

**数据挖掘实习二**

**中国地质大学（武汉）地理与信息工程学院**

**2021年 6 月 26 日**

# Lab\_2 基于PCA/LDA的图像降维

## 实习目的

1、结合实际应用理解数据降维；

2、采用PCA或LDA两种方法对数据集的数据进行降维，并进行可视化；

3、通过进一步查阅文献，了解相关研究方向的最新研究进展。

## 二、实习题目

基于PCA/LDA的图像降维

#### 【实验数据】

第一次课程实习制作的数据集。

#### 【实验要求】

1. 读取数据集中的图片，将每一个图片都调整到同样大小, 使用PCA或LDA两种降维方法对数据进行降维；

2、可视化并保存降维的数据。

#### 【过程描述】

我用的数据集是徐佳宜同学用scale与填0得到的256\*256大小的图片。我做的工作有以下四个部分。

**1、从每个视频中挑选8张图片**

1.1 挑选的原因

之所以要从视频中挑选图片，是因为视频逐帧截取的人像图数量分布的范围十分广泛，从1张到几百张不等。

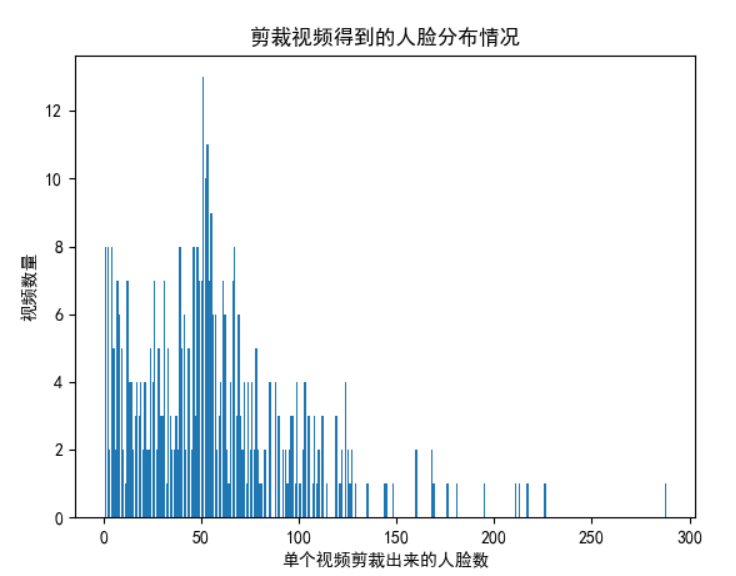


图1 人脸图像数量分布图

之所以选择8张，是因为训练集小于6张人脸的样本有21个，小于8张的有30个，我们决定删掉小于8张的，得到**训练集样本412个，测试集样本109个**。

1.2 挑选的策略

以**均匀分布**为准则的选择策略：设裁剪出的某个视频的人脸x张，以stride = int（（x - 1）/（8-1））为步长，从0开始取，得到一个图片集，但是这个图片集可能会比8张要多，因此取前8张就好了。

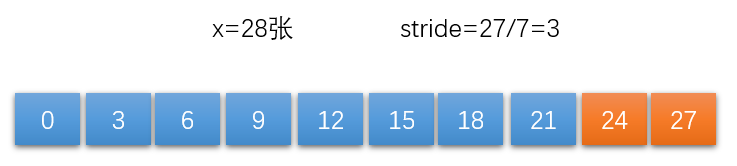


图2 样例

为什么不/8而要/7？因为/8的话，大多数视频都会挑出8张以上。为什么x要-1，因为不-1的话，有一些视频会挑出6张。

**2、降维**

由于我电脑运行内存有限，跑不动256\*256这么大的图片，尺寸缩小到64\*64仍然跑不动，最终缩小到**32\*32**大小才能运行，所以需要装入的人脸图片尺寸全缩小到了32\*32。

