高维数据流形的低维嵌入及嵌入维数研究*

赵连伟[1+], 罗四维[1], 赵艳敞[2], 刘蕴辉[1]

[1](北京交通大学 计算机与信息技术学院,北京 100044)

[2](Faculty of Information Technology, University of Technology, Sydney, Australia)

Study on the Low-Dimensional Embedding and the Embedding Dimensionality of Manifold of High-Dimensional Data

ZHAO Lian-Wei[1+], LUO Si-Wei[1], ZHAO Yan-Chang[2], LIU Yun-Hui[1]

[1](School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

[2](Faculty of Information Technology, University of Technology, Sydney, Australia)

+ Corresponding author: Phn: +86-10-51688556, E-mail: lw_zhao@hotmail.com, http://www.bjtu.edu.cn

Received 2004-07-14; Accepted 2004-09-08

Zhao LW, Luo SW, Zhao YC, Liu YH. Study on the low-dimensional embedding and the embedding dimensionality of manifold of high-dimensional data. *Journal of Software*, 2005,16(8):1423−1430. DOI: 10.1360/jos161423

Abstract: Finding meaningful low-dimensional embedded in a high-dimensional space is a classical problem. Isomap is a nonlinear dimensionality reduction method proposed and based on the theory of manifold. It not only can reveal the meaningful low-dimensional structure hidden in the high-dimensional observation data, but can recover the underlying parameter of data lying on a low-dimensional submanifold. Based on the hypothesis that there is an isometric mapping between the data space and the parameter space, Isomap works, but this hypothesis has not been proved. In this paper, the existence of isometric mapping between the manifold in the high-dimensional data space and the parameter space is proved. By distinguishing the intrinsic dimensionality of high-dimensional data space from the manifold dimensionality, and it is proved that the intrinsic dimensionality is the upper bound of the manifold dimensionality in the high-dimensional space in which there is a toroidal manifold. Finally an algorithm is proposed to find the underlying toroidal manifold and judge whether there exists one. The results of experiments on the multi-pose three-dimensional object show that the method is effective.

Key words: Isomap; toroidal manifold; isometric mapping; embedding dimensionality

摘 要: 发现高维数据空间流形中有意义的低维嵌入是一个经典难题.Isomap 是提出的一种有效的基于流形理论的非线性降维方法,它不仅能够揭示高维观察数据的内在结构,还能够发现潜在的低维参数空间.Isomap 的理论基础是假设在高维数据空间和低维参数空间存在等距映射,但并没有进行证明.首先给出了高维数据的连续流形和低维

* Supported by the National Natural Science Foundation of China under Grant No.60373029 (国家自然科学基金)

作者简介: 赵连伟(1976－),男,河南驻马店人,博士生,讲师,主要研究领域为人工神经网络,流形学习;罗四维(1943－)男,博士,教授,博士生导师,主要研究领域为人工神经网络,模式识别,并行计算;赵艳敞(1977－)男,博士,主要研究领域为模式识别,数据挖掘;刘蕴辉(1976－)女,博士生,主要研究领域为人工神经网络,信息几何.

参数空间之间的等距映射存在性证明,然后区分了嵌入空间维数、高维数据空间的固有维数和流形维数,并证明存在环状流形高维数据空间的参数空间维数小于嵌入空间维数.最后提出一种环状流形的发现算法,判断高维数据空间是否存在环状流形,进而估计其固有维数及潜在空间维数.在多姿态三维对象的实验中证明了算法的有效性,并得到正确的低维参数空间.

关键词: Isomap;环状流形;等距映射;嵌入维数
中图法分类号: TP391 文献标识码: A

在不同距离、不同方向,或在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作是以位置、尺度、姿态、光照等为参数的一个高维空间流形.人类能够感知由同一个对象产生的变化着的信号,并能够正确地识别.为了更精确地刻画图像和其他感知刺激的变化,采取数学方法是非常必要的.如果每一个像素都对应于空间中的一维,那么一幅图像就可以看作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.文献[1]认为,流形是感知的基础,经过自然界长期进化的人脑能够用流形的方法表示对外界对象的感知.大量神经元对信息的编码方法成为我们对人脑表示方法研究的基础,如果一个神经元的触发率对应于一维,那么图像信息就能够由与像素个数相等的神经元来表示.神经生理学家已经发现,群体中神经元的点火率都能够表示为几个变量的连续函数,比如人眼转动的角度和头旋转的方向,这说明群体活动被限定在低维空间光滑流形上,所以在理解人脑如何从神经动力学中产生感知时,流形的低维嵌入起到非常重要的作用.

