

|  |  |
| --- | --- |
| **题 目：** | 数据洞悉美食：机器学习在食物领域的应 |
|  | 用研究 |
| **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级数据科学与大数据技术1班 |
| **学 号：** | 421470112 |
| **学生姓名**： | 符秀惠 |

**提交日期：** 2024 **年** 11 月

**目录**

摘要1

绪论1

训练食物图像数据集的意义1

数据描述2

数据清洗2

机器学习3

数据可视化6

机器学习结论与建议7

**摘 要**

本论文致力于探究实用机器学习在训练食物图像数据集方面的应用。我们首先对收集到的食物图像数据进行详细描述，包括数据规模、种类分布以及标注信息等。随后，通过一系列数据处理步骤，如图像清洗、标注和增强，以提高数据质量和模型训练效果。在机器学习部分，我们利用卷积神经网络等模型进行特征提取和分类任务，并对模型进行参数优化以提升性能。最后，通过数据可视化技术，我们直观地展示了模型在食物图像分类任务上的表现，以及不同食物种类的分布特征。本论文的研究取得了一定成果，所训练的机器学习模型在食物图像识别方面展现出有效性，并为食品行业的数据分析和可视化提供了有益的参考与支持。我们期望这些初步探索能为相关领域的研究和应用提供一定帮助。

关键词：数据集、数据可视化、机器学习

**绪 论**

在当今日益数据驱动的时代，机器学习已经渗透到众多领域，包括与人们日常生活息息相关的食品行业。实用机器学习在食品领域的应用为食品科学、营养学以及食品加工等多个方面带来了革命性的进步。训练食物模型是其中一个重要环节，食物图像数据集的训练有助于我们更深入地了解食物的特征、识别食物种类、分析食物的营养成分以及评估食物的质量。下面将详细探讨训练食物图像数据集的意义以及如何用多种算法对食物图像数据进行分类。

**训练食物图像数据集的意义**

**食物种类识别**

通过训练机器学习模型，我们可以实现对食物种类的自动识别和分类。这对于食品生产、餐饮行业以及消费者来说都是十分有益的。例如，在餐厅点餐时，通过拍摄食物图片并上传至识别系统，系统可以自动识别出食物种类，为消费者提供详细的营养信息和推荐搭配。

**营养成分分析**

食物图像数据集的训练还可以帮助我们分析食物的营养成分。通过对食物图像进行特征提取和模式识别，机器学习模型可以预测食物中的蛋白质、脂肪、碳水化合物等营养成分含量。这对于制定健康饮食计划、评估食品营养价值以及开发功能性食品具有重要意义。

**食物质量评估**

食物图像数据集的训练有助于我们评估食物的质量。通过分析食物的外观、颜色、纹理等特征，机器学习模型可以预测食物的新鲜度、口感以及可能存在的瑕疵。这对于食品生产商、零售商以及消费者来说都至关重要，可以帮助他们及时发现和处理问题，确保食品的质量和安全。

**数据描述**

本次实验我将使用“food-101”图像数据集进行机器学习。该数据来源于OpenDataLab平台，通过搜索食物类别获得多个与食物有关的图像数据集，根据平台提示选取下载了“food-101”的图像数据集。该数据集共包含了101个食物类别，每种类别分别各有1000张图片，总共有101k个图像，可以通过对这些数据的分类实验实现其训练的意义。

**数据清洗**

获取到数据集后，首要做的是数据清洗，图像数据的清洗是数据预处理的一个重要步骤，它可以提高图像识别和图像处理的准确性。

在数据清洗的过程中，我通过对数据进行去重确保了数据集中无重复项，提高数据质量。然后对其进行了数据过滤，根据特定条件筛选出符合需求的数据，减少干扰信息。对于数据中出现的噪声，我进行了数据修复，保证数据的准确性。此外，我对数据进行了数据分割（如图1），将数据划分为训练集和测试集，以便于模型训练和评估。最后，我将清洗后的数据保存本地文件系统中确保后续使用的便捷性和安全性。

