ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.11, November 2008, pp.2908−2920
DOI: 10.3724/SP.J.1001.2008.02908

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

针对环状流形数据的非线性降维*

孟德宇[1], 古楠楠[1], 徐宗本[1+], 梁 怡[2]

[1](西安交通大学 信息与系统科学研究所,陕西 西安 710049)

[2](香港中文大学 地理与资源管理学系,香港)

Nonlinear Dimensionality Reduction for Data on Manifold with Rings

MENG De-Yu[1], GU Nan-Nan[1], XU Zong-Ben[1+], LEUNG Yee[2]

[1](Institute for Information and System Science, Xi'an Jiaotong University, Xi'an 710049, China)

[2](Department of Geography & Resource Management, The Chinese University of Hong Kong, Hong Kong, China)

+ Corresponding author: E-mail: zbxu@mail.xjtu.edu.cn

Meng DY, Gu NN, Xu ZB, Leung Y. Nonlinear dimensionality reduction for data on manifold with rings. *Journal of Software*, 2008,19(11):2908−2920. http://www.jos.org.cn/1000-9825/19/2908.htm

Abstract: Isomap has attracted attentions recently due to its prominent performance on nonlinear dimensionality reduction. However, how to implement effective learning for data on manifold with rings is still a remaining problem. To solve this problem, a systemic strategy is presented in this study. Based on the intrinsic implementation principle of Isomap, a theorem is presented which gives a sufficient and necessary condition to judge whether a manifold is with rings. Besides, an algorithm for detecting ring structures in the manifold is constructed and a nonlinear dimensionality reduction strategy is developed through polar coordinates transformation. A series of simulation results implemented on a series of synthetic and real-world data sets generated by manifolds with or without rings verify the prominent performance of the new method.

Key words: manifold with rings; manifold learning; nonlinear dimensionality reduction

摘 要: 近年来出现了多种新型的非线性降维方法,且在一些应用中体现出良好的效果.然而,当面对球体、柱体等环状流形产生的非线性流形数据时,这些方法往往会失效.针对这一问题,提出了针对环状流形数据的环结构检测算法与非线性降维方法.理论上,基于目前极受关注的 Isomap 降维方法的运行原理,给出了一个判断环状流形的充要条件;算法上利用所得的判断定理,制订了基于数据的环状流形检测算法;最后基于所找到的环结构,利用极坐标展开的思想设计了针对环状流形数据的非线性降维策略.针对一系列典型环状流形数据集的仿真实验结果表明,与其他流形学习降维方法相比,该方法对环状流形数据进行降维具有显著优势.

关键词: 环状流形;流形学习;非线性降维

中图法分类号: TP18 **文献标识码**: A

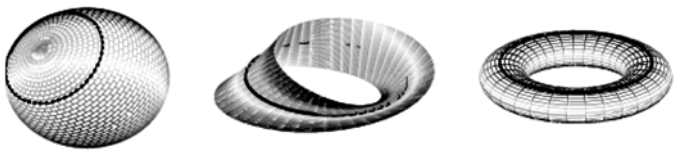
在数据挖掘、人工智能及信息获取等研究领域中,常常面对位于高维表示空间却能被本质低维描述的数据

* Supported by the National Natural Science Foundation of China under Grant Nos.60575045, 70531030 (国家自然科学基金); the Hong Kong Research Grants Council of China under Grant No.4701/06H (香港 RGC 基金)

Received 2007-06-05; Accepted 2007-08-03

集.对于这样的数据集,易知其所处的高维表示空间为可由少数变量表示的低维流形.近年来,针对具有上述特征的高维数据集寻找其对应本质低维嵌入表示的研究问题已引起相关研究领域极大的关注.近年来对此研究问题提出的典型方法(称为流形学习降维方法)包括等距特征映射(isometric feature mapping,简称 Isomap[1])、局部线性嵌入(locally linear embedding,简称 LLE[2])、拉普拉斯特征映射(laplacian eigenmap[3])等.流形学习降维方法已在人脸识别、手写数字辨识、文本分类等领域得到初步成功应用[1−3],使其成为近期的研究热点之一.

然而,当面对环状流形数据(如位于球体、Mobius 环、镯状等流形上的数据集,流形形状如图 1 所示)时,目前的流形学习降维方法一般均会失效,即无法一致准确地找到环状流形数据集的对应低维嵌入表示.如何基于流形数据集自动地检测出流形中是否存在环结构与如何对存在环结构的流形数据集进行有效降维均为流形学习降维研究领域代表性的公开问题[4,5].目前,国内外针对此问题还缺乏深入的理论研究成果,而现有的少数可行算法普适性亦尚待提高(如文献[6,7]中所提出的算法无法对球体流形进行有效降维).



Sphere manifold Mobius manifold Torus manifold

Fig.1 Several typical manifolds with rings

图 1 几种典型的环状流形

本文针对一种典型的流形学习降维方法——Isomap 方法展开研究,目的是通过 Isomap 方法的内在有效性机理深入探索其对环状数据流形失效的本质原因,从而提出环状流形的概念与理论,进而发展出针对环状流形数据集的有效检测算法与非线性降维策略.特别是,在理论上提出了一个合理的环状流形定义,并基于 Isomap 的内在机理分析了此方法对环状流形数据降维失效的本质原因,且提出一个环状流形判断定理;利用所得判断定理设计了一种有效的环状流形检测算法,此算法可直接基于数据判断出其所处流形是否存在环结构,并可获得足以描述流形环结构的数据环,且可近似求得数据集的最大无环子集;应用检测算法求得的环结构,根据极坐标展开的思想建立了针对环状数据的非线性降维策略.对球体、Mobius 环、镯状流形、兵马俑旋转图像等典型环状流形数据集的仿真实验结果表明,所提出的环状流形检测算法不仅可以对其环形结构是否存在进行有效判断,而且可以较完美地描述出数据内在的环状结构;所提出的环状流形降维方法也可以对环状流形数据实行稳健、有效的降维,与原 Isomap 方法相比,实验结果具有明显的优势.

本文第 1 节简要回顾 Isomap 方法.第 2 节给出环状流形的定义与环状流形判断定理,这一节内容是算法的理论部分.第 3 节介绍针对环状数据的检测算法,第 4 节提出针对环状数据的非线性降维方法,第 5 节展示新方法在一系列典型环状流形数据集上的仿真实验结果,这 3 节是算法的构造与实现部分.最后总结全文并提出需要进一步研究的问题.

为了避免符号的混淆,首先统一文中出现的记号如下:记数据所处的高维流形为 $\Omega_n \subset R^n$, 对应的低维表示空间为 $\Omega_d \subset R^d$;隐含的流形映射为 $f:\Omega_d \to \Omega_n$(此映射默认为满射);任意两点 $x_a, x_b \in \Omega_n$, 记其间流形测地距离为 $G(x_a, x_b)$; 初始给定的流形数据记为 $D_l = \{x_i\}_{i=1}^{l} \subset \Omega_n$;基于 Isomap 方法计算的 $x_i, x_j \in D_l$ 之间的近似测地距离记为 $\tilde{G}(x_i, x_j)$.

1 Isomap概述

Isomap 方法是一种利用全局数据信息实现数据降维的流形学习降维方法,已在降维特征描述、聚类与数据可视化等很多应用领域展示出其强大的数据降维功能,这也使得此方法成为目前最受关注的非线性降维方法之一.其主要实现思想是利用局部邻域距离近似计算数据点之间的全局流形测地线距离,通过建立原数据间的测地线距离与降维数据间的空间距离的对等关系从而实现数据降维.由于测地距离一般能够内在地反映数据

的本质流形几何特征,Isomap 常可以成功地找到高维数据本质对应的低维嵌入[1,8].其主要实现步骤如下:

步骤 I(建立邻域图):定义 V 为原数据集合,E 为连接所有邻域数据对的边集合(一般取ε-邻域或 k-邻域),从而建立邻域图(V,E);

步骤 II(计算测地距离):计算 V 中任意两节点在邻域图(V,E)中的最短路径,将此最短路径值作为对应节点间的近似测地距离;

步骤 III(数据嵌入):将步骤 II 中计算得到的 V 中的测地距离矩阵作为输入,应用经典的 MDS 方法[9]可以计算得出数据最终低维嵌入表示.