很多科学家都在寻求发现嵌入在高维数据中有意义低维结构的方法,对流形学习算法的研究引起了广泛的兴趣.对于由一个对象在不同参数(如不同光照和姿态)下的数字图像组成的流形 M ,其参数的个数未知,相应的参数值也未知.但是对于图像理解和图像编码这样的问题,学习图像流形的结构和发现潜在的参数又是非常有用的,比如人脸识别中不同表情的人脸和目标检测中目标的姿态等.利用分散样本进行流形学习一直是一个令人关注的难题,现在也已经有了一些高维数据低维表示方法,比如主成分分析(PCA)、独立分量分析(ICA)、Fisher 判别分析(FDA)、多维尺度分析(MDS)等.这些大都是线性的方法,所以对于那些非线性结构的数据就无能为力,而非线性降维技术则能产生较好的结果.LLE[2]和 Isomap[3]是两种有代表性的非线性降维方法.Roweis 和 Saul 提出的 LLE 算法能够实现高维输入数据点映射到一个全局低维坐标系,同时保留了邻接点之间的关系,这样,固有的几何结构就能够得到保留.此算法不仅能够有效地发现数据的非线性结构,同时还具有平移、旋转等不变特性.Tenenbaum 等人提出的 Isomap 算法首先使用最近邻图中的最短路径得到近似的测地线距离,代替不能表示内在流形结构的 Euclidean 距离,然后输入到多维尺度分析(MDS)中处理,进而发现嵌入在高维空间的低维坐标.在人脸和手势的实验中,Isomap 发现了存在于高维空间中的潜在低维参数空间.Donoho 等人[4]用人工合成(实验者可以事先知道其潜在的参数,比如平移、旋转等)的数据用 Isomap 算法进行测试实验,实验结果表明,Isomap 能够准确地发现图像流形潜在的参数空间,并在自然图像(人脸图像)中不同姿态和亮度等潜在的未知参数下也可得到较好的结果.Donoho 等人还拓展了 LLE 算法,提出 HLLE 算法[5],能够发现流形上局部的潜在等距映射参数.张长水等人[6]在 LLE 的基础上提出一种从低维嵌入空间向高维空间映射的方法,并在多姿态人脸图像的重构实验中得到有效的验证,进一步完善了非线性降维方法.虽然这些算法都要求知道嵌入空间的维数,但很少有文献对它进行分析和估计.文献[6]使用的是文献[3]中的方法,而在文献[3]中只是通过剩余方差与维数的关系来估计 d 值的范围.Marzia Polito 和 Pietro Perona[7]提出了应该首先知道嵌入空间维数,但没有给出一个有效的方法.

本文首先介绍了 Isomap 算法,并通过两个典型流形的实验结果,提出嵌入空间的维数问题.第 2 节给出了连续流形与其低维参数空间等距映射的存在性证明,完善了 Isomap 的理论基础,并指出在圆筒形曲面实验中之所以没有能够发现潜在的结构,是因为没有能够正确估计嵌入空间的维数.然后区分了嵌入空间维数、高维数据的固有维数与流形维数,并且证明如果数据空间存在环状流形,则流形维数要小于数据的固有维数,从而说明了并非任何情况下二维流形都能够嵌入在二维空间.第 3 节给出一种环状流形发现算法.根据此算法,能够判断数据空间是否存在环状流形.第 4 节在多姿态三维对象的实验中证明了算法的有效性,并得到正确的低维参数空

间.最后总结全文.

1 Isomap——非线性降维算法

1.1 Isomap的主要思想及算法步骤

Tenenbaum 等人提出的 Isomap 算法[3]的主要思想就是首先计算流形上的测地线距离,然后应用 MDS 算法,发现嵌入在高维空间的低维坐标,这样 Isomap 就通过数据间的测地线距离,保留了数据固有的几何分布结构.下面给出标准 Isomap 算法,共 3 步:

Step 1. 构建输入空间 X 中流形 M 上所有数据点 $x_i,i=1,2,\ldots,N,X_i\in R^D$ 的邻接图,距离定义为 Euclidean 距离 $d_x(i,j)$,邻接关系定义为 ε 球或 K 最近邻.

