

|  |  |
| --- | --- |
| **题 目：** | 基于深度学习的图像识别系统：“MNIST |
|  | 手写数字识别项目” |
| **学 院：** | 数据科学学院 |
| **专 业：** | 数据科学与大数据技术 |
| **年级班别：** | 2021级（1）班 |
| **学 号：** | 421470165 |
| **学生姓名**： | 朱晓瑜 |

**提交日期：** 2024 **年** 11 月

# 1 绪论

手写数字识别是机器学习领域中的一个重要应用，它可以应用于自动化数据输入、智能检测等领域。MNIST手写数字数据库是机器学习领域中的一个经典数据集，包含了一系列手写数字图片和对应的标签，是进行手写数字识别算法研究和性能评估的标准数据集之一。下面我将介绍基于CNN深度学习网络的MNIST手写数字数据库训练和识别的实现方法和步骤。近年来，机器视觉领域发展迅速，受到全球研究人员的广泛关注。尽管在数字识别的实际应用中希望实现对每个手写数字的准确识别，但当前的算法识别率仍无法达到100%。实际环境中存在各种复杂因素，会干扰机器人对手写数字的准确识别，因此针对不同复杂环境下的数字识别仍然具有挑战性。因此，研究如何在复杂环境中准确识别手写数字具有重要意义

## 1.1研究背景和研究意义

MNIST是一个手写体数字的图片数据集，该数据集来由美国国家标准与技术研究所（National Institute of Standards and Technology (NIST)）发起整理，一共统计了来自250个不同的人手写数字图片，其中50%是高中生，50%来自人口普查局的工作人员。该数据集的收集目的是希望通过算法，实现对手写数字的识别。1998年，Yan LeCun 等人发表了论文《Gradient-Based Learning Applied to Document Recognition》，首次提出了LeNet-5 网络，利用上述数据集实现了手写字体的识别。

## 1.2研究内容

1.本项目为分别使用CNN、SVM和KNN方法进行手写数字识别，并比较三种方法的准确率。

2.将图像数据集进行数据清洗，以及数据归类。

3.使用不同的分类算法，设计三款分类器，对图像进行分类

4.使用分类指标，对以上三款分类器的分类结果进行比较

# 2 数据描述

## 2.1 数据集简介

MNIST数据集是一个常用的手写数字识别数据集，包含了60000张训练图片和10000张测试图片。每张图片大小为28x28像素，灰度级别为0~255，表示了0~9之间的数字。其中，训练集用于训练模型，测试集用于测试模型的准确率。MNIST数据集的图像数据文件为"train-images-idx3-ubyte.gz"和"t10k-images-idx3-ubyte.gz"，其中前者存储了训练数据，后者存储了测试数据。

**数据获取：** Mnist数据集官网：<http://yann.lecun.com/exdb/mnist/>

它包含四个部分：

- train-images-idx3-ubyte.gz：训练集图像（9912422字节）

- 训练集图像，共60000张图像

- train-labels-idx1-ubyte.gz：训练集标签（28881 字节）

训练标签集

- t10k-images-idx3-ubyte.gz：测试集图像（1648877字节）

测试集图像，共10000张图像

- t10k-labels-idx1-ubyte.gz：测试集标签（4542字节）

测试标签集

**标签信息：** 标签是用于标识图像中手写数字的数字，范围从0到9。例如，标签0表示图像中的手写数字是数字0，标签1表示数字1，以此类推。

**数据的应用：**MNIST数据集（Mixed National Institute of Standards and Technology database）是一个用来训练各种图像处理系统的二进制图像数据集，广泛应用于机器学习中的训练和测试。 MNIST数据集常被用作数字识别问题的基准数据集，尤其是在机器学习领域。研究者和开发者可以使用MNIST数据集来验证算法、模型或方法的性能。例如，可以将MNIST数据集用于训练和测试各种图像分类算法，如卷积神经网络（CNN）、支持向量机（SVM）等。

**归一化：**图像像素值的范围是（0，255），这导致图像的各像素值的范围差异大，这 会增加模型训练和优化的挑战，并且使像素之间难以直接进行比较。通过对图像 数据进行归一化处理，可以有效地解决这个问题。 归一化先计算每个像素在整个数据集中的平均值，然后从原始像素中减去这 个平均值实现中心化，再除以像素标准差实现规范化，使所有的像素都落在零均 值单位标准差的范围内，实现了范围和规模的统一。具体公式如下：

normalized\_output = output − min(output) max(output) − min(output)