Isomap 方法最重要的隐含要求是,当流形上的两节点足够近时,它们之间的距离与其低维嵌入之间的距离近似等同.这一要求是其算法模型构造的本质内在机理.我们可以给出此要求的数学描述如下:

假设 1(局部保距假设). $\lim\limits_{y\to y^*}\dfrac{\|f(y)-f(y^*)\|}{\|y-y^*\|}=1$,其中 $y,y^*\in\Omega_d$.

另外,Isomap 方法要求 $D_l\in D_n$ 中任意两数据之间的测地距离能够通过邻域点扩展的方法近似求得.根据算法的运行原理,可推出 D_l 对应低维嵌入表示数据集中任意两数据之间的距离亦可由邻域点扩展的方法近似求得,即嵌入集中两数据之间的直线段可由邻域数据连线近似构成.由此分析可知,嵌入集中任意两数据之间的直线段应近似位于Ω_d 中,这就引出了如下的凸性假设:

假设 2(凸性假设). Ω_d 为有界闭凸集.

基于 Isomap 方法的这两条隐含假设,我们可以对环状流形进行一般性定义,并进一步探索 Isomap 方法针对环状流形数据失效的本质原因.

2 环状流形的定义及其判断定理

可由上一节提出的假设推导出流形映射 f 具有如下重要性质.

定理 1(弧长对等定理)[10]. 若假设 1 与假设 2 成立,则:

(1) 流形映射 f 一致连续;

(2) 记Ω_d 上一条连续曲线为 $\Gamma_{y_s}^{y_e}$(y_s 为曲线起点, y_e 为曲线终点),其弧长为 $S_d(\Gamma_{y_s}^{y_e})=\int_{\Gamma_{y_s}^{y_e}}\mathrm{ds}$; Ω_n 上的对应连续曲线记为 $f(\Gamma)_{f(y_s)}^{f(y_e)}$, 对应弧长为 $S_n(f(\Gamma)_{f(y_s)}^{f(y_e)})$, 则有

$$S_n(f(\Gamma)_{f(y_s)}^{f(y_e)})=S_d(\Gamma_{y_s}^{y_e}) \tag{1}$$

由于 f 是连续映射,将一条位于Ω_d 上的连续曲线 $\Gamma_{y_s}^{y_e}$ 通过 f 映射至Ω_n,可以得到位于流形上的连续曲线 $f(\Gamma)_{f(y_s)}^{f(y_e)}$. 定理 1 给出了 $\Gamma_{y_s}^{y_e}$ 与 $f(\Gamma)_{f(y_s)}^{f(y_e)}$ 曲线弧长的对等关系.

反之,若"Ω_n上两点间的任意连续路径对应于Ω_d上的连续路径"这一结论成立,则由定理 1 可知,计算Ω_n上两点间的高维测地距离(即两点间沿流形的最短路径长度)的问题等价于计算Ω_d 上对应两点间的欧氏距离.此时用高维数据间的测地距离矩阵作为低维数据间的欧氏距离矩阵的近似,应用 MDS 方法即可找到高维数据集的对应低维嵌入数据集,从而保证了 Isomap 方法运行的有效性.但是,对于一些流形映射有可能出现这样的情况:Ω_n 上两点间的连续路径在Ω_d 上的对应路径为若干断开的连续曲线,在这种情况下,两高维点间的测地距离与其相应低维嵌入点间的欧氏距离并不等同,这导致了 Isomap 方法的失效.而出现这种问题的本质原因就是流形中存在环结构.我们首先给出具有环结构流形(即环状流形)的定义,然后再对此问题进行深入分析.

定义 1(环状流形定义). 对于流形Ω_n,有界闭凸集Ω_d 与流形映射 f: $\Omega_d\to\Omega_n$,记 $Y_d=\{(y_r,y_r')\,|\,y_r,y_r'\in\Omega_d,\ y_r\neq y_r',f(y_r)=f(y_r'),$ 且对任意 $0<t'<t''\le 1, f(y_r+t'(y_r'-y_r))\neq f(y_r+t''(y_r'-y_r))\}$,$Y_n=\{f(y_r)\,|\,(y_r,y_{r'})\in Y_d\}$. 若 $Y_d=\varnothing$,则称为流形无环结构;若 $Y_d=\varnothing$,则称流形为环状流形,并存在环结构 Y_d;称 Y_n 为流形的断点集,Y_n 中元素称为流形的断点.

定义 1 对环状流形与流形中的环结构进行了形式化描述.实际上,定义 1 对环状流形的定义可以按如下方

式理解:对于低维嵌入空间Ω_d上的一条线段,通过流形映射 f 转化成一条闭合的高维曲线环,这个环可以形象地看作将原线段进行首尾粘合后形成的流形环(如图 1 所示).定义 1 对流形环结构的定义方法可以通过定理 2 进行等价描述.

定理 2. 在假设 1 与假设 2 成立的前提下,若 f 为单射,则对应流形无环结构;若非单射,则流形存在环结构.

由定理 2 可知,流形映射 f 为单射是流形不存在环结构的充分必要关系.若流形中不存在环结构,则Ω_d 与 Ω_n 之间通过流形映射建立了一一对应关系(既单射又满射).因此,Ω_d 中的一条曲线 $\Gamma_{y_s}^{y_e}$ 一一对应于Ω_n 中的曲线 $f(\Gamma)_{f(y_s)}^{f(y_e)}$.根据定理 1,这两条曲线的弧长相同,因此在假设 1 与假设 2 成立的条件下,计算 $f(y_s)$与 $f(y_e)$之间的流形测地距离在理论上等同于计算 y_s 与 y_e 之间的欧氏距离.因此保证了 Isomap 方法的有效运行.

若流形中存在环结构,即流形存在断点集,则对于高维流形上经过某断点 $f(y_r)((y_r,y_r')\in Y_d)$ 的连续曲线 $f(\Gamma)_{f(y_s)}^{f(y_e)}$,其低维对应可能由低维嵌入空间上的两段曲线 $\Gamma_{y_s}^{y_r}$ 与 $\Gamma_{y_r'}^{y_e}$ 构成;若流形曲线经过更多的断点,则其对应低维嵌入曲线可能由更多段曲线构成,因此,$f(y_s)$与 $f(y_e)$之间的流形测地距离有可能并不等同于其对应低维嵌入 y_s 与 y_e 之间的欧氏距离,这导致了 Isomap 方法的失效.

以上分析可归纳为定理 3.

定理 3. 若假设 1 与假设 2 成立,则

(1) 若流形中不存在环结构,则 $G(f(y_s),f(y_e)=\|y_s-y_e\|$;

(2) 若流形中存在环结构 Y_d,则 $G(f(y_s),f(y_e)\le \min_{(y_r,y_r')\in Y_d}(\|y_s-y_r\|+\|y_r'-y_e\|,\|y_s-y_e\|)$.

定理的结论可由上面的分析直接得出.需要注意的是,当存在环结构时,若 $f(y_s)$与 $f(y_e)$之间测地线不经过断点,则其间测地距离仍等于 $\|y_s-y_e\|$;若其经过唯一断点 $f(y_r)((y_r,y_r')\in Y_d)$ 时,其测地距离等于 $\min_{(y_r,y_r')\in Y_d}(\|y_s-y_r\|+\|y_r'-y_e\|,\|y_s-y_e\|)$;而若经过多个断点时,测地距离有可能小于 $\min_{(y_r,y_r')\in Y_d}(\|y_s-y_r\|+\|y_r'-y_e\|,\|y_s-y_e\|)$.因此,严格来说,定理 3 给出的是环状流形上两点间测地距离的一个上界.