Step 2. 通过计算图 G 上两点间的最短路径 $d_G(i,j)$ 估计流形 M 上测地线距离 $d_M(i,j)$,得到的矩阵 $D_G=\{d_G(i,j)\}$ 为图 G 上任意两点间的最短路径距离.

Step 3. 应用 MDS 算法,构建 d 维 Euclidean 空间 Y 上的嵌入.详见文献[3].

Isomap 的有效性在人工合成数据和自然图像的实验中已经得到验证.

1.2 使用Isomap降维实验

在文献[3]中,使用 Swiss roll 数据集说明 Isomap 近似计算测地线距离以及降维的过程,并得到较好的结果.这里使用 Cylinder 数据集进行实验,如图 1 所示,随机选择 1000 个数据点,使用 Isomap 算法降维,其中每一数据点的最近邻连接数 $k=7$,投影到二维空间.图 2 为得到的维数和剩余方差的关系,可以看出,在维数大于 2 时,随着维数的增加,剩余方差并没有减少.

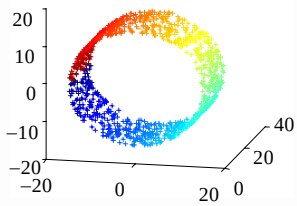

Fig.1 The Cylinder manifolds

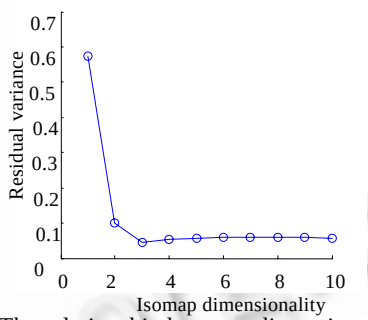

Fig.2 The relationship between dimensionality and residual variance applying Isomap

图 1 Cylinder 流形 图 2 对 Cylinder 数据集应用 Isomap 得到维数和剩余方差的关系

图 3 给出 Isomap 对二维投影结果,可以看出 Cylinder 的投影图上只保留了圆面上的距离,高度上的距离丢失,而不同于 Swiss roll 的投影图很好地保留了邻接图中的最短路径距离,这表明 Isomap 很难对 Cylinder 进行降维.这就产生一个问题:是 Isomap 降维不适用于所有的光滑流形,如像 Cylinder 之类的流形,还是另有其他原因?

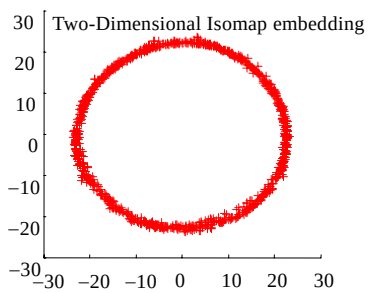

Fig.3 Two-Dimensional projections from the cylinder by Isomap

图 3 应用 Isomap 得到 Cylinder 数据集的二维映射图

2 光滑流形与低维参数空间等距映射存在性

Isomap 算法基础就是假设光滑流形 M 和参数空间(R^d 的子集)之间存在等距映射,使我们能够找到内在的映射参数.如果有满足一定条件的等距映射存在,Isomap 就适用.文献[4]给出了存在等距映射的条件(定理 1),并证明了一些特定条件下等距映射存在,但是对于一般光滑流形与低维参数空间之间等距映射的存在性没有给予证明.

定理 **1**. 假设参数图像族 $f(\theta):R^2\to R$,$f(\theta)$属于 L^2,且在 L^2 上可微,其中$\theta\in\Theta$,Θ为参数空间.如果总存在一个 $c>0$,使得 $f(\theta_0)$ 和 $f(\theta_1)$ 间的测地线距离可以由下式表示:

$$G(\theta_0,\theta_1)=c\left\|\theta_1-\theta_0\right\|_{L^2},$$

则(Θ,G)和$(\Theta,\|\cdot\|)$之间存在等距映射,且 Isomap 成立,并能发现潜在的参数空间.

