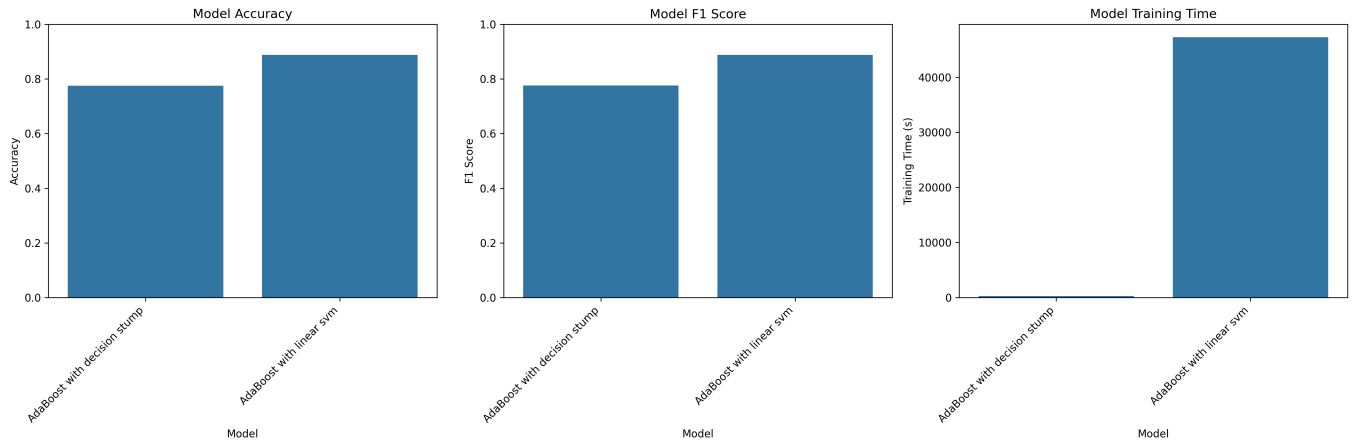
4. AdaBoost算法实现

4.1 AdaBoost多分类算法实现

代码中使用SAMME算法实现了AdaBoost多分类器，分别以决策树桩和线性SVM作为基分类器。

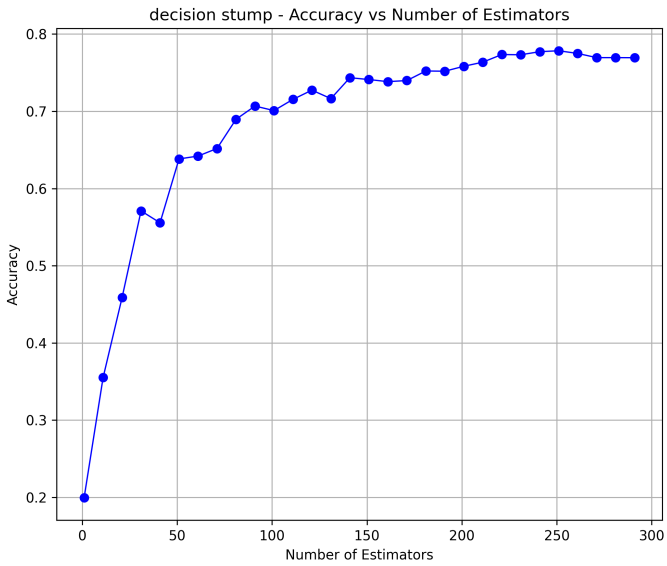
4.2 基分类器性能对比

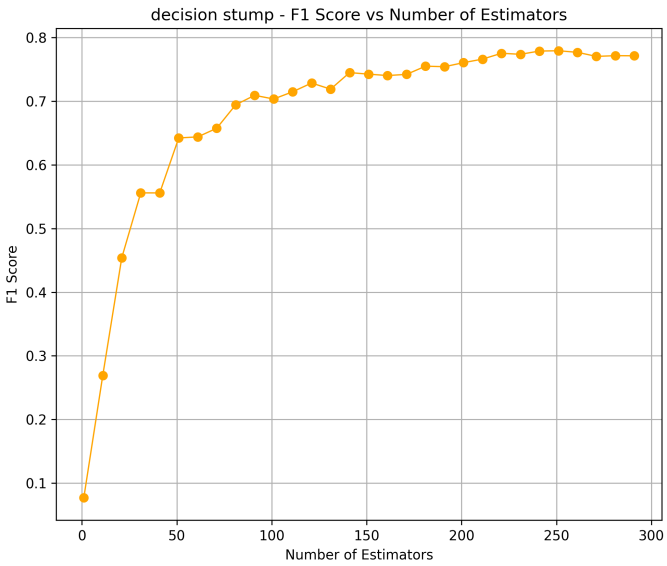
模型	准确率	F1分数	训练时间(秒)	基分类器数量
AdaBoost+决策树桩	0.7728	0.7765	299	300
AdaBoost+线性SVM	0.8885	0.8885	47259	20



