文章编号: 1001- 2486(2004) 06- 0091- 05

稳健局部线性嵌入方法*

谭 璐, 吴 翊, 易东云
(国防科技大学理学院, 湖南 长沙 410073)

摘 要: 针对局部线性嵌入方法对于噪声的敏感性, 从分析噪声对数据集局部特性的影响入手, 提出了稳健局部线性嵌入方法。通过与局部线性嵌入方法的理论分析和实例对比, 表明稳健局部线性嵌入方法不仅对噪声影响不敏感, 而且对邻域的选择有较好的适应性, 可更好地挖掘数据集的本征特性, 具有更强的数据可视化能力。

关键词: 局部线性嵌入; 高维数据; 降维; 邻域
中图分类号: O29 文献标识码: A

Robust Locally Linear Embedding

TAN Lu, WU Yi, YI Dong-yun
(College of Science, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Since the locally linear embedding method is sensitive to noise, a new method is presented to solve this problem. This new method, namely robust locally linear embedding method, is constructed by analyzing noise influence on the character of the data set' s local neighborhood. Compared with LLE, the RLLE is insensitive to noise and adaptive to the selection of neighborhood, which is verified by theoretical analysis and practical experiments. Therefore, the RLLE can discover the intrinsic structure of the data set and visualize the data better.

Key words: locally linear embedding; high dimensional data; dimensionality reduction; neighborhood

高维数据处理的研究, 因其在航天遥感数据、生物数据、网络数据以及金融市场交易数据等领域的广泛应用, 而受到普遍重视。由于众所周知的"维数灾祸"(curses of dimensionality) 问题, 通常对于高维数据的处理, 如图像分类、模式识别等, 存在着种种困难。一种常用的做法是在保持数据所含感兴趣信息的前提下, 尽可能降低数据的维数, 即降维。现已有许多可行的线性降维方法, 如主成分分析 PCA[1]、多维尺度分析 MDS[2]、线性奇异分析 LDA[3] 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[4], 它们的优势是运算快捷、方便, 并能产生简单的变换函数, 对线性结构效果好, 但是对于某些数据而言, 其所携带的信息呈现某种"非线性性", 如高维空间中的曲面数据, 使得线性方法失效。于是本世纪以来, 出现了许多新的非线性方法, 主要有局部线性嵌入 LLE[5] (Locally Linear Embedding)、拉普拉斯特征映射(Laplacian Eigenmap) [6]、基于 Hessian 矩阵的 LLE[8] 等, 它们通过在高维空间中设计数据集所在流形的拓扑、几何等特性, 很好地弥补了线性降维不能发现数据集非线性结构的不足。

LLE 方法是 2000 年由 Sam T. Roweis 和 Lawrence K. Saul 提出的一种新的非线性降维方法, 在脸谱图像数据和文档类数据处理中获得了很好的效果[7]。但在应用中仍有许多问题, 其中较突出的是在求解模型过程中可能出现的矩阵病态, 会导致求解过程对于数据点噪声十分敏感, 使结果受噪声影响很大。针对 LLE 方法的这一不足, 从分析噪声影响入手, 在此基础上提出了稳健局部线性嵌入方法, 简记为 RLLE(Robust Locally Linear Embedding) , 并通过高维数据分类的实例说明, 它很好地减弱了噪声对降维过程的影响, 可更好地揭示数据集的本征结构。

* 收稿日期: 2004- 06- 19
基金项目: 国家自然科学基金资助项目(60003013) ; 高等学校博士学科点专项科研基金资助项目(20049998008)
作者简介: 谭璐(1977—) , 男, 博士生。

1 LLE方法

LLE的具体做法为,设在高维欧氏空间 R^D 中有数据集 $\boldsymbol{X}=\{x_1, x_2, \cdots, x_N\}$。该方法希望将 $\boldsymbol{X}$ 嵌入到一个相对低维的空间 R^d 中,同时尽可能地保持原数据所携带的信息,即保持数据集的拓扑结构(通过每点的邻域关系确定)。考虑任意取定的点 $x_i \in \boldsymbol{X}$, $U(x_i)$ 表示 x_i 的 $k-$ 近邻域(即 $\boldsymbol{X}$ 中离 x_i 最近的 k 个点组成的集合),寻求 N 维向量 $\boldsymbol{w}_i=(w_{i1}, w_{i2}, \cdots, w_{iN})$,使得

$$Q(\boldsymbol{w}_i)=\min_{\boldsymbol{w}_i \in R^N}\left\|x_i-\sum_{j=1}^{N} w_{ij}x_j\right\|^2 \tag{1}$$

