实验报告

1. 实验内容

1.1 实验任务

本实验主要完成手写数字识别实验，基于SVM算法，利用经典手写识别数据集MNIST训练并验证我们的模型。在本实验中，我们实现了模型训练及验证，并测试了我们提供的测试案例，同时，我们还分析了不同核函数和分类策略对模型训练时间及分类准确度的影响。在本次实验中，我们发现采用高斯核函数是最佳的SVM分类模型，其准确度达到98%

1.2 主要方法

SVM算法、对比实验

2. 实验方法和步骤

2.1 方法原理

2.1.1 SVM原理

SVM是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM还包括核技巧，这使它成为实质上的非线性分类器。SVM的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM的的学习算法就是求解凸二次规划的最优化算法。同时，SVM算法还可用通过选择分类策略来实现多分类问题，如策略*ovo*是对于多个分类任务，两两之间构建一个二分类模型，从而达到多分类任务；而*ovr*分类策略的思想是将其中一个类和剩余类看成两个类构建二分类模型，从而实现多分类。

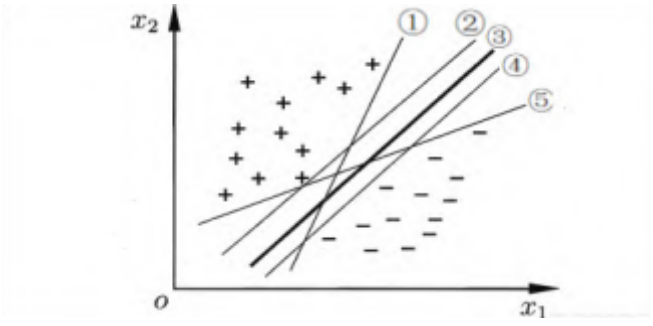

图1. 如图所示，在上面几个线性划分中，显然只有③号曲线的容错性最高

由上图启发，我们的目标就是寻找该划分直线。定义数据到该直线的最小距离为决策边界，我们上述问题就等价于求解最大决策边界的问题，定义距离该直线最近的点为支持向量(support

vector)，支持向量到决策边界的距离称为决策距离。通过上述转化，我们的优化目标就是求解最大决策距离的决策边界。

2.2 实验方案

我们首先进行训练数据的收集，本次实验我们使用的是手写数字识别的经典数据集MNIST，也可以使用自己创建的数据集。对于我们得到的数据集，给出其标签，并对其进行图像预处理，合理划分出训练集和测试集（训练集60000张图像，测试集10000张图像）。训练出我们的模型并进行测试。选择不同的核函数和分类策略，比较最终的识别准确度和训练速度并进行绘图分析，从而得出最佳模型。

2.3 实现步骤

支持向量集实现手写数字识别主要分成4步：

第一步：数据准备。本实验中我们使用经典数据集MNIST，并且已经对其进行了大小处理成20*20，按60000张训练集和10000张测试集进行划分。

第二步：数据预处理。为了提高计算机对手写数字识别的精度，需要对数据进行预处理。预处理的操作包括对数据的形状、属性等按照某一规则进行修改。为了使数据格式统一（大小、形状、维度等），需要对所有手写图像数据进行归一化、二值化等操作，以便计算机更好地学习数据，提高识别精度。

第三步：数据特征提取。每个手写数字都具有复杂多样的特征，因此特征提取是支持向量机数字识别算法的关键步骤。对于图像数据，可以将其像素统一，如设置为28×28像素，并统计像素点在该区域的百分比等特征。

第四步：数据训练及测试。在特征提取后，生成多个具有不同参数的支持向量机模型，并利用误差矩阵比较这些模型。通过比较误差矩阵，可以得到不同参数下支持向量机模型的预测精度、准确性等数值指标。这一步骤包括模型的训练和测试，旨在评估模型的性能并选择最优的参数组合。

