

**人工智能数学基础实验报告**

题 目 数据降维（PCA+RPCA）

学 院 计算机科学与技术

专 业 人工智能

学 号 2021112845

学 生 张智雄

任 课 教 师 刘绍辉

哈尔滨工业大学计算机科学与技术学院

2023.5

**实验二:** **数据降维（PCA+RPCA）**

1. **实验内容或者文献情况介绍**

**1.1 数据降维的背景**

数据降维(Dimensionality Reduction)是在机器学习和数据分析中经常使用的技术，用于处理高维数据集的复杂性和冗余性。

高维数据包含大量的特征或变量，对计算和存储资源要求高，模型的复杂性高，且缺乏一定的可解释性；同时，高维数据易造成“维度灾难”，即在高维空间中，数据点之间的距离变得非常稀疏，使得数据分布变得不均匀，导致模型的泛化能力下降，容易出现过拟合问题。

数据降维能够在从原始高维空间中提取出最相关和最重要的特征、去除冗余特征，在保留数据分布特征的同时减少数据的维度，提高计算效率，并改善模型的性能和解释性。

**1.2 实验内容**

理解主成分分析(PCA)和鲁棒主成分分析(RPCA)的基本原理，并使用PCA和RPCA用来对MNIST数据集进行分类。

1. **算法简介及其实现细节**

**2.1 算法简介**

主成分分析（Principal Component Analysis,PCA) 基于协方差矩阵进行线性变换从而将高维数据转换为低维空间，并最大程度地保留原始数据的方差。

具体而言，PCA的主要思想是从原始的空间中顺序地找一组相互正交的坐标轴，将维特征映射到全新的维正交特征向量上。而对于维正交特征向量的选择，一般选择协方差矩阵对应特征值最大的个特征向量，相当于保留个原始数据中方差最大的方向，具体伪代码如下。

**ALGORITHM 1** Principal Component Analysis (主成分分析)

1：**input** 高维数据矩阵()，降维目标；

2： //数据中心化；

3：协方差矩阵；

4：求出的特征值和特征矩阵，前大的特征值对应特征向量构成；

5：**return** //返回降维后的矩阵；

鲁棒主成分分析(Robust Principal Component Analysis,RPCA)可以将观测的高维数据分解为低秩成分和稀疏成分，能够处理含有异常值或噪声的数据的降维。

