非线性流形学习方法的分析与应用*

尹峻松** 肖 健 周宗潭 胡德文
国防科技大学机电工程与自动化学院, 长沙 410073

摘要 非线性流形学习以保持数据局部结构的方式将高维输入投影到低维空间, 发现隐藏在数据中的内在几何结构与规律性, 是近年来机器学习与认知科学中一个新的研究热点. 文中分析了几种主要的流形学习方法, 通过比较给出各方法的优缺点; 提出了基于谱分析的非线性降维的统一框架, 对于流形学习方法的研究具有重要意义; 给出了手写数字和人脸图像序列等降维的实验结果, 显示了非线性流形学习在数据约简和可视化方面的有效应用; 最后结合作者的研究探索, 总结了非线性流形学习需要解决的问题并展望其研究趋势.

关键词 非线性降维 流形学习 维数估计 统一框架

在模式识别研究中, 很多地方都需要对高维数据进行分析, 如全球气候模式、恒星光谱、人的基因分布、受光照影响的多姿态表情人脸图像序列等. 数据维数的大幅度提高给随后的数据处理工作带来了前所未有的困难, 如何从这些高维数据中找出事物的本质规律成为迫切需要解决的问题.

降维是指将样本从高维观测空间通过线性或非线性映射投影到一个低维特征空间, 从而找出隐藏在高维观测数据中有意义的低维结构. 对观测数据降维主要有四个目的[1]: (i) 压缩数据降低存储需求; (ii) 消除噪声; (iii) 便于从数据中提取用于识别的特征; (iv) 把数据投影到低维空间, 可以实现高维数据可视化. 传统的线性降维方法包括主分量分析(PCA)[2], 独立分量分析(ICA)[3], 线性判别分析(LDA)[4]等, 它们主要研究在高维空间中如何设计线性模型的特征向量, 优点是运算简便, 并能产生简单的变换函数, 对线性结构效果好[5]. 但是现实中的高维数据大多是非线性的, 这时线性方法很难发掘高维数据的几何结构和相关性, 揭示其流形分布. 针对高维数据的非线性特性, 近年来发展了非线性降维方法. 非线性降维(NLDR)方法很多, 包括神经网络[6]、遗传算法[7]、流形学习等, 其中基于流形学习的有: 等度规映射(ISOMAP)[8], 局部线性嵌入(LLE)[9], Laplace 特征映射[10], 局部保持投影(LPP)[11], 随机邻域嵌入(SNE)[12], 图册化流形(Charting a Manifold)[13], 局部线性平滑[14]等, 这些方法均能保持原始数据的拓扑结构不变, 并能较好解决数据处理中的“维数灾难”问题.

非线性降维概念提出以后, 各种非线性降维方法很快成为模式识别领域一个非常活跃的研究热点. ISOMAP 是 Tenenbaum 等在 2000 年 Science 上提出的一种非线性降维方法[8]. 该方法首先在几何空间计算成对测地距离, 再运用多元尺度分析(MDS)[15]方法把数据点从高维输入空间投影到低维非线性拓扑空间中, 获得保持样本间测地距离不变的低维流形. 在同一期杂志上, Roweis 等提出了局部线性嵌入(LLE)算法. LLE 是一种无监督的学习算法, 以保存数据局部邻域间相互关系的方式, 把高维数据映射到一个低维全局坐标系下[9]. ISOMAP 和 LLE 两种方法的核心都是流形学习, 能够展示数

2006 07 27 收稿, 2007 02 02 收修改稿
* 国家自然科学基金(批准号: 60675005, 60575044)、高校博士点专项基金(批准号: 20049998012)、国家“九七三”计划(批准号: 2003CB716104)和国家高技术研究计划(批准号: 2006AA01Z193)资助项目
** E-mail: jsyin@nudt.edu.cn

据的本质结构，并成功地运用到模拟和真实数据中. 2003 年 Belkin 等基于局部保序的思想提出了 Laplace 特征映射，获得高维观测空间和低维结构在局部几何意义下的对应[10]. 该算法利用 Laplace Beltrami 算子(定义为流形切空间上梯度向量的负散度函数)的特性，通过计算该算子的特征函数来实现流形的最优嵌入. 此后，流形学习技术逐渐兴起，很快成为近几年非线性降维中的热点问题.

与前面的谱分析方法有所不同，SNE 和图册流形都是基于数理统计的方法，保持观测空间和嵌入空间数据集中近邻数据的互信息一致[12, 13]，从而可在基于熵的意义下获得互信息构成的目标函数，前者通过梯度下降法获得低维结构，后者通过求取最小二乘的闭式解实现，两者均为迭代算法.

本文将介绍近年来研究较多的基于流形的非线性降维方法，对各方法的优缺点做出详细比较，并给出非线性降维的一个基于谱分析的统一框架. 在这个统一框架下，归纳出该领域还需要解决的问题，如数据本征维数的估计、拓扑空间存在空洞或维数跳变的情况、有监督学习用于分类的情况等.

1 非线性流形学习方法介绍

流形定义为满足 Hausdorff 公理的拓扑空间，每个点的局部都同胚于 n 维欧式空间[16]. 流形学习的主要目的就是发现高维观测数据中隐藏的嵌入子流形，找出产生数据集的内在规律.

设初始数据集为高维空间 $\mathcal{R}^D$ 中的 N 个实值向量 $\boldsymbol{X}_i$，通过非线性降维方法映射到低维空间 $\mathcal{R}^d$ ($d<D$). 各非线性流形学习算法基本原理介绍如下，并给出详细的算法分析.

1.1 等度规映射(isometric map, ISOMAP)

ISOMAP 是 Tenenbaum 等于 2000 年提出，其基本思想是当数据集的分布具有低维嵌入流形结构时，可以通过保距映射获得观测空间数据集在低维空间的表示[8]. 具体算法分以下三步：

(1) 构造邻域图 G：由点 i, j 之间的欧式距离 $d_X(i, j)$ 定义，j 在 i 的半径 ε 之内或者是 i 的 K 个最近邻点之一，则连接 i 和 j，该边的长度值等于 $d_X(i, j)$.

(2) 计算最短路径：在邻域图中使用 Floyd 算法计算任意两点 i 和 j 之间的最短路径 $d_G(i, j)$，并以此估计拓扑空间 M 的测地距离 $d_M(i, j)$.

(3) 构造 d 维嵌入：在距离矩阵 $\boldsymbol{D}_G=\{d_G(i, j)\}$ 上，采用经典 MDS 方法构造能保持拓扑空间本质结构的 d 维嵌入空间 $\boldsymbol{Y}$，坐标向量 $\boldsymbol{y}_i$ 由最小化下列误差方程得到

$$E = \| \tau(\boldsymbol{D}_G) - \tau(\boldsymbol{D}_Y) \|_2 \qquad (1)$$

其中矩阵变换算子 $\tau_D=-\boldsymbol{HSH}/2$，$\boldsymbol{S}$ 是平方距离矩阵 $\{\boldsymbol{S}_{x_ix_j}=\boldsymbol{D}^2_{x_ix_j}\}$，$H$ 是集中矩阵 $\{\boldsymbol{H}_{x_ix_j}=\delta_{x_ix_j}-1/N\}$. (1)式的最小值可通过求取矩阵 $\tau(\boldsymbol{D}_G)$ 的 d 个最大特征值对应的特征向量来实现.