其中，normalized\_output 表示归一化后的图像像素值输出，output 表示原始 图像像素值输入，max 和 min 分别表示最大像素值和最小像素值。将数据集的 R、 G、B 三通道的值分布归一化到 0 均值附近，这样便于平衡网络每层的激活点， 更全面和公平地学习到各个视觉特征，同时平衡重要像素和微小像素的影响，有 效地提升了模型在分类任务上的表现。

## 3 数据处理

**3.1数据预处理**

首先肯定是导入包：

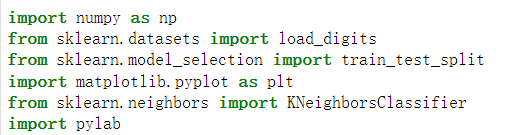


图3.1 导入包代码截图

然后将两个文件的数据分别读入数据集和测试集以及相对应的image和label，并进行归一化处理（归一化可以提高的模型的准确度和收敛速度），简单的归一化处理就是直接除以最大值255，也有其他的方法如min-max归一化和Z-score标准化等。

**3.2数据清洗**

数据清洗是一个关键的数据处理步骤，旨在提高数据质量，确保数据分析的准确性和可靠性。它涉及对数据进行重新审查和校验，发现并纠正数据文件中的错误，处理无效值和缺失值，以及确保数据的一致性和完整性.

数据清理方法是将数据库精简去除不合规的数据（如图3.2），使之转换成标准可接收格式的过程。数据清理从数据的准确性、完整性、一致性、惟一性、适时性、有效性几个方面来处理数据的丢失值、越界值、不一致、重复数据等问题。根据数据源的实际需要，不同的数据需要不同的数量清洗方法进行处理。本次实验用到的数据集已经是进行过数据清洗的数据集。



图3.2 需要进行数据清洗的数据

**3.3数据可视化**

数据可视化是以图形的方式呈现结构化或非结构化数据，从而将隐藏在数据中的信息直接呈现给人们。数据可视化最大重要性是在于它**帮助人们更快地理解数据**。寻找堆积如山的信息之间的联系并不容易，但图形和图表可以将无形的信息，转化为可见的图形符号，直接清晰地表达出来，帮助你快速发现关键点。本次实验几万个图像经过数据可视化后才能进行最后的比较，以下图3.3是数据可视化做出来的图形之一。

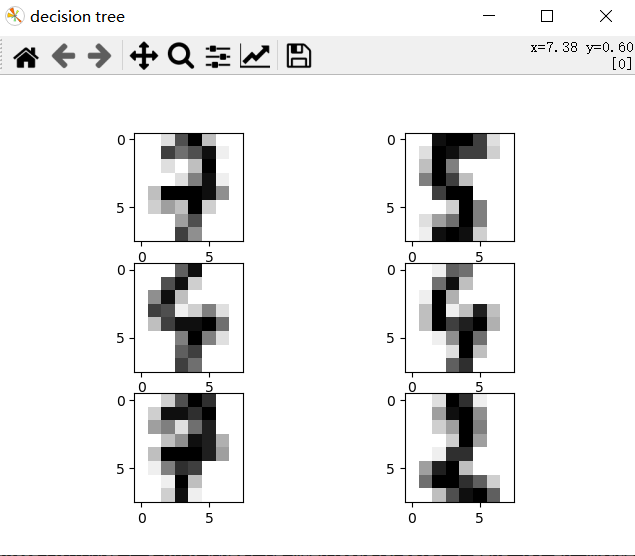


图3.3数据可视化图形

**4 模型构建**

**4.1 CNN实现MNIST手写数字识别**

**4.1.1卷积神经网络介绍**

CNN（Convolutional Neural Network，卷积神经网络）是一种深度学习网络，特别适用于图像分类和识别任务。与传统神经网络不同，CNN可以自动提取数据的特征，从而实现对图像的高效分类和识别。CNN网络的核心组件包括卷积层、池化层和全连接层。

