ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.8, August 2010, pp.1898−1907
doi: 10.3724/SP.J.1001.2010.03648

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

针对非连通流形数据降维的过渡曲线方法*

古楠楠, 孟德宇+, 徐宗本

(西安交通大学 信息与系统科学研究所,陕西 西安 710049)

Transition Curve Method for Dimensionality Reduction of Data on Disconnected Manifold

GU Nan-Nan, MENG De-Yu+, XU Zong-Ben

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

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

Gu NN, Meng DY, Xu ZB. Transition curve method for dimensionality reduction of data on disconnected manifold. *Journal of Software*, 2010,21(8):1898−1907. http://www.jos.org.cn/1000-9825/3648.htm

Abstract: Feature extraction of data lying on disconnected manifold is an open problem in the field of manifold learning, and decomposition-composition (D-C) algorithm is the most effective method so far to deal with this problem. However, the biggest limitation of D-C method is edge problem, that is when the nearest data points of different clusters are located in the inner part instead of the edge part of the corresponding cluster, D-C method always behaves poorly. To tackle this key issue, a method, called transition curve method, is presented in this paper. The main idea of the method is to make all clusters on the underlying manifold connect more effectively by constructing smooth transition curves which connect the nearest edge points of different clusters, and in this way the global shape of the data can be preserved better in the low-dimensional space. Experimental results on a series of synthetic and image data sets verify that the transition curve method performs evidently better than D-C method. Particularlly, the edge problem is alleviated. In this way, the application scope of D-C method is expanded remarkably.

Key words: data on disconnected manifold; dimensionality reduction; edge problem; manifold learning

摘 要: 针对位于非连通流形上的数据的特征提取是流形学习领域的一个公开问题,分解-整合算法是目前处理此问题的最有效的方法.然而,此算法的最大局限是边缘问题,即当不同类间的最短距数据对位于相应类内而非类边缘时,算法往往表现异常.针对这一关键问题,提出了一种解决方法——过渡曲线方法.其主要思想为,通过构建连接不同类边缘最短距数据对间的平滑过渡曲线以使流形类间的连接关系更为有效,进而使得数据的全局形态在低维空间中能够更好地保持.一系列人工与图像数据集上的实验结果表明,过渡曲线方法的表现明显优于分解-整合算法,特别是,边缘问题得到了解决,这极大地扩展了分解-整合算法的应用范围.

关键词: 非连通流形数据;数据降维;边缘问题;流形学习

中图法分类号: TP391 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.60905003 (国家自然科学基金); the National Basic Research Program of China under Grant No.2007CB31102 (国家重点基础研究发展计划(973))

Received 2008-09-10; Revised 2009-03-30; Accepted 2009-04-29

1 Isomap 概述

数据特征提取(或数据降维)是数据建模与数据挖掘的基本问题.流形学习是近年来兴起的数据特征提取(特别是低维特征表示)最受关注的方法[1−4].典型的流形学习方法包括等距特征映射(isometric feature mapping, 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]、局部切空间校准(local tangent space alignment,简称 LTSA)[4]等.这些流形学习降维方法已在人脸识别、手写数字辨识及文本分类等领域得到初步成功应用[1,2,5,6],从而使相关的理论与算法研究成为数据挖掘、人工智能及信息获取等领域新的研究热点.

流形学习方法在所考虑的数据位于高维空间的一个低维流形(即原高维数据空间是由少数独立特征共同作用所张成的低维流形)的假设基础上工作.目前构造的流形学习方法都是针对位于连通流形上的数据集构造的,而当面对分布于非连通(或多类)流形上的数据(简称非连通数据),即通过邻域连接后数据体现为多个互不连通的数据邻域图时,这些方法所得降维结果一般不能成功地保持原数据集的类间或类内拓扑结构,如降维结果往往出现高度收缩的点或线的形态,或仅获得其中一部分数据的降维结果.此时,流形学习方法的降维结果十分不理想(从保持原数据结构与特征的意义上).因此,针对非连通(或多类)流形数据的数据特征提取问题是流形学习领域的公开问题之一.