ISOMAP 的一个优点是对于单一流形结构，在降维过程中求(1)式的方差可产生“elbow”现象，可由此判断流形的本征维数，在人脸和手数据集中均取得一定的研究成果[8]. ISOMAP 的缺陷在于没有定义样本空间到嵌入空间的映射，对于一个未知点不能直接投影到嵌入空间；另外，当数据集存在过大噪声时，ISOMAP 很难恢复其内在结构，这也是其他非线性降维方法需要改进的问题[17].

由于ISOMAP 在估计测地距离时计算复杂度很高($O(N^3)$). 为减小该计算复杂度，Silva 等提出一种带标记点(landmark)的 ISOMAP 方法[18]. 该方法选取 n($n\ll N$)个标记点来估计测地距离，使计算量减小到 $O(knN\log(N))$. 另一方面，Odest 等基于时空关系对 ISOMAP 进行扩展，提出了改进算法 ST ISOMAP[19]，在连续和分离的数据集中分别得到很好的应用.

1.2 局部线性嵌入(locally linear embedding, LLE)

2000 年 Roweis 等在“Science”上提出了 LLE[9]，基本思想为保存原流形中局部邻域间相互关系，将高维数据映射到低维全局坐标系中，具体算法分以下三步：

(1) 邻域点搜索：计算出每个向量 $\boldsymbol{X}_i$ 的邻域点，每个点都是 D 维(取欧式距离最小的 K 个点为邻域或者固定半径 ε 的球状邻域).

(2) 在 $\boldsymbol{X}_i$ 的邻域中，计算能最好地重构每个 $\boldsymbol{X}_i$ 的权值 $\boldsymbol{W}_{ij}$，使重构误差最小. 定义重构误差为：

$$\varepsilon_I(\boldsymbol{W}) = \sum_i \left\| \boldsymbol{X}_i - \sum_j \boldsymbol{W}_{ij}\boldsymbol{X}_j \right\|^2 \qquad (2)$$

重构权值 W_{ij} 满足两个条件：当 X_i 不属于 X_j 的邻域时 $W_{ij}=0$，以及 $\sum_j W_{ij}=1$. 求解 W 的过程就是求解带约束的最小二乘问题.

(3) 映射到低维嵌入空间 $\mathcal{R}^d$($d<D$)：嵌入空间的代价误差定义为(3)式，与前面定义的代价误差(2)式类似，都是基于局部线性重构误差，但这里是固定 W_{ij}，优化 d 维坐标系下 Y_i，使代价误差(3)式最小. 对任何一数据点 i，W_{ij} 具有旋转、尺度和全局变换不变性，因此对 Y 的求解也就是一定约束条件下求解稀疏矩阵的特征向量问题.

$$\varepsilon_{II}(\boldsymbol{Y})=\sum_i \left\| \boldsymbol{Y}_i-\sum_j \boldsymbol{W}_{ij}\boldsymbol{Y}_j \right\|^2 \tag{3}$$

与 ISOMAP 算法不同，LLE 是通过局部线性拟合来获得内在的全局线性结构. LLE 的一个优点是不需要计算成对的距离矩阵，嵌入向量的求解是求解稀疏矩阵的特征向量，大大减少了计算量. LLE 算法强调观测空间近邻数据间的序应该在嵌入空间中同样保持，由此形成了求取近邻数据间权值的闭式求解法，但对于未知数据如何应用这一权值矩阵，没有给出一般性答案.

LLE 用于高维数据的可视化和统计描述非常有效，在特征提取中还应使用数据的类别信息. 因此，一种有监督的 LLE 算法被提出，并取得较满意的结果[5, 20, 21]. 文献[20]中把 LLE 和有监督的线性 Fisher 判别映射结合起来，通过实验证明其具有较好的分类和判别能力.

1.3 Laplace 特征映射(Laplace eigenmaps)

Laplace 特征映射[10]也是一种使用特征向量求解的方法，采用与前面类似的方法构造邻域图，基于谱图理论，可以构造相应嵌入空间目标函数为：

$$\sum_{i,j}(\boldsymbol{y}_i-\boldsymbol{y}_j)^2\boldsymbol{W}_{ij} \tag{4}$$

其中权值矩阵 W_{ij} 采用 $\boldsymbol{W}_{ij}=\exp\left(-\|x_i-x_j\|^2/t\right)$，$t\in R$ 的核函数，在满足低维结构对域的约束 $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{y}=1$($\boldsymbol{D}$ 为对角矩阵，$\boldsymbol{D}_{ii}=\mathrm{sum}_j \boldsymbol{W}_{ij}$)及防止数据集收缩至单点的约束 $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{I}=0$，最小化误差方程可对应于求解下式的最小特征向量：

$$\boldsymbol{L}\boldsymbol{f}=\lambda\boldsymbol{D}\boldsymbol{f} \tag{5}$$

其中 $\boldsymbol{D}$ 为对角权矩阵，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$ 为 Laplace(对称、半正定)矩阵.

可以证明(5)式能够近似对应于 Laplace Beltrami 的特征向量求解，因而能够寻找流形的最优嵌入. 在算法上，流形结构的描述由相邻图来近似，选择适当的权值，Laplace Beltrami 算子可以通过相邻图的加权热核 Laplace 来近似，数据集的嵌入映射可以近似估计定义在整个流形上的 Laplace Beltrami 算子的内在特征映射.

Laplace 特征映射也是基于局部邻域，矩阵表现为稀疏矩阵，因此，可以通过对稀疏矩阵的处理来加速谱分析的算法实现.

1.4 随机邻域嵌入(stochastic neighbor embedding)

SNE[12]在高维空间数据点的欧式距离基础上，定义了邻域概率函数. 在高维观测空间中，点 j 属于点 i 邻域的概率函数 p_{ij}(非对称)由(6)式确定：

$$p_{ij}=\frac{\exp(-d_{ij}^2)}{\sum_{k\neq i}\exp(-d_{ik}^2)} \tag{6}$$

其中 $d_{ij}^2=\frac{\|x_i-x_j\|^2}{2\sigma_i^2}$，标准差 σ_i 由人为依据经验给定.

在低维空间，我们期望使本来临近的数据点接近，本来远离的数据点彼此远离，因此，低维嵌入空间与高维观测空间应该有相似的概率分布. 与 p 类似，低维空间中的邻域概率函数 q 定义为：

$$q_{ij}=\frac{\exp(-\|y_i-y_j\|^2)}{\sum_{k\neq i}\exp(-\|y_i-y_k\|^2)} \tag{7}$$

低维嵌入的目标是使上面两个分布尽可能匹配，可利用 Kull Leibler 散度和来构造损失函数：

$$C=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}}=\sum_i KL(P_i\|Q_i) \tag{8}$$

最小化(8)式，相当于对(8)式求微分：

$$\frac{\partial C}{\partial y_i}=2\sum_j(y_i-y_j)(p_{ij}-q_{ij}+p_{ji}-q_{ji}) \tag{9}$$

(9)式可以理解为让远点尽可能分开同时使近点尽可能接近的动力总和. 因此, 我们可以通过梯度方法来调整低维空间数据点集的相应位置. 这种方法是依赖于观测空间的数据集来调整内在的低维结构, 而迭代法使其损失函数的最小化易陷入局部极小, 这一方面需要改善.