**卷积层：**卷积层是CNN网络中的核心组件，它可以自动提取输入图像的特征。在卷积层中，使用一组可学习的卷积核对输入图像进行卷积操作，得到一组特征图。卷积核是一个小的矩阵，可以通过反向传播算法进行训练，以提取输入图像的不同特征。

**池化层：**池化层是CNN网络中的另一个核心组件，它用于对特征图进行降维操作。在池化层中，通常使用最大池化或平均池化等操作，将每个特征图中的一定区域进行压缩，从而减少网络中的参数数量和计算量。

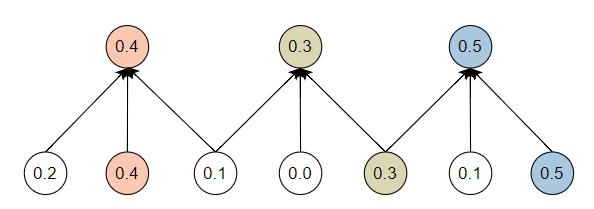


图4.1 池化（一维表示）

**全连接层：**全连接层是CNN网络中的最后一层，它用于对特征图进行分类。在全连接层中，可以使用softmax函数将特征图映射到0~1之间的概率值，从而得到对输入图像的分类结果。

**4.1.2使用CNN实现MNIST手写数字识别**

机器识图的过程：机器识别图像并不是一下子将一个复杂的图片完整识别出来，而是将一个完整的图片分割成许多个小部分，把每个小部分里具有的特征提取出来，再将这些小部分具有的特征汇总到一起，从而完成机器识别整个图像。

采用深度学习中的卷积神经网络实现对手写数字的识别，卷积神经网络是被设计用来处理多维组数据的，如常见的彩色图像就是三个颜色通道组合。手写数字图片是典型的2D型图像数据，使用卷积神经网络可以有效通过训练提取去手写提数字的特征。

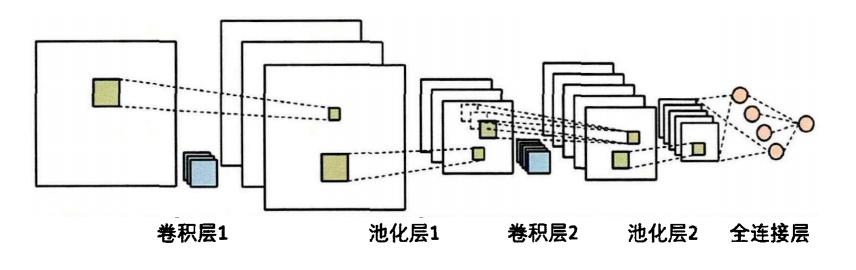


图4.2 典型的CNN结构

**4.1.3 CNN部分代码以及实验结果**

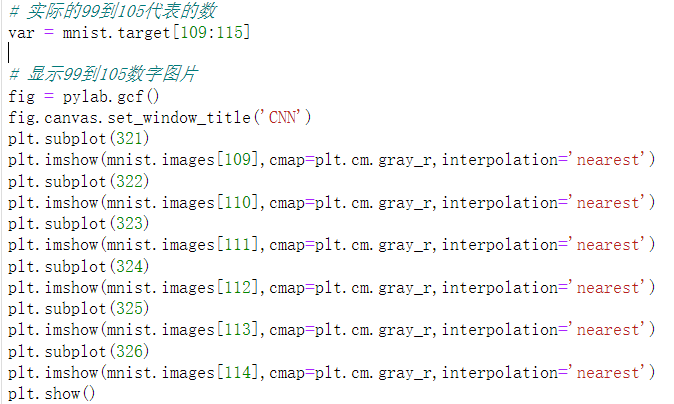


图4.3 CNN部分代码截图

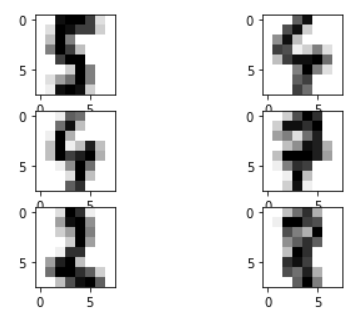


图4.4 CNN代码运行结果



图4.5 CNN代码运行结果

**总结：**上面基于MNIST数据集，构建了卷积神经网络识别模型，经过多个epochs训练，识别率成功达到了惊人的99.25%（如图4.5），充分展示了卷积神经网络在图片识别中的作用。

**4.2 SVM实现MNIST手写数字识别**

**4.2.1 SVM 基本原理介绍**

SVM是一种机器学习方法，通过寻找超平面将训练样本进行分类，尤其适用于处理非线性和小样本分类问题。SVM在模式识别和回归等领域得到广泛应用，并在统计学习理论基础上发展而来，具有解决小样本、非线性和高维问题的优势。