2.1 PCA

PCA的降维我是把训练数据都装入进来，计算协方差矩阵，选取其特征值从大到小排列的前800个对应的特征向量来把所有数据从8192（32\*32\*8）降到**800维**。

2.1 LDA

由于训练集的**Sw为奇异矩阵，不可逆**，因此无法直接用LDA对训练集进行降维。这是因为max rank（Sw）= n-c，此时的（样本数）n=412，（类数）c=11，故而Sw的最大秩为401，而Sw为8192\*8192的矩阵，所以Sw奇异，不可逆。这种现象被称为**小样本问题**。

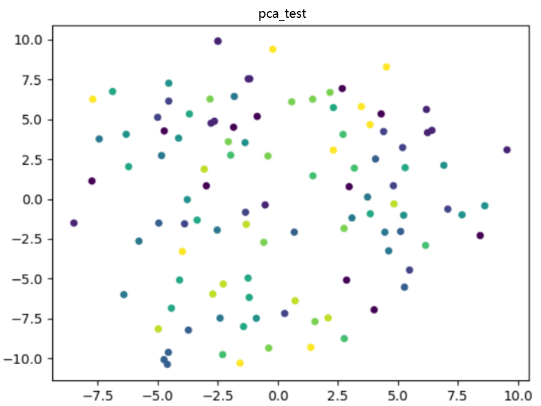
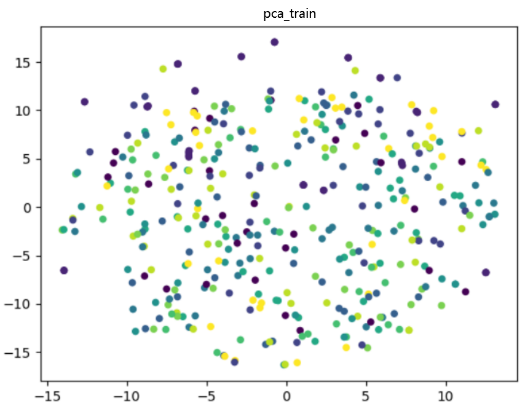
我的解决办法是**用PCA把所有数据从8192维降到400维**，因为训练集Sw的秩为400，然后用训练集计算的10个特征向量，把所有数据降到**10维**。

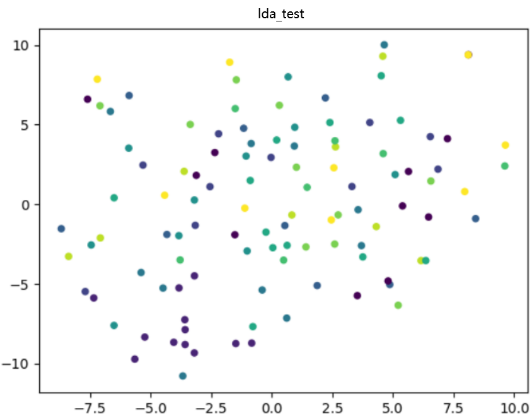
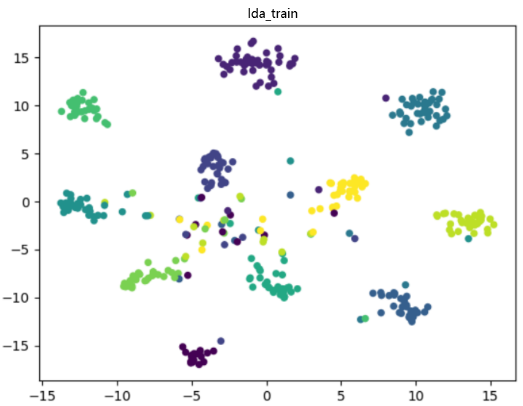
为什么是10维？因为max rank（Sb）= c-1，rank（Sw^-1\*Sb）<= min{rank(Sw^-1)，rank(Sb)} = c-1。

