第一次作业报告

杨辉韬　徐培文　黄彦澄

2024 年 10 月 20 日

一、 实验内容

1. 实验背景

基于基因表达对细胞进行分类在医学领域有着广泛的应用和重要背景，尤其是在个性化医疗、疾病诊断和药物开发方面。例如以下一些关键应用：癌症分类和诊断、个性化医疗、免疫疗法的开发、疾病的早期诊断和预后评估、干细胞研究和再生医学。在实际的临床医学诊断中，往往得到的是大量的细胞数据，而基因的表达又是千变万化的，常常以数万维度出现。因此，根据基因的表达情况来快速将不同细胞分类将是前文所述技术的应用基础。此报告就此项统计方法进行研究。

2. 具体问题

一组细胞的基因组成为 $X \in \mathbb{R}^{N\times d}$, 其中 N 为细胞的数量，d 为不同基因的数量。即每个细胞被一个 d 维欧式空间的向量所表达。我们的目标是利用映射 Φ：

$$\Phi : \mathbb{R}^{N\times d} \longrightarrow S = \bigcup_{n\in\mathbb{N}} n \times [n]^N \subset \mathbb{N} \times \mathbb{N}^N$$

$$\Phi : X \longmapsto a \times [a]^N$$

找到 a 为总的类数，将 $[a] = \{1, 2, \ldots, a\}$ 指标赋予给每个细胞。

Cluster. 聚类方法（Cluster）的函数形式为 $\mathbb{R}^{N\times d} \longrightarrow S$，通过每个样本点的信息将其分类为若干组。

Dimension Reduction. 函数形式为 $\mathbb{R}^{N\times d} \longrightarrow \mathbb{R}^{N\times d'}$ 其中 $d' \ll d$。目的是在不失去大部分信息的情况下对数据降维。

Dimension Reduction + Cluster. 主流的聚类方法，都需要利用样本点之间的距离信息，无论是直接利用欧式距离 $||x_i - x_j||$，还是利用模型 $f(||x_i - x_j||)$ eg. $f(\cdot)$ 为某个概率分布的密度函数，以达到将距离相近的点分在一起的目的。因此，从计算范数 $||\cdot||$ 复杂度的角度，常常需要较低的数据维数 d。另一方面，高维数据向量一般较为稀疏（实验数据集中大约只有不到 25% 为非零元素），距离会比较难以衡量相对位置关系。因此为了实现利用聚类方法对高维细胞分类的目的，我们需要先降低维数。

二、 实验设计与步骤

我们对于三个不同基因维数的细胞组 $X_1(\text{Deng}), X_2(\text{Seger}), X_3(\text{Zeisel})$ 进行实验。在预处理数据后，我们选取一种 PCA 方法，将数据 X_i 嵌入到 $\bar{X}_i \in \mathbb{R}^{N\times d'}$，观察图像以及答案在该降维后数据上的轮廓系数来判断是否选取该 PCA，接着选取合适的聚类方法，观察最终的得分（ARI/NMI）。

数据预处理. 我们在原始数据集上作用对数预处理，即 $X \longmapsto \log(X + \mathbf{1}_{N\times d})$。目的是为了减小作为整数矩阵的 X 的数据范围，将一些数据的非线形关系转换为线形关系。另一方面，如果着眼于每个细胞的单一基因的分布，且该分布是 Heavy-tailed 的，这一举措减小了该分布的方差。实验上看，这一举措大幅降低了降维后的分类难度，让样本之间的关系更加显著。我们在将数据标准化时使用 sklearn 的 StandardScaler 函数，并且认为数据矩阵稀疏，标准化后减少数据原本量纲的影响。

PCA 的选择. 我们选择了 $d' = 3$，目的在于可以利用 3D 可视化来增强分类的可解释性，以及在不依赖于如轮廓系数这类参数，分辨哪一种 PCA 达到的提取主成分效果最好。我们使用的定量分析方法是对于 PCA 后的数据 $\bar{X} \in \mathbb{R}^{N\times d'}$，以及参考分类 $Y \in S = \bigcup_{n\in\mathbb{N}} n \times [n]^N$，计算轮廓参数 $\text{Silhouette}(\bar{X}, Y)$ 以确定 PCA 是否有效展现了样本细胞间真实的关系。但定量分析利用了真实分类信息，因此只能作为分类后分析该 PCA 方法是否有效的一种方法，而不能作为判断选择 PCA 的依据。最为理想的情形为 $\text{Silhouette}(\bar{X}, Y) = \max_{s\in S} \text{Silhouette}(\bar{X}, s)$，即 PCA 没有改变样本点的相对分布。