其中, $\sum_{j=1}^{N} w_{ij}=1$, $w_{ii}=0$,以及当 $x_j \notin U(x_i)$ 时, $w_{ij}=0$。w_i 记录了 x_i 点的邻域信息,即每点的局部拓扑结构,记它们的全体构成的矩阵为 $\boldsymbol{W}$。此时选择适当的维数 d,做 $\boldsymbol{X}$ 的 d 维嵌入,即求 $\boldsymbol{Y}=(y_1, y_2, \cdots, y_N) \in R^{d\times N}$,使其满足

$$L(y_1^*, y_2^*, \cdots y_N^*)=\min_{\boldsymbol{Y}}\sum_{i=1}^{N}\| y_i-\boldsymbol{Y}\boldsymbol{W}_i^{\mathrm{T}}\|^2=\min_{\boldsymbol{Y}}(\mathrm{tr}(\boldsymbol{Y}\boldsymbol{M}\boldsymbol{Y}^{\mathrm{T}}))$$

为了求解的惟一性,施加规范化限制,即令 $\sum_{i=1}^{N} y_i=0$, $\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}=\boldsymbol{I}_{d\times d}$, $\boldsymbol{I}_{d\times d}$ 为单位矩阵。易知求解上述数据集 $\boldsymbol{Y}$ 等价于求 $\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W}^{\mathrm{T}})$ 的特征向量问题。由于 $\boldsymbol{W}$ 是由数据集 $\boldsymbol{X}$ 解出的,故当数据集 $\boldsymbol{X}$ 受噪声污染时,计算对噪声非常敏感,特别是当 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的特征值较小时,甚至得不到需要的结果。

2 噪声影响分析

通过上一节的分析可以看出,在LLE降维过程中,矩阵 $\boldsymbol{W}$ 是否准确将直接影响低维嵌入的最终效果。下面研究当数据集受噪声污染时, w_i 的变化情况。为简化记号,令 x_0 代表 x_i, $U(x_0)$ 代表 x_0 点的邻域,不妨设 $x_1, x_2, \cdots, x_k \in U(x_0)$,则

$$x_0=\sum_{i=1}^{k} w^i x_i, \quad \sum_{i=1}^{k} w^i=1$$

令 $x'_i=x_i+\varepsilon(i=0, 1, 2, \cdots, k)$,代表相应的受噪声污染的点,以及相应的

$$x'_0=\sum_{i=1}^{k} w^{i'} x'_i, \quad \sum_{i=1}^{k} w^{i'}=1$$

若再令 $\boldsymbol{X}^0=(x_1, x_2, \cdots, x_k)$, $\boldsymbol{X}^{0'}=(x'_1, x'_2, \cdots, x'_k)$,以及 $\boldsymbol{w}=(w^1, w^2, \cdots, w^k)^{\mathrm{T}}$, $\boldsymbol{w}'=(w^{1'} w^{2'}, \cdots, w^{k'})^{\mathrm{T}}$,那么有

$$x_0=\boldsymbol{X}^0\boldsymbol{w}, \quad x'_0=\boldsymbol{X}^{0'}\boldsymbol{w}'$$

定理 在上述记号下,若各点噪声之间,不同维数之间,以及 $\boldsymbol{w}'$ 与噪声之间是相互独立的,各点噪声是同均值,为0,同方差的。那么对于 $\delta\boldsymbol{w}=\boldsymbol{w}'-\boldsymbol{w}$ 有如下估计

$$\mathrm{E}\|\delta\boldsymbol{w}\|^2 \leqslant \frac{k(k+1)\sigma^2}{\lambda_{\min} l}\mathrm{E}\|\boldsymbol{w}'\|^2$$

