哈爾濱工業大學

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。

2、 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n维特征映射到全新的k维正交特征向量上。而对于k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k个特征向量，相当于保留k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** $X \leftarrow$高维数据矩阵$(m \times n)$，$k \leftarrow$降维目标；

2：$X \leftarrow (X - \bar{X})$ //数据中心化；

3：协方差矩阵$C \leftarrow 1/m(X \cdot X^T)$；

4：求出C的特征值和特征矩阵，$W \leftarrow C$前k大的特征值对应特征向量构成；

5：**return** $W \cdot X$ //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：

$$\min_{L,N} rank(L) + \lambda\|N\|_0, \qquad s.t.\ M = L + N$$

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$T_\varepsilon(M)$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** M ←高维观测矩阵，$iter$ ←迭代次数，ε ←收敛阈值；

2：$L = \mathbf{0}$，$N = \mathbf{0}$，$Y, \mu > 0$，$\rho > 1$；

3：**while** $i < iter$ **do**

4：　　$U, \Sigma, V = SVD(X - N + Y/\mu)$；

5：　　$L = U \cdot T_{1/\mu}(\Sigma) \cdot V^T$，$N = T_{\lambda/\mu}(M - L + Y/\mu)$；

6：　　$Y = Y + \mu(M - L - N)$；

7：　　$\mu = \rho\mu$；

8：　　**if** $\|M - L - N\|_F \le \varepsilon$ **then goto** 10；//达到收敛条件

9：**end while**

10：**return** L, N //返回低秩矩阵L和稀疏矩阵N；

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$\{x^1, x^2, \dots, x^m\}$，每个样本的特征数为n，那么我们可以用一个$n \times m$矩阵X来表示这个样本集。

$$X = (x^1, x^2, \dots, x^m) = \begin{pmatrix} x_1^1 & \cdots & x_1^m \\ \vdots & \ddots & \vdots \\ x_n^1 & \cdots & x_n^m \end{pmatrix}$$

那么我们希望能够找到一个$k \times n$的投影矩阵$W = [w^1, w^2, \dots, w^k]^T$，$w^i$为$1 \times n$的行向量，使得$W \cdot X = Z(k \times m) \leftarrow \{z^1, z^2, \dots, z^m\}$，实现对$X$的降维。

而对于X中的任意一个样本x，经过W投影过后得到$z = Wx$，z_i表示第i维新的特征，z_i^j表示第j个样本x^j经降维投影后的第i维特征，则有$z_i^j = w^i \cdot x^j$，于是第i维新特征的样本均值$\overline{z_\iota}$为

$$\overline{z_\iota} = \frac{1}{m}\sum\nolimits_{j=1}^{m} z_i^j = \frac{1}{m}\sum\nolimits_{j=1}^{m} w^i \cdot x^j = w^i \cdot \frac{1}{m}\sum\nolimits_{j=1}^{m} x^j = w^i \cdot \bar{x}$$

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

$$max\ Var(z_i)=\frac{1}{m}\sum_{j=1}^{m}\left(z_i^j-\overline{z}_i\right)^2,\ \left\|w^i\right\|_2=1\ 且(w^i)^T\cdot w^j=0$$

则$Var(z_i)$可等价变形为

$$\begin{aligned}Var(z_i)&=\frac{1}{m}\sum_{j=1}^{m}\left(z_i^j-\overline{z}_i\right)^2=\frac{1}{m}\sum_{j=1}^{m}\left(w^i\cdot x^j-w^i\cdot\bar{x}\right)^2\\&=\frac{1}{m}\sum_{j=1}^{m}\left(w^i\cdot\left(x^j-\bar{x}\right)\right)^2=\frac{1}{m}\sum_{j=1}^{m}\left(w^i\right)^T\left(x^j-\bar{x}\right)\left(x^j-\bar{x}\right)^Tw^i\\&=\left(w^i\right)^T\frac{1}{m}\sum_{j=1}^{m}\left(x^j-\bar{x}\right)\left(x^j-\bar{x}\right)^Tw^i=\left(w^i\right)^TCov(x)w^i\end{aligned}$$

令$S=\ Cov(x)$，则问题形式可简化为，

$$max\ \left(w^i\right)^TSw^i,\ \left\|w^i\right\|_2=\left(w^i\right)^Tw^i=1\ 且(w^i)^T\cdot w^j=0$$

