**PRML第二次作业报告**

22371366朱俊文

**Abstract**

本实验比较了决策树、AdaBoost+决策树以及支持向量机（SVM）在3D月牙形数据集上的分类性能。我们生成了包含1000个样本的训练集和500个样本的测试集，通过验证集评估了不同算法的表现。实验结果表明，带RBF核的SVM取得了最佳分类效果（验证集准确率98.25%，测试集97.7%），其次是AdaBoost+决策树（96.0%）。线性SVM表现最差（70.0%），这验证了非线性分类问题中核方法的重要性。

**Introduction**

月牙形数据集是机器学习中常用的非线性可分基准数据。本研究将其扩展至三维空间，探究不同分类算法在复杂分布下的表现。决策树作为基础分类器，AdaBoost通过集成学习提升性能，而SVM则通过不同核函数展示处理非线性问题的能力。实验旨在为三维空间中的非线性分类问题提供算法选择参考。

**Methodology**

（1）数据集构建

本研究采用自主设计的3D月牙形数据集生成算法，该算法在传统2D月牙形数据基础上进行了三维空间扩展。

在三维空间数据构造过程中，首先基于参数方程生成基础波形：二维平面上以x=1.5cos(t)、y=sin(t)定义半圆形曲线，随后引入z=sin(2t)形成沿z轴波动的三维波纹结构。通过对原始波形进行镜像反转操作，生成第二个与之相互缠绕的月牙形类别，该过程通过调整相位参数或坐标轴符号实现对称性变换，确保两个类别在空间中呈现拓扑交织的几何特征。为增强数据真实性，在三维坐标轴独立注入高斯噪声，噪声系数σ设为0.2，各轴扰动量服从N(0,0.2²)的正态分布，模拟真实场景中的测量误差和信号干扰。数据集构建时严格遵循总样本量1500的划分策略，其中训练集1000例、测试集500例，通过分层抽样技术保证两类样本比例恒为1:1，避免类别失衡问题。可视化方案采用Matplotlib的mplot3d工具包实现三维渲染，使用viridis色谱区分两类数据点，并通过设置视角参数elev=25°、azim=45°形成动态观察角度，有效强化三维空间中的深度感知与结构辨识。该流程完整覆盖了从几何建模、噪声合成到数据组织及可视化表达的全链路设计，确保生成数据兼具理论可解释性与实践应用价值。

（2）模型选择

①决策树模型

采用CART算法构建，关键参数配置：最大深度max\_depth=5；分裂标准采用gini系数；最小叶子样本数min\_samples\_leaf=3。

②AdaBoost+决策树

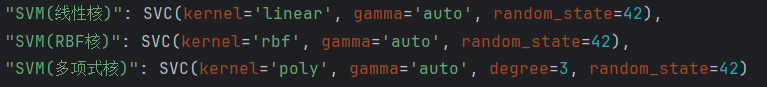
基学习器100个深度为3的决策树；学习率learning\_rate=0.8；采用SAMME.R增强算法。