其中, $\|\cdot\|$ 取为欧几里得范数; $\sigma^2=\sum_{i=1}^{D}\sigma_i^2$, $\sigma_i^2=\mathrm{Var}(\varepsilon^i)(i=1, 2, \cdots, D)$, ε^i 代表 ε 的第 i 个分量, $l=\mathrm{rank}(\boldsymbol{X}^0)$, $\lambda_{\min}$ 为 $\boldsymbol{X}^{0^{\mathrm{T}}}\boldsymbol{X}^0$ 的最小非零特征值。

证明 由 $x'_i=x_i+\varepsilon(i=0, 1, \cdots, k)$ 可见

$$\boldsymbol{x}'_0=\boldsymbol{X}'_0\boldsymbol{w}'=\boldsymbol{X}^0\boldsymbol{w}'+\sum_{i=1}^{k} w'_i\varepsilon_i \Rightarrow x_0=\boldsymbol{X}^0\boldsymbol{w}'+\sum_{i=1}^{k} w'_i\varepsilon_i-\varepsilon_0$$

进一步有

$$\boldsymbol{X}^0(\boldsymbol{w}'-\boldsymbol{w})=\sum_{i=1}^{k} w'_i(\varepsilon_0-\varepsilon_i), \quad \sum_{i=1}^{k} w'_i=\sum_{i=1}^{k} w_i=1$$

即

$$\boldsymbol{X}^0\delta\boldsymbol{w}=\sum_{i=1}^{k}w'_i(\varepsilon_0-\varepsilon_i)$$

由于噪声是独立的, 则有

$$\mathrm{E}(\delta\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}^{0^{\mathrm{T}}}\boldsymbol{X}^0\delta\boldsymbol{w})=\mathrm{E}\Big[\sum_{i=1}^{k}w'_i(\varepsilon_0-\varepsilon)^{\mathrm{T}}\sum_{i=1}^{k}w'_i(\varepsilon_0-\varepsilon_i)\Big]$$

$$=\sum_{i=1}^{k}\mathrm{E}w_i'^2\sigma^2+\sum_{i,j}^{k}\mathrm{E}(w'_iw'_j)\sigma^2\leqslant(k+1)\sigma^2\sum_{i=1}^{k}\mathrm{E}w_i'^2=(k+1)\sigma^2\mathrm{E}\|\boldsymbol{w}'\|^2 \tag{2}$$

记 $\boldsymbol{X}^{0^{\mathrm{T}}}\boldsymbol{X}^0$ 的正交分解为 $\boldsymbol{X}^{0^{\mathrm{T}}}\boldsymbol{X}^0=\boldsymbol{A}^{\mathrm{T}}\Lambda\boldsymbol{A}$, 其中, $\boldsymbol{A}$ 为正交矩阵, $\Lambda=\begin{pmatrix}\lambda_1&\cdots&0\\ \vdots&\vdots&\vdots\\ 0&\cdots&\lambda_k\end{pmatrix}$。由于 $\mathrm{rank}(\boldsymbol{X}^{0^{\mathrm{T}}}\boldsymbol{X}^0)=\mathrm{rank}(X^0)$, 故

$$\mathrm{E}(\delta\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}^{0^{\mathrm{T}}}\boldsymbol{X}^0\delta\boldsymbol{w})=\mathrm{E}(\delta\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\Lambda\boldsymbol{A}\delta\boldsymbol{w})\geqslant\lambda_{\min}\sum_{i=1,\lambda_i\neq0}^{k}\mathrm{E}(\delta w_i^2) \tag{3}$$

通过对 $\boldsymbol{A}$ 进行行初等变换(改变两行的位置), 然后计算相应的(3) 式, 将所得结果相加可得

$$\mathrm{E}(\delta\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}^{0^{\mathrm{T}}}\boldsymbol{X}^0\delta\boldsymbol{w})=\mathrm{E}(\delta\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\Lambda\boldsymbol{A}\delta\boldsymbol{w})\geqslant\lambda_{\min}\mathrm{E}\|\delta\boldsymbol{w}\|^2\frac{l}{k}$$

其中, $\lambda_{\min}=\min_{1\leqslant i\leqslant k}\{\lambda_i>0\}$。

结合(2) 式可知

$$\mathrm{E}\|\delta\boldsymbol{w}\|^2\leqslant\frac{k(k+1)}{\lambda_{\min}l}\sum_{i=1}^{k}\mathrm{E}\boldsymbol{w}_i'^2\sigma^2$$