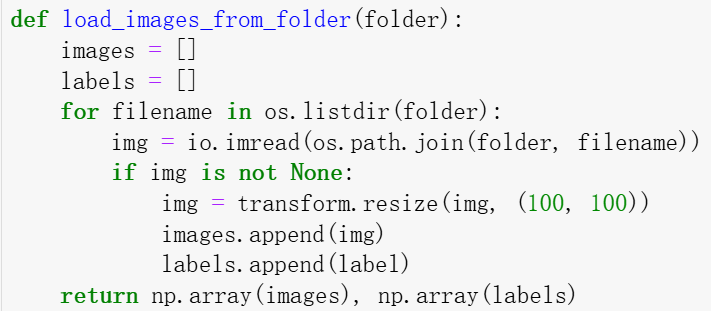


图1 数据分割

**机器学习**

**K近邻（KNN）分类器**

机器学习中，K近邻（KNN）分类器是一种简单且常用的监督学习方法。在图像数据集分类任务中，KNN分类器表现出色，能够有效地对图像进行分类。下面将讲述使用了n\_neighbors=3参数的SVM分类器对图像数据集进行分类的原理

这段代码（如图2）的思路是使用K近邻分类器来对图像进行分类。首先，从指定的文件夹中加载图像，并将每张图像调整为统一大小（100x100像素）。然后，将四维的图像数组（样本数，高度，宽度，通道数）重塑为二维数组（样本数，特征数），以便能够作为KNN分类器的输入。接着，使用训练数据来拟合KNN模型，并用测试数据来评估模型的性能。

这之中的KNN分类器基本原理是：对于一个新的待分类图像样本，计算它与训练集中所已标记样本的距离（通常使用欧氏距离或曼哈顿距离等度量方式）。然后，根据距离大小排序，选择前K个最近的邻居。最后，根据这K个邻居的类别分布，采用多数表决的方式确定新样本的类别。这种“少数服从多数”的原则使得KNN分类器能够有效地处理分类问题。

在图像分类任务中，每个图像样本通常被表示为一系列特征，这些特征可能包括颜色、纹理、形状等。通过提取这些特征，可以将图像数据转换为数值型数据，进而应用KNN分类器进行分类。

使用KNN分类器对图像数据集进行分类后，最终得到的结果是每张测试图像所属的预测类别。这些预测类别是根据K个最近邻居的类别分布通过多数投票确定的。通过对比预测类别与实际类别，可以计算出分类器的性能指标，如F1分数（图3）等，以评估分类器的性能。

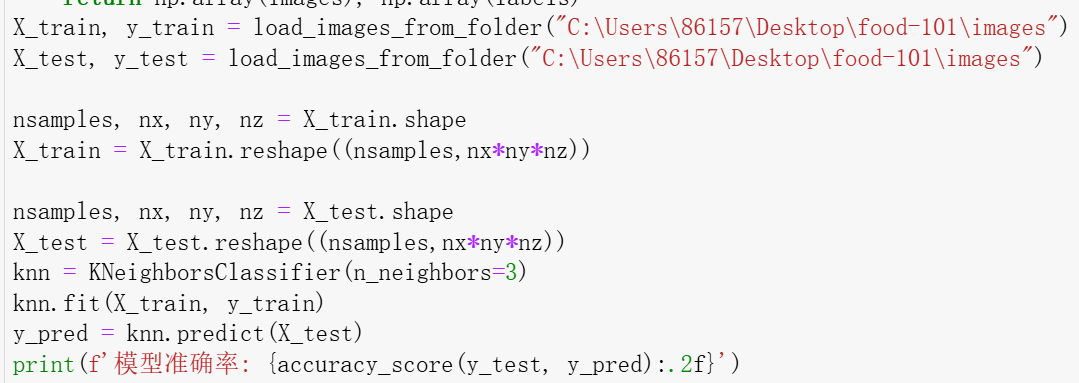


图2



图3

**支持向量机（SVM）分类器**

在机器学习中，支持向量机（SVM）是一种广泛应用的分类算法，特别适用于高维数据的分类问题。当使用SVM分类器对图像数据集进行分类时，可以通过调整不同的参数来优化分类性能。下面将讲述使用了kernel='rbf'、C=100、gamma='scale'参数的SVM分类器对图像数据集进行分类的原理。

这个代码（如图4）通过使用支持向量机分类器对“food-101”图像数据集进行分类。我们利用它从文件夹中加载图像数据，并将每个图像调整为统一的大小（100x100像素），然后将图像数据展平为一维数组，以便可以作为SVM的输入特征。接着，使用StandardScaler对数据进行标准化处理，然后划分数据集为训练集和测试集。最后，使用训练集训练SVM分类器，并用测试集评估分类器的性能。