则使用拉格朗日乘子法构造函数组$g(w)$如下

$$\begin{cases}g(w^1)=(w^1)^TSw^1-\alpha((w^1)^Tw^1-1)\\g(w^2)=(w^2)^TSw^2-\alpha((w^2)^Tw^2-1)-\beta((w^2)^Tw^1-0)\\\dots\dots\\g(w^k)=(w^2)^TSw^2-\alpha((w^2)^Tw^2-1)-\sum_{j=1}^{k-1}\beta_j\left((w^k)^Tw^j-0\right)\end{cases}$$

对w^i内各元素$w_1^i,w_2^i,\dots,w_n^i$求偏导得到

$$\partial g(w^1)/\partial w_1^1=0\,,\partial g(w^1)/\partial w_2^1=0\,,\dots,\partial g(w^k)/\partial w_n^k=0$$

则对于w^1，解得$Sw^1-\alpha w^1=0$，则两边同乘$(w^1)^T$可以得到等式

$$(w^1)^TSw^1=\alpha(w^1)^Tw^1=\alpha$$

由此推出w^1为协方差矩阵S对应的最大特征值λ_1的特征向量。

而对于w^2，解得$Sw^2-\alpha w^2-\beta w^1=0$，则两边同乘$(w^1)^T$可以得到等式

$$(w^1)^TSw^2-\alpha(w^1)^Tw^2-\beta(w^1)^Tw^1=0$$

而由于前面两项正交，可得到

$$((w^1)^TSw^2)^T=(w^2)^TS^Tw^1=(w^2)^TSw^1=\lambda_1(w^2)^Tw^1=0;$$

$$\alpha(w^1)^Tw^2=0$$

所以$\beta=0$，所以$Sw^2-\alpha w^2=0$，由此推出w^2为协方差矩阵S对应的第二大特征值λ_2的特征向量。

同理$w^3,\dots,w^k$分别对应协方差矩阵S的前k大特征值$\lambda_3,\dots,\lambda_k$的特征向量，组合即可得到投影矩阵$W$。

将投影矩阵与输入矩阵做矩阵乘法$W\cdot X=Z$即可得到降维后的目标主成分矩阵Z，包含X中方差最大的k个特征。

2.2.2 鲁棒主成分分析 RPCA

设观测矩阵为$X(m\times n)$，N为满足稀疏约束的噪声矩阵，L为X的低秩矩阵，则目标函数为$min\,rank(L)+\lambda\|N\|_0$，其中$\lambda=1/\sqrt{max(m,n)}$。由于秩函数和$l_0$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，l_1范数是l_0范数的最佳凸松弛，而矩阵核范数是$rank(\cdot)$函数的最佳凸松弛，因此上述 NP 问题可以转化为

$$\min_{L,N}||L||_*+\lambda||N||_1,\qquad s.t.\ X=L+N$$

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

$$L(L,N,Y)=||L||_*+\lambda||N||_1+<Y,X-L-N>$$

其中，Y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

$$L(L,N,Y,\mu)=||L||_*+\lambda||N||_1+<Y,X-L-N>+\frac{\mu}{2}\left||X-L-N|\right|_F^2(\mu>0)$$

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$L(L,N,Y,\mu)$可以得到关于L和N的化简的函数如下

$$\begin{cases} L=arg\min_{L}\frac{1}{\mu}||L||_1+\frac{1}{2}||L-(X-A+Y/\mu)||_F^2 \\ N=arg\min_{N}\frac{\lambda}{\mu}||N||_1+\frac{1}{2}||N-(X-A+Y/\mu)||_F^2 \end{cases}$$

查阅资料发现软阈值函数$T_\varepsilon(M)$可以求解$arg\min_{X}\varepsilon||X||_1+\frac{1}{2}||X-M||_F^2$这类优化问题，得到稀疏矩阵$X$。$T_\varepsilon(x)$的具体定义如下

$$T_\varepsilon(x)=\begin{cases} x-\varepsilon & \text{if } x>\varepsilon, \\ x+\varepsilon & \text{if } x<-\varepsilon, \\ 0 & \text{otherwise} \end{cases}$$

