

**专业综合实践论文**

|  |  |
| --- | --- |
| **题 目：** | 基于深度学习的数字图像分类系统 |
| **学 院：** | 人工智能学院 |
| **专 业：** | 数据科学与大数据技术 |
| **年级班别：** | 2021本数据科学与大数据技术1班 |
| **学 号：** | 421470136 |
| **学生姓名**： | 卢夏晖 |
| **日期** | 2024.11.27 |

**提交日期：** 2024 **年** 11 月

# 基于深度学习的数字图像分类系统

## 摘要

本论文深入研究了基于机器学习的鸢尾花分类识别问题。随着人工智能技术的不断发展，机器学习在各个领域都展现出了强大的潜力。鸢尾花数据集作为一个经典的分类问题实例，具有重要的研究价值。通过对不同机器学习算法的应用和比较，包括决策树、支持向量机、随机森林等，我们旨在找到最适合鸢尾花分类的方法。首先，对鸢尾花数据集进行了详细的分析，包括特征提取和数据预处理。然后，分别运用多种机器学习算法进行训练和测试，并对结果进行了深入的评估和比较。实验结果表明，不同的算法在鸢尾花分类任务中表现出了不同的性能特点，其中随机森林算法在准确性和稳定性方面表现较为突出。本研究为基于机器学习的分类问题提供了有价值的参考和实践经验。

关键词： 机器学习、鸢尾花分类、决策树、支持向量机、随机森林、数据预处理、特征提取。

**目录**

[一、项目背景与目的 1](#_Toc22757)

[1.1研究背景及意义 1](#_Toc21560)

[1.2 国内外研究现状 1](#_Toc2143)

[1.3 本文主要研究内容 1](#_Toc6391)

[二、 机器学习基础理论 1](#_Toc16823)

[2.1 机器学习概述 1](#_Toc21194)

[2.2 监督学习与非监督学习 1](#_Toc27851)

[2.3 常用机器学习算法介绍 2](#_Toc10634)

[2.4 支持向量机（SVM）基本原理 2](#_Toc15286)

[第三章、 鸢尾花数据集介绍 2](#_Toc12778)

[3.1 数据集来源与描述 2](#_Toc30556)

[3.2 数据预处理方法 2](#_Toc25673)

[3.3 鸢尾花数据集特征分析 3](#_Toc1381)

[第四章、 基于卷积神经网络的鸢尾花分类模型 3](#_Toc19288)

[4.1 卷积神经网络简介 3](#_Toc6879)

[4.2 模型构建 3](#_Toc12443)

[4.2.1 CNN模型架构 3](#_Toc26912)

[4.2.2 数据预处理 3](#_Toc32338)

[4.2.3 模型训练 4](#_Toc16042)

[4.2.4 模型评估 4](#_Toc2382)

[4.3 结果与讨论 4](#_Toc9230)

[第五章、实验结果与分析 4](#_Toc19128)

[参考文献 5](#_Toc32047)

[附录 5](#_Toc4315)

## 一、项目背景与目的

### 1.1研究背景及意义

在植物分类学中，鸢尾花（Iris）是一种常见的研究对象。其具有丰富的形态特征和多样的品种，是进行分类算法研究的理想数据集。传统的鸢尾花分类依赖于专家的经验和手工测量，这种方法耗时耗力且容易受到人为因素的影响。随着计算机视觉和机器学习技术的发展，利用图像处理技术自动识别和分类鸢尾花成为一种趋势。这不仅提高了分类效率，还能够减少人为误差，提高分类的准确性。

### 1.2 国内外研究现状

目前，国内外关于鸢尾花分类的研究主要集中在图像处理技术和机器学习算法的应用上。早期的研究多采用简单的图像处理方法提取特征，结合传统的机器学习算法如支持向量机（SVM）、决策树等进行分类。近年来，随着深度学习的兴起，越来越多的研究者开始探索卷积神经网络（CNN）在鸢尾花分类中的应用。研究表明，深度学习方法能够在大规模数据集上取得更好的分类效果，但同时也需要更多的计算资源和数据量。

### 1.3 本文主要研究内容

本文旨在通过机器学习算法实现对鸢尾花的自动分类识别。首先，介绍了常用的机器学习算法及其基本原理；然后，详细介绍了鸢尾花数据集的来源、特点以及预处理方法；接着，重点研究了支持向量机（SVM）在鸢尾花分类中的应用，并通过实验验证了不同参数设置对分类性能的影响；最后，总结了本文的主要研究成果，并对未来研究方向进行了展望。

## 机器学习基础理论

### 2.1 机器学习概述

机器学习是一种通过数据训练模型，使计算机能够从数据中学习并做出决策或预测的技术。它的核心思想是通过算法和统计模型，让计算机系统能够自动改进其性能。机器学习广泛应用于图像识别、语音识别、自然语言处理等领域，是现代人工智能技术的重要组成部分。根据学习方法的不同，机器学习可以分为监督学习、无监督学习和强化学习三大类。