表现指标我们利用轮廓系数 $\text{Silhouette}(\cdot,\cdot) : \mathbb{R}^{N\times d} \times S \longmapsto \mathbb{R}$ 衡量聚类的好坏。轮廓系数的输入为分类 $s \in S$ 以及数据矩阵 $X \in \mathbb{R}^{N\times d}$，不依赖于细胞的真实分类。ARI 以及 NMI 的函数形式为 $S \times S \longmapsto \mathbb{R}$，需要输入细胞的真实分类以及我们通过非监督学习方法得到的分类，用于计算所得分类以及真实分类的一致性。ARI/NMI/Silhouette 均取值于 $[-1, 1]$，且值越大意味着表现越好。

三、 实验数据

1. PCA 降维

对 Deng，Seger，Zeisel 数据集进行线性 PCA 降维，降维后的三维数据散点分布如图 1 所示

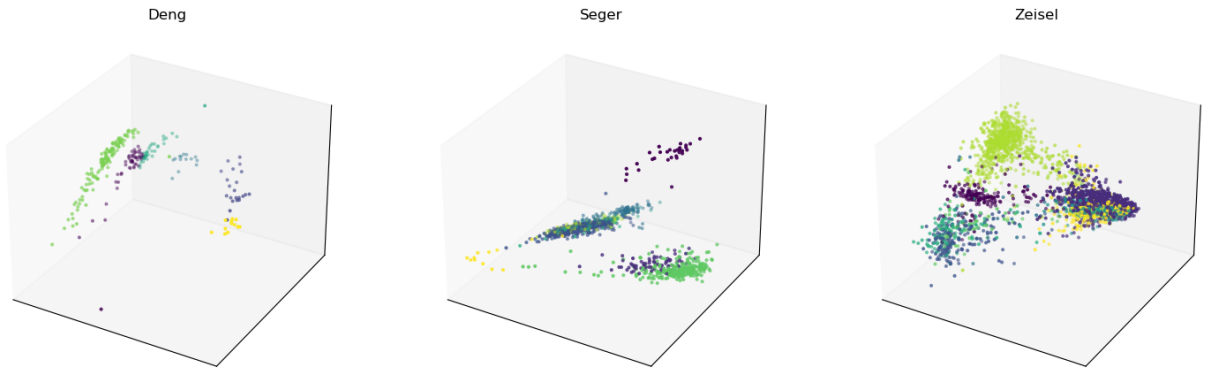

图 1: 各数据线性 PCA 降维

我们继续尝试了使用非线性的方法来进行降维，以 Deng 数据集为例，数据经过 kpca、lle、isomap 等方法降维后如图 2 所示