**3、可视化**

对降维后的数据可视化用的是**t-SNE方法**。用了手写的t-SNE与sklearn库中的做对比，数据来源为三种：一种是上述PCA降维，另一种是sklearn的LDA降维（对数据不进行PCA降维），最后一种是上述LDA降维方法。

3.1 t-SNE（DIY）





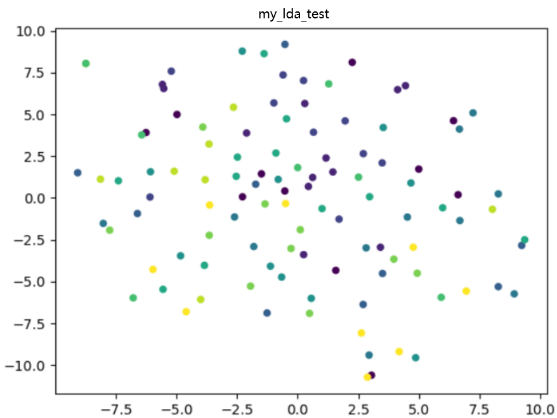
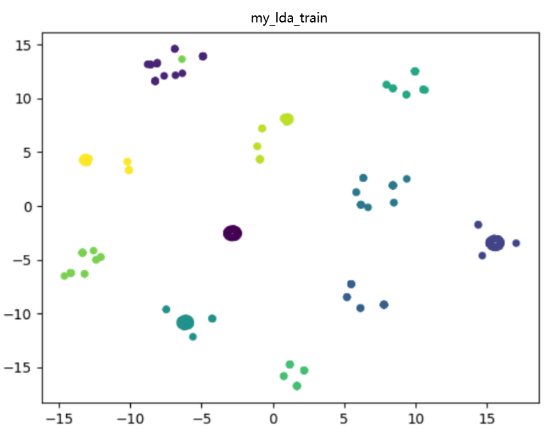
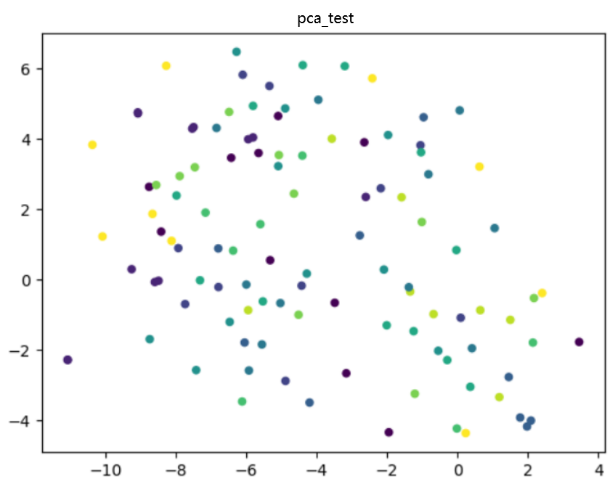
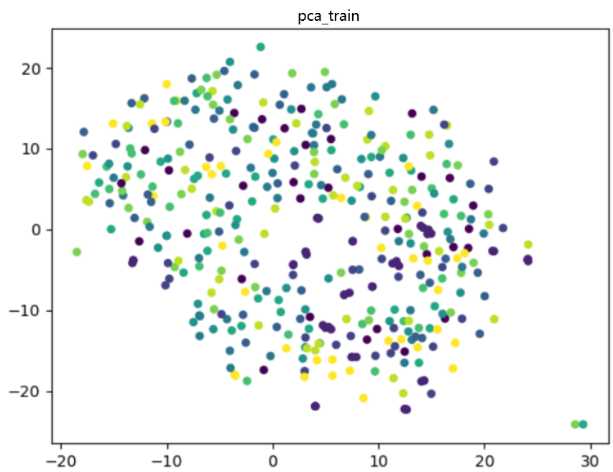
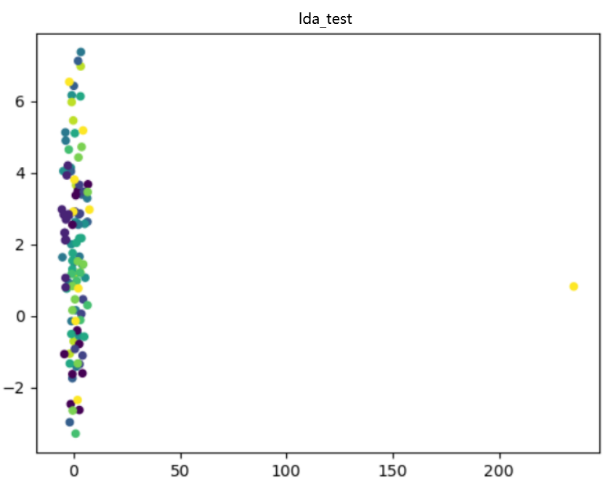
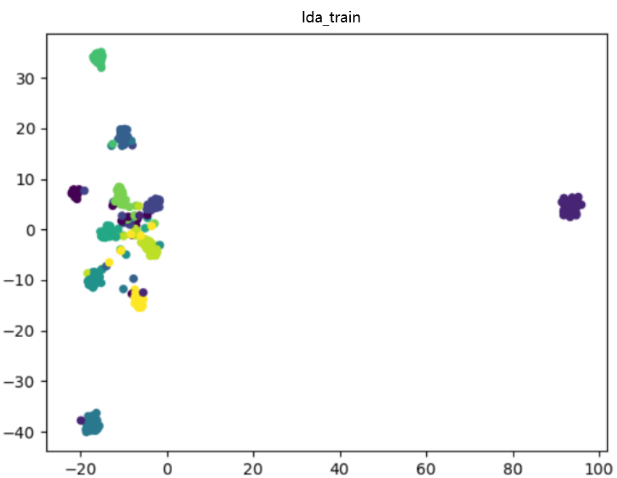