这里, $f(\theta)$ 可以看作是光滑流形 M 上的点,所以此定理对于一般光滑流形情况同样适用.这样,只要能够证明一般光滑流形与其低维参数空间存在等距映射,就可以得到 Isomap 对一般光滑流形适用的结论.

光滑流形 M 上两点 y_0, y_1 之间的测地线距离为连接两点最短的曲线长度,表示为

$$d(y_0,y_1;M)=\inf\{l(\gamma):\gamma(0)=y_0,\gamma(1)=y_1\} \tag{1}$$

命题. 对于任意光滑流形 $M\subseteq R^d$,$\Theta\subseteq R^m$ 为其低维参数空间,则(Θ,G)和$(\Theta,\|\cdot\|)$之间存在等距映射.

在证明之前,我们首先给出共形映射以及等距映射的定义.

定义 **1**. $\varphi:\Theta\to M$ 称为共形映射,如果 φ 是双可微映射,如果对于任意的 $\theta\in\Theta$,具有保角性和伸缩不变性,即对于Θ上任意的切向量 v 和 w,都有

$$(d\varphi_\theta v)^T(d\varphi_\theta w)=\left|\varphi'(\theta)\right|v^Tw,$$

其中, $\left|\varphi'(\theta)\right|>0$ 称为伸缩率.如果对所有的$\theta\in\Theta$,都有 $\left|\varphi'(\theta)\right|=1$,则 φ 称为等距映射.

证明:对高维流形 $M\subseteq R^d$,令$\Theta\subseteq R^m$为其参数空间,则存在映射 $\varphi:\Theta\to M$,即 $M=\varphi(\Theta)$.

流形 M 上测地线距离可以表示为

$$l(\gamma)=\int_0^1\left\|\gamma'(t)\right\|_{L^2}\mathrm{d}t,$$

其中, $\gamma:[0,1]\to M$.令 $\Gamma:[0,1]\to\Theta$ 是 R^m 上光滑曲线,则任意光滑曲线 $\gamma:[0,1]\to M$ 能够表示为 $\gamma(t)=\varphi(\Gamma(t))$,那么,曲线的长度

$$l(\gamma)=\int_0^1\left\|\varphi'(\Gamma(t))\right\|\mathrm{d}t=\int_0^1\left\|J_\varphi(\Gamma(t))\Gamma'(t)\right\|\mathrm{d}t \tag{2}$$

考虑非线性共形映射φ,由共形映射的定义可知,在曲面上的切线向量之间的夹角和参数空间中相应的向量之间的夹角相等,所以无论空间Θ经映射φ在 M 上如何变形,M 上的测地线距离和Θ上的 Euclidean 距离都保持一定的关系.

又因为 $\varphi:\Theta\to M$ 为两个流形上的共形映射,Γ为Θ上的曲线,则对任意的点 $x\in\Theta$,切向量为 v,切映射为 $d\varphi_x$,那么在 M 上点 $\varphi(x)$ 的切向量为 $d\varphi_x v$.如果 v 是Θ上Γ的方向, $d\varphi_x v$ 就是 M 上曲线 $\varphi(x)$ 的方向.因为 $\varphi:\Theta\to M$ 为共形映射,所以有

$$J_\varphi(\theta)^TJ_\varphi(\theta)=\varphi'(\theta)I_m,$$

这里, I_m 是一个 m 阶单位矩阵.代入式(2),曲线长度可以表示为

$$l(\gamma)=\int_0^1\sqrt{\varphi'(\Gamma(t))}\left\|\Gamma'(t)\right\|\mathrm{d}t \tag{3}$$

因为在 R^m 中任意两点之间的最短路径等于连接它们的直线长度,若Θ为开的凸集,则在光滑曲线上有 $\Gamma(t)=\theta_0+t(\theta_1-\theta_0)$,其中 θ_0 为起点, θ_1 为终点, $t\in[0,1]$.代入式(3)有

$$l(\gamma)=\int_0^1\sqrt{\varphi'(\Gamma(t))}\left\|\theta_1-\theta_0\right\|\mathrm{d}t=\int_0^1\sqrt{\varphi'(\Gamma(t))}\,\mathrm{d}t\left\|\theta_1-\theta_0\right\| \tag{4}$$