1.5 图册化流形[1] (charting a manifold)

Charting a manifold 要解决的问题是, 在低维坐标系与高维样本空间中的数据集[13] 之间建立一个光滑的连续映射. 假定采样空间数据分布满足一个低维流形, 且流形与低维坐标之间存在一个同胚平滑非线性变换, 其目标是找到一个基于核函数的线性混合投影, 使采样点密度和相对位置的信息损失达到最小.

局部线性尺度 r 的意义是, 在以 r 为半径定义的邻域尺度范围内, 原始空间到低维空间的映射是线性的. 图册数据是要寻找一个软分割方法, 把数据分割为局部线性的低维邻域, 每个图册由 Gauss 混合模型(Gauss mixture model, GMM)定义. 局部相关性用来估计两个图册的距离:

$$m_i(\mu_j) = N(\mu_j : \mu_i, \sigma^2) \tag{10}$$

令每个数据点与一个 Gauss 邻域相关联, 即 $\mu_i = x_i$. μ_i, $m_i(\mu_j)$ 固定.

连接每个图时, 为使损失信息最小, 投影到每个子空间的数据点和它的每个邻域点应该满足: (i) 局部方差损失最小; (ii) 原来相近的点到邻域的投影满足最大一致性. 为满足第一条原则, 在每个图上使用 PCA 得到一个低维局部坐标系统, 每个初始数据点在低维坐标中有不同的投影点. 为满足第二条原则, 需要把每个局部坐标映射到一个全局坐标下, 使映射到全局空间的数据点与原来的误差达到最小. 每个数据点(i)由(11)式投影到相邻的局部坐标系(j)中:

$$u_{ji} = l_j x_i \tag{11}$$

最后, 每个低维坐标系中的点由(12)式映射到一个全局坐标系下:

$$y_i = \sum_j \boldsymbol{G}_j u_{ji} p_{j|x}(x_i) \tag{12}$$

其中 $\boldsymbol{G}_j$ 是第 j 个图到全局空间的映射. 使全局空间的数据点与原来的误差最小, 即解决(13)式中的加权最小二乘距离问题.

$$\boldsymbol{G} = [\boldsymbol{G}_1, \cdots, \boldsymbol{G}_K] = \arg\min_{\boldsymbol{G}_k, \boldsymbol{G}_j} \sum p_{k|x}(x_i) p_{j|x}(x_i) \left\| \boldsymbol{G}_k \begin{bmatrix} u^{ki} \\ 1 \end{bmatrix} - \boldsymbol{G}_j \begin{bmatrix} u^{ji} \\ 1 \end{bmatrix} \right\|_F^2 \tag{13}$$

2 非线性流形学习统一框架

前面介绍的几种非线性降维方法都是采用局部几何线性的理论, 假定隐藏在高维数据中的流形结构是紧致连续的, 只是降维准则的选取有所不同. ISOMAP 引入了测地距离, 并运用经典 MDS 发掘非线性结构, 他能够保持数据的全局几何结构, 并用局部线性理论来得到成对的距离矩阵. LLE 采取局部线性原理把原始数据空间分为相互交叠的邻域, 每个数据点由其相应邻域中的点线性重构, 从而保持全局集合结构. Laplace 特征映射保持局部几何结构, 流形结构的描述由邻域图给出, 再选取适当的权值, Laplace-Beltrami 算子可以通过相邻图的加权 Laplace 来近似. 这三种方法都是把降维过程看成是在一定误差准则条件下特征向量的求解问题, 具有全局最优解, 并且不陷入局部极小值.

与上面三种定义邻域的方法(k-NN)不同, SNE 和图册化流形是采用概率方程来定义邻域. 这两种方法较之其他方法具有较好的抑制噪声能力, 但是迭代算法有时会使损失函数陷入局部极小值.

把无监督的非线性流形学习方法综合到一个统一的图形嵌入框架, 归纳如下:

(1) 在原始高维空间中设包含 N 个样本的数据集 $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, 构造一个 $N \times N$ 的邻域矩阵或相似性矩阵 $\boldsymbol{M}$. 设 $\boldsymbol{D}(\cdot, \cdot)$ 为定义在样本空间的二元函数, 则 $\boldsymbol{M}_{ij} = \boldsymbol{D}(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

(2) 根据每个算法相应变换矩阵 $\boldsymbol{M}$, 产生一个规范化谱矩阵 $\boldsymbol{M}$. 这也相当于在成对样本点($\boldsymbol{x}_i$, $\boldsymbol{x}_j$)中运用一个对称二元函数 $\boldsymbol{D}$, 得到 $\boldsymbol{M}_{ij}$.

(3) 这样, 嵌入向量均可以归纳到框架:

1) 令 X 为一 Hausdorff 空间, 一个从开子集 $\boldsymbol{U} \subset \boldsymbol{X}$ 到 C 的子集的同胚称为图册(Chart)

$$\boldsymbol{y} = \arg\min \boldsymbol{y}^{\mathrm{T}} \boldsymbol{M} \boldsymbol{y} \qquad (14)$$

计算矩阵 $\boldsymbol{M}$ 最大的 m 个特征值 λ_j 和特征向量 $\boldsymbol{v}_j$，只考虑正的特征值.

(4) 每个样本点 $\boldsymbol{x}_i$ 的嵌入值为向量 $\boldsymbol{y}_i$，$\boldsymbol{y}_{ij}$ 代表矩阵 $\boldsymbol{M}$ 第 j 个主特征向量 $\boldsymbol{v}_j$ 的第 i 个元素.

大部分谱分析的非线性降维方法都遵循这一框架，可见，高维非线性数据集的降维离不开局部理论来保持原始数据间的相互关系. 但是，统一框架在解决不同的实际问题时，还需要针对问题做出相应的改动. 如在 LLE 中我们给出 $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{M}$，而在 ISOMAP 中我们令 $\boldsymbol{M}_{ij} = \boldsymbol{D}^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 且 $\boldsymbol{S} = \sum_j \boldsymbol{M}_{ij}$，则谱矩阵计算如下：

$$\boldsymbol{M}_{ij} = -\frac{1}{2}\left(\boldsymbol{M}_{ij} - \frac{1}{N}\boldsymbol{S}_i - \frac{1}{N}\boldsymbol{S}_j + \frac{1}{N^2}\boldsymbol{S}_i\boldsymbol{S}_j\right) \qquad (15)$$

最近，颜水成在这一工作的基础上，糅合线性化与张量化，提出了更一般的框架[22].