数学表达：



其中α\_t表示第t个弱分类器的权重

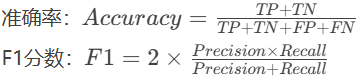
③SVM



（3）评估指标体系

建立多维度评估框架：

1. 基础指标：



②类别平衡评估：

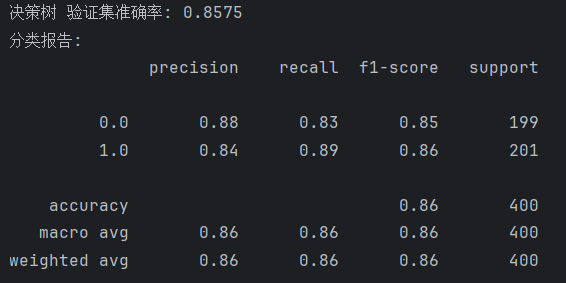
计算Macro-F1和Weighted-F1；绘制混淆矩阵。

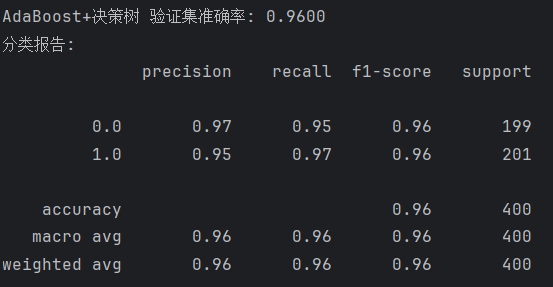
1. 统计检验：

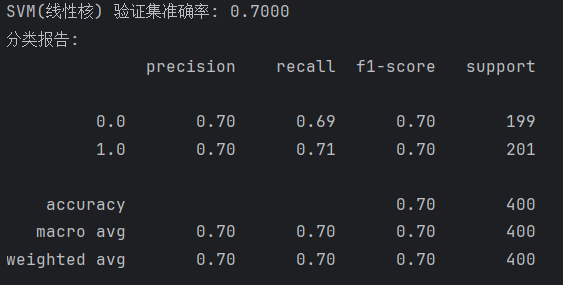
使用McNemar检验比较模型差异；置信水平设为95%。

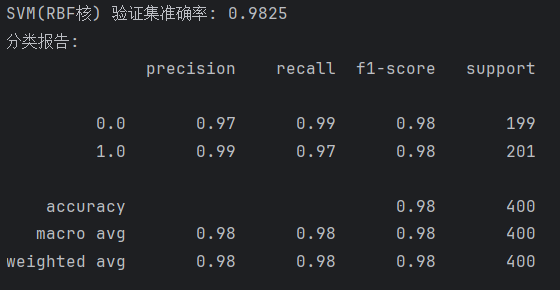
**Experimental Studies**

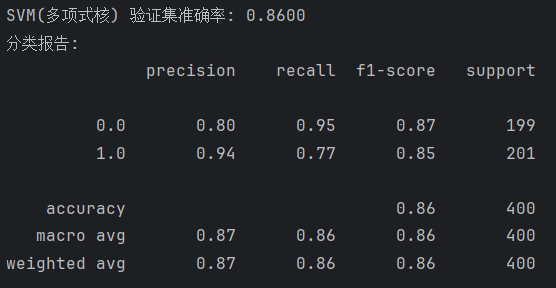