即

$$\mathrm{E}\|\delta\boldsymbol{w}\|^2\leqslant\frac{k(k+1)\sigma^2}{\lambda_{\min}l}\mathrm{E}\|\boldsymbol{w}'\|^2 \qquad \square$$

由上述定理可以看出, 在邻域大小 k 已知的情形之下, $\boldsymbol{w}$ 的误差主要由 3 个因素决定: ① 噪声的影响, 即 σ 的大小; ② 邻域的影响, 即 $\lambda_{\min}$ 和秩 l 的大小; ③ 权重能量的影响, 即 $\|\boldsymbol{w}'\|$ 的大小。

①的影响来自于采样本身, 不是本文考虑的范围; ②中使得 $\lambda_{\min}$ 和秩 l 尽可能地大, 要求搜索全体数据点, 过于繁琐和耗时, 所以本文考虑 ③, 即通过减少权重能量来减弱噪声的影响。

此外, 在各种局部嵌入方法中, 邻域大小 k 的选取一般都是经验性的, 所以人们总是希望方法对于邻域的选择不是很敏感的, 而最后的实例表明, RLLE 方法与 LLE 方法相比, 对邻域具有一定的适应性, 可较好地克服 LLE 对邻域的敏感性问题。

3　RLLE 方法

通过上节的分析可以看出, 对 LLE 方法的改善关键是在 LLE 方法的第一步, 即求(1) 式时将权重的能量加入极小化的目标函数 $Q(\boldsymbol{w}_0)$, 即使得

$$Q(\boldsymbol{w}_0)=\|x_0-\boldsymbol{X}^0\boldsymbol{w}_0\|^2+\chi(\|\boldsymbol{w}_0\|)\quad\left(\lambda\geqslant0,\ \sum_{i=1}^{k}w_0^i=1\right) \tag{4}$$

达到最小, 来寻求 x_0 点的拓扑结构向量 $\boldsymbol{w}_0$, 相应可得拓扑结构矩阵 $\boldsymbol{W}$, 进一步可得 $\boldsymbol{M}$。

这里 λ 为稳健因子, $\boldsymbol{X}^0$ 表示 x_0 点邻域中点组成的矩阵; 然后在 $\boldsymbol{W}$ 不变意义下的降维, 即同 LLE 方法一样, 是关于矩阵 $\boldsymbol{M}$ 的特征向量问题。

可以看出, RLLE 与 LLE 是相似的, 在邻域给定的情形下, 二者的区别只在于局部拓扑结构矩阵 $\boldsymbol{W}$ 的寻求方式不同。RLLE 通过引入正则项 $f(\|\boldsymbol{w}\|)$, 避免了 LLE 的病态性, 减弱了噪声的影响, 同时又保持了 LLE 原有的特性。此外当 $\lambda=0$ 时, RLLE 即为 LLE。当然 λ 的选取与邻域点数 k、数据集的维数 D, 以及数据集的尺度有关。如若 $k<D$, 则当邻域取大时, (4) 式第 1 项趋向于小, λ 应趋于取小些; 当

数据集的尺度取小时,(4) 式第 1 项趋向于大, λ应趋于取大些。通过实际计算,发现 λ的选择对结果的影响不是很大,从而 λ的选取通常不是最关键的问题。本文选取 $\lambda = 2C_{\max}k/N$, $C_{\max}$ 为 $\boldsymbol{X}$ 的协方差矩阵 var($\boldsymbol{X}$) 的最大特征值。此外正则项 $f(\|\boldsymbol{w}\|)$ 的选择要满足两点,一是使得(4) 式的求解稳健;二是使得(4) 式的 $\|\boldsymbol{w}\|$ 尽可能地小。为便于处理,简单取 $f(x) = x^2$。

采用第 1 节的记号,RLLE 的求解过程如下:

(1) 选择邻域

这里使用固定邻域容量的 $k-$ 近邻方法, $x_j \in U_i \Leftrightarrow x_j$ 属于 x_i 的 k 个最近邻点之一。

(2) 选择权重

对每点 x_i,寻求 $\boldsymbol{w}_i^*$ 使得

$$Q(\boldsymbol{w}_i^*) = \min\{\|x_i - \boldsymbol{X}\boldsymbol{w}_i^{\mathrm{T}}\|^2 + \lambda\|\boldsymbol{w}_i\|^2\}$$