基于如上准备,我们提出判断环结构的关键定理.此定理将对下节基于数据的环状流形检测策略的建立提供理论指导.在介绍此定理之前,首先需要构造一类重要的距离函数如下:

以点 $y_c\in\Omega_d$ 为中心构造线段,线段的起点与终点分别为

$$y_{start}(t)=y_c+t\vec{v},y_{end}(t)=y_c-t\vec{v} \tag{2}$$

其中,$t\in[0,\theta],\theta=\min(\arg\max_t(y_{start}(t)\in\Omega_d),\arg\max_t(y_{end}(t)\in\Omega_d)),\vec{v}\in R^d$ 为具有单位长度的方向向量.构造如下距离函数:

$$Dis_{y_c,\vec{v}}(t)=G(f(y_{start}(t)),f(y_{end}(t))),t\in[0,\theta] \tag{3}$$

基于函数 $Dis_{y_c,\vec{v}}(t)$,可给出流形环结构判断定理.

定理 4(环结构判断定理). 在假设 1 与假设 2 成立的条件下,流形中存在环结构的充分必要条件为:存在 $y_c\in\Omega_d$ 与单位方向向量 $\vec{v}\in R^d$,使 $Dis_{y_c,\vec{v}}(t)$ 在$[0,\theta]$中非单调递增.

定理 4 获得了如流形环结构判断原则:若对任意 $y_c\in\Omega_d$ 与单位方向向量 $\vec{v}\in R^d$,$Dis_{y_c,\vec{v}}(t)$ 在$[0,\theta]$中单调递增,则流形中不存在环结构;若存在 $y_c\in\Omega_d$ 与单位方向向量 $\vec{v}\in R^d$,$Dis_{y_c,\vec{v}}(t)$ 在$[0,\theta]$中非单调递增,则流形中存在环结构.

评注 1. 构造序列 $0\le t_1<t_2<\ldots<t_n\le\theta$,根据定理 4,可以通过判断 $G(f(y_{start}(t_i)),f(y_{end}(t_i)))$ 的增减变化判断环结构.实际上,易证当 $y_{start}(t_i)$ 与 $y_{end}(t_j)$ 交叉前进时,即前进序列为 $(t_1,t_1),(t_1,t_2),(t_2,t_2),(t_2,t_3),\ldots,(t_{n-1},t_n),(t_n,t_n)$ 时,通过判断对应距离函数

$$Dis_{y_c,\vec{v}}(t_i,t_j)=G(f(y_{start}(t_i)),f(y_{end}(t_j))) \tag{4}$$

的增减变化,同样可以对环结构进行判断.

评注 2. 实际上,当流形存在环结构时,使 $Dis_{y_c,\vec{v}}(t)$ 出现非单调递增区间的 y_c 应选在具有最小距离的环结构

(y_r, y_r') 的中心位置,即 $\frac{y_r + y_r'}{2}$ 处.实际上,可在更宽松的条件下对 y_c 进行选择:当 y_c 处于以 $\frac{y_r + y_r'}{2}$ 为中心的开球 $B_{\frac{\|y_r - y_r'\|}{6}}\left(\frac{y_r + y_r'}{2}\right) = \left\{ y \in \Omega_d \middle\| y - \frac{y_r - y_r'}{2} \right\| < \frac{\|y_r - y_r'\|}{6} \}$ 内时,对 $\vec{v} = \frac{y_r - y_r'}{\|y_r - y_r'\|}$, $Dis_{y_c,\vec{v}}(t)$ 仍非单调递增.此结论可用与定理 4 类似的证明思路加以证明,限于篇幅,我们略去相关证明步骤.

这样,我们就得到了流形的环结构判断定理.基于此定理,我们可以构造流形的环结构检测算法.

3 基于数据的环状流形检测算法

上节提出的流形环结构判断定理启示我们构造可行、有效的自动化流形环结构检测算法.然而,注意到此定理中两个构造性的变量 y_c 与 $\vec{v}$ 分别在低维空间 Ω_d 与 R_d 中取得,而构造算法可用的信息仅为给定的有限流形数据集 $D_l = \{x_i\}_{i=1}^l \subset \Omega_n$,因此,构造可行的环结构检测算法必须仅基于流形数据集 D_l 建立变量 y_c 与 $\vec{v}$ 的选择策略(或相应 y_c 与 $\vec{v}$ 下 $Dis_{y_c,\vec{v}}(t)$ 的计算方法).本节中,我们将利用上节得出的理论结果对其建立合理的选择策略.

首先讨论 y_c(或 $f(y_c)$)的选择方法.由定理 4 与评注 1 可知,选取的 y_c 最好在任意环结构 $(y_r, y_r') \in Y_d$ 的中心或中心附近的位置.在没有任何先验知识的情况下,一个合理的 y_c 选择策略是将 y_c 选在整个凸集 Ω_d 的中心位置,即

$$y_c = \arg\min_{y_a \in \Omega_d} \left(\max_{y_b \in \Omega_d} (\|y_a - y_b\|) \right).$$

若流形无环结构,则根据定理 3,y_c 对应的高维流形数据为

$$x_c = f(y_c) = \arg\min_{x_a \in \Omega_n} \left(\max_{x_b \in \Omega_n} (G(x_a, x_b)) \right).$$

基于上式,利用已知信息 $D_l = \{x_i\}_{i=1}^l \subset \Omega_n$,可构造一个近似的中心选择公式如下:

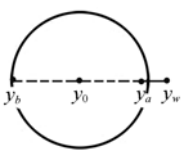
$$x_c^* = f(y_c^*) = \arg\min_{x_i \in D_l} \left(\max_{x_j \in \Omega_l} (\tilde{G}(x_i, x_j)) \right) \tag{5}$$

实际上,对于存在环结构的流形,中心选择公式(5)选取的中心点仍然是合理的.这是由于对流形中的对应任一环结构的环形,可构造不同的流形映射与低维嵌入,使环形上任意一点成为断点,即理论上说,针对某一环形,其上任意一点均可认为是近似中心.因此,公式(5)对有环结构流形依然合理、有效.

下面介绍选取方向 $\vec{v}$ (或计算以 $y_c, \vec{v}$ 为变量的 $Dis_{y_c,\vec{v}}(t)$)的方法.在选择了合理的流形中心点 $x_c^* = f(y_c^*)$ 之后,我们需要基于 D_l 构造足够多的方向 $\vec{v}$,并由此得出 $Dis_{y_c,\vec{v}}(t)$ 的近似表达式,从而根据定理 4(或评注 1)对环结构情况进行检测.

注意到,任意构造闭球 $B_r(y_0) = \{ y \| \|y - y_0\| \le r, y_0, y \in R^d \}$,对任意一个闭球外的点 y_ω,有如下结论成立:

结论 1. 分别记 $B_r(y_0)$ 中距离 y_ω 最近的点为 y_a,最远的点为 y_b,则 y_b 也是 $B_r(y_0)$ 中距离 y_a 最远的点,且 y_a, y_n, y_ω 及圆心 y_0 四点共线.



$Br(y_0)$

Fig.2 Demonstration for Conclusion 1

图 2 结论 1 的图示

此结论可由图 2 明显地观察到.