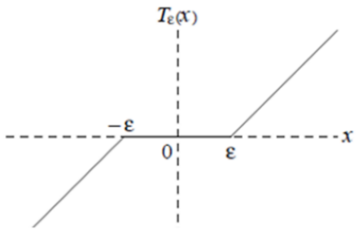

Figure 1 软阈值函数

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

$$L=\ U\cdot T_{1/\mu}(\Sigma)\cdot V^T,N=T_{\lambda/\mu}(X-N+Y/\mu)$$

最后更新拉格朗日乘子矩阵$Y=\ Y\ +\ \mu(X-L-N)$，重复L和N的计算直至达到收敛条件$\|X-L-N\|_F\le\varepsilon$即为最终的低秩矩阵$L$和稀疏矩阵$N$。

3、 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 28×28 像素的灰度图像。

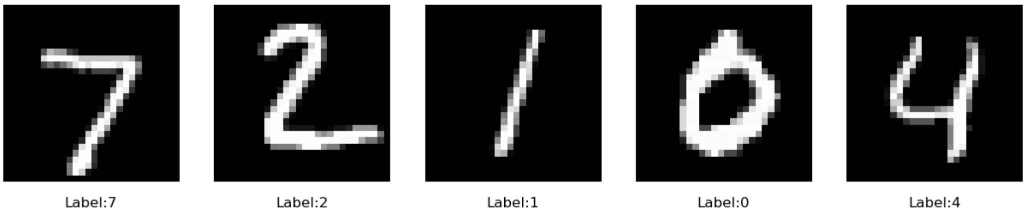

Figure 2 MNIST 数据集

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28×28的像素展开为一维的行向量，拼接得到60000×784的训练集矩阵X_{train}和10000×784的测试集矩阵X_{test}。

3.1 主成分分析 PCA

实验通过 PCA 对X_{train}进行降维处理，得到投影矩阵W和降维后的Z_{train}。将W作用到X_{test}上，得到降维后的Z_{test}。选择W中区分度最大的两个或三个方向(即Z_{test}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。

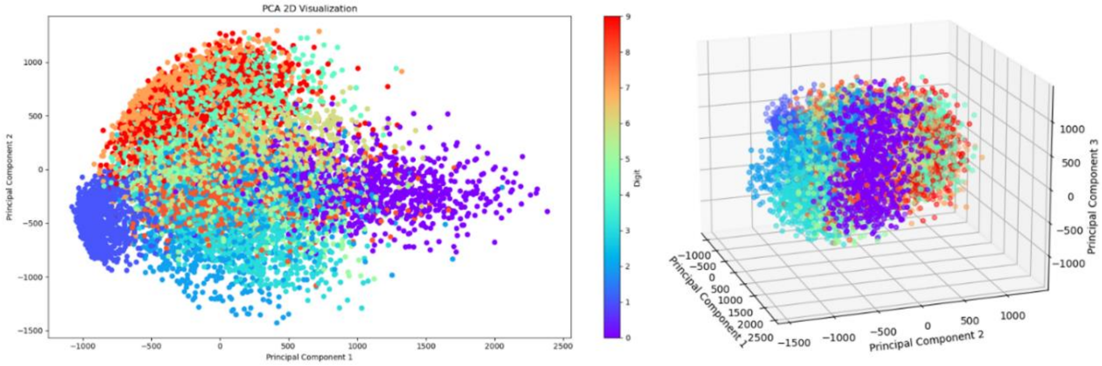

Figure 3 数据空间可视化散点图

（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k的增长，分类正确率先迅速增长而后趋于稳定，在$k = 8$时，正确率达到 90%；在$k = 16$时，正确率达到 97%，而