图3 手写t-SNE可视化图（左边是训练集，右边是测试集）

3.2 t-SNE（sklearn库）





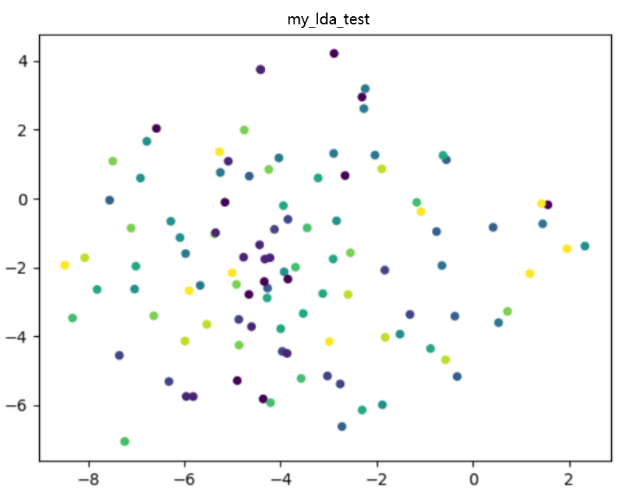
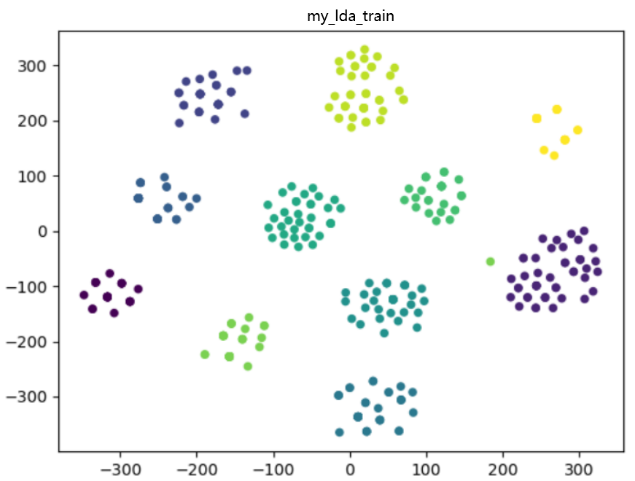