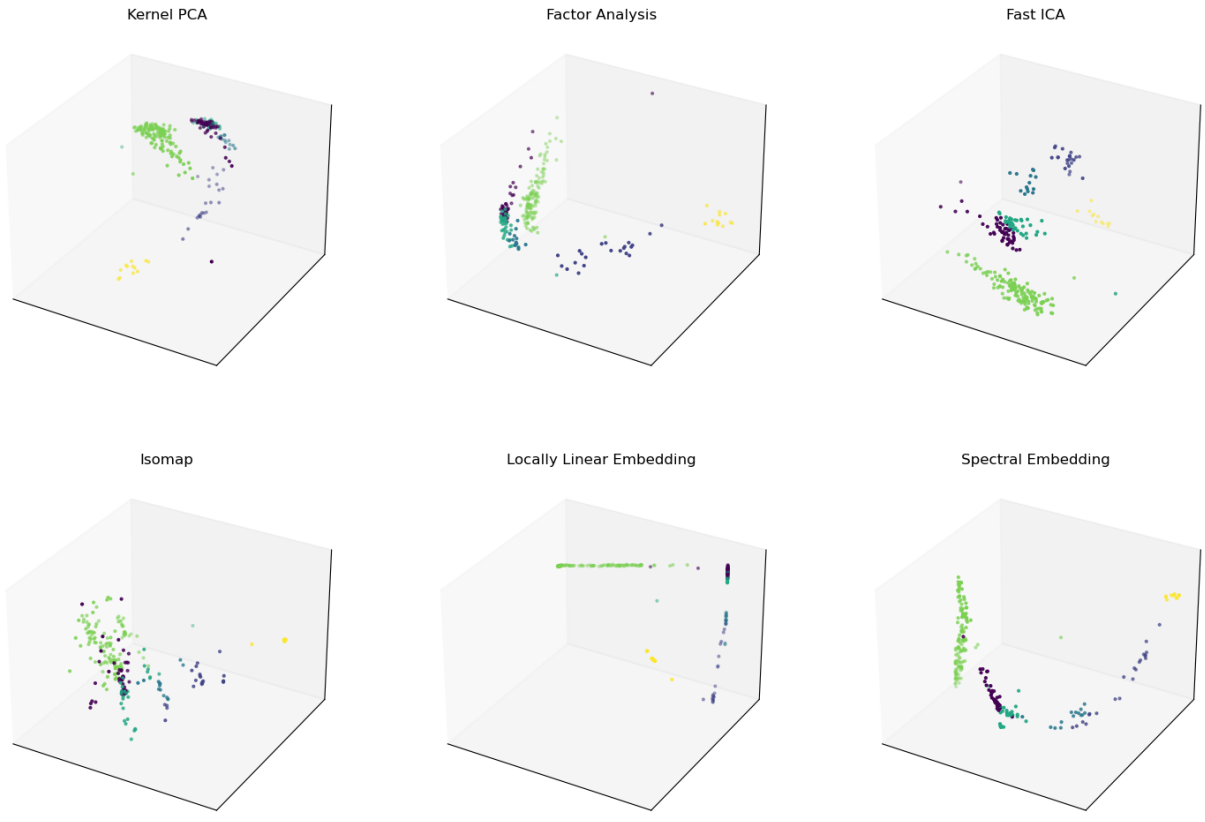

图 2: Deng 数据非线性降维后

对线性与非线性降维后的 3 维数据计算真实细胞分类标签下的轮廓系数，得到

	pca	kpca	fa	fica	isomap	lle	se
Silhouette score	0.36	0.33	0.33	0.33	0.34	0.51	0.32

表 1: Deng 数据降维后在真实细胞分类下的轮廓系数

可以看到，lle 在真实分类下的轮廓系数指标优于其他降维方法，于是选定 lle 作为降维方法。实验中，进行 lle 降维需要选定的邻点数参数，我们固定 mclust 聚类方法，分析邻点数 (默认为 10) 这一关键对于 ARI、NMI 指标的影响，最终得到邻点数为 79(总样本数为 268) 时，分数达到最优。对于 Seger 与 Zeisel 数据集，我们进行了类似的分析，均选择了 lle 作为降维方法，由于这两个数据集样本数量大，严格遍历邻点参数消耗大量时长，我们选择在较小的邻点参数区间上做参数分析并得到较佳的结果。