目前,针对非连通数据的降维主要有 3 种方法:建立类间最短连接方法[7]、构造连通图的图论方法[8−16]以及最近提出的分解-整合算法[17].前两种方法的实现思想本质上是一致的,都是通过增加类间连接从而在非连通流形数据上构造出连通邻域图,然后利用传统流形学习方法实现数据降维.具体来说,第 1 种方法首先在合适的邻域尺寸下构造数据的 k 或 ε近邻邻域图,然后通过建立少量类间数据连接强行建立非连通类间的关联;第 2 种方法主要运用图论技术直接在合理的邻域尺寸 k 下建立数据集上的 k 点连接或 k 边连接图.这两种方法尽管保证了流形学习方法在整个数据集上的降维可行性,然而其所达到的降维效果一般不能令人满意,特别是往往出现降维类收缩问题[17].这主要是由于这两种方法建立的类间连接具有很大程度的粗糙性,往往从整体上破坏了数据流形整体的平滑结构与本质维形态,从而对流形学习方法的有效运行对产生了负面影响.

分解-整合算法通过分别利用每个非连通类内数据的流形结构与数据类间的拓扑关系,有效地避免了以往算法中的类间连接不可靠性问题.具体来说,此算法首先利用非连通数据类内的可靠流形结构对每类分别进行降维(分解步骤),然后利用数据类间的拓扑关系对每个降维类分别进行准确的全局定向与定位(整合步骤),从而最终获得原数据集的降维表示.实验结果表明,分解-整合算法明显地改进了以往非连通流形学习方法的降维效果,特别是避免了降维类收缩的异常表现[17].

然而,分解-整合算法存在的最大缺陷是边缘问题,即当不同数据类间最短距数据对位于类的内部而不是边缘时,其所建立的类间连接关系不能准确地反映数据的全局拓扑结构,从而导致算法的失效.这一问题的解决将极大地扩展这一方法的有效应用范围,因而无论是从算法完善角度,还是从应用拓展的角度,都具有重要的意义.

针对这一问题,我们在文中提出了过渡曲线算法.此算法能够基于数据有效构建各非连通类边缘最近点间的平滑过渡曲线,从而能够更准确地反映类间的流形拓扑关联,能够更合适地进行整合步骤中的定向与定位步骤,并最终获得数据本质的低维表示.在一系列人工与图像数据集上的实验结果表明,新算法明显地改进了原方法的计算效果,特别是使边缘问题得到明显的改善.

本文第 2 节对分解-整合算法进行概述.第 3 节提出过渡曲线算法.第 4 节将展示新算法在一系列人工与图像非连通数据集上的实验效果.第 5 节对全文进行总结.

2 分解-整合算法概述

分解-整合算法是孟德宇等人近期提出的一种针对非连通数据的流形学习的有效策略.其基本实现思想是:首先对数据进行分解获得其非连通子类,并对每个子类数据分别进行降维;然后通过对每个降维子类进行全局

定向与定位,从而获得整体数据集的低维表示[17].该算法由分解与整合两个步骤构成,具体描述如下:

分解步骤.对非连通流形数据集 $X=\{x_i\}_{i=1}^n$, 该步骤的目标是按其本质聚类对数据进行分解归类,并对每类数据分别计算其低维表示.这一目标通过如下 3 个子步骤实现:首先建立分布在数据集 X 上的邻域图 $G=(V,E)$,然后通过寻找图 G 的连通分支对原数据集进行自然聚类,聚类结果记为 $X^i(i=1,\ldots,s)$,最后采用 Isomap 方法对每类数据集 X^i 计算其低维表示 $Y^i(i=1,\ldots,s)$.

整合步骤.该步骤的目标是对分解步骤所得的降维子类实施旋转与平移变换,以对其进行准确定位与定向,从而整合形成整个数据集的本质维表示.具体来说,这一目标通过如下子步骤实现:首先估计类中心点集 CX 间的测地距离矩阵 $\tilde{D}$,通过对 $\tilde{D}$ 运行 Isomap 方法来将 CX 在本质维空间中进行定位;再建立每个数据类的旋转/平移参照点集,通过构造合理的旋转/平移变换,将每类降维表示 $Y^i(i=1,2,\ldots,s)$中对应的标记点向其参照点进行定位与定向,进而实现对整个类的定位与定向,并最终获得原非连通数据集的本质维表示.