若𝑁为满足稀疏约束的噪声矩阵，问题可以形式化为：

从而可以通过拉格朗日乘子法、矩阵奇异值分解(SVD)以及软阈值函数进行迭代求解，具体推导过程后续给出，下给出伪代码。

**ALGORITHM 2** Robust Principal Component Analysis (鲁棒主成分分析)

1：**input** 高维观测矩阵，迭代次数,收敛阈值；

2：，；

3：**while** **do**

4： ；

5： ，；

6： ；

7： ；

8： **if**  **then goto** 10；//达到收敛条件

9：**end while**

10：**return** //返回低秩矩阵和稀疏矩阵；

**2.2 理论推导**

2.2.1 主成分分析PCA

假设我们有一个样本集，每个样本的特征数为，那么我们可以用一个矩阵来表示这个样本集。

那么我们希望能够找到一个的投影矩阵，为的行向量，使得，实现对的降维。

而对于中的任意一个样本，经过投影过后得到，表示第维新的特征，表示第个样本经降维投影后的第维特征，则有，于是第维新特征的样本均值为

要使在矩阵中的个维度最大程度保留的数据特征，则等价于此个方向上方差最大，由此可将问题形式化为

则可等价变形为

令，则问题形式可简化为，

则使用拉格朗日乘子法构造函数组如下

对内各元素求偏导得到

则对于，解得，则两边同乘可以得到等式

由此推出为协方差矩阵对应的最大特征值的特征向量。

而对于，解得，则两边同乘可以得到等式

而由于前面两项正交，可得到

；

所以，所以，由此推出为协方差矩阵对应的第二大特征值的特征向量。

同理分别对应协方差矩阵的前大特征值的特征向量，组合即可得到投影矩阵。

将投影矩阵与输入矩阵做矩阵乘法即可得到降维后的目标主成分矩阵，包含中方差最大的个特征。

2.2.2 鲁棒主成分分析RPCA

设观测矩阵为()，𝑁为满足稀疏约束的噪声矩阵，为的低秩矩阵，则目标函数为，其中。由于秩函数和范数均为非凸，所以此问题是一个NP-hard问题。

而在稀疏建模中，范数是范数的最佳凸松弛，而矩阵核范数是𝑟𝑎𝑛𝑘(⋅)函数的最佳凸松弛，因此上述NP问题可以转化为

从而可以使用增广拉格朗日方法ALM和交替方向法ADM对问题进行求解，首先构造拉格朗日函数

其中，为拉格朗日乘子。而后增加惩罚项，将有约束问题转化为无约束问题

接下来使用交替方向乘子法，在每个迭代周期内，每一步只更新一个变量而固定另外其余变量，如此交替重复更新，由此化简可以得到关于和的化简的函数如下

查阅资料发现软阈值函数可以求解这类优化问题，得到稀疏矩阵。的具体定义如下

手机屏幕截图

描述已自动生成

Figure 软阈值函数

则稀疏矩阵可以由软阈值函数直接解出，而对于低秩矩阵，我们需要首先对进行奇异值的分解，将分解得到的那个对角矩阵进行稀疏求解，稀疏即表明的低秩，由此可以给出的表达式为

最后更新拉格朗日乘子矩阵，重复和的计算直至达到收敛条件即为最终的低秩矩阵和稀疏矩阵。

1. **实验设置及结果分析（包括实验数据集）**

MNIST数据集（Mixed National Institute of Standards and Technology database）是美国国家标准与技术研究院收集整理的大型手写数字数据集，包含了60,000个样本的训练集以及10,000个样本的测试集。其中包括0到9的数字。每个图像是2828像素的灰度图像。

图形用户界面, 应用程序

描述已自动生成

Figure MNIST数据集

实验将对MNIST数据集分别通过PCA和Robust-PCA方法进行降维处理，并对降维后的数据进行分类。读取数据时，首先将每张图片的像素展开为一维的行向量，拼接得到的训练集矩阵和的测试集矩阵。

**3.1 主成分分析PCA**

实验通过PCA对进行降维处理，得到投影矩阵和降维后的。将作用到上，得到降维后的。选择中区分度最大的两个或三个方向(即的前2或3个特征)进行可视化，在空间中分布的散点图如下。

图表, 散点图

描述已自动生成

Figure 数据空间可视化散点图

（左图为2维空间，右图为3维空间）

通过直接观察可以发现，降至2或3维已经能区分开部分数据，但是大部分数据仍不能准确分类，因而需要增加目标主成分的维数。

将降维后的训练集在SVM分类器模型上进行训练，并在测试集上进行分类正确率检验。实验发现，随目标主成分维数的增长，分类正确率先迅速增长而后趋于稳定，在时，正确率达到90%；在时，正确率达到97%，而当后，正确率几乎稳定在98%左右，说明此时增加的成分对分类影响很小，在实际应用中可以忽略。

图表

描述已自动生成

Figure 分类正确率随主成分维数的变化

取时，将降维得到的主成分输出为图像如下，发现几乎只保留了模糊的数字轮廓主体特征，数字基本可辨识，此时预测正确率为97.56%。

手机屏幕截图

低可信度描述已自动生成

Figure 主成分图像（=20）

而当取时，将降维得到的主成分输出为图像如下，发现保留的轮廓特征更加模糊，肉眼难以直接辨识，此时机器分类预测正确率只有47.78%。

图片包含 游戏机

描述已自动生成

Figure 主成分图像（=2）

对比和时的具体分类结果发现，当目标主成分维数较低时，主要保留的是粗粒度层面上的数字特征，对于4、7和9，5、6和8，2和3等结构相似的数字不能较好地区分。而随维数增加，更多细节层面的特征加入，分类解耦的效果也逐渐变好。