支持向量机（Support Vector Machine，SVM）是一种机器学习算法，用于解决二分类和多分类问题。其基本原理是通过寻找一个超平面，将不同类别的样本正确地分割开来。对于线性可分的情况，SVM追求最大化支持向量到超平面的距离；对于线性不可分的情况，SVM引入松弛变量和惩罚项，允许一些错误分类的样本点存在。通过核函数，SVM可以将样本映射到更高维的特征空间，从而处理线性不可分的问题。SVM具有处理高维和小样本数据的能力，并对噪声和过拟合具有一定的鲁棒性。它在求解过程中只涉及到支持向量，因此对于大规模数据集计算开销相对较小。总之，SVM是一种强大的分类算法，具有广泛的应用和优越的性能。

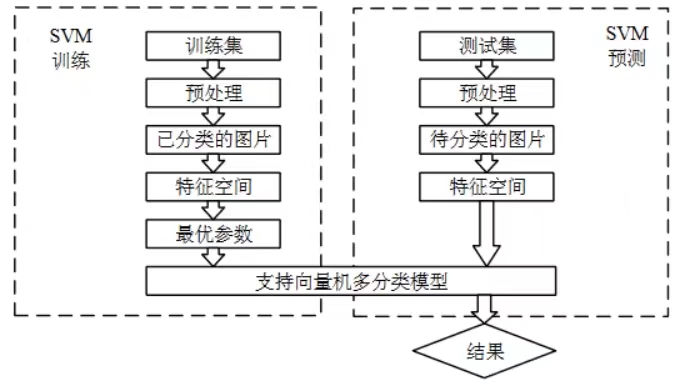


图4.6 SVM训练和预测

支持向量机（Support Vector Machine）作为机器学习中最常用的算法之一，有着非常强大的性能。SVM既可以用来分类，即SVC（Support Vector Classifier）；也可以用来预测（回归），那就是SVR（Support Vector Regression）。sklearn中的svm模块中同时集成了SVC和SVR。SVC，中文可称做支持向量机分类器，也被称做最大化类间间隔分类器（maximum margin classifier），这是一种较为直观的理解（如下图的L2），也是最为人们所熟悉的对于SVM的理解。通俗的来说，SVM在拟合过程中，会根据训练样本的分布，搜索所有可能的线性分类器中最佳的那个，进一步观察我们可以发现，决定其直线位置的样本并不是所有的训练数据，而是其中两个空间间隔最小的两个不同类别的数据点，我们把这种可以用来真正帮助决策最优线性分类模型的数据点叫做“支持向量”。