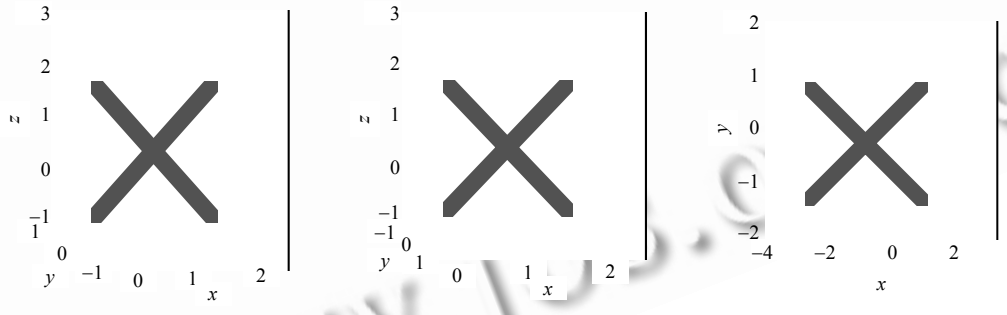
分解-整合算法与以往的非连通数据的降维方法的最大不同在于其分别利用了非连通流形类内流形结构与类间拓扑关系,这一特性使得非连通流形类间的不可靠信息对每类的局部降维表示不产生任何负面影响,而且通过合理利用非连通类间关系,每类的降维表示集(视为刚体)能够被准确地定位与定向,从而能够恢复出原非连通流形全局的本质低维形态.

3 过渡曲线算法

本节将提出过渡曲线算法.它是针对分解-整合算法中出现的边缘问题所提出的.下面首先对边缘问题进行介绍与分析.

3.1 边缘问题

分解-整合算法所面临的最大缺陷是边缘问题.具体来说,当不同子类间最近距数据对位于类的内部而不是边缘时,该算法的整合步骤所建立的类间连接关系有可能不能准确地反映数据的全局拓扑结构,从而导致算法的失效.图 1 直观地展示了分解-整合算法的这一问题.在如图 1(b)所示的 3 类非连通流形数据上运行分解-整合算法,获得的降维表示(如图 1(c)所示)存在明显的混叠现象.这正是由于数据集不同子类间最短距数据对位于类的内部所造成的.下面,我们对出现这一问题的本质原因进行分析.



(a) 3-Cluster manifold formed by three semi-spheres
(a) 3 个半球面构成的 3 类流形

(b) Data set generated from the manifold, containing 1 500 data points
(b) 分布在此流形上的包含 1 500 个点的数据集

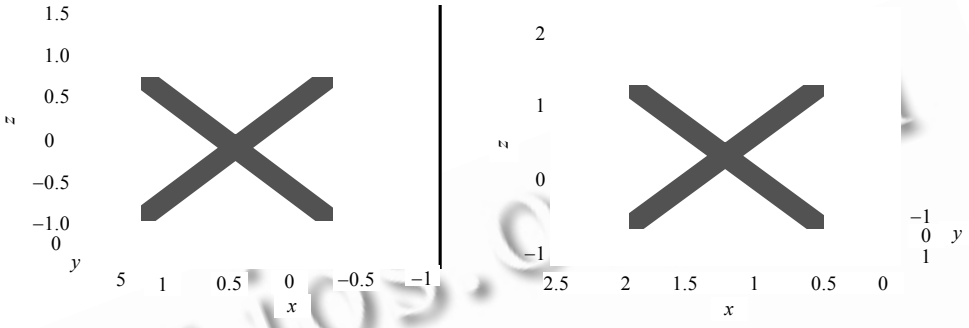
(c) 2-D embedding of the data set calculated by the decomposition-composition algorithm
(c) 分解-整合算法所得的此数据集的 2-维嵌入

Fig.1 Performance of decomposition-composition algorithm on the data set generated from 3-cluster manifold composed by three semi-spheres

