

流形学习中的算法研究

周 红, 吴 炜, 滕奇志, 杨晓敏, 李 旻, 陶德元
(四川大学 电子信息学院 图像信息研究所, 四川 成都 610064)

摘 要: 详细介绍了一种新的机器学习的方法——流形学习。流形学习是一种新的非监督学习方法, 可以有效地发现高维非线性数据集的内在维数并进行维数约简, 近年来越来越受到机器学习和认知科学领域的研究者的重视。目前已经出现了很多有效的流形学习算法, 如等度规映射 (ISOMAP)、局部线性嵌套 (Locally Linear Embedding, LLE) 等。详细讲述了当前常用的几种流形学习算法以及在流形方面已经取得的研究成果, 并对流形学习目前在各方面的应用作了较为细致的阐述。最后展望了流形学习的研究发展趋势, 且提出了流形学习中仍需解决的关键问题。

关键词: 流形学习; 主流形; 局部线性嵌套; 等度规映射; 变分法; 互信息

中图分类号: TP301 文献标志码: A 文章编号: 1001-3695(2007)07-0214-04

Research of Manifold Learning Algorithm

ZHOU Hong, WU Wei, TENG Qi-zhi, YANG Xiao-min, LI Min, TAO De-yuan

(Institute of Image Information, College of Electronics & Information Engineering, Sichuan University, Chengdu Sichuan 610064, China)

Abstract: In the development of machine learning that to discover and study the orderliness from the data is always the main object of the machine learning. This paper illustrated the study achievement gained through a new method of machine learning—manifold learning in details. As a new unsupervised learning method, manifold learning has captured the attention of many researchers in the field of machine learning and cognitive science. Currently, the major algorithms include isometric mapping (ISOMAP) and Locally Linear Embedding (LLE), and so on. The approaches can be used for discovering the intrinsic dimensions of nonlinear high-dimensional data effectively and aim researchers to analyze the data better. It also gave a prospect of the development trend of the manifold learning research and proposed the crucial problems which need to be resolved in further work in manifold learning.

Key words: manifold learning; principal manifold; locally linear embedding; isometric mapping; variational approach; mutual information

0 引言

机器学习过程中, 往往面临庞大的数据量, 要在保证数据信息足够完整的条件下合理地约简数据集, 是对其提出的严峻挑战。以往的系统大多使用线性方法 (如维数约简中的线性主成分分析 (PCA) 等), 通过特征的线性组合来降维, 本质上是把数据投影到线性子空间, 这种方法相对简单且容易计算。但由于现实中有用特征往往不是特征的线性组合, 线性方法不能有效地处理庞大的高维数据。经试验发现, 许多高维采样数据均由少数隐含变量所决定, 这些隐含变量是以嵌套在高维欧式空间的组合型非线性流形存在的。为此, 本文提出了一种新的机器学习方法——流形学习算法。

流形学习的数学定义: 在 R^D 空间 ($N \gg D$) 中存在由某个随机过程生成的数据 $\{y_i\} \subset Y$, 经过某个函数 f 可以映射形成 R^N 空间中的观测数据 $\{x_i = f(y_i)\}$ 。流形学习的目标是要在观测数据 $\{x_i\}$ 中重构 f 和数据 $\{y_i\}$, 以达到数据压缩和降维的目的。

的。与以往机器学习的不同之处是, 流形学习强调整体结构, 要通过局部与整体相结合来发现和重建数据的内在规律。

目前, 在流形学习上存在大量的研究方法。按照对观测空间数据几何结构的分析将其分成五个主要的研究领域, 即神经网络、主流形、谱分析、变分法和互信息。本文从上述几方面对流形学习的一些方法进行了总结。

1 主要研究方法 with 成果

1.1 主流形

1.1.1 主流形框架

主流形思想假定数据存在于嵌套流形中, 试图寻找数据集的中间结构, 将其描述为流形拟合问题。具体形式化描述为: 在 n 维空间中, x 通过光滑嵌入 f 来近似逼近 x_1 。其中, x 为无界的光滑紧致流形; y 为包含在 x_1 的光滑管状邻域内的点集。所谓的流形拟合问题就是在最小二乘意义下来寻找 f , 使得