各数据集的线性 PCA 与实验所得最佳降维比如图 3 所示

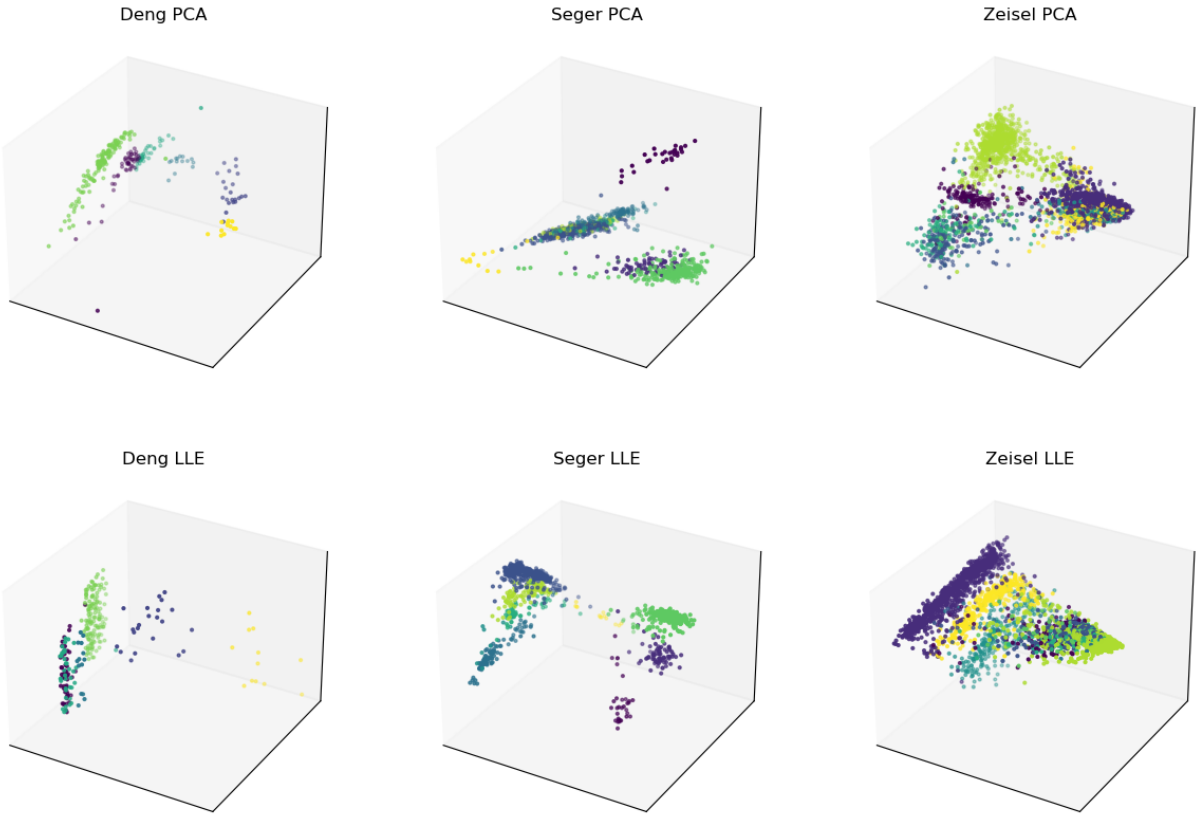

图 3: 线性 PCA 与实验所得最佳降维对比

各数据集线性 PCA 降维与实验最佳降维的指标对比

Silhoette score	Deng	Seger	Zeisel
PCA	0.36	0.04	0.10
LLE	0.30	0.28	0.15

表 2: 线性 PCA 降维与实验最佳降维在真实细胞分类下轮廓系数对比

Silhoette score	Deng	Seger	Zeisel
PCA	0.48	0.37	0.42
LLE	0.43	0.46	0.43

表 3: 线性 PCA 降维与实验最佳降维在 Mclust 聚类下轮廓系数对比

ARI | *NMI*	Deng	Seger	Zeisel
PCA	0.79|0.77	0.52|0.58	0.40|0.50
LLE	0.83|0.82	0.62|0.62	0.51|0.53

表 4: 线性 PCA 降维与实验最佳降维 Mclust 聚类后分数对比

四、 实验结果与分析

1. 聚类方法选择

Silhouette score	$k-$ Means	Leiden	Mclust
Deng	**0.56**	0.36	0.48
Seger	**0.43**	0.23	0.37
Zeisel	0.32	0.19	**0.42**

表 5: 线性 PCA 后各种聚类方法对应轮廓系数

ARI | NMI	$k-$ Means	Leiden	Mclust
Deng	0.62 | 0.73	0.25 | 0.64	**0.78 | 0.77**
Seger	0.32 | 0.50	0.14 | 0.43	**0.52 | 0.59**
Zeisel	0.34 | 0.49	0.14 | 0.41	**0.40 | 0.50**

表 6: 线性 PCA 后各种聚类方法对应最终得分

我们发现在三个 PCA 后的数据集上 cluster 分类本身的好坏（轮廓系数）与最终分类结果与真实分类的吻合程度并不一致，甚至相差甚远（如 K-means 在线性 PCA 得到的数据上得分很高，但分类结果对应于原始细胞而言并不合理）。这就说明线性 PCA 没有有效地保留原始细胞基因表达的信息。我们的 Kmeans 在选取 k 的时候使用真实分类数，所以分数较高，但分类结果对应于原始细胞而言并不合理）。这就说明线性 PCA 没有有效地保留原始细胞基因表达的信息。

ARI | *NMI*	$k-$ Means	Leiden	Mclust
Deng	0.46 | 0.60	0.76 | 0.74	**0.83 | 0.82**
Seger	0.49 | 0.65	0.60 | 0.62	**0.63 | 0.63**
Zeisel	0.40 | 0.52	0.48 | **0.58**	**0.51** | 0.54

表 7: (不固定降维方法) 不同聚类方法在各数据集上得到的最好结果

由于在 kmeans 聚类中，我们根据真实分类数设定 k，所以 kmeans 的分数仅仅作为参考，主要结果关注于 Leiden 和 Mclust。由以上表格，Mclust 在所有数据集上都表现更优异。

2. 实验结果

	ARI	NMI	Method
Deng	0.83	0.81	LLE (79 neighbors) & Mclust(GMM)
Seger	0.63	0.63	LLE (216 neighbors)& Mclust(BGMM)
Zeisel	0.51	0.53	LLE (166 neighbors) & Mclust(BGMM)

表 8: 各个数据集得到的最好结果以及对应的方法

我们发现在三个数据集上，如果限定降低维数到 3 维，利用 LLE 降维再进行聚类得到了最好的结果。说明在基于基因表达进行细胞分类的任务中，流形嵌入是一种不错的降维方式。