如果对于任意的 $\theta\in\Theta$,都有 $\varphi'(\Gamma(t))=c$ 为常数,那么点 y_0, $y_1\in M$ 之间的测地线距离为

$$d(y_0,y_1;M)=\sqrt{c}\left|\theta_1-\theta_0\right|.$$

由定理 1 可知, (Θ,G) 和$(\Theta,\|\cdot\|)$之间存在等距映射.所以对于任意光滑流形 M 可以通过计算 M 上点之间的

测地线距离,计算Θ上点之间的 Euclidean 距离. □

从上面的讨论可以看出,测地线距离对于研究高维空间中的流形是非常重要的.(Θ,G)和(Θ,$\|\cdot\|$)之间如果存在一个等距映射,那么就可以从 M 中获得其潜在的参数空间Θ和参数值 θ,并重新描述参数空间.然而,计算测地线要经由 φ 及其 Jacobian 矩阵,但一般情况下 φ 很难求出,这里我们只是证明了其存在性.

在证明中,需要假设Θ是一个开的凸集,原因在于,如果流形上有一个洞,测地线曲线需要绕这个洞,即使有

$$J_\varphi(\theta)^T J_\varphi(\theta)=c(\theta)I_m,$$

$$d(y_0,y_1;M)=\sqrt{c}\left|\theta_1-\theta_0\right|$$

也不一定成立.虽然在非凸的情况下,等距依然成立,但是成比例的性质不再成立.当 c 为任意函数时,就可以进行任意的拓扑映射.只是恒等于 1 时,φ 为等距映射,要求更为严格.这也说明了为什么虽然 Cylinder 数据集在拓扑上和二维是同胚的,但二者不存在等距映射,所以不能利用等距映射投影到二维空间.

测地线距离是流形的全局性质,而等距映射则是每个点附近的局部性质.测地线距离和参数空间中的 Euclidean 距离成比例是等距的结果,所以 Isomap 使用等距映射,得到高维流形的低维嵌入空间,一个前提条件就是要能够覆盖其全局性质,即要知道低维嵌入空间的维数.

在很多算法中都要求预设低维嵌入空间维数作为参数,对其参数分析和估计却很少涉及.能够决定嵌入空间维数的一个基本概念是高维数据集的固有维数,反映的是流形的固有性质,对固有维数的研究也有很多算法[8−11].如果一个数据集能够完全嵌入在一个 d 维子空间中,而不损失信息,则认为其嵌入维数等于 d.流形维数大多是指在损失较少信息的情况下其子流形的维数.准确地讲,固有维数是流形维数的上界,是嵌入空间维数的下界.

由此可见,Isomap 能够发现光滑流形上的潜在参数空间,但在 Cylinder 数据集的实验中,失败的原因在于混淆了流形维数和嵌入空间维数,使用了未能覆盖其全局性质的流形维数.如何利用拓扑方法研究固有维数,然后确定嵌入空间维数呢?

3 一种环状流形发现算法

拓扑方法是估计数据流形维数的常用的重要方法之一.一般情况下,R^m 中的正则曲线γ:$(a,b)\to R^m$ 是一维嵌入子流形,同样,R^m 中的正则曲面是 R^m 的二维嵌入子流形.而 Whitney 定理同时表明,任意高维 Euclidean 空间的嵌入子流形囊括了所有可能的 m 维光滑流形,所以嵌入子流形的状态是十分复杂的.

定义 **2**. n 维球面 $S^n=\{x:x\in R^{n+1},|x|=1\}$为 n 维光滑流形,一维单位球面 S^1 就称为一维光滑流形,r 维环面 T^r 定义为 r 个 S^1 的积流形 $T^r=S^1\times\ldots\times S^1$.