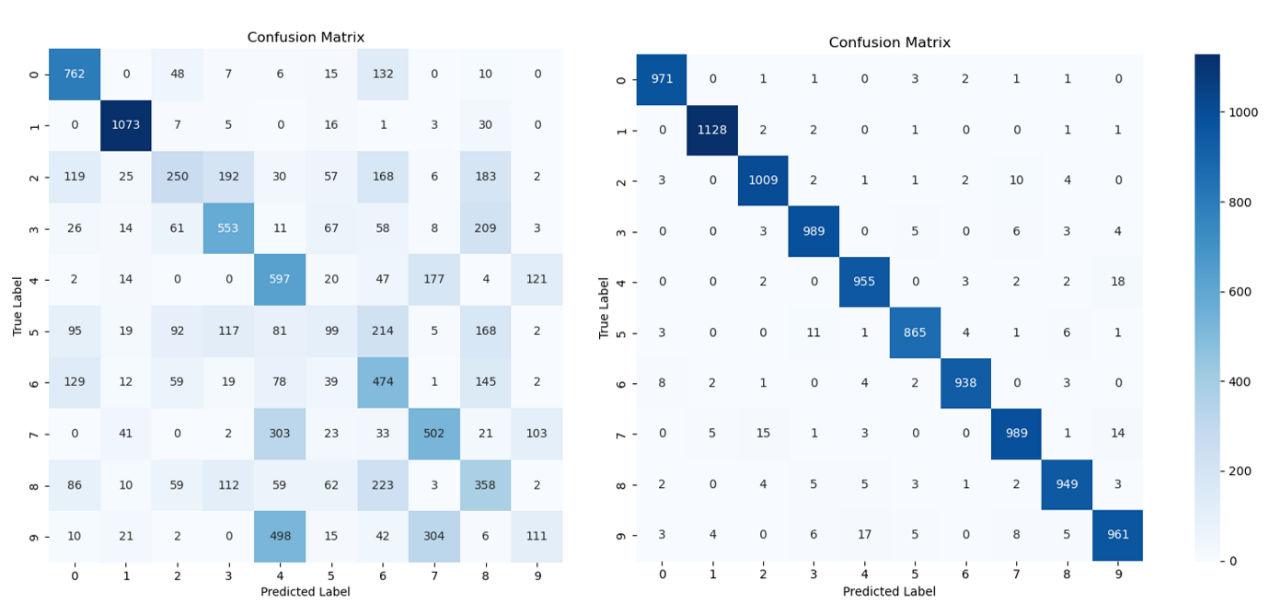


Figure 分类结果混淆矩阵

（左图=2，右图=20）

实际应用中，常常希望保留90%以上的信息量，在本例中对应为，此时分类正确率为98.44%，此时得到的图像更为清晰，并且保留了数字周围的部分背景，与原图的相似度更高。

图片包含 游戏机

描述已自动生成

Figure 主成分图像（）

**3.2 鲁棒主成分分析RPCA**

鲁棒主成分分析可以将观测的高维数据分解为低秩成分和稀疏成分，能够处理含有异常值或噪声的数据的降维。直接在MNIST数据集上使用RPCA得到的图像仅去掉了简单的数字轮廓，数字结构基本保留，具体结果如下。此时分类的正确率为97.18%。

