模式识别 KNN

22009200601汤栋文

2024年10月31日

目录

[摘要 2](#_Toc181302157)

[1. 方法 2](#_Toc181302158)

[1.1 数据集 2](#_Toc181302159)

[1.2 验证方式 2](#_Toc181302160)

[2. 实验 2](#_Toc181302161)

[1.1 将原数据直接进行KNN 2](#_Toc181302162)

[1.2 将原数据降维后进行KNN 3](#_Toc181302163)

[3. 结论 3](#_Toc181302164)

# 摘要

本研究使用MNIST和CIFAR10数据集，评估了K最近邻（KNN）算法在不同数据集上的分类性能，并在GPU并行计算的情况下，评估了计算成本。

1. 分析了KNN算法在多个数据集上的准确率。
2. 分析了KNN算法在不同规模数据集和距离度量下的计算成本。

# 方法

## 数据集

本研究使用了MNIST和CIFAR10数据集。MNIST数据集包含了70,000个手写数字的灰度图像（其中训练集60,000个，测试集10,000个），每个图像的大小为28x28像素，标记为0到9的数字类别。CIFAR-10数据集包含了60,000张彩色图像（其中训练集50,000张，测试集10,000张），每张图像的大小为32x32像素，共有10个类别，包括飞机、汽车等。

为了消除特征间的量纲影响，我们对原始图像数据进行了标准化处理，即对每个像素值减去平均值并除以标准差，使得每个特征（像素值）具有零均值和单位方差。

## 验证方式

使用MNIST和CIFAR的官方数据集划分，划分为训练集和测试集。使用KNN算法分类后评估准确率。以及在相同的代码下的运行时间。

# 实验

## 将原数据直接进行KNN

|  |  |
| --- | --- |
| **K** | **acc (%)** |
| 1 | 96.9 |
| 3 | 97.2 |
| 10 | 96.8 |
| 50 | 95.4 |
| 200 | 92.9 |

|  |  |
| --- | --- |
| **K** | **acc (%)** |
| 1 | 35.4 |
| 3 | 35.6 |
| 10 | 34.7 |
| 50 | 32.7 |
| 200 | 29.7 |

MNIST数据集 CIFAR10数据集

|  |  |  |
| --- | --- | --- |
| **metrics** | **acc (%)** | **time (s)** |
| Euclidean | 97.2 | 29.0 |
| Manhattan | 96.4 | 6.4 |
| Cosine | 97.4 | 21.9 |

|  |  |  |
| --- | --- | --- |
| **metrics** | **acc (%)** | **time (s)** |
| Euclidean | 35.6 | 87.2 |
| Manhattan | 39.2 | 15.2 |
| Cosine | 37.7 | 58.7 |

## 将原数据降维后进行KNN

CIFAR10 (K=3)

|  |  |  |  |
| --- | --- | --- | --- |
| **method** | **dim** | **acc (%)** | **time (s)** |
| PCA | 32 | 40.5 | 4.4 |
| PCA | 128 | 38.5 | 7.3 |
| PCA | 1024 | 35.8 | 30.3 |
| LDA | 8 | 31.0 | 11.8 |

**结论：**

1. KNN算法简单直接，分类有效能够在简单的数据集上得到不错的结果。
2. KNN算法的时间消耗较多，特别是当数据量较大时。

# 结论

本研究表明，KNN算法在简单数据集上表现良好，能够得到较高的分类准确率。然而，KNN算法的计算时间较长，特别是在数据量较大时。我们在不同的距离度量下测试了KNN算法，发现欧几里得距离和余弦距离在MNIST数据集上的表现较优，而在CIFAR10数据集上曼哈顿距离的准确率更高。利用PCA和t-SNE对数据进行降维处理后，KNN算法的准确率有所下降，但计算时间明显减少。综上所述，KNN算法在某些场景下具有实用性，但其计算时间需在实际应用中加以考虑。