拓扑学已经证明了 S^m 不可能与 R^m 同胚,比如圆不可能与直线同胚,球面也不可能与平面区域同胚,存在环面的低维流形其固有维数大于拓扑流形维数的.Robert Pless 和 Ian Simon[12]对环状流形进行了研究,并针对球形、柱形、环形等流形,利用测地线距离,分别给出了拓展的 MDS 算法,嵌入到低维空间中,但其要求首先要知道流形的形状,但是否存在环状流形以及如何判断流形形状却没有提及.本节在流形定向理论的基础上提出一种环状流形发现算法,根据此算法能够判断高维数据空间中是否存在潜在的环状流形,并且可以根据流形上存在的环状,通过拓扑维数进一步估计其固有维数.

定义 **3**. 设 M 是 m 维的光滑流形,如果存在 M 的一个允许的坐标卡集 $A_0=\{(U_\alpha,\varphi_\alpha)\}$,使得 $\{U_\alpha\}$ 构成 M 的开覆盖,并且当 $U_\alpha\cap U_\beta\neq\varnothing$(不为空)时,坐标变换 $\varphi_\beta\circ\varphi_\alpha^{-1}:\varphi(U_\alpha\bigcap U_\beta)\to\varphi_\beta(U_\alpha\bigcap U_\beta)$ 的 Jacobi 行列式

$$\det\left(\frac{\partial(\varphi_\beta\circ\varphi_\alpha^{-1})^i}{\partial x_\alpha^j}\right)>0 \tag{2}$$

则称 M 是可定向的 m 维光滑流形.

若 M 是一个可定向的连通光滑流形,在任意一点 $p\in M$ 的切空间 T_pM 上指定一个定向,则通过该定向沿着从点 p 出发的任意一条路径的传播在每一点 $q\in M$ 的切空间 T_qM 上唯一地确定了一个定向.对于 M 中任意一条闭路径 $\gamma:[0,1]\to M$,即 $\gamma(0)=\gamma(1)$,使得在 $T_{\gamma(0)}M$ 中的一个定向 λ 沿着 γ 的传播在 $T_{\gamma(1)}M=T_{\gamma(0)}M$ 上都能够获

得相同的方向,这样就能够发现流形上的环状.对于空间曲线(面),取得标架场 $\{p;e\}$,使得 e 为曲线(面)的切向量.很明显,这个标架场给出了切空间的定向沿着曲线 EF 的连续延拓.如果点 p 沿着该曲线从 E 到 F,再回到 E 时,e 和原来的方向是一致的,则存在环状流形.

下面给出一种环状流形发现算法的步骤:

设输入空间 X 中流形 M 上所有数据点为 $x_i, i=1,2,...,N, X_i \in R^D$,

Step 1. 构建邻接图.方法同 Isomap 算法 Step 1.对于所有数据点构成的图 G,找到每个点的邻接点.

Step 2. 选择 $p = X_i$ 为起始点.

Step 3. 选取 $X_i k$ 个近邻中的一个 X_j,计算空间标架场 $\{p;e\}$,计算为流形方向 $e=X_j-X_i$,并令 $q=X_j$.

Step 4. 发现 X_j 的 k 个近邻中与切空间方向相同的方向 X_k,并令 $q = X_k$;如果不存在,则执行 Step 3,选择下一个近邻.

Step 5. 如果 q 能够沿着一定的路径回到起始的样本点 X_i,则存在从 X_i 开始的环状流形.

Step 6. 选择下一个 $p = X_i$ 为起始点,重复 Step 2~Step 5,直到选遍所有 n 个数据点为止.

用 n 表示样本点的个数,k 表示每个节点的近邻数.以其中一点为起始点进行一趟循环在最坏情况下的 $O(kn)$时间来完成,所以算法选遍 n 个样本点最坏情况下的时间复杂度为 $O(kn^2)$.另外,本文提出的环状流形发现算法能够发现高维数据空间中的低维环状流形,其理论基础是取得定向流形上某处的标架场 $\{p;e\}$,而标架场的取得并不受维数的限制.但是对于高维流形上的复杂数据来说,无论是从数值算法还是从实际应用来讲都有一定的困难,一是因为算法使用图的最短路径逼近测地线距离,需要大样本;二是因为随着维数的增加,对样本量的需求也呈指数增加.本文提出的算法主要适用于高维观察数据嵌入的低维子流形情况.