当$k>25$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。

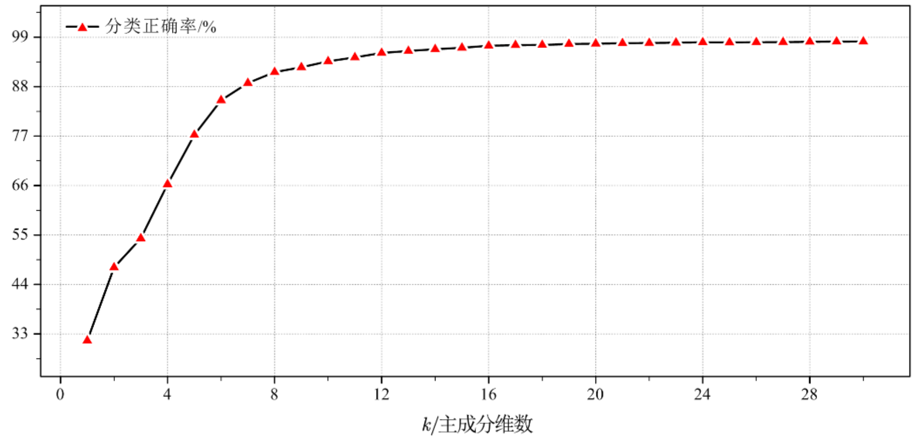

Figure 4 分类正确率随主成分维数的变化

取$k=20$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%。

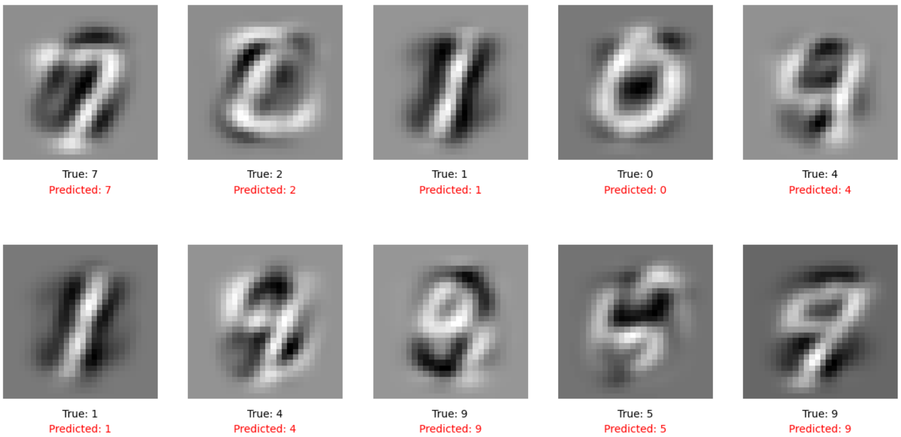

Figure 5 主成分图像（k=20）

而当取$k=2$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%。

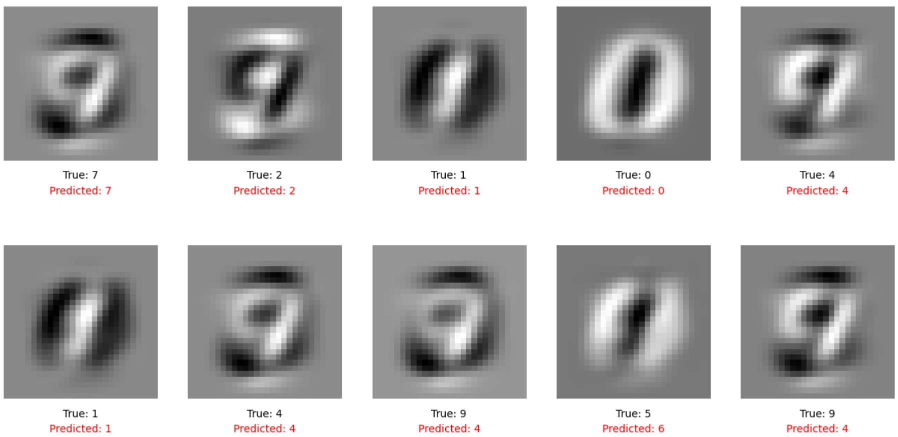

Figure 6 主成分图像（k=2）

对比$k = 2$和$k = 20$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。

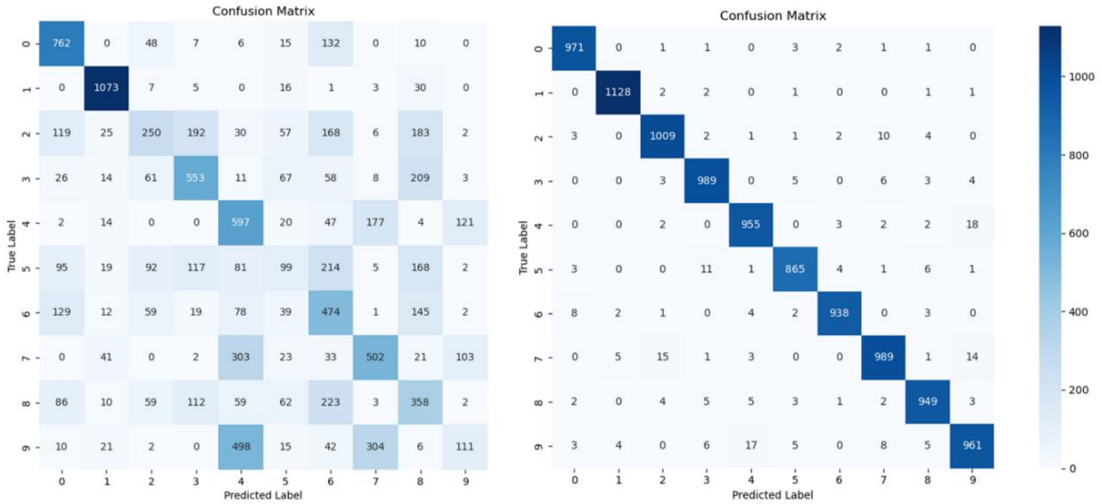

Figure 7 分类结果混淆矩阵