图4 sklearn库中t-SNE可视化图（左为训练集，右为测试集）

3.3 分析

从图中可看出，**PCA仅仅就是降维，而LDA在训练集上的降维比较出色，让类内聚、类之间分散**，但**LDA在测试集上的应用效果不明显**。

我试图用测试集来算LDA的特征向量，但也没能让测试集内聚。分析这其中的原因，我觉得有两种可能：**一是测试集数据本身比较分散；二是图片压缩损失了数据的精度**。

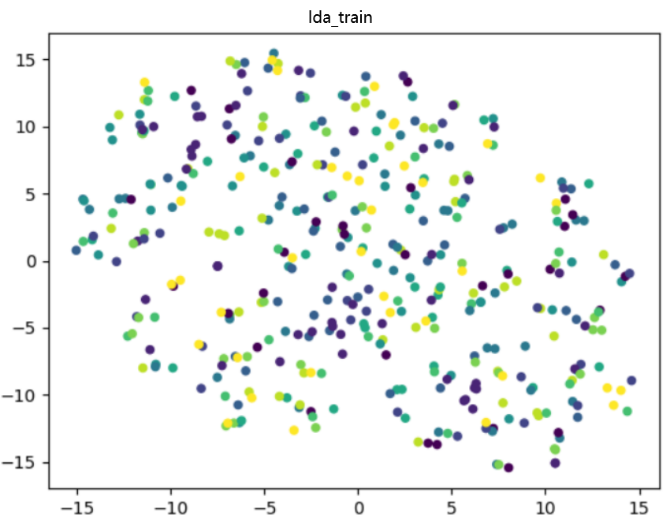
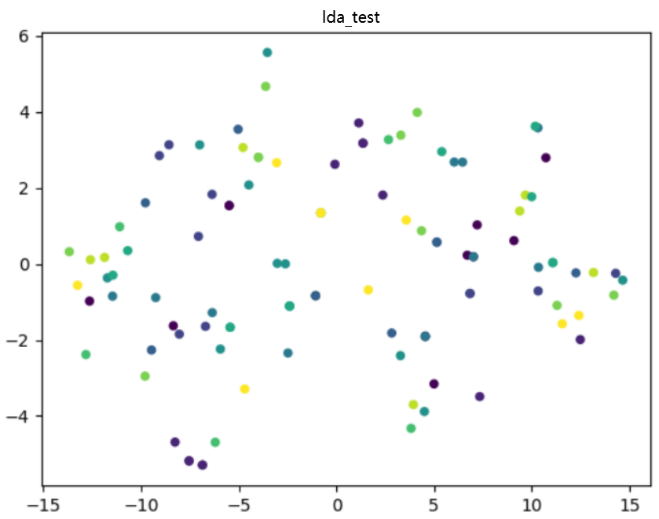


图5 用测试集进行LDA降维，再应用到训练集上的可视化图

**4、一个案例**

这是我学习PCA与LDA算法，手写的一个案例。用十个带标签的二维散点，画出散点图、降维的特征向量直线、散点在特征向量上的投影。一开始我画出来的如图6，**散点和投影的连线不与直线垂直**。