收稿日期: 2006-04-15; 修返日期: 2006-07-31

作者简介: 周红 (1982-), 女, 河北邯郸人, 硕士研究生, 主要研究方向为模式识别与智能系统方面研究 (lnzh2000@126.com); 吴炜, 男, 讲师, 主要研究方向为图像处理 and 图像通信; 滕奇志, 女, 教授, 博导, 主要研究方向为模式识别与智能系统; 杨晓敏 (1963-), 女, 博士研究生, 主要研究方向为图像处理和图像通信; 李旻, 男, 硕士研究生, 主要研究方向为模式识别与智能系统; 陶德元, 男, 教授, 博导, 主要研究方向为图像处理和图像通信。

$f(x)$ 最逼近 x_1 。嵌入的拟合度可以用两者距离的期望平方来度量。

1. 1. 2 主流形的发展过程

Hastie 和 Stuetzle 在主成分分析的基础上最早提出了主曲线(一维主流形)的概念^[1]。直观地讲,主曲线就是通过数据中间分布的光滑自相合曲线,是很好的一维非线性数据总和,形象化描述如图 1 所示。在 R^d 空间中, $f(t) = (f_1(t), \dots, f_d(t))$ 是光滑曲线,参数 $t \in R$, 且对任意 $x \in R^d$, $t_f(x)$ 表示当 x 和 $f(t)$ 直接距离最小时所取的最大参数值 t , 公式化为

$$t_f(x) = \max\{t: \|x - f(t)\| = \min_{t'} \|x - f(t')\|\}$$
 (1)

通过 HS 主曲线的定义,如果满足以下条件,那么光滑曲线 $f(t)$ 就是主曲线。条件如下: f 自身不相交;在 R^d 的任意限制子集中 f 都有有限长度,且自相合,如

$$f(t) = E(X | t_f(X) = t)$$
 (2)

自相合就是对于 f 的每个点(在 X 分布下)都是所有点的集合。这样, Hastie 强调主曲线的非参数方法,即曲线类型未知,在曲线簇中选择满足自相合的有中间性的曲线。但实验证明这种算法存在模型偏差和估计偏差,并且在实际应用过程中经常采用局部光滑子或者样条函数来近似寻优,这偏离了原有的自相合性质。

之后的 BR 主曲线^[2]改进了 HS 主曲线的算法,减小了估计偏差对实际曲线的影响,但其产生的数据不稳定,有可能得到光滑但不正确的主曲线。T 主曲线^[3]引入了半参数法,算法过程中利用 EM 算法来估计主曲线,且假定噪声是正态分布的,这与 HS 主曲线的无参数原则相背离。K 主曲线^[4]证明了在理论分布下定义的 K 主曲线的存在性和唯一性,利用统计学方法用多边形线算法来估计 K 主曲线。这是第一次可以证明主曲线存在,但是限制条件是曲线的长度必须预先固定。Smola 提出了试图向高维推广的基于统计学习理论的正则主流形,采用监督学习中的量化误差最小理论下的正则化来寻找具有多种正则项的正则主流形^[5]。Delicado 给出了主定向点概念和基于定向点的 D 主曲线^[6]理论,这保持了自相合特点,并且利用参数模型可以向高维推广。Verbeek 提出了 K 主曲线^[7]算法,即用局部主成分算法来构造 K 段线段,再连接成光滑的主曲线,不足之处是无法向高维推广。

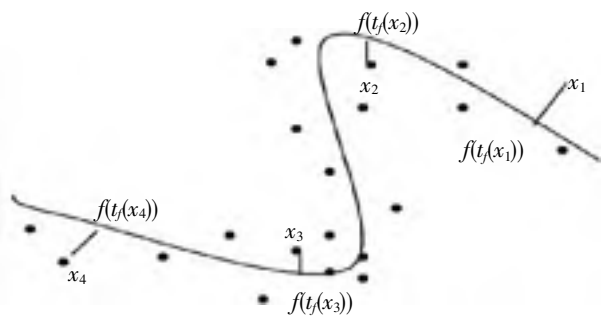


图 1 一组数据集的主曲线

1. 1. 3 主流形的应用

主曲线最初应用于斯坦福线性加速器上,其目标函数可由磁铁之间的光滑度来控制。之后,在其他方面应用也很广泛,如生态学、语音识别、智能交通^[8]等。

1. 2 谱分析

谱分析是一种经典的数学分析方法。与经典的谱分析不

同的是,流形学习中的谱分析是利用局部结构来描述整体的,不具备全局线性结构。其目的是寻找最优基函数的组合来构造数据集内在低维嵌套结构,实质上最终目标还是期望获得具有全局线性坐标的数据结构。采用谱分析来完成流形学习的方法包括等度规映射(ISOMAP)、局部线性嵌套(LLE)、Laplacian 特征映射和核主成分分析(KPCA)。