最终可求得拓扑结构矩阵 $\boldsymbol{W}^*$,为方便仍记为 $\boldsymbol{W}$。

(3) 降维映射

最小化损失函数

$$L(y_1, y_2, \cdots, y_N) = \min_{Y}\left\{\sum_{i=1}^{N}\|y_i - \boldsymbol{Y}\boldsymbol{W}_i^{\mathrm{T}}\|^2\right\} = \min_{Y}\mathrm{tr}[\boldsymbol{Y}\boldsymbol{M}\boldsymbol{T}^{\mathrm{T}}]$$

其中, $\boldsymbol{Y} = (y_1, y_2, \cdots, y_N)$, $y_i \in R^d (i = 1, 2, \cdots, N)$,满足第 1 节中的限制,得 d 维降维结果 Y。

4 实 例

例 **1** 抛物线数据,图 1 中左一为原抛物线数据,由 200 个点组成,左二为施加噪声后的数据,右二为采用 LLE 方法的 2 维结果,右一为使用 RLLE 的 2 维结果。

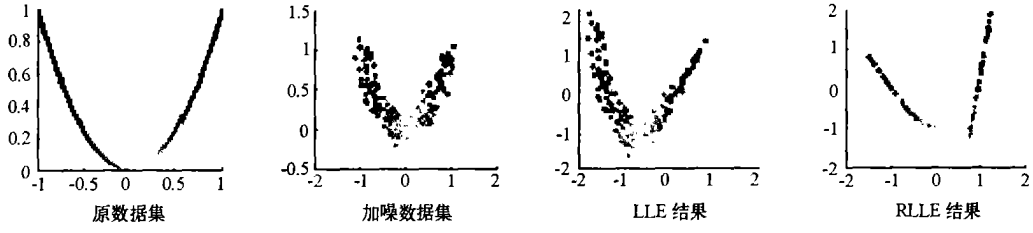

图 1 抛物线数据(10 邻域, λ= 0.0349)

Fig. 1 Parabola data (10-neighborhood, λ= 0.0349)

由上图可以看出,LLE 在保持原数据集特性的同时,由于数据集受到了噪声的污染,使得结果产生了一定的厚度,即将本征 1 维的数据变成了本征两维的数据,这显然有悖于处理的初衷。相反,RLLE 则在很好地保持数据集特性的基础上,抑制了噪声的影响,可更加清晰地揭示数据集的本征结构。

例 **2** (乳腺癌数据集)* 每组数据由 30 个属性值组成,即 $D = 30$,用来描述一个乳腺癌病例。用 LLE 和 RLLE 方法将其嵌入到 2 维空间,用以区分 98 位疑似病例中,对检测结果呈阳性(“ ”)和阴性(“•”)的两类患者。图 2、3 给出了使用 LLE、RLLE 的结果。

由图 2, 3 可以看出,(1) RLLE 与 LLE 相比可更好地分离两类数据,如图 2;(2) RLLE 对邻域的选取具有一定的适应性,如图 2、3 的右一,而 LLE 受邻域选取的影响很大,如图 2 左一, $k = 15$ 时,尚可分离两类数据,而对于图 3 左一,当 $k = 20$ 时,方法将失效;(3) 通过图 2、3 还可以看出,RLLE 结果显示数据集的本征维数为 1, LLE 结果看起来更像是 2 维的,而实际上通过对数据进一步的分析可以看出,数据点之间的差异由第 4, 24 属性值基本确定,而这两者之间是强相关的(相关系数为 0.9858),从而数据集的本征维数应为 1,使用 LLE 方法显然是无法获得相应结论的。