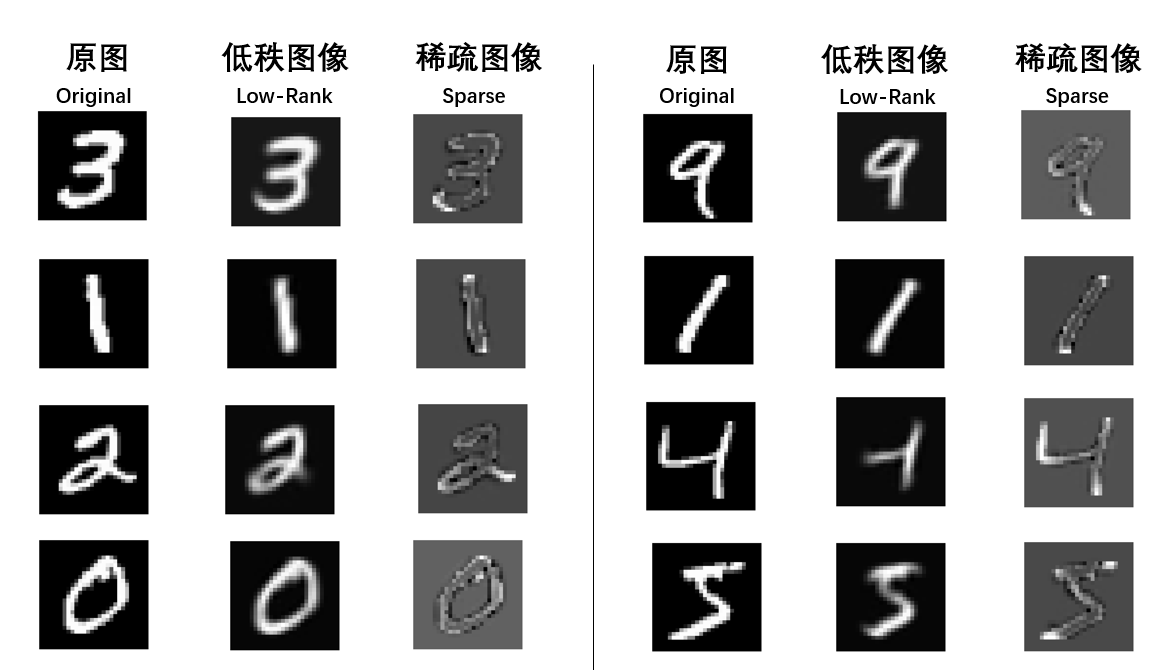


Figure 不加噪声的RPCA处理

而为了验证算法的鲁棒性，实验对数据集图像进行随机噪声处理，分别对图像随机加入服从分布的高斯噪声和10%的掩膜噪声（即随机选择10%的像素点置为白色），而后进行RPCA降维处理，同样通过训练SVM模型进行分类，返回模型在测试集上的预测正确率，观察得到图像以及正确率。

3.2.1 高斯噪声下RPCA

加入服从分布的高斯噪声后，经过RPCA降维处理后能一定程度上处理噪声，保留数字主体结构，分离后的低秩图像Low-Rand和稀疏图像Sparse与原图对比如下。此时分类的正确率为96.94%。

图片包含 QR 代码

描述已自动生成

Figure 高斯噪声的RPCA处理

3.2.2 掩膜噪声下RPCA

随机选择10%的像素点置为白色后，经过RPCA降维处理后能够较好地处理干扰，保留或还原数字结构，分离后的低秩图像Low-Rand和稀疏图像Sparse与原图对比如下。此时分类的正确率为96.86%。