在流形空间上,根据定理 1 与定理 3,可以得到结论 1 的对应结论.更具体地,以 $x_0 \in \Omega_n$ 为圆心,在流形上构造流形球 $B_r'(x_0) = \{x | G(x, x_0) \le r, x \in \Omega_n\}$, 对 Ω_n 上 $B_r'(x_0)$ 外一点 x_w,如下结论成立:

结论 2. 记 $B_r'(x_0)$ 中与 x_w 测地距离最近与最远的点分别为 x_a, x_b,另记 x_w, x_0 间的流形测地距离为 $G(x_w, x_0) = r + \varepsilon$.若 $B_{r+\varepsilon}'(x_0)$ 中不存在环结构点对,则 x_a, x_b, x_w 与球心 x_0 位于同一条测地线上.

在结论 2 中,x_a, x_b, x_0 共测地线的原因是其低维对应 y_a, y_b 与 y_0 位于同一条直线,即 y_a, y_b 可分别表示为 $y_a = y_0 + t\vec{v}, y_b = y_0 - t\vec{v}$, 其中 $\vec{v} = \frac{y_a - y_b}{\|y_a - y_b\|}$,

$t=\|y_a-y_0\|=\|y_b-y_0\|$, 此时,式(3)中的 $Dis_{y_0,\vec{v}}(t)=G(x_a,x_b)$. 若构造包围球形 $B'_r(x_0)$ 的球壳 $HULL=B'_{r+\varepsilon}(x_0)/B'_r(x_0)$, 则通过取遍所有 $x_w\in HULL$, 并分别求得 $B'_r(x_0)$ 中与之测地距离最近、最远的点 x_a,x_b 与其间流形测地距离 $G(x_a,x_b)$,可获得任意可行方向 $\vec{v}$ 下 $Dis_{y_c,\vec{v}}(t)$ 的值.

基于已知信息 $D_l=\{x_i\}_{i=1}^l$, 根据上述分析,可以构造可行的环结构检测策略.构造流形球 $B_r(x_0)(x_0\in D_l)$ 的近似为

$$\tilde{B}_r(x_0)=\{x_i\in D_l\,|\,\tilde{G}(x_i,x_0)\le r,i=1,2,...,l\} \tag{6}$$

其球壳 *HULL* 的近似为

$$HULL=\{x_i\in D_l\,|\,x_i\text{与}\tilde{B}_r(x_0)\text{中某数据}x_j\text{互为领域},x_i\notin\tilde{B}_r(x_0)\} \tag{7}$$

球壳中任意点 x_j 在流形球中的最近点与最远点为

$$\tilde{x}_a=\arg\min_{x_k\in\tilde{B}_r(x_0)}\{\tilde{G}(x_k,x_j)\},\quad \tilde{x}_b=\arg\max_{x_k\in\tilde{B}_r(x_0)}\{\tilde{G}(x_k,x_j)\} \tag{8}$$

则任意方向 $\vec{v}$ 下的距离函数 $Dis_{y_c,\vec{v}}(t)$ 值可通过任意球壳元素 x_j 下的 $\tilde{G}(\tilde{x}_a,\tilde{x}_b)$ 近似求得.

基于上述思想,可根据已知流形数据集 D_l 分别计算出 y_c 的近似(y_c^*)与任意方向 $\vec{v}$ 下 $Dis_{y_c,\vec{v}}(t)$ 的近似 $(\tilde{G}(\tilde{x}_a,\tilde{x}_b))$. 根据定理 4 与评注 1,可构造出完整的基于数据的环状流形检测算法.在给出算法流程之前,需要说明以下几点:

1. 算法的 Step 1 中,调用 Isomap 方法的输出分别为数据邻域距离矩阵 *D*、路径矩阵 *P* 和近似测地距离矩阵 *G*.其中,若 x_i 与 x_j 位于同一邻域内,则 $D_{i,j}=\|x_i-x_j\|$;否则,D_{ij}=0.路径矩阵 $P=\{P_{i,j}\}_{l\times l}$ 中元素 $P_{i,j}=\{i,i_1,i_2,...,j\}$ 为 Isomap 方法得出的 x_i 到 x_j 的最短路径标号序列.

2. 集合 *N* 为无环集,*E* 为工作集,*C* 为环集;*flagN* 为无环集数据标号,*flagE* 为工作集数据标号,*flagNE* 为无环集与工作集并集(非环集)标号,*flagC* 为环集标号;*Record*={*Record*(*i*,*j*)}$_{3\times l}$ 为数据信息记录矩阵,其中第 1 行表示数据由谁扩充而得,第 2 行代表当前无环集中与该数据测地距离最远的点标号,第 3 行为该数据当前递减次数.

3. 符号 *length*(*A*)表示 *A* 集合中所包含元素的个数;

4. 符号 *S*/*a* 表示若集合 *S* 中存在 *a* 元素,则从 *S* 中将其删除;[*S a*]表示将 *a* 按序并入集合 *S*.

算法 1. 基于数据的环状流形检测算法.

已知:流形数据集 $D_l=\{x_i\}_{i=1}^l$; 邻域个数参数 *k*;距离函数最多下降阈值 *Reducetime*.

求:流形数据的最大无环子集 *N*;流形数据的环子集 *C*;流形数据的环形结构 *RingPath*.

Step 1. 对 D_l 运行标准 Isomap 计算得出邻域距离矩阵 *D*、测地距离矩阵 *G* 与路径矩阵 *P*.

Step 2. 调用式(5)计算流形的近似中心位置 *StartP*.

Step 3. 初始化如下矩阵与向量:3×*l* 的全零矩阵 *Record*;*l* 维全零向量 *flagNE*,*flagN*,*flagE*,*flagC*.初始化空集 *Label* 与 *C*,初始化以 *StartP* 为唯一元素的集合 *N*,*E*,*NE*;令 *flagN*(*StartP*)=1,*flagE*(*Label*)=1,*flagNE*(*StartP*)=1,*flagNE*(*Label*)=1,*Record*(1,*Label*)=*StartP*,*Record*(2,*Label*)=*StartP*,*Record*(1,*StartP*)=*StartP*,*Record*(2,*StartP*)=*StartP*.

Step 4. 若 *E* 为空集,则转入 Step 14;否则,转入 Step 5.

Step 5. 令 *CurrentLabel*=*B*(argmin(*D*(*E*,*StartP*))),将 *CurrentLabel* 标号数据邻域中不属于 *NE* 与 *C* 集合中的数据标号集赋为 *Label* 集,将当前无环集 *N* 中距离 *CurrentLabel* 标号数据最近的数据标号赋值 *NearestLabel*;若 *G*(*NearestLabel*,*Record*(2,*NearestLabel*))>*G*(*CurrentLabel*,*Record*(2,*NearestLabel*)),则转入 Step 9;否则,转入 Step 6.

Step 6. 将当前无环集 *G* 中与 *CurrentLabel* 标号数据测地距离最远的数据标号赋为 *Farlabel* 值;*Record*(1,*CurrentLabel*)=*NearestLabel*;*Record*(2,*CurrentLabel*)=*Farlabel*;*Record*(3,*CurrentLabel*)=0. 若 *length*(*Label*)=0, 则转入 Step 8;否则,转入 Step 7.

Step 7. *E*=[*E Label*],*NE*=[*NE Label*],*E*=*E*/*CurrentLabel*,*N*=[*N*,*CurrentLabel*];*flagE*(*Label*)=1,*flagNE*(*Label*)=1,*flagE*(*CurrentLabel*)=0,*flagN*(*CurrentLabel*)=1;转入 Step 4.

Step 8. N=[N,*CurrentLabel*],E=E/*CurrentLabel*;*flagE*(*CurrentLabel*)=0,*flagN*(*CurrentLabel*)=1;转入 Step 4.

Step 9. *Record*(3,*CurrentLabel*)=*Record*(3,*NearestLabel*)+1;若 *Record*(3,*CurrentLabel*)<*ReduceTime*,则转入 Step 11;否则,转入 Step 10.