4 仿真实验

我们进行对象实验的对象数据集为 COIL-20 (Columbia object image library)数据库.数据库中共有 20 个对象,对每一个对象从 0°~360°进行水平方向的旋转,每隔 5°采样一幅图像,这样每一对象共有 72 幅图像.整个数据库共有 1440 幅图像,图像大小为 64×64,向量化图像以后,观察数据的维数 D=4096.在这样一个高维空间中,使用稀疏样本很难描述数据分布.在对象识别过程中,这种多姿态的对象识别还是非常困难的,特别是姿态估计.对象旋转时,图像的变化是光滑的,我们可以把它看作是连续的;又因为它是由一个自由度变化产生的,所以又是一维的.所以说这个流形可以看作是嵌入在高维图像空间中的一维光滑流形.图 4 给出一个对象的图像部分样本(每 30°取一个样本).

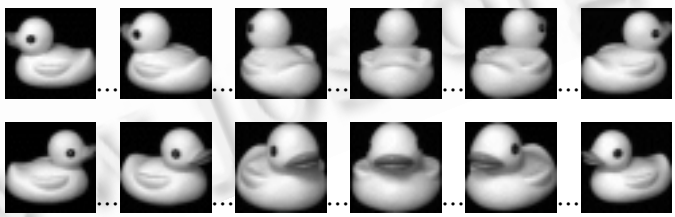

Fig.4 Example of multi-pose object images

图 4 多姿态对象图像示例

实验 1. 选定一个对象从 0°~180°共 36 幅图像样本,首先使用环状流形发现算法,结果没有能够发现环状流形,这时映射维数等于拓扑维数,所以能够投影到一维空间.使用 Isomap 在一维和二维空间的投影结果如图 5 所示.实验中我们发现,投影在一维空间和二维空间剩余方差的变化并不大,所以剩余方差和维数的关系不能作为估计嵌入空间维数的标准.又因为可以投影在一维空间,所以可以认为图像流形的变化由一个参数引起——旋转的角度,从图中也可以看出,从左到右,随着旋转角度的变大,在横轴的投影也越来越大.

实验 2.选定一个对象从 0°~360°全部的 72 幅图像样本,首先使用环状流形发现算法,结果发现存在一条环状路径,所以不能投影到一维空间.这时考虑投影到更高维的空间——二维.使用 Isomap 算法,投影结果如图 6

所示.图中发现旋转一周的图像流形投影在二维空间形成一个近似于圆的流形.

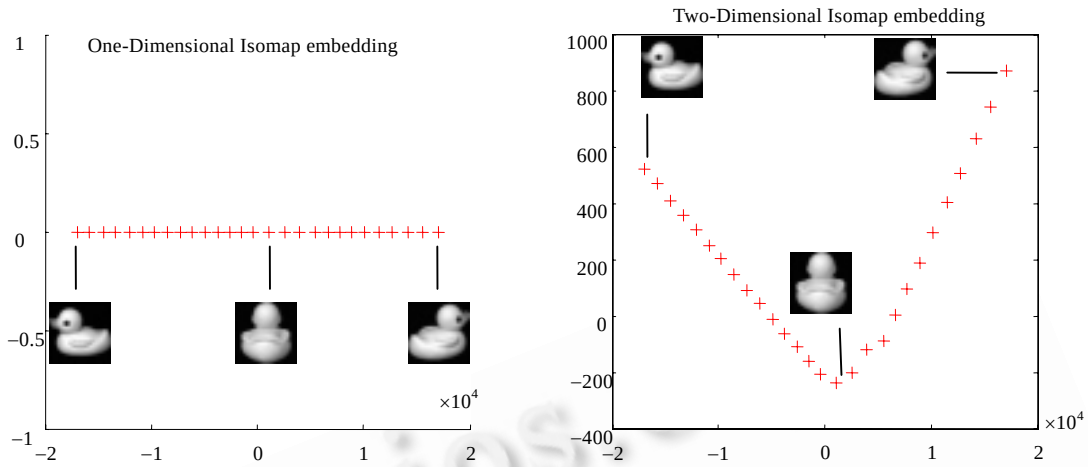

Fig.5 Output data points in one-dimensional (left) and two-dimensional (right) embedded space and the corresponding images respectively