SVM的基本思想是通过找到一个超平面来将不同类别的数据分开，同时最大化超平面与最近数据点之间的距离，即间隔。这个间隔最大化原则有助于增强分类器的泛化能力。

在SVM分类器的训练过程中，算法会寻找一个最优的超平面来最大化间隔，并通过调整C和gamma等参数来控制模型的复杂度和泛化能力。对于图像数据集，每张图像的特征向量将被作为输入数据，通过SVM分类器进行训练和分类。最终，分类器会根据学习到的模型对新的图像数据进行预测，并输出其所属的类别。

运行代码后，得到一个准确率分数（如图5）和F1（如图6），表示SVM分类器在测试集上的性能。

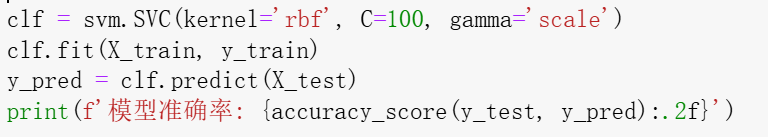


图4



图5



图6

**随机森林分类器**

随机森是一种集成学方法，它在分类和回归任务中都表现得相当出色。在机器学习中，特别是处理图像数据集时，随机森林可以作为一个强大的工具，对图像进行分类。下面讲述使用了n\_estimators和random\_state参数的随机森林分类器对图像数据集进行分类的原理。

随机森林利用“集成学习”的原理，即结合多个模型的预测结果来提高整体性能。单个决策树可能容易受到训练数据的噪声或特定样本的影响，从而产生过拟合。但是，当多个树被组合在一起时，它们的预测错误通常会相互抵消，从而提高整体的预测精度。增加树的数量通常会使模型更加稳定，但也会增加计算成本。

该代码（如图7）先使用训练集来训练随机森林模型，并使用测试集进行预测。通过比较测试集的预测结果和真实标签，我们可以计算出模型的准确率。假设我们的数据集是平衡的（即每个类别的样本数量大致相等），并且模型在测试集上的表现良好，那么F1（如图8）的结果可以作为一个合理的性能评估指标。

使用随机森林分类器对图像数据集进行分类时，n\_estimators 和 random\_state 是两个关键的参数。n\_estimators 控制了随机森林中树的数量，通过增加树的数量可以提高模型的稳定性和准确性；而 random\_state 则用于控制随机数生成器的种子，确保模型的复现性和探索不同随机性对模型性能的影响。通过合理调整这两个参数，可以获得更好的分类效果和模型性能。

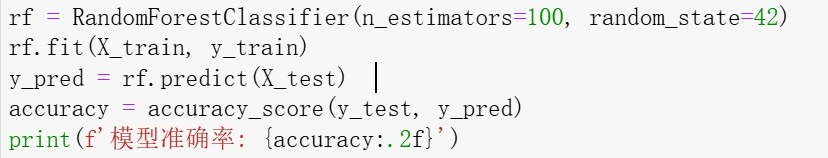


图7



图8

**数据可视化**

数据可视化能直观展示数据的分布、趋势和关联，帮助快速理解复杂数据，发现潜在规律。它提高了数据分析效率，有助于决策制定，对于科学研究和业务应用都具有重要价值。

本代码（如图9）的目的是绘制一个柱状图，展示不同水果的数量。首先对三个水果的类别（categories）和对应的数量（quantities）进行了定义。然后，使用matplotlib.pyplot库的subplots函数创建一个新的图形和坐标轴对象。接着，使用bar函数在坐标轴上绘制柱状图，并通过set\_title、set\_xlabel和set\_ylabel函数分别设置图形的标题、x轴和y轴的标签。最后，使用plt.show()函数显示图形（如图10）。