Step 10. *Record*(1,*CurrentLabel*)=*NearestLabel*,*Record*(2,*CurrentLabel*)=*Record*(2,*NearestLabel*),C=[C *CurrentLabel*],E=E/*CurrentLabel*;NE=NE/*CurrentLabel*,*flagC*(*CurrentLabel*)=1,*flagE*(*CurrentLabel*)=0,*flagE*(*CurrentLabel*)=0,转入 Step 4.

Step 11. *Record*(1,*CurrentLabel*)=*NearestLabel*,*Record*(2,*CurrentLabel*)=*Record*(2,*NearestLabel*);若 *length*(*Label*)=0,则转入 Step 13;否则,转入 Step 12.

Step 12. E=[E *Label*],NE=[NE *Label*],*flagE*(*Label*)=1,*flagNE*(*Label*)=1,N=[N,*CurrentLabel*],E=E/*CurrentLabel*;*flagE*(*CurrentLabel*)=0,*flagN*(*CurrentLabel*)=1;转入 Step 4.

Step 13. N=[N,*CurrentLabel*],E=E/*CurrentLabel*;*flagE*(*CurrentLabel*)=0,*flagN*(*CurrentLabel*)=1,转入 Step 4.

Step 14. *TempC*=C,若 *length*(C)=0,则转入 Step 18;否则,i=1,然后转入 Step 15.

Step 15. 若 i>*length*(C),则转入 Step 18;否则, *RingPath*(i)=$P_{C(i),Record(2,C(i))}$;j=1,转入 Step 16.

Step 16. 若 j>*Reducetime*,则转入 Step 17;否则,*RingPath*(i)=[*Record*(1,*TempC*(i)),*RingPath*(i)],*TempC*(i)=*Record*(1,*TempC*(i)),j=j+1,转入 Step 16.

Step 17. *RingPath*(i)=[$P_{Record(2,TempC(i)),TempC(i)}$,*RingPath*($i$)],$i$=$i$+1,转入 Step 15.

Step 18. 输出 N 为流形数据的最大无环子集;C 为流形数据的环子集;*RingPath* 为流形数据的环形结构.

所提环状流形检测算法的实现步骤如下:算法每步迭代均保证无环集 N 中数据近似构成球形结构,对应以上(6)中的流形球概念;工作集 E 中数据近似构成流形上包围球形的一圈环结构,对应(7)中的球壳概念;每步产生的当前处理对象 *CurrentLabel* 为工作集 E 中与近似流形中心 *StartP* 测地距离最近的点,计算在 N 中与 *CurrentLabel* 测地距离最近(*NearestLabel*)与最远的点(*Record*(2,*CurrentLabel*)或 *Record*(2,*NearestLabel*)),分别对应(8)中 $\tilde{x}_a$ 与 $\tilde{x}_b$ 概念.

算法执行过程中,首先判断 *CurrentLabel* 与 *Record*(2,*NearestLabel*))间的测地距离相对 *NearestLabel* 与 *Record*(2,*NearestLabel*))间的测地距离是否有所递增.在递增与递减情况下算法的执行方式分别为:

(1) 若递增,依据定理 4 与评注 1,不能判断出环信息,则首先将当前递减次数 *Record*(3,*CurrentLabel*)清零;此时若 *CurrentLabel* 不可通过邻域向外扩充(向 $N \cup E \cup C$ 外),则将 *CurrentLabel* 并入无环集,并将其从工作集中删除,并计算当前无环集中与此点距离最远的数据标号,赋入 *Record*(2,*CurrentLabel*);若可扩充,则将扩充点标号并入工作集,保证工作集对无环集的包围状态,然后同样更新标号 *Record*(2,*CurrentLabel*).

(2) 若递减,根据定理 4 与评注 1,有可能存在环结构.为了保证算法的稳健性,我们采取如下策略:首先将前扩充点下降次数 *Record*(3,*NearestLabel*)增加 1 后赋入 *Record*(3,*CurrentLabel*),然后判断递减次数 *Record*(3,*CurrentLabel*)是否已超过预设阈值 *Reducetime*.若已超过,即说明 *CurrentLabel* 所处的路径距离函数值已连续下降 *Reducetime* 次,故判断 *CurrentLabel* 为环点,将其并入环集,并从工作集中删除,将 *Record*(2,*NearestLabel*)赋入 *Record*(2,*CurrentLabel*).此时由于 *CurrentLabel* 所处路径为环形已获判断,因此对其不再进行扩充.若下降次数 *Record*(3,*CurrentLabel*) 仍小于 *Reducetime*, 即 *CurrentLabel* 为值得怀疑的环形路径点,则将 *Record*(2,*NearestLabel*)直接赋入 *Record*(2,*CurrentLabel*),并将其从工作集中删除,并入无环集,并对工作集进行与(1)步同样的扩充过程.

注意到(2)与(1)的一个重要区别是对标号 *Record*(2,*CurrentLabel*)的更新方法.在(2)中,只是简单地把 *CurrentLabel* 的前扩充点 *NearestLabel* 的 *Record*(2,*NearestLabel*)赋入 *Record*(2,*CurrentLabel*),而并未像(1)中那样计算 *CurrentLabel* 与当前无环集中距离最远的数据标号.此策略的目的在于,经过了 *Reducetime* 次连续判断递减后,每次处理的 *CurrentLabel* 对应的 *Record*(2,*CurrentLabel*)均相同,这保证了最后对应的环集点 *CurrentLabel* 可产生一条从 *Record*(2,*CurrentLabel*)到 *CurrentLabel* 再到 *Record*(2,*CurrentLabel*)的环形路径.根据定理 4 与评注 1,此判断环形的策略是合理的,且可基于环集数据产生的所有环形路径 *RingPath* 实现对环状

流形整体环结构的具体描述.

综上可知,基于数据的环状流形检测算法至少挖掘出了流形数据中的 3 个信息:

1. 流形数据集的一个近似最大无环子集 N.

2. 若 $C\neq\varnothing$,则判断出流形存在环结构;否则,判断出流形不存在环结构.

3. 若 C 非空,则对应任意一个 $x\in C$,可输出一条环形路径,这些环形路径的集合 *RingPath* 可对流形环结构进行整体描述.

因此,运用所提出的环状流形检测算法,完全基于流形数据集 D_l 就可以挖掘出数据所在流形的环结构信息.利用这些信息,就能构造环状流形数据合理、可行的非线性降维策略.

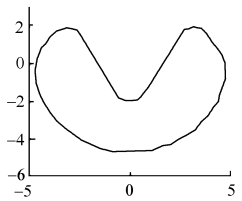
4 针对环状流形数据的非线性降维策略

根据上述分析可知,Isomap 对由流形 Ω_n 产生的数据进行非线性降维失效的本质原因在于 Ω_n 中存在环结构.因此,只有构造方法消除 Ω_n 中的环结构,才能有效地利用 Isomap 寻找其对应低维嵌入.我们采取的策略如下:构造连续映射 $T:\Omega_n\to\Omega_n'$,使 Ω_n' 中不存在环结构,则此时流形映射由 $f:\Omega_d\to\Omega_n$ 转换为 $T(f):\Omega_d\to\Omega_n'$,因此,对 Ω_n' 计算其低维嵌入等价于寻找原流形 Ω_n 的低维嵌入(同为 Ω_d).此时,对无环结构的流形 Ω_n' 的 Isomap 方法就会成为有效的降维方法.因此,如何构造转换函数 T 就成为针对环状流形数据进行可行、有效降维的关键.