1. 2. 1 等度规映射(Isometric Feature Mapping, ISOMAP)^[9]

多维尺度变换(MDS) 是一种非监督的维数约简方法。其基本思想是:约简后的低维空间中任意两点间的距离应该与它们在原始空间中的距离相同。等距映射(ISOMAP) 算法的基本思想是在多维尺度变换的基础上,力求保持数据点的内在几何性质,即保持两点间的测地距离,如图 2 所示。ISOMAP 算法的核心就是要估计两点间的测地距离:保证离得很近的点间的测地距离用欧氏距离代替;离得较远的点间的测地距离用最短路径来逼近。

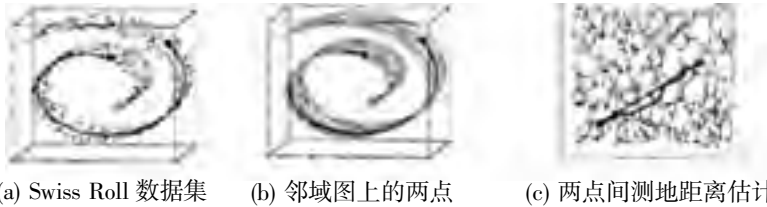


图 2 Swill Roll 数据集的测地距离估计

具体算法:假设已知观测数据为 $(x_i, x_j), i = 1, \dots, N$ 。其中 $x_i \in R^q$, 计算每个点的近邻点(用 K 近邻或 ϵ 邻域)在样本集上定义一个赋权无向图。如果 x_i 和 x_j 互为近邻点,则边的权值为 $dx(i, j)$ 。计算图中两点间的最短距离,记所得的距离矩阵为 $Dx = dx(i, j)$ 。用 MDS 求低维嵌入流形,代价函数为

$$Dy = dy(i, j) = \|y_i - y_j\|$$
 (3)

令 $S = (S_{ij}) = (D_{ij}^2)$, $H = (H_{ij}) = (\delta_{ij} - 1/N)$, $\tau(D) = -HSH/2$, δ_{ij} 是克罗内克符号, N 为样本点 x 的个数, τ 参数可以将距离转换为 MDS 所需的内积, H 是集中矩阵。低维嵌入 $\tau(D)$ 是的第 $2 \sim d + 1$ 个最大的特征值所对应的特征向量。ISOMAP 在确定人脸和手的图像的内在维数上有一定的成果。此算法的缺陷是,一旦图像数据包含过大噪声, ISOMAP 算法很难恢复内在结构。

1. 2. 2 监督等度规映射(Supervised ISOMAP, S-ISOMAP)

ISOMAP 在大噪声情况下是不适用的,因此在其基础上提出了具有鲁棒性的 S-ISOMAP 算法。它是一种监督的维数约简方法,对输入的数据进行分类标记来完成流形学习。改进 ISOMAP 算法中的代价函数为

$$D(x_i, x_j) = \begin{cases} \sqrt{1 - e^{-d^2(x_i, x_j)/\beta}} & y_i = y_j \\ \sqrt{e^{-d^2(x_i, x_j)/\beta} - \alpha} & y_i \neq y_j \end{cases}$$
 (4)

也可以改进为
$$D(x_i, x_j) = \begin{cases} \sqrt{1 - e^{-d^2(x_i, x_j)/\beta}} & y_i = y_j \\ \sqrt{e^{-d^2(x_i, x_j)/\beta}} & y_i \neq y_j \end{cases}$$
 (5)

其中, y_i 表示 x_i 所属分类; $d(x_i, x_j)$ 表示 x_i 和 x_j 之间的欧式距离, $D(x_i, x_j)$ 如图 3 所示; $d(x_i, x_j)$ 在指数位置上; 参数 β 在 $d(x_i, x_j)$ 相当大时用于防止 α 增长过快。 β 的值取决于数据点的密度,通常被设为数据点之间的平均欧式距离。参数 α 使得不同类的数据点之间更接近。从图 3 与 4 的对比中可以看出,用于保证类间距离比类内距离大。

S-ISOMAP 算法的优点: 当欧式距离 $d(x_i, x_j)$ 相同时, 类间距离比类内距离大, 适用于分类。类间距离大于或者等于 $1 - \alpha$ 类内距离小于 1。无论噪声影响多大, 在一定范围内, 类间和类内的距离均可以得到控制, 这使得 $D(x_i, x_j)$ 适用于含噪声的数据。