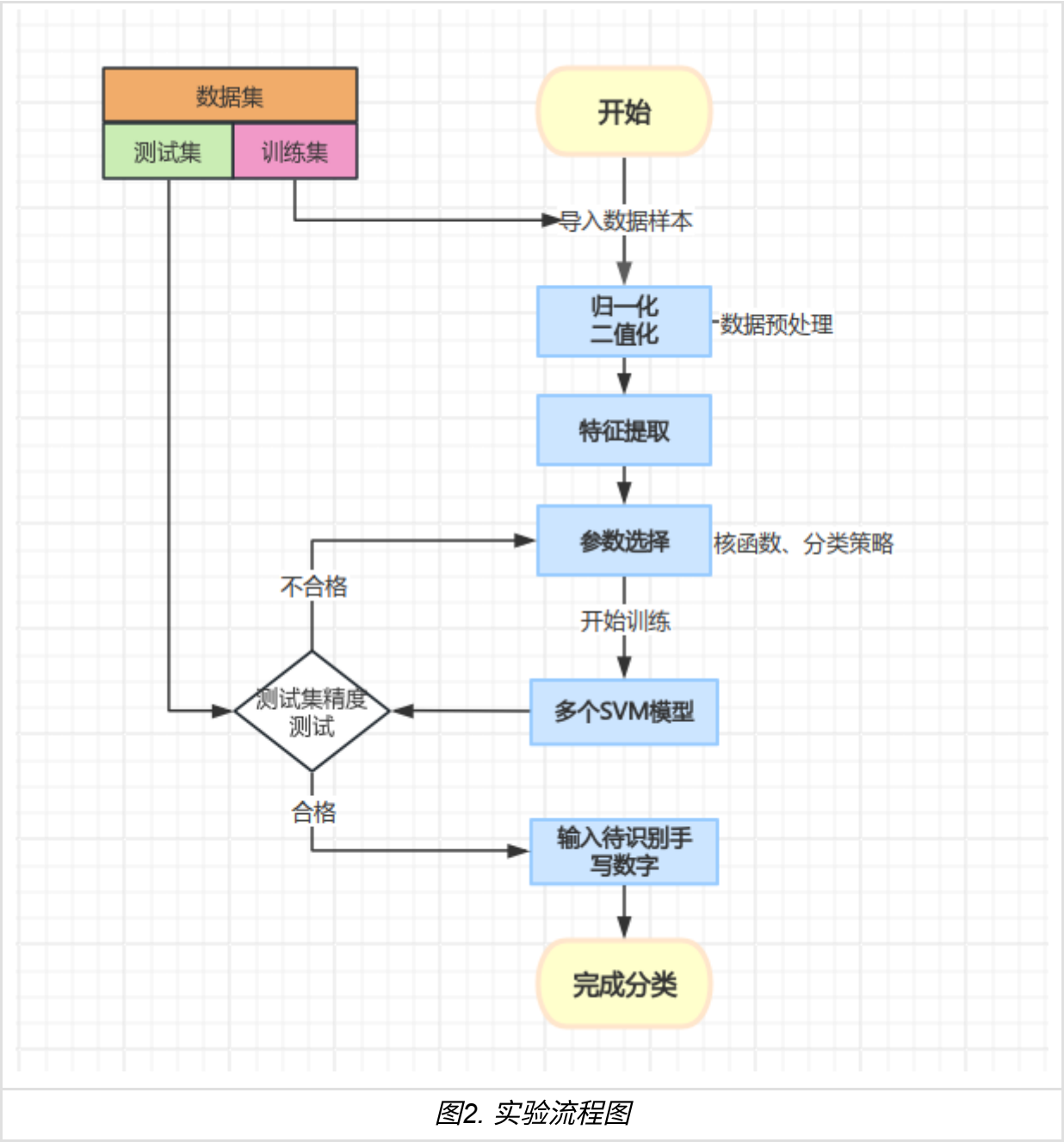

图2. 实验流程图

3. 实验结果

3.1 实验数据详情

本实验采用经典数据集MNIST，其中训练集包含60000张图片，测试集10000张图片，大小为20*20。具体如下：

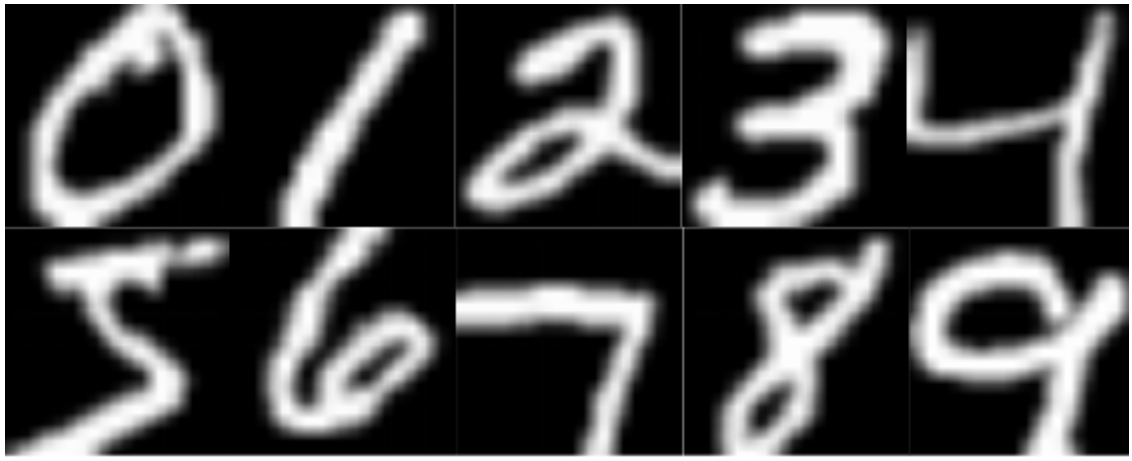

图3. MNIST数据集图片示例

3.2 实验环境设置

本实验使用具体环境如下：

操作系统	**Ubuntu22.04**
开发语言	python3.8
所使用第三方模块	numpy、matplotlib、scikit-learn、Pillow、···

3.3 实验结果及分析

具体数据记录见附表

我们首先分析不同核花费的训练时间和测试时间，从图4我们能看出，多项式核(poly)在训练时间和测试时间上花费较少。

在图中，我们还看到了一处反常现象，即图中不同决策方式的所需时间几乎相等，甚至有些地方ovr决策方式竟然比ovo决策方式所需时间多，这与理论分析相悖。我们知道，由于ovr方式是将多分类看成一种和剩下一种，即对于一个n分类，其需要n-1个二分类器；而对于ovo方式则是在任意两个之间构建一个二分类器，即需要$\frac{n(n-1)}{2}$个二分类器，这可能与分类总数和训练集大小有关。