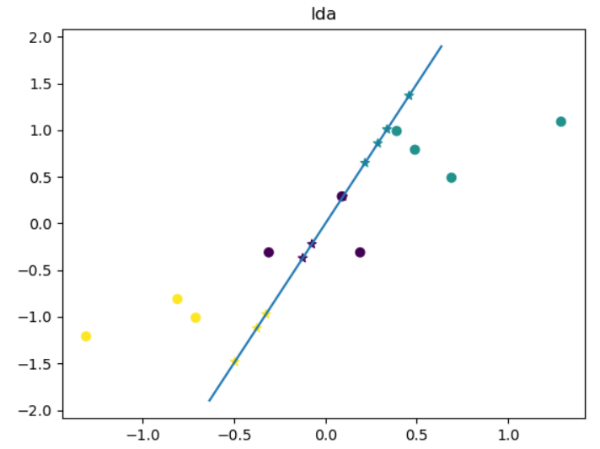
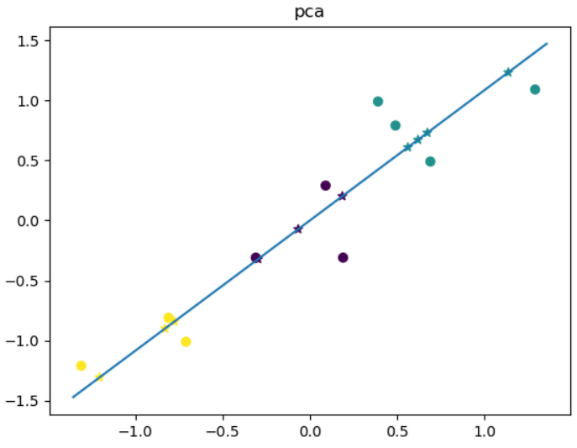


图6 可视化图

**散点和投影的连线与直线成一定斜度**启发了我，是**横纵坐标的单位不统一**，统一后的可视化如图7。

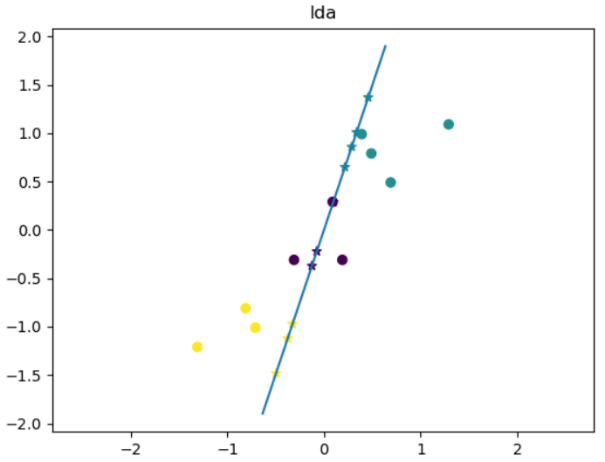
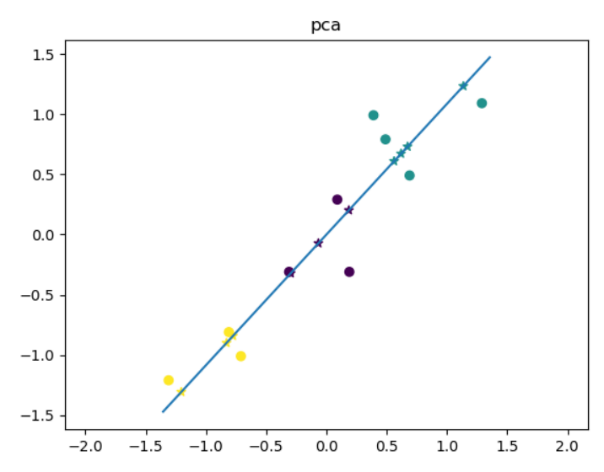


图7 横纵坐标统一单位后的可视化

仔细看可以发现，虽然PCA与LDA都实现了降维，都让这三个类别可分，但是**LDA更内聚**。

#### 【改进设想】

无

## 三、本次实习小结

本次实习让我深入理解了PCA、LDA算法，对它们的局限性也有了新的认识，其次是玩了一把t-SNE，本来是在刚结束的产学研中要实验的一个东西，由于时间和样本量的问题，没能实现可视化。本次实习，可谓收获满满。

=======================================================

我的联系方式：

电话：

邮箱：