（左图k=2，右图k=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$k = 87$，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。

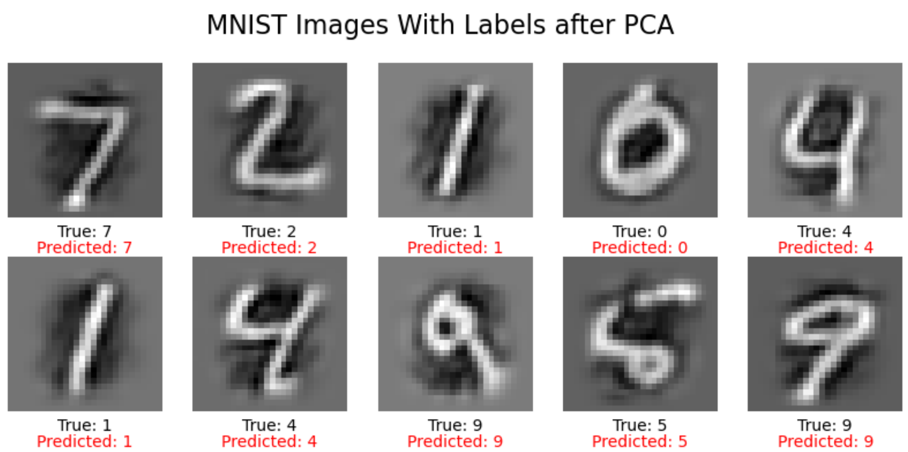

Figure 8 主成分图像（$n\_compoents = 0.9$）

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%。

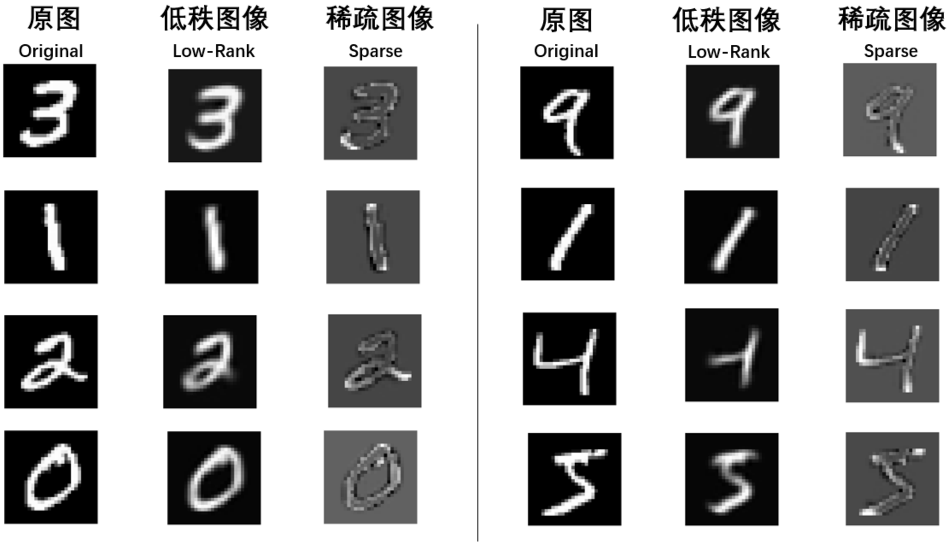

Figure 9 不加噪声的 RPCA 处理

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$N(0,30)$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$N(0,30)$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%。