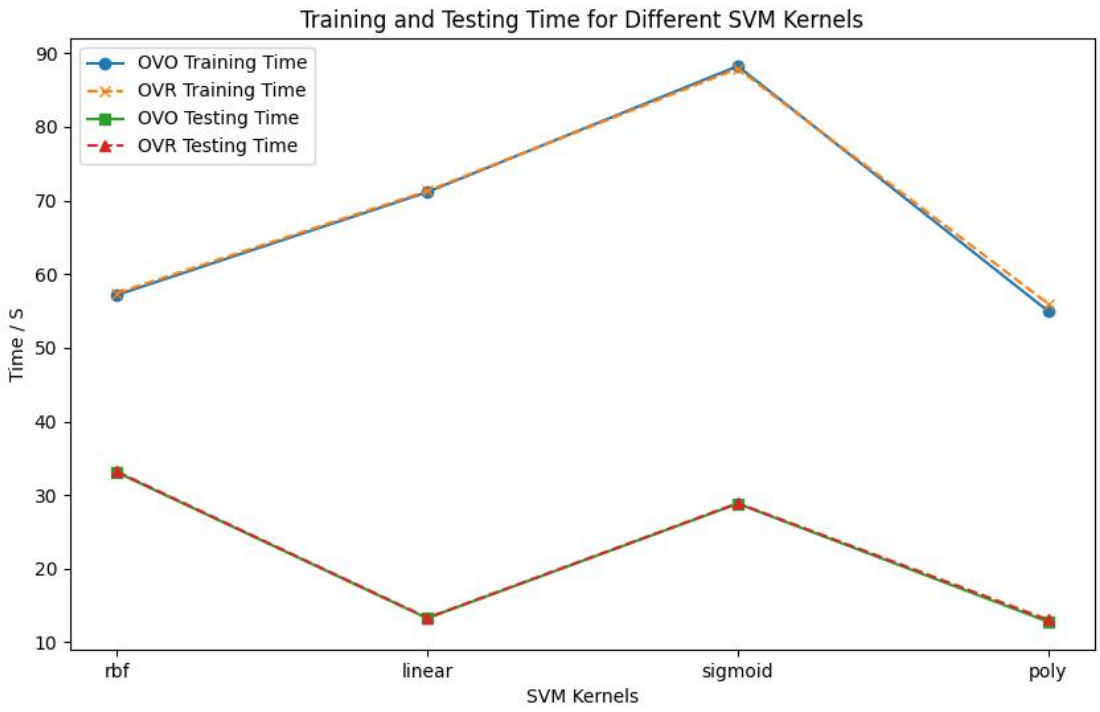

图4.不同核对应的训练时间和测试时间。从图中可以看出多项式核(poly)所花费的训练时间和测试时间都优于其他核

我们再来比较不同核函数的分类正确率，由于不同决策方式对分类正确率影响较小，故我们全部比较ovo决策方式下不同核函数的分类准确率。从图中我们可以看出，多项式核函数和高斯核函数在准确率上领先其他两个核函数，并且都达到了98%以上，具有较高的准确性。

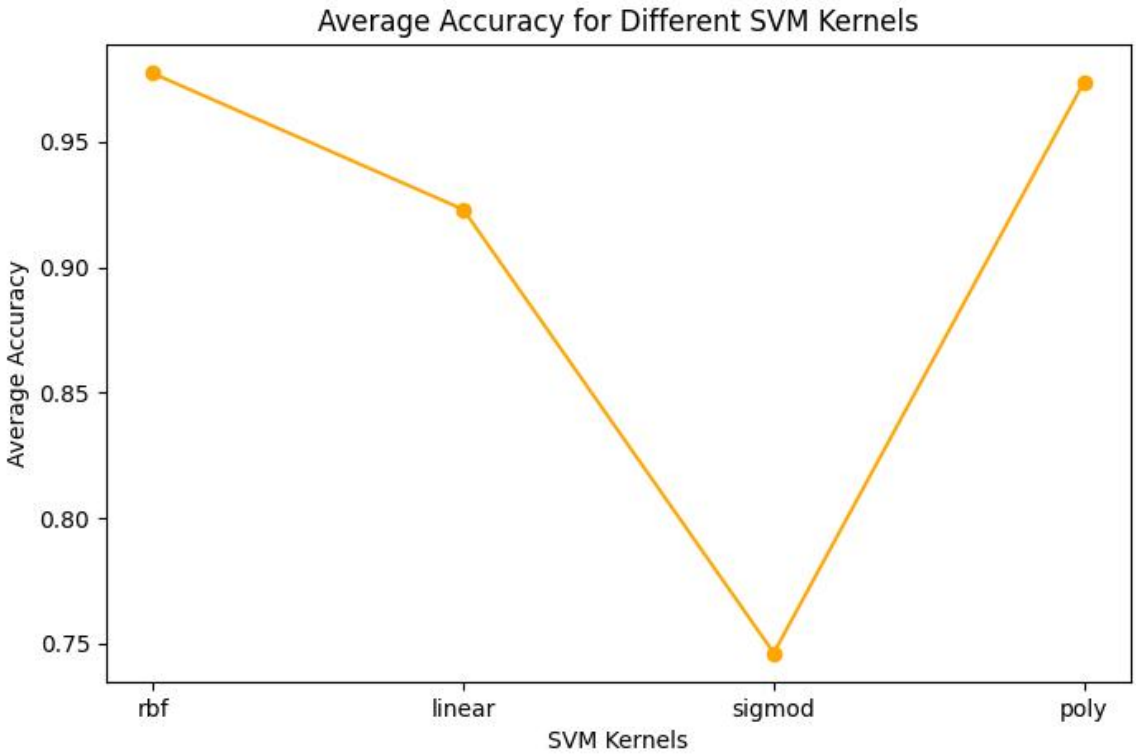

图5. 不同核函数的平均识别准确率（已归一化）

综上所述，多项式核函数poly构建的SVM分类其具有最好的识别效果，能构建出训练速度快且识别准确率高的模型。