图 1 分解-整合算法在分布于 3 类半球面流形的数据集上的表现

上述边缘问题出现的本质原因是,分解-整合算法在整合步骤中计算类中心点集 CX 间的测地距离矩阵

$\tilde{D}$ 时可能会出现错误估计.由于此测地距离直接决定了类中心点集在低维空间中的定位框架,并进而会影响每个降维子类的定位信息,因此,类中心间测地距离的估计会对算法的有效性产生直接的影响.原算法采用的估计策略如下:对非连通数据子类 X^i 与 X^j(如图 2(a)中两类黑点所示)的中心点 cx_i 与 cx_j(如图 2(a)中方点所示)间的测地距离 $\tilde{D}_{ij}$,寻找非连通类间的最短距数据对 $nx_j^i \in X^i$ 与 $nx_i^j \in X^j$(如图 2(a)中圆点所示),它们之间的距离记为 d_{ij}^0(图 2(a)中虚线的长度),并分别计算 cx_i 与 nx_j^i,cx_j 与 nx_i^j 间的测地距离 d_{ij},d_{ji}(图 2(a)中实线的长度),则 $\tilde{D}_{ij} = d_{ij} + d_{ij}^0 + d_{ji}$.当子类间最短距数据对 nx_j^i 与 nx_i^j 位于类边缘时,中心点间的测地距离得到较为精确的估计(如图 2(a)所示).然而,当数据类间的最短距数据对位于类的内部时(如图 2(b)所示),类间的连接关系通过最短距数据对建立在类的内部,由上述估计方法所得的 $\tilde{D}_{ij}$ 往往偏低地估计了类中心点间的真实测地距离,从而在定位分解步骤获得的每个降维子类时使其距离过近,进而导致降维子类出现混叠的异常现象.



(a) Nearest data points of two clusters are located in the edge part of the corresponding cluster
(a) 两类间的最短距数据点位于相应类的边缘

(b) Nearest data points of two clusters are located in the inner part of the corresponding cluster
(b) 两类间的最短距数据点位于相应类的内部

Fig.2 Demonstration for the process of estimating the geodesic distance of centers corresponding to different clusters in decomposition-composition algorithm

图 2 分解-整合算法估计不同类中心间测地距离的过程示意图

针对这一问题,我们构造了一种解决方法——过渡曲线方法.其实现思想是,构建不同数据类间的平滑过渡曲线,以期较为准确地估计类中心点间的测地距离,从而建立准确的类间关系,以期对每个降维子类进行准确的定向与定位,进而解决边缘问题.

3.2 过渡曲线算法

过渡曲线算法的基本构造原则是,建立流形非连通类间平滑的过渡曲线,从而对非连通类中心点间的测地距离进行准确估计,进而指导相应降维子类的准确定位与定向.算法通过如下策略估计类中心间的测地距离:首先寻找各非连通流形类的近似边缘表示集(如图 3(a)所示),然后建立流形类边缘的最短距数据对之间的平滑过渡曲线(如图 3(b)所示),计算该曲线的长度和最短距数据对与其相应类中心间的测地距离之和,以此估计类中心间的测地距离.

下面将更为细致地阐述该算法的两个关键步骤,即寻找流形类边缘与建立平滑过渡曲线的实现过程.

(a) 寻找流形类边缘

寻找流形的近似类边缘表示集是过渡曲线算法的第 1 阶段.该阶段共包括 3 个子步骤:

(1) 对每个流形数据类产生一个最短路径树.最短路径树是指对于一个连通图上指定的根节点(这里取为流形的中心点),连接其与其他所有数据间的最短路径而构成的树.

(2) 计算每个最短路径树的主干结构.对于一个流形类的边缘点来说,沿着中心位置到该点的方向,流形类中不存在比该点更远的点.因此,我们可以计算该流形类最短路径树的主干结构(即具有较大长度的树

分支),通过提取这些主干的叶节点获得边缘点候选集.