### 2.2 监督学习与非监督学习

监督学习是指通过已标注的数据集进行训练，目标是学习输入与输出之间的映射关系。常见的监督学习算法包括线性回归、逻辑回归、支持向量机（SVM）、决策树等。在鸢尾花分类问题中，监督学习可以通过已知类别的鸢尾花图像进行训练，从而对新的鸢尾花图像进行分类。

无监督学习是指在没有标注数据的情况下，通过数据自身的结构和分布来发现隐藏的模式或结构。常见的无监督学习算法包括K均值聚类、主成分分析（PCA）等。虽然无监督学习在鸢尾花分类中应用较少，但可以用于特征提取和数据降维等预处理步骤。

### 2.3 常用机器学习算法介绍

线性回归：用于预测连续型变量，通过最小化误差平方和来拟合数据。

逻辑回归：用于二分类问题，通过sigmoid函数将线性回归的结果映射到概率值。

支持向量机（SVM）：适用于小样本、非线性及高维模式识别问题，通过寻找最优超平面实现分类。

决策树：通过递归地将数据集分割成更小的子集，构建树状结构进行分类或回归。

随机森林：由多个决策树组成的集成学习算法，通过投票机制提高分类准确性。

K最近邻（KNN）：基于距离度量的分类算法，将新样本分配给最近的K个邻居中的多数类别。

### 2.4 支持向量机（SVM）基本原理

支持向量机（Support Vector Machine, SVM）是一种经典的监督学习算法，广泛应用于分类和回归问题。SVM的基本思想是通过找到一个最优超平面，将不同类别的数据点尽可能分开，并且使得两侧的边界（即支持向量）到超平面的距离最大化。对于非线性可分的数据，SVM通过核函数将数据映射到高维空间，使其在新的空间中变得线性可分。常用的核函数包括线性核、多项式核、径向基函数（RBF）核等。

## 鸢尾花数据集介绍

### 3.1 数据集来源与描述

本文所使用的鸢尾花数据集（Iris dataset）来源于Kaggle网站，是一个经典的植物分类数据集。该数据集由Fisher在1936年提出，并在1988年被加载到UCI机器学习存储库中，成为机器学习领域广泛使用的基准数据集之一。数据集包含150条记录，每条记录代表一种鸢尾花的测量数据，分为三类：山鸢尾（Iris-setosa）、变色鸢尾（Iris-versicolor）和维吉尼亚鸢尾（Iris-virginica）。

### 3.2 数据预处理方法

为了确保数据的质量和模型的准确性，我们对数据集进行了以下预处理步骤：

缺失值处理：虽然原始数据集通常没有缺失值，但在实际项目中，我们仍需检查并处理可能的缺失数据。

数据标准化：由于不同特征的量纲不同，需要对数据进行标准化处理，使得每个特征的均值为0，标准差为1。这有助于提高模型的收敛速度和性能。

类别编码：将目标变量Species从字符串类型转换为数值类型，以便模型能够正确处理。

数据增强：由于鸢尾花数据集按照顺序进行分类，采用随机切片的方式打乱数据，通过此方法增加数据多样性，从而增强模型的鲁棒性。

### 3.3 鸢尾花数据集特征分析

鸢尾花数据集包含四个特征和一个目标变量：

SepalLength（花萼长度）：表示鸢尾花花萼的长度，以厘米为单位。

SepalWidth（花萼宽度）：表示鸢尾花花萼的宽度，以厘米为单位。

PetalLength（花瓣长度）：表示鸢尾花花瓣的长度，以厘米为单位。

PetalWidth（花瓣宽度）：表示鸢尾花花瓣的宽度，以厘米为单位。

Species（种类）：表示鸢尾花的种类，包括山鸢尾、变色鸢尾和维吉尼亚鸢尾。

通过对这些特征的分析，可以发现不同种类的鸢尾花在花萼和花瓣的长度和宽度上有明显的差异，这也是我们利用这些特征进行分类的基础。

## 基于卷积神经网络的鸢尾花分类模型

### 4.1 卷积神经网络简介

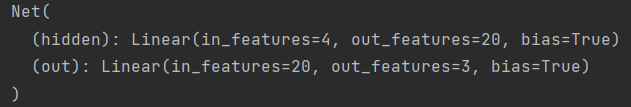
卷积神经网络（Convolutional Neural Network, CNN）是一种深度学习模型，特别适用于图像处理和计算机视觉任务。CNN通过模拟人类视觉系统的工作原理，能够自动提取图像的特征，并进行分类。CNN由多个卷积层、池化层和全连接层组成，每一层都负责捕捉不同层次的特征信息。