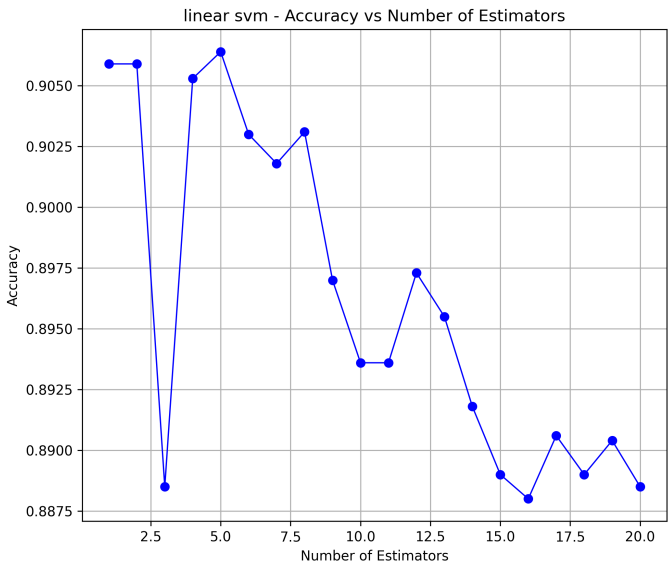
使用线性SVM作为基分类器的AdaBoost模型在准确率和F1分数上均优于使用决策树桩作为基分类器的模型，但训练时间显著更长。

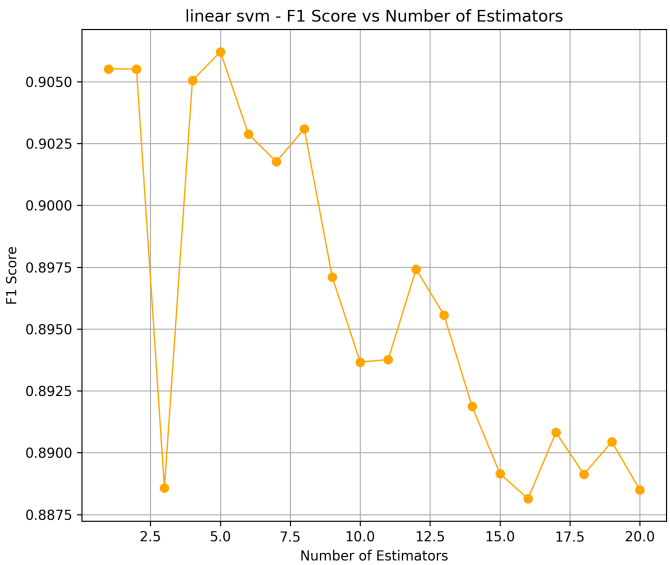
4.3 学习曲线分析





以决策树桩为基分类器的准确率和F1分数总体上随着基分类器数量的增加而提高，但在250个基分类器后，性能提升趋于平稳，说明模型已经接近最佳状态。





以线性SVM为基分类器时，准确率和F1分数随着基分类器数量的增加的波动较大，在数量为5时达到最高点，之后总体上有所下降，说明线性SVM作为基分类器在AdaBoost中可能存在过拟合现象。这可能是由于线性SVM本身的复杂度较高，导致在增加基分类器数量时模型过于复杂。

4.4 基分类器优缺点分析

决策树桩作为基分类器：

- 优点：
  - 模型结构极其简单，单棵树只考虑一个特征的划分，训练速度非常快，计算资源消耗低。
  - 不易过拟合，泛化能力较强，尤其适合在AdaBoost等集成框架下作为弱分类器反复迭代提升整体性能。
  - 易于解释和可视化，便于理解每一步的决策过程。
  - 对异常值和噪声数据不敏感，鲁棒性较好。
- 缺点：
  - 单个决策树桩的表达能力有限，分类性能较弱，依赖大量集成才能获得较好效果。
  - 对于特征之间存在复杂关系的数据，单棵树桩难以捕捉高阶特征。

线性SVM作为基分类器：

- 优点：
  - 具有较强的判别能力，能够有效处理高维数据，分类性能通常优于简单的决策树桩。
  - 适合于特征空间线性可分或近似线性可分的数据。
  - 在AdaBoost框架下，少量SVM基分类器即可获得较高的准确率和F1分数。
- 缺点：
  - 训练时间较长，尤其是在样本量大或特征维度高时，计算资源消耗明显增加。
  - 由于SVM本身属于强分类器，作为AdaBoost的基分类器时，容易导致模型复杂度过高，出现过拟合现象，特别是在基分类器数量较多时。

5. 总结与讨论

本项目系统地实现并比较了基于SVM的AdaBoost算法在手写数字分类任务中的表现，主要结论如下：

1. **SVM模型对比**：RBF核SVM在准确率和F1分数上均优于线性核SVM，且训练时间相近。这表明RBF核能够更好地捕捉数据的非线性特征，适用于更复杂的数据分布。

2. **AdaBoost基分类器选择**：

   - 使用决策树桩作为基分类器时，虽然单个模型性能有限，但通过集成大量弱分类器，整体模型表现稳定且训练效率高，适合大规模数据和对训练速度有要求的场景。
   - 使用线性SVM作为基分类器时，模型在准确率和F1分数上表现更优，且在集成数量较少时即可达到较高性能，但训练时间显著增加，且随着基分类器数量增加，容易出现过拟合现象。

3. **学习曲线分析**：决策树桩AdaBoost的性能随基分类器数量增加而提升，但在一定数量后趋于平稳，说明过多的基分类器带来的收益有限。线性SVM AdaBoost在基分类器数量较少时性

能提升明显，但过多时反而可能导致性能下降，需合理控制集成规模。

4. **实际应用建议**：
    - 若对训练效率和模型可解释性要求较高，推荐选择决策树桩作为基分类器，并适当增加基分类器数量以提升性能。
    - 若追求更高的分类准确率且计算资源充足，可尝试线性SVM作为基分类器，但需防止过拟合，并合理设置基分类器数量。