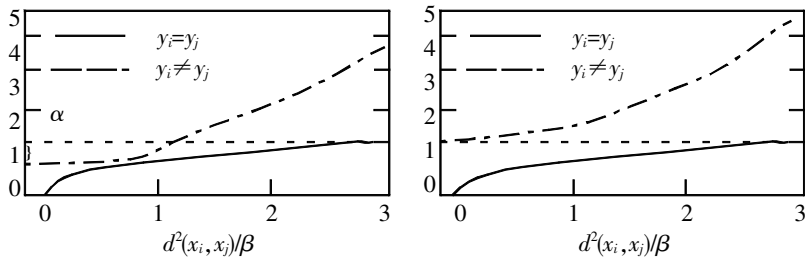


图 3 $D(x_i, x_j)$ 曲线图

图 4 $D'(x_i, x_j)$ 曲线图

1.2.3 局部线性嵌套(LLE)^[10]

LLE 算法的主要原理就是保持在嵌入空间和本质空间中的数据的数据的局部顺序关系。观察空间中的每个样本点被假定是由其邻域内的点加权平均组成, 所有点的加权值可以在高维空间中形成一个权矩阵。基于权矩阵, 通过一个闭式可以求出数据点集在低维结构上的有效描述。在高维空间 R^L 中给定数据集 $(x_1, \dots, x_k)$, 目的是要找到低维数据集 $(y_1, \dots, y_m)$, 它存在于 R^m 空间, 且 $m \ll k$ 。

具体算法步骤如下:

(1) 寻找邻域信息。对于每个 x_i , 在数据集中找到它的几个最近邻 $x_{i1}, \dots, x_{in}$ 。 $x_{i1}, \dots, x_{in}$ 可以是以 ε 为距离包围 x_i 的球体内的数据点。构造近似值矩阵: 选取 W_{ij} , 使 $D = \sum_{i=1}^L P x_i - \sum_{j=1}^n W_{ij} x_j P^2$ 最小。其中 $\sum_j W_{ij} = 1 (i = 1 \dots R)$ 。

(2) 再次利用最小二乘思想, 将步骤(1)中得到的满足条件的权值代入 D 中, 得到解:

$$Y^* = \sum_{i=1}^L P y_i - \sum_{j=1}^n W_{ij} y_j P^2 = \sum_{i=1}^L \arg \min_Y P (I - W) Y P^2 = \sum_{i=1}^L \arg \min_Y Y^T (I - W)^T (I - W) Y$$

(6)

则可求得最优解为矩阵 $(I - W)^T (I - W)$ 的最小特征向量。为了防止数据集坍塌至一个点, 所求特征向量中要去掉特征值 0 所对应的特征向量。用在 S 形曲线上的效果如图 5 所示。



图 5 S 形曲线的效果图

1.2.4 Laplacian 特征映射(Laplacian Eigenmap)^[11,12]

Laplacian 特性映射的基本思想是在高维空间中离得很近的点投影到低维空间中的像也应该离得很近。通过使用两点间的加权距离作为损失函数, 可求得相应的降维结果。它与 LLE 算法的不同之处在于它采用了 Laplacian-Beltrami(流形切空间上梯度向量的负散度函数)算子。

具体算法: 前面步骤与 LLE 算法一致。在计算嵌套时, 要通过获取相应于矩阵的 k 个最小特征值的特征向量来计算嵌套 $E = (I - W)^T (I - W)$ 。注意: E 是对称的半正定矩阵。 E 作

用于由数据点集确定的函数上。在一定条件下可以得出 $Ef \approx 1/2 L^2 f$ 。推断步骤如下:

$$[(I - W) f]_i \approx -1/2 \sum_j W_{ij} (x_i - x_{ij})^T H (x_i - x_{ij})$$

(7)

其中, f 是作用于流形的函数, 由数据点 x_i 确定; H 是 f 在 x_i 点的赫森矩阵。为了简化分析, 假设邻域点 x_{ij} 位于 x_i 周围的局部线性流形空隙中。考虑以 $o = x_i$ 为中心的切平面坐标系, 令 $v_j = x_{ij} - x_i$, 因为两点之间的差距可以由从第二个点出发的一个矢量来表示, 就可以知道切平面中这些 v_j 矢量都是从 O 点起始的矢量。令 $\alpha_j = W_{ij}$, 因为 x_i 属于它邻域的仿射度, 要重构矩阵 W , 则有 $o = x_i = \sum \alpha_j v_j$ 。其中 $\sum \alpha_j = 1$, 若 f 为光滑函数 $\alpha_j = 1$ 。它的泰勒近似表达式为

$$f(v) = f(o) + v^T \nabla f + 1/2 (v^T H v) + O(P v P^2)$$

(8)