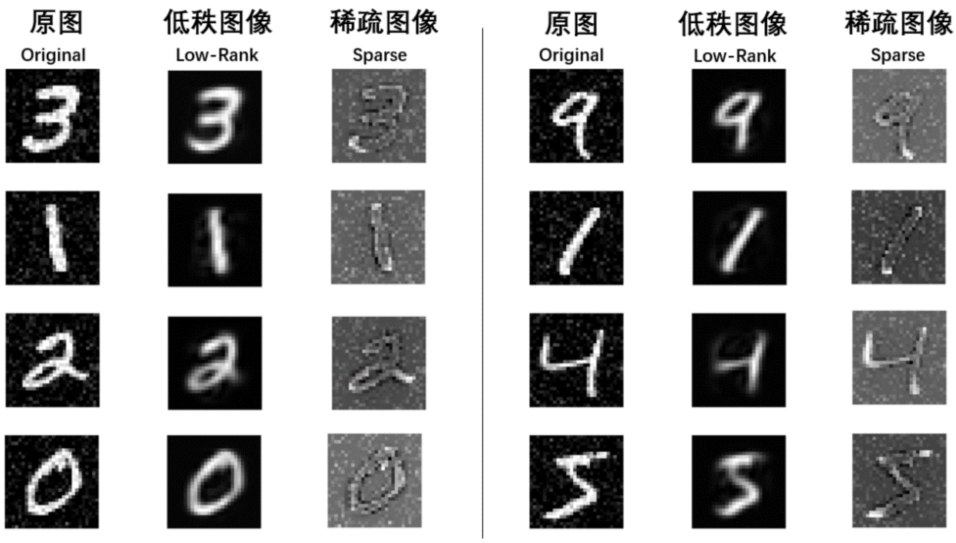

Figure 10 高斯噪声的 RPCA 处理

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%。

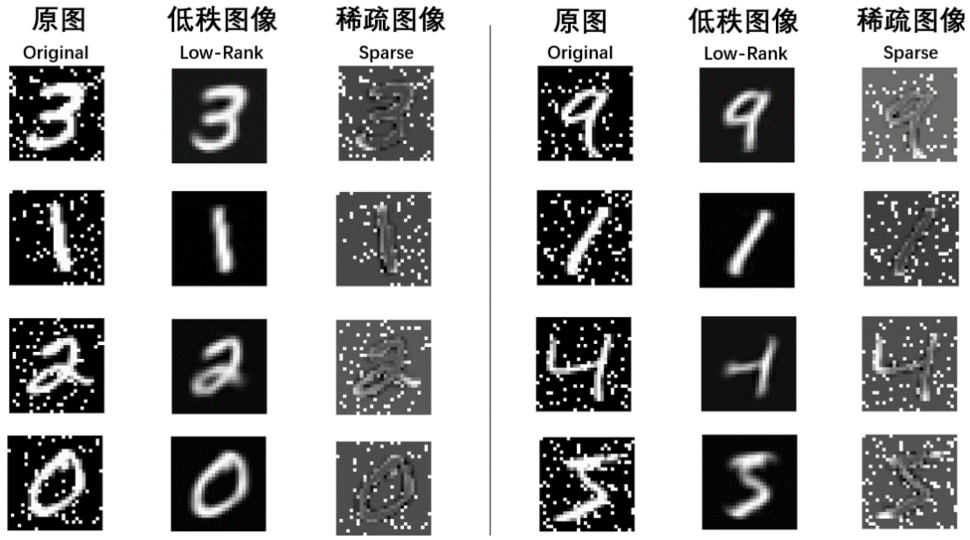

Figure 11 掩膜噪声的 RPCA 处理

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 ($k = 20$)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。

4、 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。

5、 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