对于 2 维空间的环状流形(即一条闭合曲线,如图 3(a)所示),一种简单、直观的消除环结构的策略为极坐标方法.具体来说,选取极坐标中心为闭合曲线中任意一点,记为 (c_x,c_y)(可直接基于观察选取,一般选取闭合区域中较为中心的点),然后将流形的原坐标变换为极坐标,即将原流形点的坐标 (x,y) 转换为 (r,θ),其中,

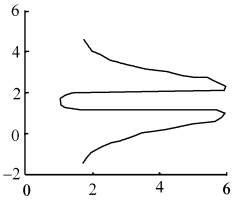
$$r=\sqrt{(x-c_x)^2+(y-c_y)^2};\quad \theta=\begin{cases}\tan^{-1}\left(\dfrac{y-c_y}{x-c_x}\right), & x-c_x\geq 0\\ \tan^{-1}\left(\dfrac{y-c_y}{x-c_x}\right)+\pi, & x-c_x<0\end{cases} \tag{9}$$

由图 3(b)可以观察到,极坐标方法将闭合曲线变换为一条连续而无环结构的曲线.因此,对于 2 维的环状流形,可以利用极坐标展开的方法对其进行环结构消除.



(a) A 1-D ring curve in 2-D space

(a) 二维空间中一个带环结构的 1-D 流形曲线



(b) Transferred curve by applying polar coordinates transformation to Fig.3(a) curve with star point as the center

(b) 以五星点为中心对图 3(a)曲线极坐标转换后所得流形曲线

Fig.3

图 3

然而,在更多情况下,流形 Ω_n 处于远大于 2 维的高维空间中.对于一条位于 Ω_n 上对应环结构 (y_r,y_r') 的闭合曲线 $\Gamma_{x_r}^{x_r}{}'(x_r=f(y_r)=f(y_r')=x_r')$,可构造如下方法对其环结构进行消除:运用经典的 PCA 方法[11]计算此曲线的主方向,然后将此曲线投影至其第一与第二主方向构成的二维空间上,可获得曲线 $\Gamma_{x_2}^{x_2}{}'$(其中 $x_2=x_2'$ 为 x_r,x_r' 在两主方向构成空间上的二维投影向量).由于 $\Gamma_{x_r}^{x_r}{}'$ 闭合,因此 $\Gamma_{x_2}^{x_2}{}'$ 也闭合.可通过视觉观察取得此曲线包围区域中的一点作为中心,从而将数据对应两方向的投影坐标用式(9)进行极坐标变换.由图 3 可获得无环结构的曲线 $\Gamma_{x_2^*}^{x_2^*}{}'(x_2^*\neq x_2^{*\prime})$,此时高维的环曲线 $\Gamma_{x_r}^{x_r}{}'$ 也相应地转换为 $\Gamma_{x_r^*}^{x_r^*}{}'$,显然有 $x_r^*\neq x_r^{*\prime}$.通过这样的过程即可完成对此曲线

环结构的消除.

以环状流形Ω_n上一条闭合曲线为目标对整体流形数据坐标进行上述极坐标变换后可以获得新的流形Ω_n',由上述分析可知,Ω_n'减少了Ω_n的环结构信息.因此,不断迭代进行极坐标展开步骤,可最终实现流形环结构的消除,进而可使 Isomap 方法进行有效的非线性降维.

另外,还有一个重要的问题是每次迭代中,如何选取当前环状流形上的一条闭合曲线来进行极坐标展开.一个合理的策略是在环状流形检测算法所求得的所有环形路径(即闭合曲线)集中挑选路径长度最大的一条.

基于以上阐述,可构造环状流形数据的非线性降维方法如下:

算法 2. 环状流形数据的非线性降维方法.

已知:流形数据集$D_l=\{x_i\}_{i=1}^{l}$; 迭代阈值 m;初始迭代次数 k=0.

求:D_l的低维嵌入数据集$\{y_i\}_{i=1}^{l}$.

Step 1. 对 D_l运行环状流形检测算法.若判断出存在环结构,并检测出流形环形路径集 $RingPath=\{P_i\}_{i=1}^{s}$, 或者,若 $k>m$,则转入 Step 2;若判断出不存在环结构,则转入 Step 6.

Step 2. 在 $RingPath$ 中寻找最长的环形路径 $P_i=\{x_{i1},x_{i2},...,x_{im}\}$.

Step 3. 运用 PCA 计算 P_i数据集的主方向,依次记为 $v_1,v_2,...,v_n$; 构造数据集 D_l与路径集 P_i在 v_1,v_2方向的二维投影数据集,分别记为 $D_2=\{(v_1,v_2)^T\cdot x_i\}_{i=1}^{l}$与 $D_{P_i}=\{(v_1,v_2)^T\cdot x_{i_j}\}_{j=1}^{m}$.

Step 4. 利用视觉观察的方法选择二维投影闭合路径 D_{P_i}的中心点 c,以 c 为中心按照式(9)计算 D_2的极坐标变换 $D_l^*=\{(r_i,\theta_i)\}_{i=1}^{l}$.

Step 5. 构造新的高维数据集: $D_l=\{(r_i,\theta_i,v_3^Tx_i,v_4^Tx_i,...,v_n^Tx_i)\}_{i=1}^{l}\subset R^n$; k=k+1,转入 Step 1.

Step 6. 对 D_l运行 Isomap 程序,得出其低维嵌入数据集$\{y_i\}_{i=1}^{l}$,算法结束.

应用以上方法,便可以对包括环状流形在内的任何流形数据实行有效的非线性降维.

5 仿真实验

为了验证本文提出的环状流形检测算法及非线性降维方法的有效性,我们分别在无环结构的 Swissroll 流形数据集,包含环结构的球体、Mobius 环、镯状流形数据集及旋转兵马俑图像流形数据集上进行了模拟仿真实验.程序均在 Matlab7.0 平台运行,所有图像均由 Matlab 程序绘制而成.

在图 4 中,“.”表示原 SwissRoll 流形数据,被圆环包围的“.”表示环状流形检测算法获得的最大无环集.由此图可以观察到,本文的环结构检测算法判断出的 SwissRoll 流形数据集的环集为空集,无环子集为原数据集,因而准确地判断出此流形数据集中不存在环结构.这解释了对 SwissRoll 形数据可直接利用 Isomap 方法进行有效非线性降维的内在原因.

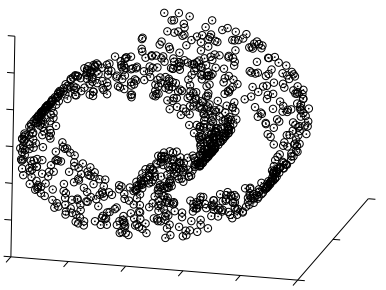


Fig.4 Performance of the proposed ring manifold detection algorithm on Swissroll data set

图 4 本文的环状流形检测算法在 Swissroll 流形数据集

在图 5 的左图中,“.”为相应流形数据集,“*”为算法获得的流形中心;在图 5 的中图中,细线连线为算法所获得的所有环形路径,粗线连线为最长的环形路径;在图 5 的右图中,“.”为算法所获的最大无环子集数据,菱形点为算法所获得的环集数据.通过此图可以观察到,本文所提出的环状流形检测算法判断出了球体、Mobius 环、镯状环形流形数据集中存在环结构的正确结论(环集非空).另外,算法分别找到了球体的 85 个环形路径、Mobius 环的 12 个环形路径与镯状数据的 17 个环形路径.由图 5(a)~图 5(c)的中图可以看出,这些环形路径较为全面地描述出了流形的环

结构特征.再由图 5(a)~图 5(c)的右图可以观察到,算法所获的最大无环集(分别呈半球形、半环形与半镯形形状)也是较为准确的.通过提取最长环形路径(如图 5(a)~图 5(c)的中图所示)对数据集进行极坐标展开,运用本文所提出的非线性降维方法对数据集进行降维,所得结果如图 6(a)~图 6(c)的上图所示.与对数据集直接进行 Isomap 方法进行降维的结果(如图 6(a)~图 6(c)的下图所示)进行比较,可观察到本文所提出的方法具有明显优势.特别地,原 Isomap 方法对环状流形数据中的环结构非常敏感,所得降维结果受数据集环结构影响,形状扭曲变形;而本文所提出的方法有效地消除了环结构,能够较为准确地获得环状流形数据的非线性降维嵌入表示.