* http://www.ics.uci.edu/~mlearn/NLRepository.html

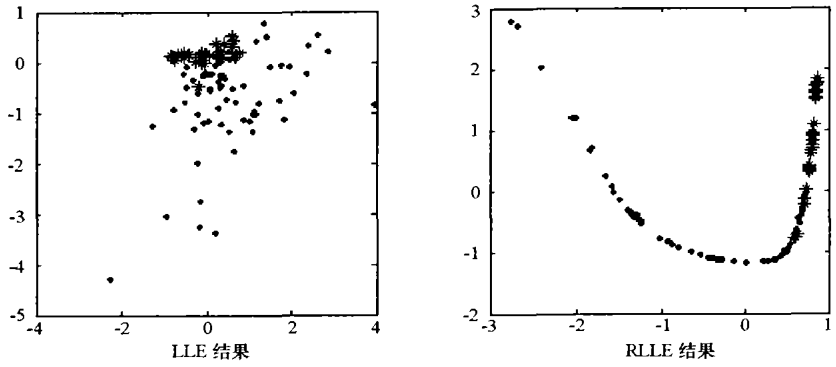

图 2　乳腺癌数据(15 邻域, $\lambda = 9.2247 \times 10^4$)

Fig. 2　Breast-cancer data (15-neighborhood, $\lambda = 9.2247 \times 10^4$)

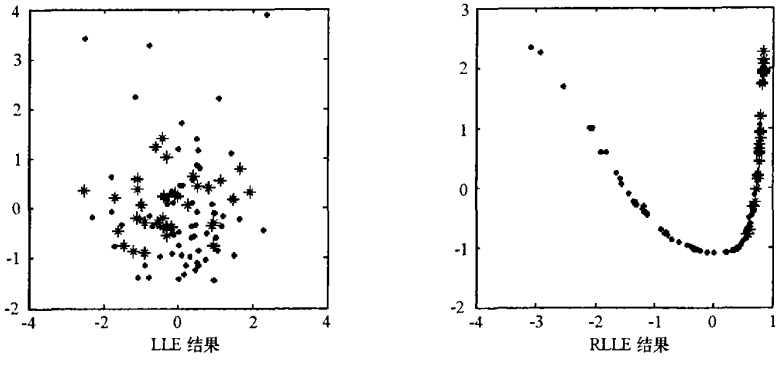

图 3　乳腺癌数据(20 邻域, $\lambda = 1.23 \times 10^5$)

Fig. 3　Breast-cancer data (20-neighborhood, $\lambda = 1.23 \times 10^5$)

5　结 论

从噪声对数据集的影响分析入手, 发现了影响数据集拓扑结构的 3 因素, 通过将其中第 3 个因素 $\|\boldsymbol{w}\|$ 加入极小化过程, 即(4) 式, 来改善拓扑结构矩阵 $\boldsymbol{W}$ 的病态性, 据此提出了 RLLE 降维方法, 并与 LLE 方法进行了实例相比。结果表明 RLLE 很好地克服了 LLE 对噪声的敏感性, 使得降维结果对邻域大小具有一定的适应性, 可更好地反映数据集的本征特性。

参 考 文 献:

[1] Jolliffe I T. Principal Component Analysis [M]. Springer-Verlag, New York, 1989.

[2] Cox T, Cox M. Multidimensional Scalling [M]. Chapman & Hall, London, 1994.

[3] Balakrishnama S, Ganapathiraju A. Linear Discriminant Analysis—A Brief Tutorial [A]. Institute for Signal and information Processing, March 2, 1998. Http:// www. isip. msstate. edu/ publications/ reports/ isip _ internal/ 1998/ linear _ discrim _ analysis/ lda _ theory. pdf.

[4] He X F, Niyogi P. Locality Preserving Projections (LPP) [R]. Technical Report, TR- 2002- 09, Computer Science Department, the University of Chicago.

[5] Roweis S T, Saul L K. Nonlmensionality Reduction by Locally Linear Embedding [J]. Science, 2000, 290(22).

[6] Bekin M, Niyogi P. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering [A]. Advances in Neural Information Processing Systems 15, Vancouver, British Columbia, Canada, 2001.

[7] Roweis S T, Saul L K. Think Globally, Fit Locally: Unsupervised Learning of Nonlinear Manifolds [R]. Technical Report MS CIS- 02- 18, and University of Pennsylvania, 2002.

[8] Donoho D L, Grimes C. Hessian Eigenmaps: New Locally Linear Embedding Techniques for High-dimensional Data [R]. Technical Report, Department of Statistics, Stanford University, 2003.