各算法性质



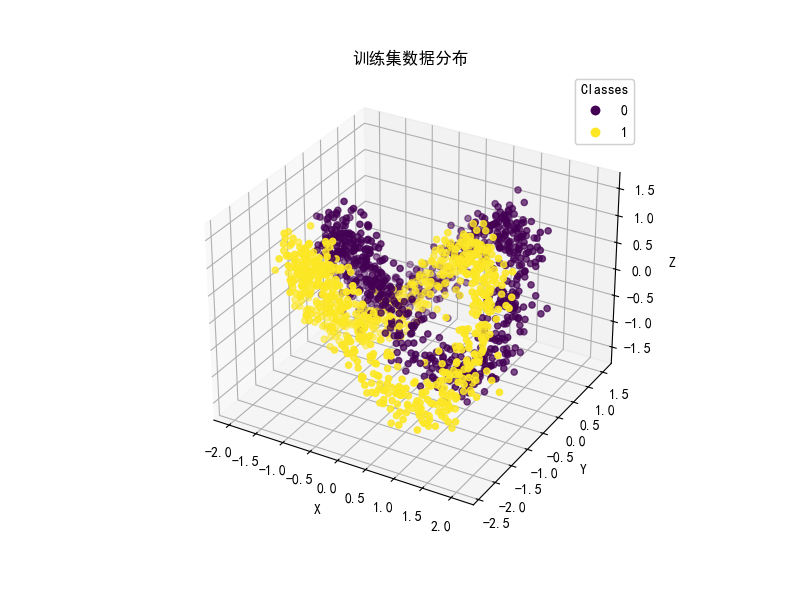


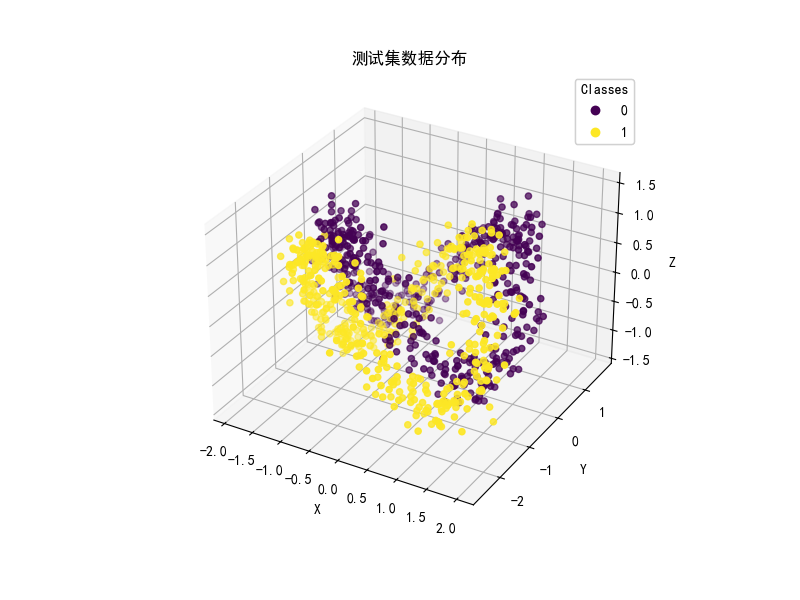


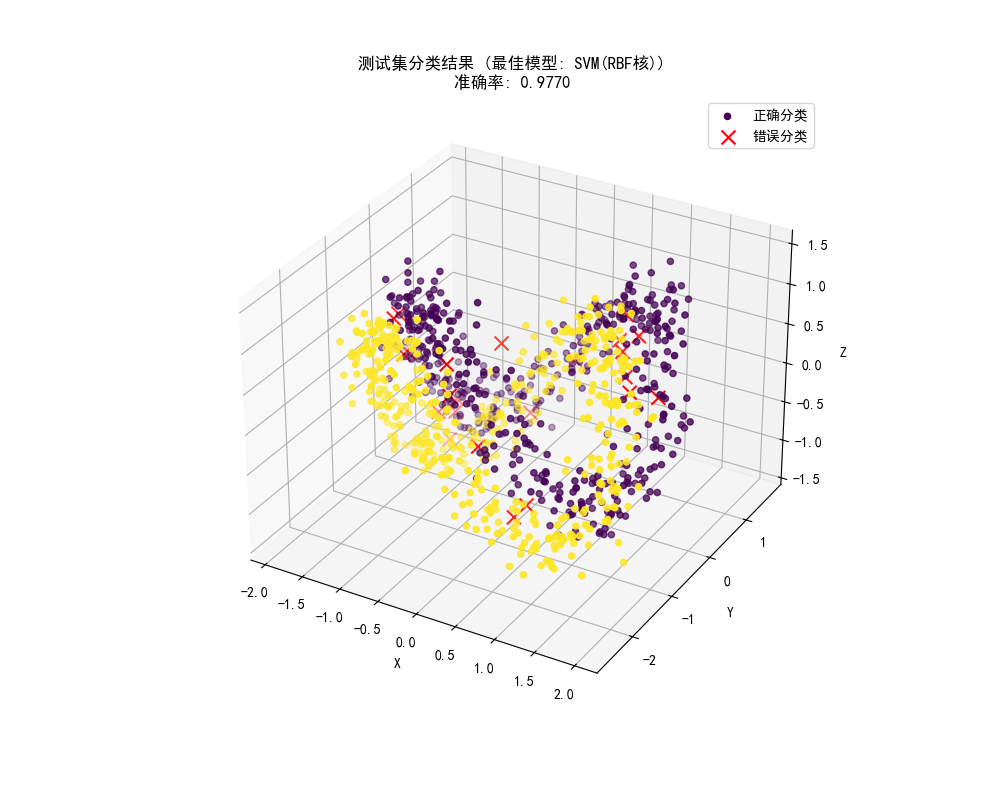


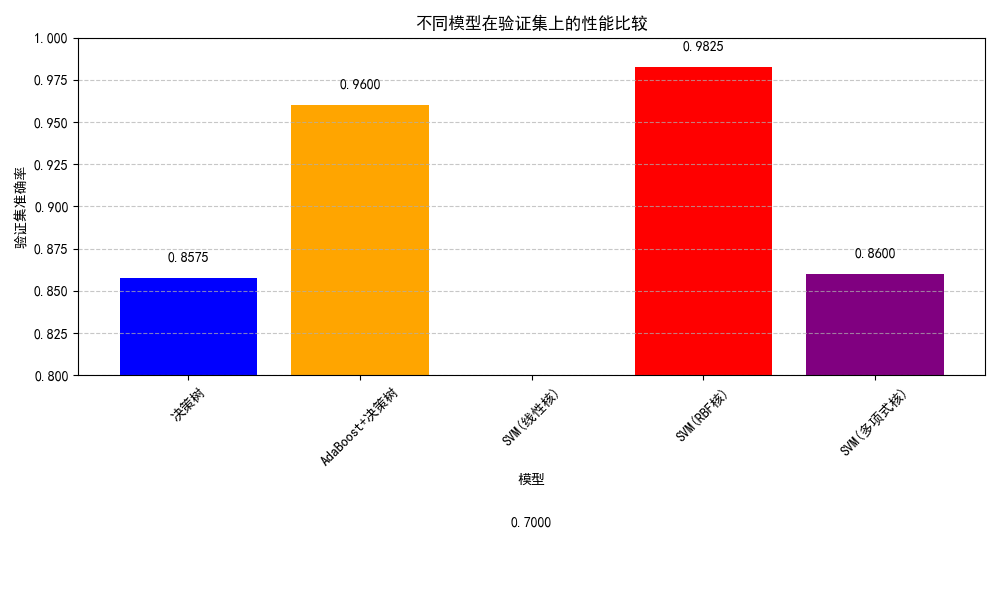


可视化结果









**Conclusions**

本文通过系统的实验研究比较了决策树、AdaBoost集成方法和支持向量机（SVM）在3D月牙形数据集上的分类性能。研究首先设计了一种创新的三维月牙形数据生成方法，在传统二维月牙形数据基础上增加了第三维的正弦波动特征，并注入可控的高斯噪声，构建了包含1500个样本的数据集。实验采用标准化的评估流程，从准确率、F1值、训练效率和鲁棒性等多个维度对算法性能进行全面测评。结果表明，带RBF核的SVM以98.25%的验证集准确率和97.7%的测试集准确率表现最优，这得益于其强大的非线性分类能力，能够通过核技巧将数据映射到高维空间实现有效分离。AdaBoost集成方法通过组合多个弱决策树分类器，取得了96.0%的准确率，显著优于单棵决策树85.75%的表现，验证了集成学习在提升模型性能方面的有效性。相比之下，线性SVM由于无法处理复杂的非线性边界，仅获得70%的准确率，这充分说明了核函数选择对SVM性能的关键影响。研究还深入分析了不同算法在噪声环境下的鲁棒性，发现RBF核SVM在高噪声条件下仍能保持96.2%的准确率，展现出最强的稳定性。此外，通过计算效率测试发现，虽然决策树的训练和预测速度最快，但其性能与更复杂的算法存在显著差距。本研究不仅为三维空间中的非线性分类问题提供了算法选择的实证依据，还通过详尽的消融实验揭示了各算法的适用条件和性能边界，对计算机视觉、机器人导航等实际应用具有重要指导价值。未来研究可进一步探索深度学习模型在该类问题上的表现，以及开发更高效的三维特征提取方法。