(3) 依据上一步得到的候选集中点的深度值来确定流形类边缘集.这里采用 L_1-数据深度算法[18],即对数据集 $X=\{x_1,x_2,\ldots,x_n\}$ 及 $y\in X$,y 在 X 中的深度值为

$$D_n(y)=1-\max\left(0,\left\|\sum_{x_i\neq y}e(x_i-y)/n\right\|-\sum_{x_i=y}\frac{1}{n}\right) \tag{1}$$

其中,$e(x_i-y)=(x_i-y)/\|x_i-y\|$.依此深度度量,数据集中最内部点的深度值接近 1,边缘点的深度值接近 0.于是,可选取深度值接近 0 的点来得到最终的流形类边缘集.

(b) 构建类间平滑过渡曲线

此步骤的目标是建立非连通流形类 X^i 与 X^j 边缘的最短距数据对间的平滑过渡曲线.

记步骤(a)中找到的类 X^i 与 X^j 的边缘集分别为 EX^i 与 EX^j,在 EX^i 与 EX^j 中寻找欧氏距离最近的数据对 $e_i\in EX^i$ 与 $e_j\in EX^j$,即

$$\|e_i-e_j\|=\min_{\substack{ex_i\in EX^i\\ex_j\in EX^j}}\|ex_i-ex_j\| \tag{2}$$

则 $e_i\in X^i$ 与 $e_j\in X^j$ 即为流形类 X^i 与 X^j 的边缘间的最短距数据对.然后,在此数据对上建立类 X^i 与 X^j 间的平滑过渡曲线.实际上,易证由如下数学表达式得到的曲线构成了 e_i 与 e_j 之间的一条平滑过渡路径:

$$f(t)=At^3+Bt^2+Ct+D,t\in[0,1] \tag{3}$$

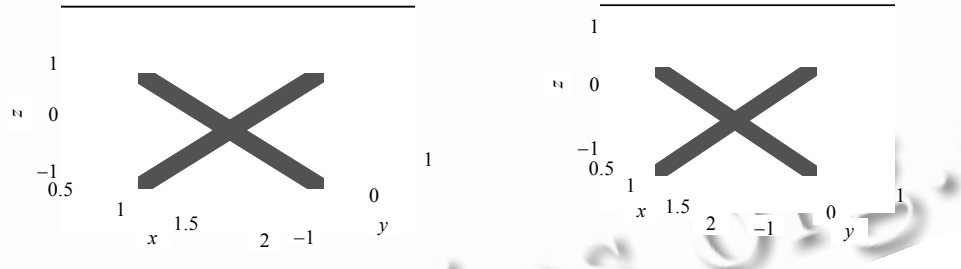
其中,$A=2e_i-2e_j+v_i-v_j$,$B=-3e_i+3e_j-2v_i+v_j$,$C=v_i$,$D=e_i$,v_i,v_j 分别为 e_i,e_j 处的切方向(可由 e_i,e_j 的邻域数据在相应最短路径树中通向根节点的下一个结点指向该数据点的方向取均值而得).

类 X^i 与 X^j 的中心点 cx_i 与 cx_j 间的测地距离可由如下方法估计:

$$\tilde{D}_{ij}=d_{ij}+d'_{ij}+d_{ji} \tag{4}$$

其中,d_{ij},d_{ji} 分别表示 cx_i 与 e_i,cx_j 与 e_j 间的测地距离,d'_{ij} 表示 e_i 与 e_j 间的平滑过渡曲线的长度.

过渡曲线算法建立了非连通流形类间的平滑过渡关系,从而能够使类中心点之间的测地距离得到更准确的估计,进而使边缘问题得到有效的解决.



(a) Finding cluster edges (a) 寻找类边缘

(b) Constructing intra-cluster smooth transition curves (b) 构建类间平滑过渡曲线

Fig.3 Demonstration for the process of the transition curve method

图 3 过渡曲线算法执行过程示意图

3.3 算法复杂度分析

下面,我们对所提出算法的计算复杂度进行分析.

