

需要解析：F1-scores, accuracy, recall and precision

使用ML算法和方法进行分类

# 摘要

# 引言

机器学习是人工智能的一个重要分支，它的目标是开发和应用算法，使计算机可以从数据中学习并做出预测或决策。机器学习中的分类问题是一类常见的监督学习任务，其主要目标是根据给定的输入数据，将其划分到不同的类别中。在分类问题中，算法通过学习输入数据与其对应的标签之间的关系，从而能够对新的、未标记的数据进行分类。分类问题在实际应用中非常广泛，涉及到图像识别、自然语言处理、医学诊断、金融欺诈检测等多个领域。通过不断改进算法和优化模型参数，研究者们致力于提高分类模型的性能和适应性。

在众多的分类算法中，随机森林和K近邻（KNN）算法是两种广泛使用的方法。随机森林是一种集成学习方法，它通过构建并结合多个决策树来做出预测。KNN算法则是一种基于实例的学习方法，它根据输入数据在特征空间中的邻近样本的类别来做出预测。

本报告的目标是使用随机森林和KNN分类器对Fashion MNIST和Cifar10数据集进行分类并评估我们将首先对数据进行预处理，然后使用随机森林和KNN分类器进行模型训练，最后使用混淆矩阵和分类报告来评估模型的性能。

本文的结构如下︰第二部分介绍了相关工作的最新进展，第三部分是主体部分，将对实验流程进行详细描述。第四部分展示了我们获得的结果，并对每种情况分别进行了讨论。最后，我们以结论和参考文献结束本文。

# 相关工作

本节总结了利用随机森林和KNN进行分类的相关工作。

Goswami [2]（Bosch, A., Zisserman, A., & Munoz, X. (2007). Image Classification using Random Forests and Ferns. In proceedings of IEEE 11th International Conference on Computer Vision.） 在他的⼯作中使⽤ K 最近邻在 CIFAR-10 数据 集上进⾏图像分类，该数据集是⼀个包含 10 个类别的⼩型数 据集。他发现 KNN 对所考虑的数据进⾏分类不是很准确。

王等⼈。[8]（Wang, S., Dou, A., Yuan, X., & Zhang, X. (2016). The airborne hyperspectral image classification based on the random forest algorithm. In proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS).） 在他们的研究中，研究了使⽤随机森林算法的机 载⾼光谱图像分类⽅法，并根据他们的研究开发了⼀个 C++ 程序。他们的数据集包括张掖市的⾼光谱图像。

Krithika 和 Selvarani [9]（Krithika, N., & Selvarani, A. G. (2017). An individual grape leaf disease identification using leaf skeletons and KNN classification. In proceedings of International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).） 在他们的⼯作中使⽤ K 最近邻分类 来对葡萄叶病进⾏分类并识别叶⻣架。使⽤葡萄图像识别叶 ⼦⻣架。他们发现通过使⽤ KNN 分类可以有效地对葡萄叶部 病害进⾏分类。

Gardner 和 Dorling [5]（Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment, 32(14-15), 2627-2636.） 在他们的⼯作中使⽤了多层感知器作 为⼤⽓科学中的应⽤。