3 非线性流形学习中需要解决的问题

尽管非线性降维方法较之线性降维法能更好发掘隐藏在高维数据中的流形分布，但是其各种算法本身还存在很多局限性，如流形本征维数的估计、流形本身带有孔洞或维数不固定的情况、缺少高维观测数据与低维数据间的映射、半监督或有监督学习等问题，需要在现有的非线性降维方法上予以改进. 我们在以下几方面分析各种非线性降维方法的不足，以及相应的改进方法.

3.1 本征维数的估计

在非线性降维过程中，原始数据本征维数 d 都是由经验已知或人为设定的，其设定值的大小对低维空间的映射结果有很大影响. d 值过大使映射结果含有过多噪声；d 值过小，本来不同的点在低维空间可能会彼此交叠. 在 ISOMAP[8] 中，本征维数 d 可以由(1)式中嵌入向量 Y 的重构误差来估计得到，见图 1. 其他方法包括估计“Packing Number”[23] 或基于分形的方法.

目前，本征维数的估计方法主要分为以下三组[24]：

(1) 特征值或映射方法. 这种方法从早期[25] 到现在[26]，都是基于全局或局部 PCA 方法，本征维数由特征值大于一给定阈值的个数决定 $\left(\left(\sum_{j=1}^{n}\lambda_j / \sum_{i=1}^{d}\lambda_i\right) \geqslant v\right.$，其中 d 为判断维数，v 为选取的阈值). 但是由于全局 PCA 对于非线性拓扑结构可能会失效，局部方法很大程度取决于邻域和阈值的选择[27]，所以特征值方法虽然是数据分析中一个简便实用的工具，但不能提供本征维数的可靠估计.

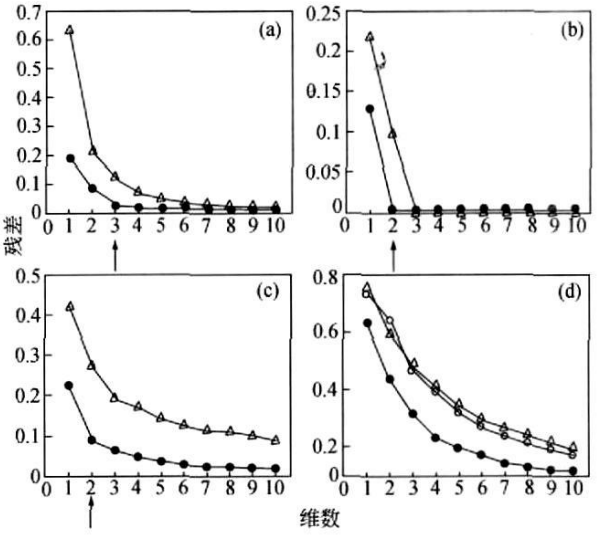

图 1 以不同维数 d 为横坐标，以残差为列坐标，得到残差的曲线图

通过曲线的“拐点”可以得出嵌入空间的维数 d，用箭头指示[8]. 四幅图的示例分别为：(a) 姿态、光照变化的人脸图像；(b) Swiss roll 数据集；(c) 人手图像；(d) 手写数字“2”

(2) 几何学习方法. 发掘数据的本质几何特征，这种方法大多是基于不规则小块的维数[27] (fractal dimensions)或最近邻点距离(NN distances). 不规则小块划分方法的选择对结果也会产生较大影响.

(3) 统计学习方法. Levina 和 Bickel 提出了一种新的估计本征维数的方法. 它在邻域点间的距离中采用最大似然理论，设计分类器，并在理论上和仿真中证明了其可行性[28].

我们在研究中采用了 DGSOM 来估计本征维数[29]. 在原始数据上运用 DGSOM 方法，得到数据点减少的拓扑图，其中相邻点之间的拓扑连接能合理有效地反映原始数据的本质结构. 通过计算网络中节点的平均连接节点数目可以估计网络的维数，

即数据点的本征维数. 图 2 是通过 Monte Carlo 方法得到的 DGSOM 网络中输入数据的维数和平均连接节点数目的关系图[30].

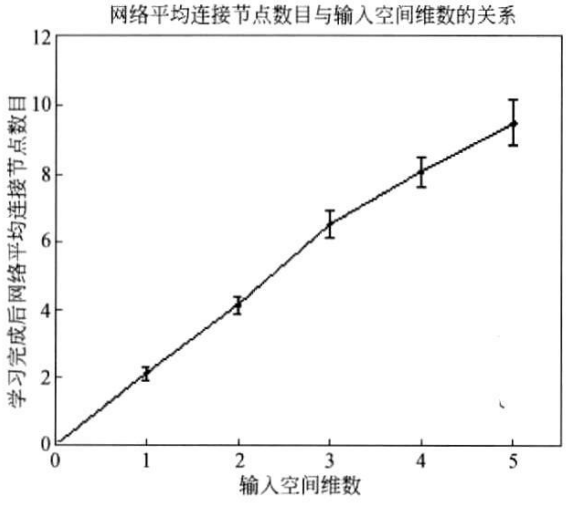

图 2 网络平均连接节点数目与输入空间维数的关系曲线
dot 表示多次使用 DGSOM 模型针对不同维数的数据映射得到的网络平均连接节点数目, bar 表示统计方差

3.2 封闭、带有孔洞或维数跳变的非线性降维

目前的非线性降维方法大都假设信任数据集分布在一个平滑非闭合的低维流形, 对于闭合曲面、不连续流形以及各处曲率相差较大(维数不固定)的情况, 现有算法就很难解决, 需要在相应地方予以改进. 例如, ISOMAP 不能发掘非凸流形结构[8], LLE 要求流形不存在孔洞[9]等.

针对有孔流形问题, Donoho 等提出一种基于 Hessian 矩阵的 LLE 算法(HLLE)[31], 该方法把原始欧式空间的局部邻域看作是 Riemann 子流形, 嵌入向量可以转化为 Hessian 矩阵的特征求解问题, 如图 3(c)所示. 由于局部子流形不要求是凸的, 因此算法可以解决更多的降维问题.

对于高维观测空间是闭合的情况(特别是连通而非单连通, 如圆柱面), 尽管整个流形也是光滑连续的, 但由于难以确定起始点与结束点, 很难找出低维嵌入结构. Lee 尝试将流形剪开的方法较好地解决了这一问题[32]. 另外, 有些高维观测空间的本征结构在某个地方发生跳变, 通过非线性降维方法很难映射到一个统一的低维坐标下, 这些都需要进一步改进.