其中, H 是赫森矩阵; $H_{ij} = \partial^2 f / \partial x_i \partial x_j$ 。因此

$$f(o) - \sum_j \alpha_j f(v_j) \approx f(o) - \sum_j \alpha_j f(o) - \sum_j \alpha_j v_j^T \nabla f - 1/2 \sum_j \alpha_j v_j^T H v_j$$

(9)

对 $f(v_j)$ 用泰勒近似:

$$f(o) - \sum_j \alpha_j f(o) - \sum_j \alpha_j v_j^T \nabla f - 1/2 \sum_j \alpha_j v_j^T H v_j$$

(10)

若 $\sum_j \alpha_j = 1$ 且 $\sum_j \alpha_j v_j = 0$, 则 $f(o) - \sum_j \alpha_j f(v_j) \approx -1/2 \sum_j \alpha_j v_j^T H v_j$

(11)

若 $\sqrt{\alpha_j} v_j$ 构成一标准正交基, 则

$$\sum_j W_{ij} v_j^T H v_j = \text{tr}(H) = L f$$

(12)

最终可得 $(I - W)^T (I - W) f \approx 1/2 L^2 f$

(13)

拉普拉斯算法试图使 $f^T (I - W)^T (I - W) f$ 最小, 这样就可以使找 $(I - W)^T (I - W)$ 的特征向量函数转变为找 L^2 的特征函数, L^2 的特征函数与 L 一致。

1.2.5 核主成分分析(Kernel Principal Component Analysis, KPCA)^[13]

核主成分分析是利用积分算子核函数的特性来分析观测空间数据集的结构。核函数能使数据集在特征空间的度量性质在原空间得以实现, 这样, 利用核函数来构造数据集在特征空间的协方差矩阵, 并求取在映射空间的结构。这使得可能在数据集的观测空间维数较低时, 获得较高维淹没子流形的特征分析。

1.2.6 谱分析的应用

谱分析的主要作用是用于维数约简, 可以应用在数据可视化、分类等方面。在不同距离、不同方向或不同姿态和光照强度下, 同一个对象能够形成多种不同的图像。一个对象所有图像的集合可以看做是以位置、尺度、姿态、光照等为参数的一个高维空间流形。如果每一个像素均对应于空间中的一维, 那么一幅图像就可以看做高维图像抽象空间中的一个点, 一个对象在不同方向上所有图像的集合就是图像空间中的一个连续流形。利用流形学习算法, 可以给出以流形存在的图像的内在低维结构, 使得降维后的图像特征能够有效地得到识别^[14]。

1.3 变分法

从物理学的角度来看, 任何一种数据集的最终存在形式应该是一个稳态, 此稳态具有最小能量。Gomes 提出采用变分微积分思想来求解偏微分方程^[15]。在算法上通过能量最小化迭代使数据点的邻域关系趋近真实, 从而获得对孤立邻域点集

的全局连通, 最终获得流形结构的有效描述。产生最小能量的目标函数包括四个方面, 即包含进了不属于流形的结构域; 没有包含全部流形; 引入的光滑性和连续性的约束; 要求结构为凸性的约束。

1.4 互信息

互信息^[16]就是在最近邻区域和流形图上, 假设交互信息是对被观察空间与被嵌入空间之间的不同概率分布的测量。从信息论的角度来看, 在观测空间数据集和嵌套空间数据集中近邻数据间的概率的信息熵应该最小。

随机邻域映射(SNE) 假设高维观测空间每个数据点具有高斯分布, 邻域意义下概率分布定义如下:

$$p_{ij} = \exp(-d_{ij}) / \sum_{k \neq i} \exp(-d_{ik})$$

其中, $d_{ij} = (Px_i - x_jP^2) / 2\sigma_i^2$, 方差和邻域值由经验或用户指定。

在低维嵌套空间, 假定内在邻域方差为固定值, 数据集在邻域意义下概率分布:

$$q_{ij} = \exp(-Py_i - y_jP^2) / \sum_{k \neq i} \exp(-Py_i - y_kP^2) \tag{14}$$