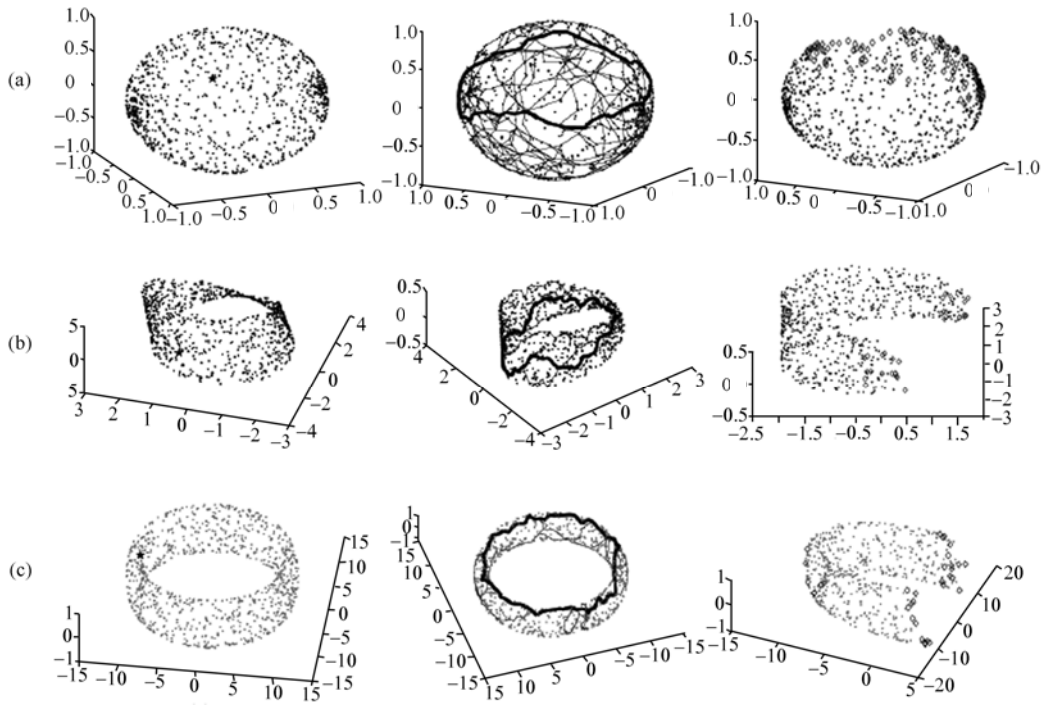


Fig.5 Performance of the presented ring manifold detection algorithm on sphere, Mobius, totus data set

图 5 本文环状流形检测算法在球体、Mobius 环、镯状数据集上的表现

特别需要说明的是,对于球体与 Mobius 环流形数据,本文所提非线性降维方法只迭代了 1 次便展开了环结构;而对于镯状流形数据,所提出的方法需要迭代 2 次方可完全消除环结构.这表明了本文的方法对于多层次环状流形同样具有较为稳健的执行效果.

为了进一步验证新方法的有效性和可用性,我们构造了一组实际的兵马俑旋转图像仿真实验.数据集共包含 105 张像素为 110×80 的兵马俑跪俑图像,分别通过将图像采集设备绕兵马俑旋转不同的角度拍摄而得.运行环状流形检测算法,其计算结果如图 7 所示.其中,图 7(a)为部分兵马俑图像数据,带黑框的数据为算法所得流形中心;图 7(b)为兵马俑流形数据在其第一与第二主方向构成的二维平面上的投影,“*”点表示算法获得的无环集,菱形点代表算法获得的环集;图 7(c)为将图 7(b)中闭合路径进行极坐标展开的结果.由图 7 可以观察到,所提出的算法能够准确测得所有的数据图像,构成了一个闭合的环形路径,并能较为准确地得出数据集的一个最大无环子集.将所获得的环形路径投影至路径数据集第一与第二主方向构成的二维平面,可得到一个闭合的二维曲线;运行本文非线性降维方法的 Step 3 与 Step 4,可对其进行极坐标展开,从而消除其内在环形结构.利用本文方法与原 Isomap 方法分别对此兵马俑图像数据集实行非线性降维,实验效果如图 8(a)与图 8(b)所示,其中图 8(a)为本文方法所获得的结果,图 8(b)为 Isomap 方法所获结果.其中上方点集为相应方法求得的原数据集二维嵌入点,下方点集为求得的原数据集一维嵌入点.被圆环包围的点为随机选取的嵌入序列数据,图 8(a)和图 8(b)最下方为此序列数据对应的兵马俑图像.可观察到在其本质的 1 维特征方向上,原 Isomap 方法所获嵌入表示杂乱无

序,特征意义并未得以较好地体现;而本文方法所获得的 1 维嵌入数据具有明显维数特征(兵马俑旋转特征).这进一步表明了本文的方法对环状流形数据的显著有效性.

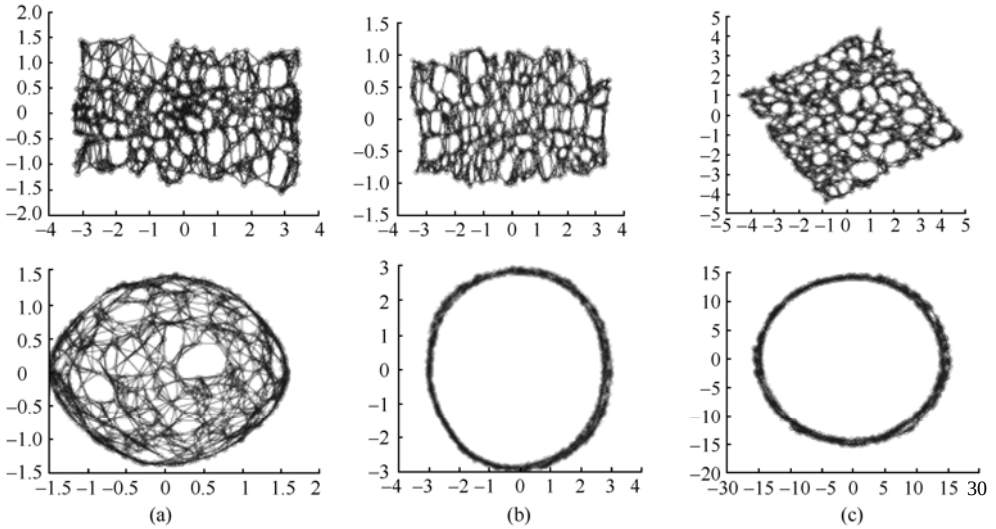


Fig.6 The comparative performance of the presented nonlinear dimensionality reduction method and Isomap on sphere, Mobius and torus data set. The upper figures are the results obtained from the proposed method on the corresponding data sets; the lower figures are the ones from Isomap

图 6 本文非线性降维方法与 Isomap 方法在球体、Mobius 环、镯状数据集上的算法表现比较.
图 6(a)~图 6(c)的上图为本文方法在相应数据集上的实验结果,
图 6(a)~图 6(c)的下图为 Isomap 方法的实验结果