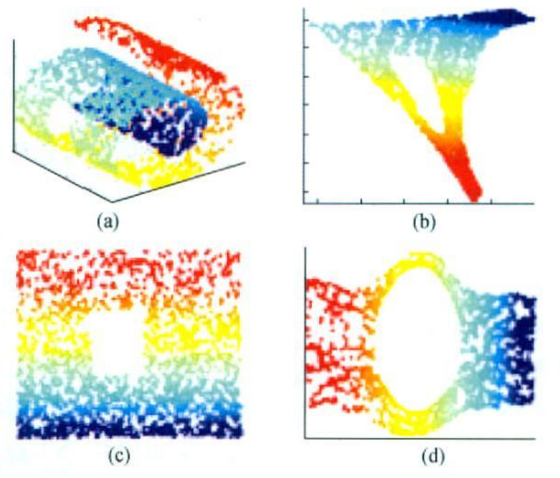

图 3 带孔洞 Swiss roll 数据集的展开
(a) 带孔洞的 Swiss roll 数据集; (b) 原始 LLE, 邻域值 $K=12$; (c) Hessian LLE, $K=12$; (d) ISOMAP, $K=7$

3.3 样本泛化(out of sample)学习能力

在流形学习中, 缺少高维观测空间与低维嵌入空间的映射关系是非线性降维普遍存在的问题. 以特征向量求解方法为例, 当引入新的样本时, 需要重新构造权值矩阵, 并计算相应矩阵的特征向量求得低维嵌入值, 这一缺点使现有的非线性降维方法不适应动态变化的样本空间, 很难实现增量学习. 在特征提取和分类过程中, 高维与低维数据间的映射关系能够提高对新样本的分类能力[33], 这就需要寻找新方法构建映射关系.

Charting 算法能够给出高维样本点与低维坐标空间的映射关系[13], 但这种方法需要大量样本作为训练集, 在实际应用中很多时候样本数量是有限的. 张长水提出一种基于表情的解析映射法, 这种方法在小数据集中就能够建立映射关系, 并在人脸识别中得到较好的验证[1]. Bengio 等在分析几种无监督非线性降维算法基础上, 提出了对这些算法的 out of sample 学习能力扩展[34], 使用一个训练好的模型用于 out of sample 数据的学习, 而无需重新计算特征向量. 图 4 展现了对不同降维方法扩展后, 训练集的嵌入变异性与 out of sample 误差. 可以看出, 在大多数情况下, out of sample 误差小于训练集的嵌入变化量或与之相比拟.

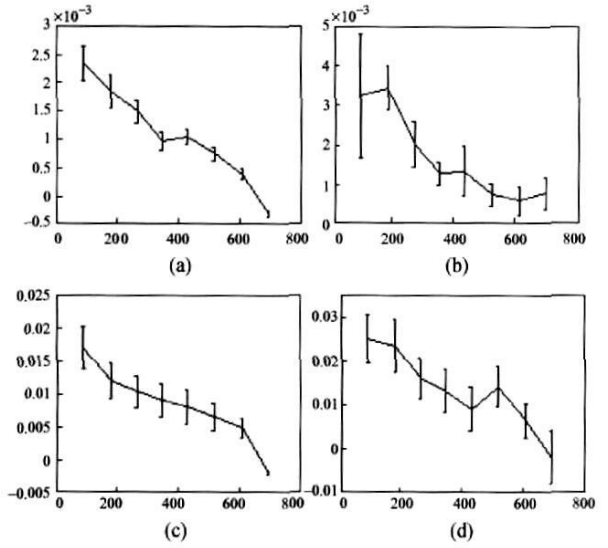

图 4 随着训练集点数(***X*** 轴)的增加，训练集的嵌入变异性与 **out of sample** 误差(***Y*** 轴)均下降[34]
(a) MDS；(b) Laplace 特征映射；(c) ISOMAP；(d) LLE

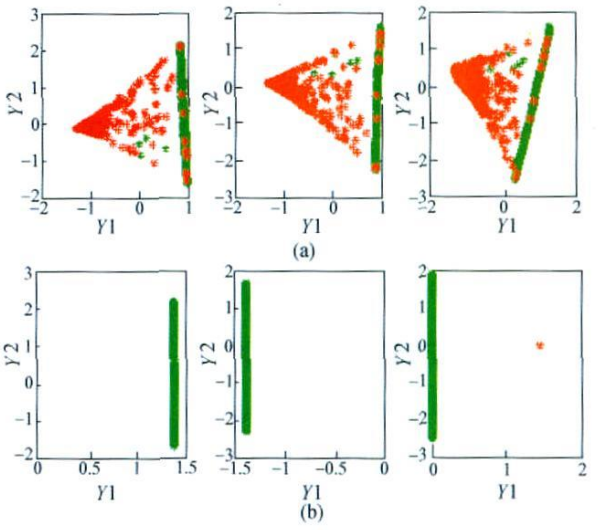

图 5 对不同的 ***K*** 值(分别为 ***K*** = 5，***K*** = 7，***K*** = 9)，把手写数字图像“1”和“7”映射到 2 D 空间[36]
(a)无监督 LLE 的映射；(b)有监督 LLE 的映射

3.4 有监督学习

当我们不知道数据点的分类情况以及类间数据点之间的相互关系时，采用无监督的算法可以看清原始数据的本质结构；而当原始数据具有一定类别信息时，采用有监督的算法可以增强可视化和聚类分类能力. 因此，有监督与无监督方法的应用领域不同，前者从分类的观点考虑，是把属于同一类的高维数据点映射到嵌入空间的同一点下；后者从可视化与聚类观点考虑，目的是发掘数据点间的内部结构.

以 LLE 算法为例，有监督 LLE(SLLE)用到了类别信息. 对任一数据点，其邻域点仅在同类的数据点中选取. SLLE 由 Ridder 和 Kouropteva 提出[35, 36]，根据先验知识把初始数据集 Ψ 分为 m 个子集$\{\Psi_1, \Psi_2, \cdots, \Psi_m\}$，改变邻域点影响权值矩阵 W 的构造，进而改变矩阵 M，使得降维后类内间距明显减小而类间间距明显增大，如图 5 对手写数字图像的降维结果；也有引入参数 α 来控制算法中引入的类别信息程度，如半监督学习. 文献[37]中给出无监督和有监督 LLE 算法统一框架，并以大量数据集验证了算法的有效性. 近年来，有监督与半监督学习受到越来越多的重视，其在模式识别与高维信息处理领域的应用也越来越广泛.

3.5 噪声流形的学习

当观测数据是对一个光滑流形较好的采样时，使用非线性降维可以找出其内在本质的流形分布. 但是，在实际的高维采样数据中由于各种因素经常存在噪声，使得映射到低维空间后会出现对原始数据结构的扭曲和变形. 图 6 表明了当三维 S 曲线加进 SNR 为 10 dB 的 Gauss 白噪声时，降维产生的结果发生了很严重的变形.