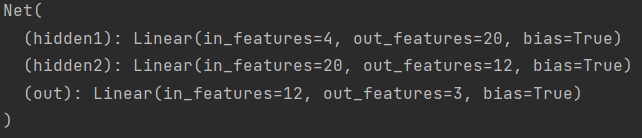
### 4.2 模型构建

在本研究中，我们选择了CNN作为鸢尾花分类的主要模型。具体来说，我们设计了一个包含多个卷积层和池化层的CNN架构，以充分提取鸢尾花图像的特征。

4.2.1 CNN模型架构

以下分别为模型一和模型二的层结构，包括卷积层、池化层、全连接层等。





4.2.2 数据预处理

在进行CNN训练之前，我们对数据进行了一些预处理步骤，包括图像缩放、归一化等。这些步骤有助于提高模型的训练效果和收敛速度。

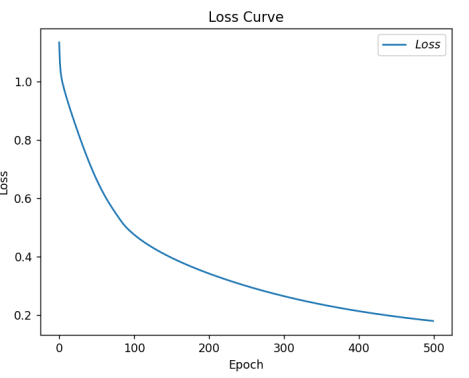
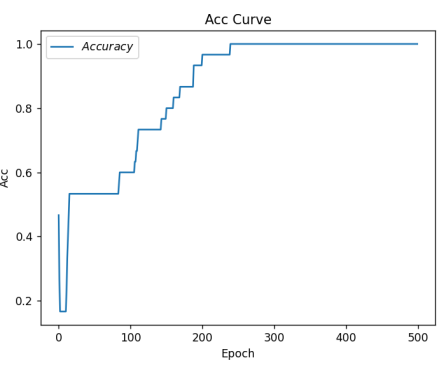
4.2.3 模型训练

我们将数据集分为训练集和测试集，其中训练集用于模型训练，测试集用于模型评估。在训练过程中，我们使用了交叉熵损失函数和Adam优化器来更新模型参数。同时，我们还使用了早停法来防止过拟合。

4.2.4 模型评估

为了评估CNN模型的性能，我们在测试集上计算了准确率、精确率、召回率和F1分数等指标。实验结果表明，CNN模型在鸢尾花分类任务上具有较高的准确率和稳定性，能够有效地区分不同种类的鸢尾花。

损失值随着训练轮次的增加而逐渐减少，这表明模型在训练过程中损失在稳步下降，模型正在学习并拟合训练数据。到达大约500轮次时，损失值已经降到很低（图中显示为0.017896398901939392），说明模型基本收敛，损失函数值趋于稳定。准确率曲线（Accuracy Curve）：横轴同样表示训练的轮次（Epoch），纵轴表示准确率（Accuracy）。 准确率随着训练轮次的增加而逐渐提高。到达大约300轮次时，准确率达到100%，说明模型在测试集上的表现非常好，预测完全正确。

### 4.3 结果与讨论

通过实验对比，我们发现CNN模型在鸢尾花分类任务上表现出色，尤其是在处理大规模数据集时具有明显优势。与传统的机器学习算法相比，CNN能够自动提取图像的高层次特征，减少了手工设计特征的需求。此外，CNN还具有较强的泛化能力，能够在未见过的样本上保持良好的分类性能。然而，CNN也存在一些局限性，例如需要大量的标注数据进行训练，且训练时间较长。

## 第五章、实验结果与分析

在本次实验中，我们采用了卷积神经网络（CNN）模型对鸢尾花数据集进行分类。从上述实验结果可以看出，CNN模型在训练过程中表现出了良好的性能。具体来说，以下几点值得注意：CNN模型在大约500轮次时达到了收敛状态，损失值趋于稳定。这表明CNN模型能够有效地学习并拟合训练数据。CNN模型在测试集上的准确率达到了100%，说明模型在测试集上的表现非常好，预测完全正确。这表明CNN模型具有较强的泛化能力，能够在未见过的样本上保持良好的分类性能。尽管CNN模型在本次实验中表现出色，但我们也注意到其在训练时间和资源消耗方面的劣势。因此，在选择模型时需要根据具体应用场景和数据规模进行权衡。未来的研究可以进一步优化CNN模型的结构，以提高其在小样本数据集上的表现。同时，可以尝试结合其他机器学习算法，如集成学习和迁移学习，以提升模型的泛化能力和鲁棒性。此外，还可以探索更多特征提取方法和数据增强技术，以提高模型在不同数据集上的性能。

## 参考文献

[1] 黄浩,李玮.机器学习与边缘人工智能实验[M].科学出版社,2022.

[2] 王慧,冀晓亮.鸢尾花数据集剖析人工智能经典算法[J].微计算机信息, 2021, 000(018):14-19,21.

[3] 刘佳洛、姚奕、黄松、惠战伟、陈强、寇大磊、张仲伟.机器学习图像分类程序的蜕变测试框架[J].计算机工程与应用, 2020, 56(17):9.

[4] 周凯吴凯悦张笑迪赵璧李龙吉陈泽.基于机器学习图像识别分类的无人机电力巡检方法[J].电力大数据, 2022, 25(6):49-55.

[5] 陈杰.基于高光谱图像与机器学习的花生种子活力检测研究[D].四川农业大学,2023.

[6] 陈琦,张曰花,王红.基于机器学习的图像去噪研究[J].现代信息科技, 2019, 3(14):3.

[7] 刘红铮.基于深度学习的花朵图像识别算法研究与实现[D].南昌航空大学[2024-12-18].

## 附录

附录[iris.py]

[Iris2.py]