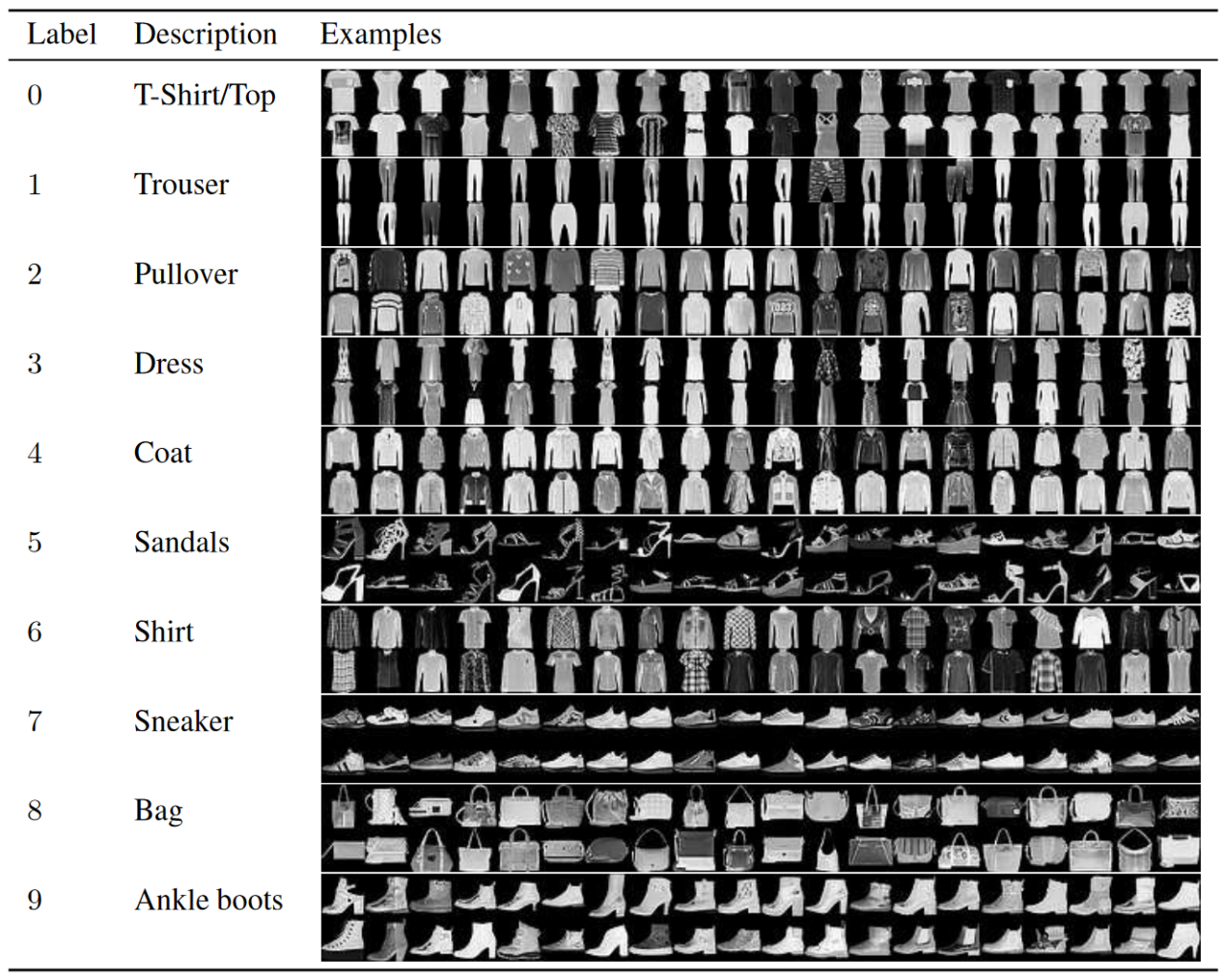
拉纳等⼈。[6]（Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. In proceedings of International Journal of Computer Applications, 67(10).）在他们的⼯作中⽐较了 使⽤三种不同距离度量的K-means算法，即欧⼏⾥德距离、 曼哈顿距离和明可夫斯基距离。他们发现欧⼏⾥得距离提供 可以使⽤各种机器学习算法对图像进⾏分类。数据科学提供 了最好的结果，⽽曼哈顿距离提供了最差的结果。

# 主体

## 数据集

本研究采用了时尚-MNIST数据集（http://arxiv.org/abs/1708.07747）该数据集包括一个由60,000张样本图像组成的训练集和一个由10,000张灰度样本图像组成的测试集。数据集中的每幅图像分辨率为28x28像素。每张训练和测试图像都属于10个类别之一，包括T恤/上衣、裤子、毛衣、连衣裙、外套、凉鞋、衬衫、运动鞋、包和靴子。每个类别有6000张样本图像。每一行都是单独的图像，第1列是类别标签，其余各列是像素编号（共784个)。

able x: Class names and example images in Fashion-MNIST dataset.