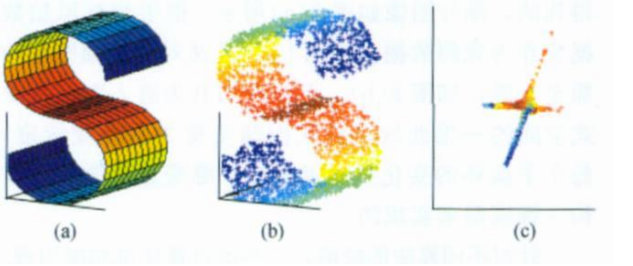

图 6 (**a**) 原始 **S** 曲线；(**b**) 加进 **Gauss** 白噪声的 **S** 曲线采样点；(**c**) 使用 **LLE** 后的低维映射结果，没有展开并且在颜色编码上发生了混乱(***K*** = 8)

针对非线性降维方法普遍存在的对噪声(奇异点)敏感的问题，出现一些改进措施. 张振跃等提出一种局部线性平滑的思想，采用加权 PCA 来构建局部小块，并用迭代方法优化权值[14]. 该方法可以用作其他非线性降维方法的预处理，平滑噪声，得

到对原始数据较好的重构. 而基于局部切线空间的主流形学习算法 LTSA (local tangent space alignment)[38], 能够发掘带有噪声的高维空间流形分布, 且重构误差具有二阶精确度. 这些改进方法也同样存在一些问题, 如迭代算法容易陷入局部极小值, 鲁棒性不够高等. 在我们的研究中也给出一种去除噪声的思想[29], 该思想对噪声的适应范围较有限, 在我们目前的研究中对此问题已有了更进一步的改善.

4 非线性流形学习的应用

非线性降维法具有线性降维法不可比拟的优点, 能够更好地应用于非线性高维数据的降维、聚类分析, 发现数据的内部结构和隐藏信息. 本节结合自己的研究, 从以下几方面对非线性降维方法的应用进行阐述.

4.1 高维数据约简与可视化

由于高维数据中存在较多的冗余信息, 同时也隐藏了数据间重要的相关性, 对其进行降维可以消除冗余信息, 突出相关性, 是高维数据处理中重要的前处理步骤.

图 7 显示了各种非线性降维方法用于 Swiss Roll 数据集的展开情况, 可以看出选择合适的参数后, 均能在二维空间中把嵌入在三维空间的 Swiss Roll 流形展开.

手旋杯数据是由一个视频序列在等间隔采样中得到的, 部分图像如图 8(a) 所示. 把手旋杯原始数据集作为观测数据, 采用 LLE 算法对 481 幅图像约简至 3 维, 如图 8(b), 不难发现其为嵌入在三维欧式空间的一维曲线, 内在控制变量为水平旋转度, 每个手旋杯的变化是通过内在一维变量的插值和重构一维模型来实现的.

针对不同算法的缺陷, 一些改进算法也相继出现. Silva 等对 ISOMAP 予以改进, 提出 e ISOMAP, 使高维空间与低维空间不仅具有保距映射, 而且具有保角映射的性质[18]. 前期工作中分析了 LLE 算法中引起重构误差的各个因素, 在空间几何流形中嵌入拓扑, 提出了具有拓扑保持能力的生长型 LLE(GLLE)[39].

4.2 在聚类分析中的应用

对于带有不同类别信息的高维数据, 非线性降维方法具有很好的聚类分析能力, 把属于不同类别的高维数据在低维空间中较明显地分开, 达到可视化目的. 聚类结果可以作为分类和识别问题的基础, 具有重要意义.

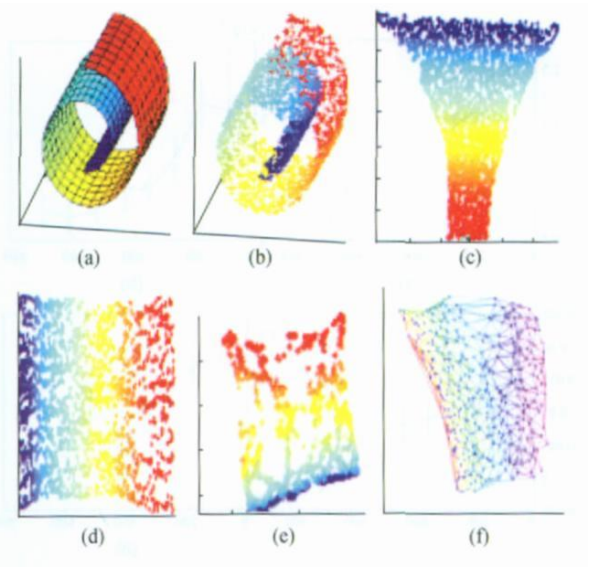

图 7 (a) 原始 Swiss Roll 曲面; (b) 二维流形(A) 的数据采样, $N=2000$; (c) LLE 二维展开结果, $K=12$; (d) ISOMAP 二维展开结果; (e) Laplace 特征映射, $K=15$, $t=5.0$; (f) Charting 二维展开结果

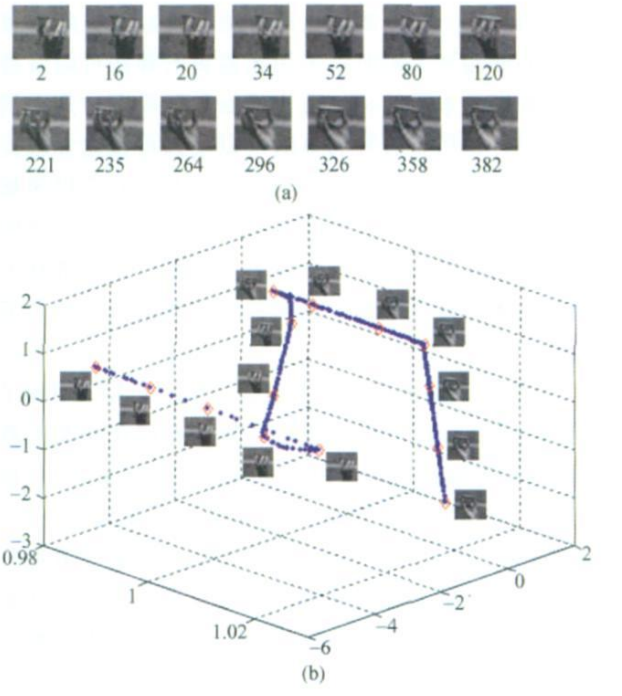

图 8 (a) 手旋杯部分图像示例; (b) 本征一维曲线

以手写体数字图像聚类分析为例，采用 USPS 数据集[40]，数字(“0”到“9”)的图像经前处理变成分辨率 16× 16(D= 256)，并把灰度值量化为 256 阶. 前处理后的图像作为 LLE 的输入数据，先用奇异值分解(SVD)把每个邻域投影到一个八维子空间，经 LLE 降维后在前两维坐标中显示如图 9. 可见 LLE 用于带类别信息数据的降维时，在低维坐标中也能把不同类别数字分开，达到聚类的目的.

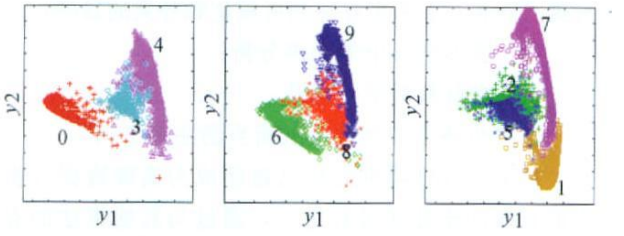

图 **9** 手写数字图像(“**0**”到“**9**”)经 **LLE** 降维后在二维空间的映射结果，N= **11000**，K= **18**，D= **256**

图 10 是相同的数据库中采用 SNE 算法降维的结果. 选取 5 类数据(0，1，2，3，4)共 3000 幅数字图像，每类有 600 幅. SNE 初始化为把所有 y_i 分配到原点附近，沿噪声梯度下降方向进行训练. SNE 在不使用类别信息的情况下，可以很清楚地把不同类别的数字分离[12]，如图 10 所示. 另外，原始数字的属性如方位、倾斜度、笔画粗细等在低维空间中的变化也趋于平滑.