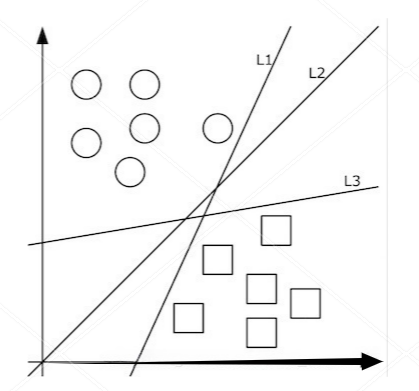


图4.7 SVM拟合

4.2.2 SVM 部分代码以及运行结果

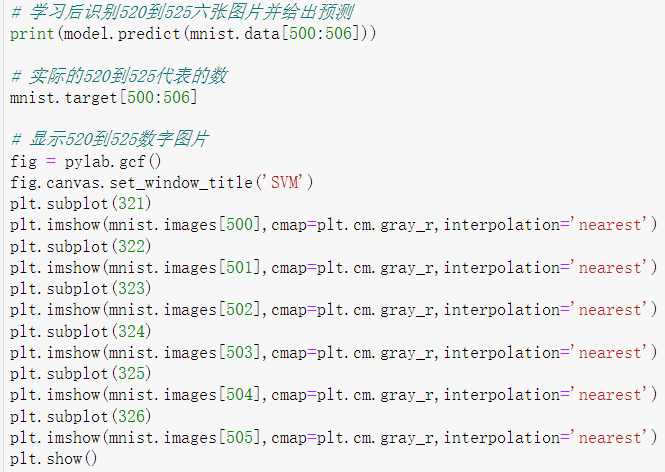


图4.8 SVM部分代码截图

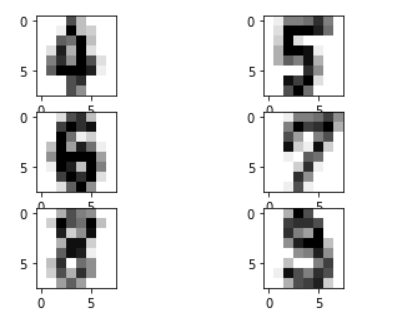


图4.9 SVM代码运行结果

00441cf82d75881b48401a4ee70b240

图4.10 SVM代码运行结果

**总结：**通过上图4.10的输出结果可以发现，SVC模型的确能够提供比较高的手写体数字识别性能，准确达到了94.66666666666667%，充分展示了支持向量机在图片识别中的作用。

**4.3 KNN实现MNIST手写数字识别**

**4.3.1 KNN基本原理介绍**

**K近邻算法：**如果一个样本在特征空间中距离最近的K个样本点的所属类别比重最大的，则该样本也属于这个类别。KNN算法的核心是样本间的距离和K的取值，主要应用于分类和回归。

**算法的一般流程如下：**

1、确定邻居数k：确定一个合适的k，是KNN算法中的一个超参数。

2、计算距离：对于待预测样本，计算它与训练集中每个样本的距离。常用的距离度量包括欧氏距离、曼哈顿距离、闵可夫斯基距离等。

3、找到最近的邻居：根据计算的距离，找到与待预测样本最近的k个训练样本。

4、投票决策：根据找到的k个最近邻样本的标签，采取多数投票或者加权投票的方式来决定待预测样本的类别。

5、输出预测结果：将投票决策的结果作为待预测样本的类别输出。

**4.3.2 部分代码和实验结果**



图4.11 KNN部分代码截图

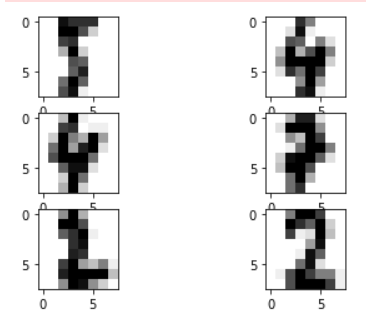


图4.12 KNN代码运行结果



图4.13 KNN代码运行结果

**总结：**通过上图4.13的输出结果可以发现，KNN模型的确能够提供比较高的手写体数字识别性能，准确达到了97.33333333333334%，充分展示了支持向量机在图片识别中的作用。KNN算法简单易懂，易于实现，但在处理大规模数据集时计算复杂度较高，手写体识别相比于用深度学习来说耗时长很多。

**5 模型评估及结果分析与优化**

**5.1 结论以及建议**

本文基于MNIST手写数字识别数据集，分别使用CNN、SVM和KNN三种分类器算法进行手写数字识别，并比较三种方法的准确率（结果如下表格5.1）。通过对比分析可以看到CNN的准确率为最高，达到99.25%，其次是KNN，准确率为97.33333333333334%，最后为SVM，准确率为94.66666666666667%。由此看出这三个模型都具有较高的手写数字识别性能，但我们可以看到此次实验CNN算法相比其他两个算法具有更好的展示效果，并且CNN算法简单易懂，易于实现。同时我在实验中也发现CNN在处理手写体识别时相比于其他两个算法来说耗时长要短。总体来说，此次试验在技术实现和功能效果上均取得了良好结果。但是，由于样本量和类别相对单一，模型在更广泛和复杂场景下的预测能力还有待提升。此 外，我还应该继续探索这三种算法关于其他方面的的性能对比，如耗时长短问题的对比，才能做出更正确的结果判断，另外我还可以增加多几种算法进行模型预测，如决策树算法，朴素贝叶斯算法等来提高实验的可信度，增加实验的合理性，正确性。

|  |  |
| --- | --- |
| 分类器 | 准确率 |
| CNN | 99.25% |
| SVM | 94.66666666666667% |
| KNN | 97.33333333333334% |

表5.1 三种分类器的准确率

**6参考文献**

[1] 张有健, 陈晨, 王再见. 深度学习算法的激活函数研究[J]. 无线电通信 技术, 2021, 47(1): 115-120.

[2] 曹丹等.手写数字识别的研究进展[J].电脑知识与技术.2009.03

[3]张捷.手写数字识别的研究与应用[D]; 西安:西安建筑科技大学;2004.