Cifar10数据集分为训练集和测试集，其中训练集包含50000张、测试集包含300000张图像。 在测试集中，10000张图像将被用于评估，而剩下的290000张图像将不会被进行评估，包含它们只是为了防止手动标记测试集并提交标记结果。 两个数据集中的图像都是png格式，高度和宽度均为32像素并有三个颜色通道（RGB）。 这些图片共涵盖10个类别：飞机、汽车、鸟类、猫、鹿、狗、青蛙、马、船和卡车。

## 数据预处理

图像分类中的预处理是一个对图像进行转换以准备分类的过程。这些步骤包括导入必要的库，然后从文件夹中检索图像。然后调整图像大小并转换为 NumPy 数组。然后将数组分割为训练数据和测试数据。 70%的数据用于训练，30%用于测试。（https://ieeexplore.ieee.org/document/9058042）

## 分类器介绍

### 随机森林

随机森林（Random Forest）是一种集成学习方法，它通过构建多个决策树并将它们整合在一起来进行分类或回归任务。随机森林是由Leo Breiman在2001年提出的（Breiman L, 2001a.Random forests.Mach.Learn., 45:5-32.），它在机器学习中被广泛应用，因为它具有良好的性能和鲁棒性。随机森林由多个决策树组成。每个决策树都是一个分类器，通过对输入数据进行逐层的判定，最终输出属于哪个类别。随机森林在构建每棵决策树时引入了随机性。对于分类任务，随机森林采用投票机制进行决策。每棵决策树对输入数据进行分类，最终的分类结果是获得最多投票的类别。随机森林通过组合多个决策树的预测结果，减小了过拟合的风险，提高了模型的鲁棒性和泛化能力。它对于处理高维数据和大规模数据集都表现出色。

随机森林中的树按以下规则生成：

　1）如果训练集大小为N，对于每棵树而言，随机且有放回地从训练集中的抽取N个训练样本（这种采样方式称为bootstrap sample方法, 为拔靴法采样），作为该树的训练集；从这里我们可以知道：每棵树的训练集都是不同的，而且里面包含重复的训练样本。

　　2）如果存在M个特征维度，则指定数量m << M，使得在每个节点处，从M中随机选择m个特征维度，并且使用这些m个特征维度中最佳特征(最大化 information gain)来分割节点。在森林生长期间，m的值保持不变。

　　3）每棵树都尽最大程度的生长，并且没有剪枝过程。

### KNN

使用KNN算法应用图像分类。在模式确认中，KNN算法是一种用于分类和回归的非参数方法。输出基于KNN是用于排序还是回归：在KNN分类中，最后一个输出是周期附属。一个对象是通过其邻居的极端投票来分类的，被分配到其k个最近邻居中最常见的类的对象，其中k是正整数，通常很小。如果k＝1，则该对象仅被选择为唯一一个最近邻居的类。在KNN回归中，最后一个乘积是对象的属性值。该特定值是其k个最近邻居的正常值。KNN是一种基于发生率的学习，或缓慢的实现，其中体积与近处相似，所有计算都被允许，直到顺序。KNN计算是所有机器学习计算中最直接的计算之一。KNN分类器的执行基本上由K的决定以及所连接的分离度量来决定。该度量受到区域度量K的确定的可影响性的影响，理由是附近区域的扫描是由与问题的第K个最近邻居的分离决定的，并且不同的K产生不同的限制性类概率。在K很小的情况下，由于信息匮乏的条件和喧闹、模糊或错误标记的焦点，附近的量规通常会非常差。最终目标是进一步平滑度量，我们可以构建K，并围绕这个问题考虑一个实质性的区域设置。令人震惊的是，对K的广泛估计有效地使规范过度平滑，并且排列执行随着不同类别异常的出现而降低。为了解决这一问题，已经进行了相关的研究工作，以提高KNN的安排执行力。(https://ieeexplore.ieee.org/document/9058042)

KNN（k近邻算法）的步骤:

1. 计算测试数据与各个训练数据之间的距离
2. 按照升序（从小到大）对距离（欧氏距离）进行排序
3. 选取距离最小的前k个点
4. 确定前k个点所在类别出现的频率
5. 返回前k个点中出现频率最高的类别作为测试数据的分类

# 结果与讨论

# 结论

# 参考文献

[Fashion Images Classification using Machine Learning, Deep Learning and Transfer Learning Models | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9786364)

# 附录（代码）