要保持邻近点在观测空间和嵌套空间概率分布相似, 损失函数用 KL 散度和来表征:

$$C = \sum_i \sum_j p_{ij} \lg(p_{ij} / q_{ij}) \tag{15}$$

为使其最小, 求微分得

$$C / y_i = 2 \sum_j (y_i - y_j) (p_{ij} - q_{ij} + p_{ji} - q_{ji}) \tag{16}$$

由此, 可以通过梯度方法来调整低维空间数据点集的相应位置。不足之处是采用的迭代法来使损失函数最小化容易陷入局部最小。

1.5 神经网络

神经生理学上发现, 在神经网络中通过邻近单元的相互竞争, 可以自适应地发展成为对不同性质信号敏感的区域。神经网络中的自组织映射^[17], 是假定输入数据可以通过神经网络映射到低维空间, 并保证在输入空间数据集的近邻关系在低维结构中可以保持。算法上假定每个输入空间的数据点在低维嵌套空间具有一定的几何位置, 通过竞争来更新数据点的位置, 然后逐渐收缩邻域, 最终获得与输入空间观测数据集最近的拓扑结构。但是, 这种方法由于邻域不断收缩, 可能造成不光滑连续, 且难以在高维空间进行扩展。

2 流形学习中有待解决的问题

尽管流形学习的算法和应用在过去的几年中已经取得了丰硕的成果, 但是由于其数学理论基础较为复杂, 以及多个学科之间交叉、融合, 对高维数据中有意义的低维结构的研究中依然有很多值得进一步探讨的问题。

流形学习本身是一种全局分析方法, 它跟局部分析方法应该是互补关系。如何有效地将两者结合起来是一个需要研究的问题。

在几何上分析高度相关、复杂分布的数据集的内在结构。流形是一种几何研究工具, 它本身作为一种发现内在规律的方法, 可以作为很多相关后续处理过程的一个基础。就目前来

看, 许多实际应用中的数据集都存在规律性, 因此研究流形与数据集的关系有可能成为一个重要领域。

流形的主要特征的个数与错识率有很大的关系。在将来的工作中, 必须找到一个行之有效的方法来估计主要特征的数目。怎样使用流形的性质来模拟未知的概念和设计新的算法是今后研究的方向。

参考文献:

[1] ASTIE T, STUETZLE W. Principal curves [J]. Journal of the American Statistical Association, 1988, 84(406) : 502-516.

[2] BANFIELD J D, RAFTERY A E. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves [J]. Journal of the American Statistical Association, 1992, 87(417) : 7-16.

[3] TIBSHIRANI R. Principal curves revisited[J]. Statistics and Computation, 1992, 2: 183-190.

[4] K GL B, KRZYZAK A, *et al.* Learning and design of principal curves [J]. Trans. on Pattern Analysis and Machine Intelligence, 2000, 22(3) : 281-297.

[5] SMOLA A J, MIKA S, SCHOLKOPE B, *et al.* Regularized principal manifolds: proc. of the 4th European Conference on Couputation Learning Theory[C]. New York: Springer, 1999: 251-256.

[6] DELICADO P. Another look at principal curves and surfaces [J]. Journal of Multivariate Analysis, 2001, 77(1) : 84-116.

[7] VERBEEK J J, VLASSIS N, KROSE B. A k-segments algorithm for finding principal curves[J]. Pattern Recognition Letters, 2000, 23(8) : 1009-1017.

[8] 张军平. 博士论文[D]. [出版地不详]: 中国科学院自动化研究所, 2003.

[9] TENENBAUM J B, VIN de SILVA, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500) : 2319-2323.

[10] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290 (5500) : 2324-2326.

[11] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data reduction and data representation[J]. Neural Computation, 2001, 151(6) : 1373-1396.

[12] MIKHAIL B, PARTHA N. Using manifold structure for partially labelled classification: proc. of Neural Information Systems: Natural and Synthetic [C]. Vancouver, Canada: MII Press, 2002: 9-14.

[13] SCH LKOPF B, SMOLA A, MULLER R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998, 10(5) : 1229-1319.

[14] 赵连伟, 罗四维, 赵艳敞, 等. 高维数据流形的低维嵌入及嵌入维数研究[J]. 软件学报, 2005, 16(8) : 1423-1430.

[15] GOMES J, MOJSILOVIC A. A variational approach to recovering a manifold from sample points: proc. of the 7th European Conference on Computer Vision-part [C]. London: Springer-Verlag, 2002: 3-17.

[16] HINTON G, ROWEIS S. Stochastic neighbor embedding: proc. of neural Information Systems: Natural and Synthetic[C]. Cambridge: MIT press, 2003: 833-840.

[17] KOHONEN T. Self-organizing maps[M]. New York: Springer, 1997.