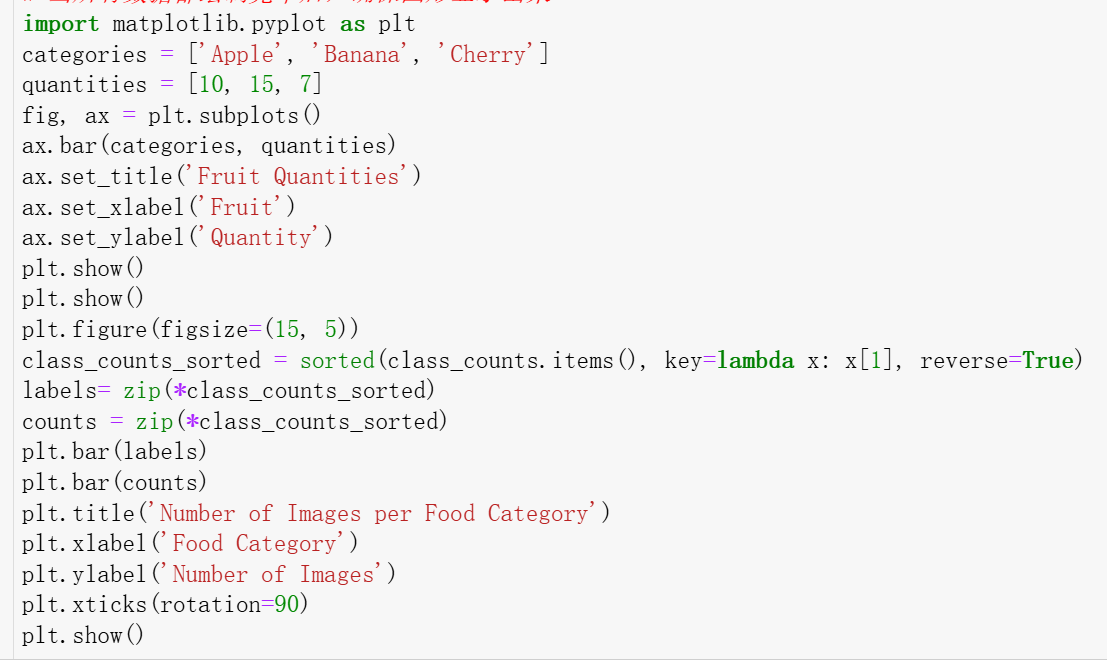


图9

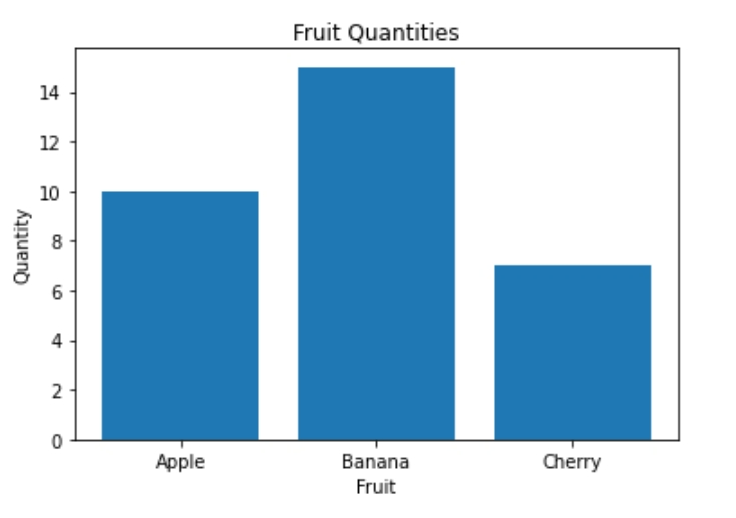
****

图10

**机器学习结论与建议**

在经过对多类别的“food-101”图像数据集进行分类分析后，我们运用KNN、随机森林和SVM三种机器学习算法进行了深入探究，得出了若干结论与建议。

我们观察到在“food-101”数据集上，三种算法均展现出了一定的分类性能，但各有特点。

KNN算法通过邻近样本的类别进行投票分类，对于图像数据中的相似性和局部特征有较好的捕捉能力。然而，在高维图像数据中，KNN可能会面临计算复杂度高和内存消耗大的问题。因此，在实际应用中，可以考虑通过特征提取或降维来减少计算负担，同时选择合适的k值以优化性能。

随机森林算法则通过集成多棵决策树来提高分类的准确性和稳定性。在“food-101”数据集上，随机森林表现出了优秀的分类效果。然而，随机森林模型相对复杂，可能导致过拟合和解释性不足的问题。因此，在构建随机森林模型时，应合理控制树的数量，同时采用适当的特征选择和重要性评估方法，以提高模型的泛化能力和可解释性。

SVM算法通过寻找最大间隔超平面来实现分类，对于非线性问题具有较好的处理能力。在“food-101”数据集上，SVM同样展现了良好的分类性能。然而，SVM对参数的选择较为敏感，且核函数的选择也会影响分类效果。因此，在使用SVM进行图像分类时，需要进行充分的参数调优和核函数选择，以提高分类精度和稳定性。

综上，对于多类别的“food-101”图像数据集，我们可以根据具体需求选择合适的机器学习算法进行分类。在实际应用中，我们需要权衡不同算法的优缺点，结合数据特点进行算法选择和参数调整，以获得最佳的分类效果。同时，随着技术的发展和数据的不断丰富，我们有望进一步提高图像分类的准确性和效率。