3. 讨论

我们只在 Deng 的数据集上得到了较好的分类结果，且主要采取了比较低效的枚举方法进行实验操作。对于实验的局限性与不足，我们认为有以下几方面的原因：

更好的数据预处理方法. 我们只尝试了作对数处理的方法。事实上通过查阅文献，对于基因表现数据的处理还有很多特定的处理方法，如针对于 Gene Duplication 现象的处理，怎样处理损失的信息等等。我们并没有对于原始的基因表达文件进行进一步分析，而只是直接采用了比较常用的对数变化，或许是后续降维不太成功的原因之一。

更好的维数选择. 我们只对于 $d' = 3$ 做了尝试，事实上可以先降低维数到 $d'' > d'$，然后利用可视化方法（如 tSNE）在三维或两维中提供解释。

跨数据集的学习. 在我们的实验中，我们将三个数据集作为完全相互独立的对象处理。但事实上作为基因表达，他们应当会具有一些类似的模式或者规律，对于得到一个统一化的方法提供思路。因此机器学习在此领域或许也存在潜力，可以利用一个模型对于基因表达整体进行学习。

附录

A 降维方法与聚类方法介绍

降维方法：我们从以下集合中选取降维方法：

$$\{\textbf{Linear PCA}, \textbf{Kernel PCA}, \textbf{ICA}, \textbf{FA}, \textbf{SE}, \textbf{LLE}, \textbf{Isomap}\}$$

线性主成分分析 (Linear PCA): PCA 通过识别数据集中方差最大的方向，将高维数据降维。首先对数据集 X 进行标准化：

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

然后计算协方差矩阵 C：

$$C = \frac{1}{n-1} X^T X$$

选择前 k 个特征向量，将数据投影到新空间：

$$X_{\text{proj}} = XV_k$$

核主成分分析 (Kernel PCA): 核 PCA 扩展了线性 PCA，通过核函数 $\phi(\cdot)$ 将数据映射到高维特征空间，以处理非线性数据：

$$X_{\text{proj}} = \Phi(X)V_k$$

在实验中我们主要采取了多项式核以及 RBF 核。对于对数预处理过后的数据，多项式核表现只有次数为 1 较好，RBF 核需要 $\gamma \approx 10^{-10}$ 总体表现较好。

独立成分分析 (ICA): 独立成分分析（ICA）将多变量信号分解为相互独立的加性子成分。它在 scikit-learn 中通过 Fast ICA 算法实现。通常，ICA 不是用于降维，而是用于分离叠加信号。由于 ICA 模型不包含噪声项，为了使模型正确，必须进行白化。这可以通过使用 whiten 参数或手动使用 PCA 的变体来完成。

因子分析 (FA): 数据集 $X = \{x_1, x_2, \ldots, x_n\}$ 可用模型

$$x_i = Wh_i + \mu + \epsilon$$

其中 $\epsilon \sim \mathcal{N}(0, \Psi)$。

矩阵表示为：

$$X = WH + M + E$$

概率解释为：

$$p(x_i|h_i) = \mathcal{N}(Wh_i + \mu, \Psi)$$

边缘分布为：

$$p(x) = \mathcal{N}(\mu, WW^T + \Psi)$$

假设：

- $\Psi = \sigma^2 I$：对应 PCA。- $\Psi = \text{diag}(\psi_1, \psi_2, \ldots, \psi_n)$：对应因子分析。

谱嵌入 (Spectral Embedding): 利用图拉普拉卡的谱分解找到数据的低维表示, 生成的图可视为低维流形在高维空间中的离散近似值。最小化图的损失函数可确保流形上相互靠近的点在低维空间中相互靠近，从而保持局部距离。

局部线性嵌入 (LLE): 局部线性嵌入（LLE）用以进行保持局部邻域内的距离关系的数据降维。它可以看作是一系列局部主成分分析，通过全局比较，找到最佳的非线性嵌入。

Isomap: 最早的流形学习方法之一，也即等距映射的简称。Isomap 可以看作是 MDS 或 KPCA 的扩展。Isomap 用以进行保持所有点之间的测地距离的数据降维

B 聚类方法

K-means 我们直接令 $k_{\text{class}} = k_{true}$，即直接预先告知了真实分类方式的分类数量。

Mclust 我们从以下集合中选取模型进行 Mclust：

$$\textbf{Models} = \{\text{GMM}, \text{BGMM}\}$$

我们直接调用了 python sklearn.mixture 函数包的函数加以实现 **Leiden** 我们利用 python 中的 leidenalg 函数包加以实现。