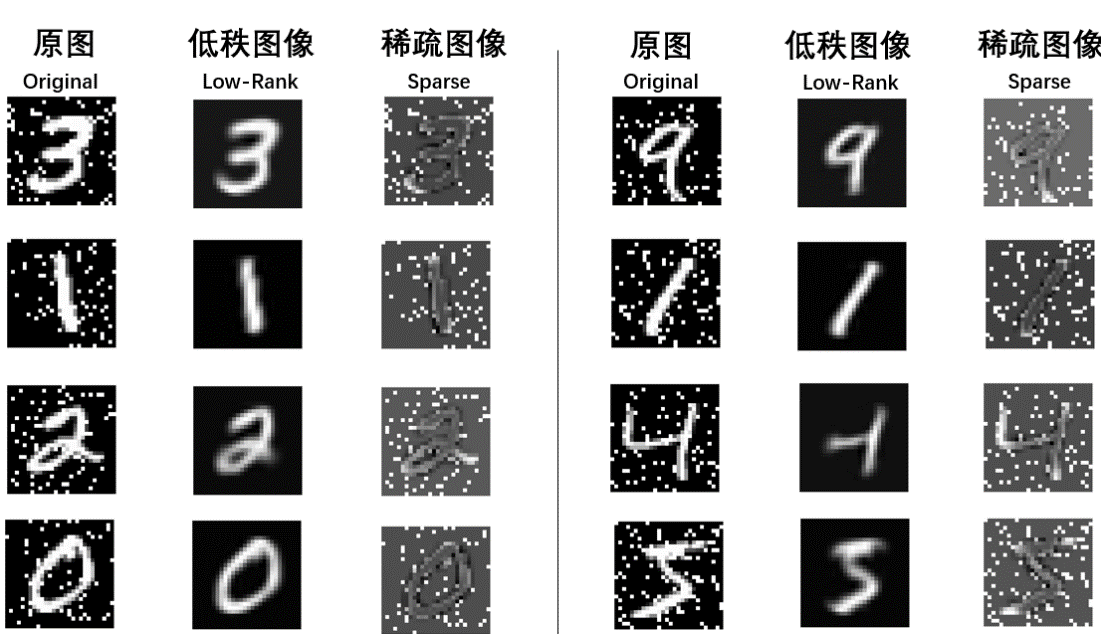


Figure 掩膜噪声的RPCA处理

3.2.3 对比总结

对比上述三次RPCA的结果发现，RPCA能够较好地保留数字具有区分度的结构，而对其余部分则当成稀疏噪声处理。如对数字4，经处理后为一个近似的字符，可能是因为对于4而言，的结构已经足够机器在此范围内识别，而如果放在不同的数据集内，可能会有不同的结果。

而RPCA图像较PCA ()更为清晰，而分类正确率却更低的原因可能是RPCA每次迭代都是选择当前最优的处理，变化可能较小，且并未包含PCA中矩阵基向量的变换步骤。对于机器而言，识别到的特征区分度不如PCA得到的方向区分度大，因而正确率会有所下降。

1. **结论**

Python运行一次PCA所用时间大约在2分钟左右，而运行一次RPCA所用时间大约在5分钟左右。就运行速度而言，PCA算法快于RPCA。且在无噪声干扰下，选择合适的目标维数，PCA分类预测正确率更高，维数更低。

PCA基于协方差矩阵的特征向量分析，从而找到数据中方差最大的方向来减少维度，保留数据中的主成分，但不专门处理噪声，对于噪声较为敏感，常用于数据的特征提取和可视化等领域。

RPCA将数据表示为低秩结构和稀疏结构的线性组合，利用噪声稀疏性的假设和凸优化方法进行求解，能够有效排除噪声对图像的影响，处理含有异常值的数据，鲁棒性较好，常用于图像去噪复原、视频分析和异常检测等领域。

总的来说，PCA和RPCA都能对数据进行降维处理，但PCA更注重提取数据的主要特征，并通过保留主成分来降低数据维度；而RPCA则更注重在包含异常值或噪声的数据中找到低秩和稀疏结构的表示，以便更好地处理异常情况。实际应用中，应根据数据的特点和分析的目标选择相应的算法。

1. **参考文献**

[1] 周志华. 机器学习[M]. 清华大学出版社, 2016.

[2] 杜子芳. 多元统计分析[M]. 清华大学出版社, 2016.

[3] RPCA原理初探 https://blog.csdn.net/qq\_41851166/article/details/108923500

[4] 主成分分析（PCA）原理和鲁棒主成分分析（RPCA）详解https://blog.csdn.net/ qq\_20199965/article/details/102657192

[5] 主成分分析（principal component analysis, PCA）公式https://blog.csdn.net/ kdazhe/article/details/104737018