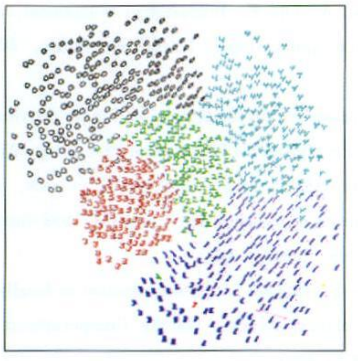

图 **10** **SNE** 用于 **3000** 幅 **256** 维手写数字灰度图像的结果[12] 与低维嵌入值 y_i 相对应的原始数据 x_i 的图像显示在相应位置

4.3 多姿态人脸图像分析

传统的人脸识别方法大都是基于静态图像的识别，这些方法对于姿态、表情、照度发生变化的人脸图像比较敏感. 当人脸发生转动或者照度等发生变化时，其相应的特征变化可以看作镶嵌在图像空间中的一个低维非线性子流形，如何从中提取出人脸图像较紧凑的分布正是非线性降维方法所要解决的问题.

采用 Frey 人脸库[1)]，经 LLE 映射后显示结果如图 11，沿图中两个坐标方向可以很清晰地看出人脸图像随姿态和表情的变化情况.

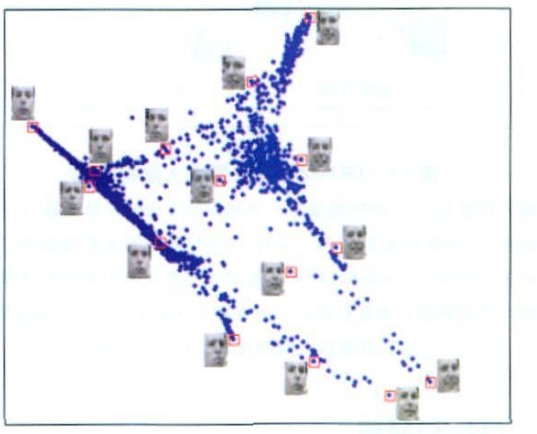

图 **11** 经 **LLE** 映射后显示人脸图像图

图像分辨率 28× 28，为同一人不同表情共 1965 幅人脸图像，作为 LLE 的输入，映射到 2 维嵌入空间，D= 560，K= 12. 不同地方的嵌入点(以小方框表示)对应的人脸图显示在相应位置

文献[6]中采用 ISOMAP 算法展示了人脸图像识别的另一个例子，见图 12. 可见 ISOMAP 能够很好地发掘隐藏在高维人脸图像中随姿势、表情发生变化的三维流形分布.

从前面的实现结果可以看出，非线性降维方法能够很好地发掘高维人脸图像中的低维流形分布，如姿态、表情、光照的变化. 但是，由于这些算法中都缺少人脸图像与低维坐标中的映射关系，给人脸识别带来一定困难，文献[1]中对 LLE 算法给出了映射关系，对此问题给出一个较好的解决办法. 另外，文献[21]把 LLE 与 LDA 相结合，在很大程度上提高了对人脸图像的分类能力. 对于不同人脸图像的识别问题，还需对目前的非线性降维方法提出进一步改进.

1) Available at http://www.cs.toronto.edu/~roweis/data.html

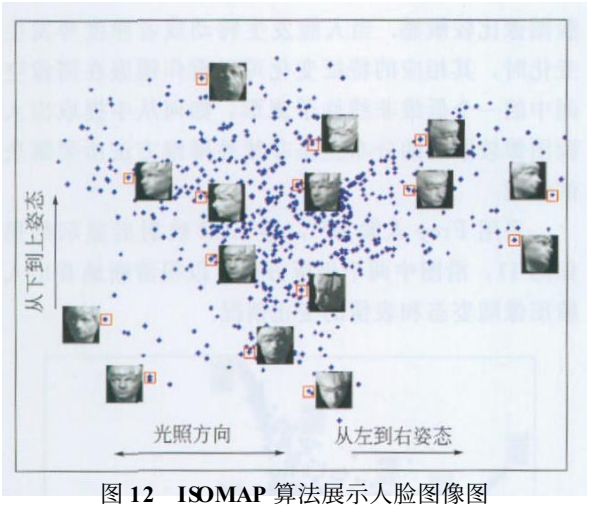

图 12 ISOMAP 算法展示人脸图像图

输入数据是一个 4096 维姿势、光照发生变化的人脸图像(3 个自由度), 分辨率为 64× 64, 运行 ISOMAP 得到本征维数是三维的嵌入空间, N= 698, K= 6. 图中为二维观测空间中的人脸分布, 所有数据点由蓝点表示, 小方框为分布在二维空间的采样点, 其高维人脸图像显示在相应位置

5 总结与展望

基于流形学习的非线性降维方法最主要的应用是数据压缩, 找出隐藏在高维观测数据中有意义的低维流形分布, 消除冗余信息, 减少运算量; 另一个主要应用是数据集的可视化[41]. 近年来流形学习算法与应用取得了丰硕的成果, 但是由于其较深的数学理论基础, 以及多学科交叉等特点, 还有很多问题需要进一步研究和改进, 这也为流形学习的发展提供了更广阔的空间.

本文介绍了目前主要的 5 种非线性流形学习方法, 包括 ISOMAP, LLE, Laplace 特征映射, SNE 和图册流形等; 系统介绍了每种方法的具体步骤, 通过比较得出各种流形学习算法优缺点, 并根据流形学习采用局部线性原理的共同点, 提出了非线性降维方法的统一框架, 对于流形学习有着重要的指导意义.

非线性降维算法自提出以来, 处理了很多传统线性降维方法解决不了的问题, 并不断得以改进[19, 21, 35, 42]. 局部保持投影[11] 基于 Laplace 特征映射给出了线性的映射关系, 为流形学习的发展带来一些新思想, 但在处理低维流形上未能保持全局拓扑关系. 实验表明在非线性数据压缩、聚类分析、多姿态人脸识别等领域中, 非线性降维方法均取得了较好的应用. 但是, 流形学习方法还需进一步完善, 主要表现在以下几方面:

(1) 原始数据本征维数的估计;

(2) 流形学习假定了数据是在流形上的稠密采集, 那么小样本条件下的流形学习如何实现;

(3) 对于不连续或维数跳变的流形降维问题;

(4) 缺少高维观测数据与低维数据间的映射关系;

(5) 半监督或有监督学习中如何利用类别信息或数据点间相互关系更好达到可视化和分类的目的;

(6) 流形学习中的噪声分析;

(7) 增量流形学习问题.

这些也将是非线性降维研究的主要方向.