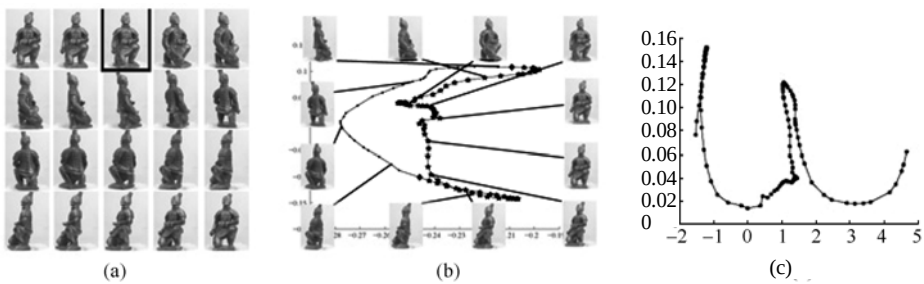


Fig.7 Performance of the new ring structure detection algorithm on Terracotta Army data

图 7 本文环状流形检测算法在兵马俑数据上的算法表现

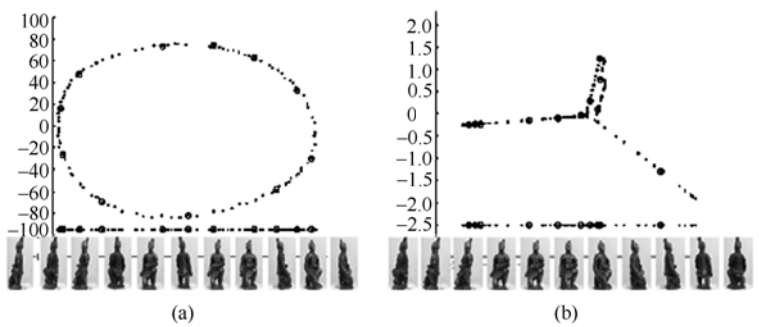


Fig.8 Comparative performance of the new method and Isomap on Terracotta Army data set

图 8 本文方法与 Isomap 方法在兵马俑数据集上的降维结果比较

6 讨 论

我们所做的大量实验表明,本文所提出的流形环结构检测算法稳健、有效,对于流形数据集能够准确地检测出其所在流形是否存在环结构,并能较为精确地利用环形路径集描述出流形的整体环结构特征.一般情况下,基于极坐标展开原理的环状数据降维策略能够在检测算法所获信息的基础上消除流形环结构,从而实现对环状数据集的有效非线性降维.然而,对于一些处于复杂形状的环状流形上的数据集,尽管仍然可以利用本文方法准确地检测出其环结构,但基于极坐标展开的非线性降维方法的可行性却可能得不到保证.

具体来说,有两种形状的环状流形会导致本文所提非线性降维方法的可行性问题.第 1 种是用环结构检测算法得到的 *RingPath* 中最长的环状路径在映射到其第一与第二主方向构成的二维平面时,得到的二维投影路径为多处交叉的闭合曲线(如图 9(a)和图 9(b)所示,其中图 9(b)为图 9(a)曲线在其第一与第二主方向所构成的二维平面上的投影).此时可能很难界定 D_{P_i} 闭合曲线所包围的区域,因而不能合理地利用视觉观察选择环中心 c 的取值.第 2 种情况是,虽然环状路径得到的二维投影 D_{P_i} 为不交叉的闭合曲线,然而曲线形式过于复杂(如图 9(c)所示),以至于无论如何选取环中心点也无法将闭合曲线利用极坐标方法展开.此时本文所提出的非线性降维的方法也会失效.

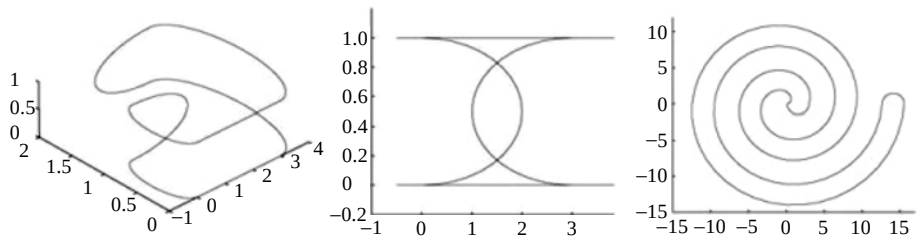


Fig.9 Demonstrations for the infeasible cases of the new method

图 9 本文方法失效的情况演示

如何充分利用流形环结构检测算法所获得的环形信息,改进基于极坐标展开的思想,发展更为普适可行的针对环状流形数据的非线性降维策略,仍然需要我们继续做更深入的探索.

7 结 论

针对流形学习降维方法(特别是 Isomap 方法)对于环状流形数据集不能进行有效非线性降维的问题,本文从理论到应用提出了一套系统的解决策略.特别地,本文对 3 个关键问题进行了解答:什么是环状流形;如何基于数据判断其所处流形是否为环状流形;如何针对环状流形数据进行有效降维.针对第 1 个问题,基于 Isomap 方法所采用的两条隐含假设,建立了环状流形的相关理论.特别地,合理地从理论上给出了环状流形的概念,并推导出判断环状流形(即流形中存在环结构)的一个充分必要条件.针对第 2 个问题,依据所推出的判断定理,构造了一种完全基于流形数据集的环状流形检测算法.此算法不仅能够判断流形中是否存在环结构,并能获得一组环形数据路径集对流形环结构进行整体描述.另外,所提出的算法还能从环状流形数据集中提取出一个无环子集,此子集是能够使 Isomap 方法生效的原数据集近似最大子集.针对第 3 个问题,利用极坐标展开的思想建立了一个针对环状流形数据的降维方法.此方法能够充分利用环状流形检测算法获得的信息,对环形数据集进行有效的非线性降维.在一系列标准的无环与环状流形数据集上的仿真实验表明,所提出的方法稳健、有效,具有广泛的理论与应用价值.

仍需研究的问题包括:基于 Isomap 的两条假设建立更完善的环状流形理论体系;利用先验信息构造更合理的环形中心 x_c^* 选择策略;提高算法对多层次环状流形数据集的稳健性等.

致谢 感谢靖稳峰教授与梁栋同学提供的图像采集实验平台与相关技术支持.感谢审稿人对文章质量的进一

步提高指出的宝贵建议与修改方案.

References:

[1] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000,290(5500):2319−2323.

[2] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290(5500):2323−2326.

[3] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 2003,15(6): 1373−1396.

[4] Saul LK, Roweis ST. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research, 2003,4(6):119−155.

[5] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006,313(5786):504−507.

[6] Lee JA, Verleysen M. How to project 'circular' manifolds using geodesic distances? In: Verleysen M, ed. Proc. of the European Symp. on Artificial Neural Networks. Bruges: IEEE Press, 2004. 223−230.

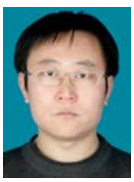
[7] Lee JA, Verleysen M. Nonlinear dimensionality reduction of data manifolds with essential loops. Neurocomputing, 2005,67(8):29−53.

[8] de Silva V, Tenenbaum JB. Global versus local methods in nonlinear dimensionality reduction. In: Becker S, Thrun S, Obermayer K, eds. Neural Information Processing Systems 15 (NIPS 2002). Cambridge: MIT Press, 2003. 705−712.

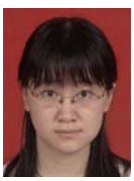
[9] Cox TF, Cox MAA. Multidimensional Scaling. London: Chapman & Hall, 1994.

[10] Zorich VA. Mathematical Analysis. New York: Springer-Verlag, 2004.

[11] Jolliffe IT. Principal Component Analysis. New York: Springer-Verlag, 1986.



孟德宇(1978－),男,山西太原人,博士,讲师,主要研究领域为非线性降维,模式识别.



古楠楠(1985－),女,硕士,主要研究领域为模式识别,流形学习.

徐宗本(1955－),男,博士,教授,博士生导师,主要研究领域为计算智能,信息科学.

梁怡(1948－),男,博士,教授,博士生导师,主要研究领域为智能决策系统,空间优化.