图 5 投影到一维(左)和二维(右)空间数据点和相应的对象图像

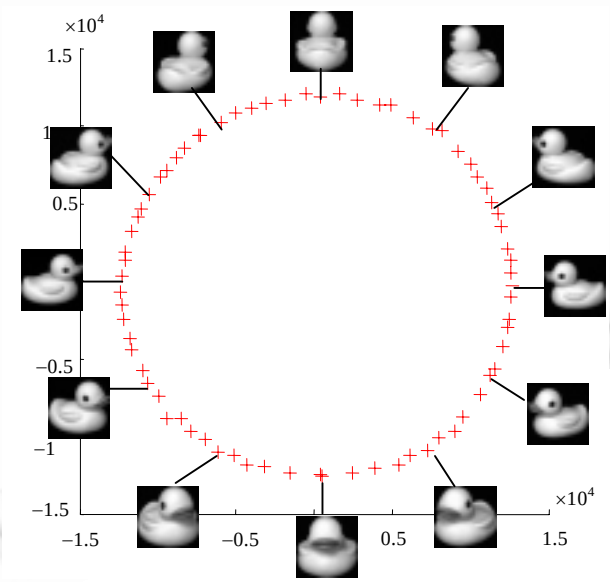

Fig.6 Output data points in two-dimensional embedded space and the corresponding images

图 6 投影到二维空间数据点和相应的对象图像

综合实验 1 的结果我们可以认为,图像流形变化是由一个参数变化引起的,而且完全可以通过流形学习的方法发现潜在的参数空间.但是即使同样是一维流形,同样的一维参数空间,却不能同样地投影到一维空间.

5 结 论

流形方法现已成为研究人类感知的一种重要方法,发现高维观察数据中有意义的低维嵌入空间是研究高维流形空间的有效途径.Isomap 是一种有效的非线性降维方法,在一些实验中也发现了潜在的低维参数空间.但是,其算法的前提是假设光滑流形 M 及其参数空间 R^d 的子集之间存在等距映射.本文从理论上对这种等距映

射的存在性进行了探讨;然后区分了高维数据空间的固有维数和嵌入在其中的低维参数空间维数这一对容易混淆的概念.三者在一些情况下是一致的;如果高维数据空间存在环状流形,流形维数则要小于嵌入空间维数.本文提出一种环状流形发现算法,能够有效地判别高维数据空间是否存在环状流形.实验结果证明了算法的有效性.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,所以对高维数据中有意义的低维结构的研究依然有很多值得进一步探讨的问题,比如对于高维数据固有维数的估计虽然已经提出很多算法,但大都要求较大的样本集.

References:

[1] Sebastian HS, Lee DD. The manifold ways of perception. Science, 2000,290(12):2268−2269.

[2] Roweis ST, Saul LK. Nonlinear dimensionality analysis by locally linear embedding. Science, 2000,290(12):2323−2326.

[3] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(12):2319−2323.

[4] Donoho DL, Grimes C. When does ISOMAP recover the natural parameterization of families of articulated images? Technical Report, 2002-27, Department of Statistics, Stanford University, 2002.

[5] Donoho DL, Grimes C. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. Proc. of the National Academy of Sciences, 2003,100(10):5591−5596.

[6] Zhang CS, Wang J, Zhao NY, Zhang D. Reconstruction and analysis of multi-pose face images based on nonlinear dimensionality reduction. Pattern Recognition, 2004,37(1):325−336.

[7] Polito M, Perona P. Grouping and dimensionality reduction by locally linear embedding. Neural Inform Process Systems, 2001, 1255−1262.

[8] Lee MD. Determining the dimensionality of multidimensional scaling models for cognitive modeling. Journal of Mathematical Psychology, 2001,45(4):149−166.

[9] Camastra F. Data dimensionality estimation methods: A survey. Pattern Recognition, 2003,36:2945−2954.

[10] Liu XW, Srivastavab A, Wang DL. Intrinsic generalization analysis of low dimensional representations. Neural Networks, 2003,16: 537−545.

[11] Camastra F, Vinciarelli A. Estimating the intrinsic dimension of data with a fractal-based method. IEEE Trans. on Pattern Analysis, 2002,24(10):1404−1407.

[12] Pless R, Simon I. Embedding images in non-flat spaces. Technical Report, WU-CS-01-43, Washington University, 2001.