目前非线性降维方法已逐渐成为高维数据分析中最有效的预处理过程之一, 通过与其他算法的有效结合可以获得聚类分类等性能的改善. 在处理多类问题中, 采用单纯的降维方法可能不会获得很好的分类结果, 如何结合拓扑学习理论, 并体现几何学习[43] 的优势, 也是非线性降维一个很好的突破点.

参 考 文 献

1 Zhang CS, Wang J, Zhao NY, et al. Reconstruction and analysis of multi pose face images based on nonlinear dimensionality reduction. Pattern Recognition, 2004, 37(2): 325—336

2 Jolliffe IT. Principal Component Analysis. New York: Springer Verlag, 1986

3 Hyvarinen A, Oja E. Independent component analysis: Algorithms and applications. Neural Networks, 2000, 13(4-5): 411—430

4 Balakrishnama S, Ganapathiraju A. Linear discriminate analysis. Institute for Signal and Information Processing, Mississippi State University, 1998. Available at http://www.isip.msstate.edu/publications/reports/isip_internal/1998/linear_discrim_analysis/lda_theory.pdf

5 Lawrence K, Roweis S. An introduction to locally linear embedding. Technical Report, Gatsby Computational Neuroscience Unit, UCL, 2001

6 Kohonen T. Self Organizing Maps (Eds. 2). Springer. 1995

7 Raymer F, Punch L, Goodman D, et al. Dimensionality reduction using genetic algorithm. IEEE Trans on Evolutionary Computation, 2000, 4: 164—171

8 Tenenbaum JB, Silva V de, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290: 2319—2323

9 Roweis S, Saul LK. Nonlinear dimensionality reduction by local

ly linear embedding. Science, 2000, 290: 2323—2326
10 Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 2003, 15(6): 1373—1396
11 He XF, Yan SC, Hu YX, et al. Face recognition using laplacianfaces. IEEE Trans on PAMI, 2005, 27(3): 328—340
12 Hinton G, Roweis S. Stochastic neighbor embedding. Neural Information Proceeding Systems: Natural and Synthetic. Vancouver, Canada, 2002
13 Brans MM. Charting a manifold. Neural Information Proceeding Systems: Natural and Synthetic. Vancouver, Canada, 2002
14 Park JH, Zhang ZY, Zha HY, et al. Local linear smoothing for nonlinear manifold learning. IEEE Computer Society Conference on CVPR. Washington DC, 2004, 2: 452—459
15 Cox T, Cox M. Multidimensional Scaling. London: Chapman & Hall, 1994
16 陈维恒, 李兴校. 黎曼几何引论. 北京: 北京大学出版社, 2002, 9—12
17 Balasubramanian M, Schwartz EL, Tenenbaum JB, et al. The isomap algorithm and topological stability. Science, 2002, 295: 7a
18 Silva VD, Tenenbaum JB. Global versus local methods in nonlinear dimensionality reduction. Neural Information Proceeding Systems: Natural and Synthetic. Vancouver, Canada, 2002
19 Odest J, Maja M. A spatio temporal extension to isomap nonlinear dimension reduction. The 21st International Conference on Machine Learning. Banff, Canada, 2004
20 Ridder D de, Loog M, Reinders MJ. Local fisher embedding. 17th ICPR. Cambridge, UK, 2004
21 Zhang JP, Shen HX, Zhou ZH. Unified locally linear embedding and linear discriminate analysis algorithm for face recognition. Sino Biometrics. 2004: 296—304
22 Yan SC, Xu D, Zhang B, et al. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans on PAMI, 2007, 29(1): 40—51
23 Kegl B. Intrinsic dimension estimation using packing numbers. Neural Information Proceeding Systems: Natural and Synthetic. Vancouver, Canada, December 2002
24 Verveer P, Duin R. An evaluation of intrinsic dimensionality estimators. IEEE Trans on PAMI, 1995, 17(1): 81—86
25 Fukunaga K, Olsen DR. An algorithm for finding intrinsic dimensionality of data. IEEE Trans on Computers, 1971, C-20: 176—183
26 Bruske J, Sommer G. Intrinsic dimensionality estimation with optimally topology preserving maps. IEEE Trans on PAMI, 1998, 20(5): 572—575
27 Camastra F, Vinciarelli A. Estimating intrinsic dimension of data with fractal-based approach. IEEE Trans on PAMI, 2002, 24(10): 1404—1407
28 Levina E, Bickel P. Maximum likelihood estimation of intrinsic dimension. In: NIPS 17. Cambridge, MA, December 2005
29 Xiao J, Zhou ZT, Hu DW, et al. Self organizing locally linear embedding for nonlinear dimensionality reduction. ICNC2005, LNCS3610. Changsha, 2005, 101—109
30 Chen S, Zhou ZT, Hu DW. Diffusion and growing self organizing map: A nitric oxide based neural model. ISNN 2004, LNCS3174. 2004, 199—204
31 Donoho D, Grimes C. Hessian eigenmaps: Locally linear embedding techniques for high dimensional data. Proceedings of the National Academy of Sciences, 2003, 100(10): 5591—5596
32 Lee JA, Michel V. Nonlinear dimensionality reduction of data manifolds with essential loops. Neurocomputing, 2005, 67: 29—53
33 Vlachos M, Domenicon C, Gunopulos D, et al. Nonlinear dimensionality reduction techniques for classification and visualization. Proc of 8th SIGKDD. Edmonton, Canada, 2002
34 Bengio Y, Paiement JF, Vincent P. Out of sample extensions for LLE, Isomap, MDS, Eigenmaps and Spectral Clustering, Technical Report 1238, University de Montreal, July, 2003
35 Ridder D de, Kouropteva O, Okun O, et al. Supervised locally linear embedding. In: Proc ICANN/ICONIP 2003, LNCS 2714. Springer Verlag, 2003, 333—341
36 Kouropteva O, Okun O, Pietikainen M. Supervised locally linear embedding algorithm for pattern recognition. In: Proc IbPRIA 2003, LNCS 2652, Springer Verlag, 2003, 386—394
37 Ridder D de, Duin RPW. Locally linear embedding for classification. Technical Report PH-2002-01, Pattern Recognition Group, Dept. of Imaging Science & Technology, Delft University of Technology, Delft, The Netherlands, 2002
38 Zhang ZY, Zha HY. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. SIAM J Scientific Computing, 2004, 26(1): 313—338
39 Yin JS, Hu DW, Zhou ZT. Growing locally linear embedding for manifold learning. Journal of Pattern Recognition Research, 2007, 2(1): 1—16
40 Hull J J. A database for handwritten text recognition research. IEEE Trans on PAMI, 1994, 16(5): 550—554
41 Lawrence K, Roweis S. Think globally, fit locally: Unsupervised learning of nonlinear manifolds. J of Machine Learning Research, 2003, 4: 119—155
42 Kouropteva O, Okun O, Hadid A, et al. Beyond locally linear embedding algorithm. Technical Report MVG-01-2002, Machine Vision Group, University of Oulu, Finland, 2002
43 Wang SJ, Lai JL. Geometrical learning, descriptive geometry, and biomimetic pattern recognition. Neurocomputing, 2005, 67: 9—28