在寻找类边缘的子步骤 1 中,采用著名的 Floyd 方法生成数据的最短路径树,所需的计算代价不超过 $O(kn^2\log(n))$;在子步骤 2 中,最多对每个点对应的最短路径树分支进行 1 次计算与比较,因此计算复杂度最多为 $O(n)$;在子步骤 3 中,采用 L_1-数据深度算法所耗费的计算复杂度最多为 $O(n\log(n))$[18].

在构造类间平滑过渡曲线的步骤中,优化公式(2)的计算本质决定了其总的计算复杂度.通过采用著名的

Heap Sorting 算法[19],计算其优化解共需要 $O(s\log(s))$的计算代价.其中,s 为各个类边缘集所含数据的最大个数.

因此,所提出算法的复杂度约为 $O(kn^2\log(n)+n+n\log(n)+s\log(s))=O(kn^2\log(n))$.与分解-整合算法的复杂度 $O\left(\sum_{i=1}^{s}(kn_i^2\log(n_i)+n_i^3)+n\log(n)\right)$相比,其计算复杂度要高一些.

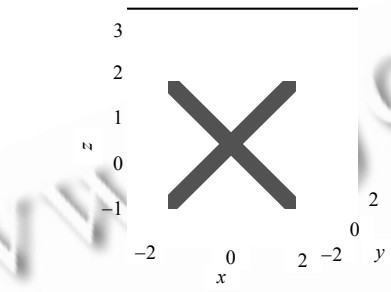
在下一节中,我们将通过比较分解-整合算法与所提出的过渡曲线算法在一系列人工与图像非连通流形数据集上的实验结果来验证过渡曲线方法的有效性.

4 数值实验

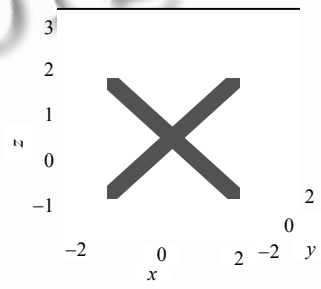
本节我们提供一系列人工及图像数据集上的数值实验结果来比较所提出的过渡曲线算法与分解-整合算法的应用性能,从而对新方法的有效性(特别在克服边缘问题方面)进行验证.

4.1 应用到4类非连通流形数据集

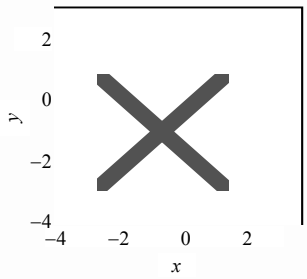
首先采用如图 4(b)所示的 4 类非连通数据集测试过渡曲线算法的性能.该数据集共包含了 2 000 个数据点,它们均匀地分布在如图 4(a)所示的 4 类非连通流形上,这些流形分别为半球面、无底的锥面、有底无盖的柱面、无左侧面的正方体面.对它们分别应用分解-整合算法与过渡曲线算法(邻域尺寸均取为 k=6),得到如图 4(c)、图 4(e)所示的降维结果.图 4(d)中深色点则展示了过渡曲线算法得到的近似类边缘表示集.



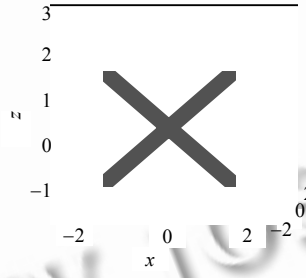
(a) 4-Cluster disconnected manifold
(a) 4 类非连通流形



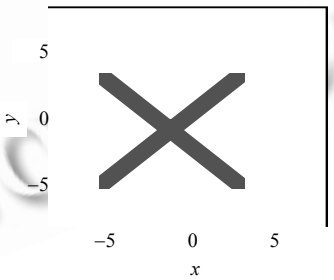
(b) 2 000 data points randomly distributed on the manifold
(b) 随机分布在此流形上的 2 000 个数据点



(c) Performance of D-C algorithm
(c) 分解-整合算法的表现



(d) Edge points found by transition curve method
(d) 过渡曲线方法寻找到的边缘点



(e) Performance of the transition curve method
(e) 过渡曲线方法的表现

Fig.4 Comparative performance of transition curve method and D-C algorithm for data sets lying on 4-cluster disconnected manifold

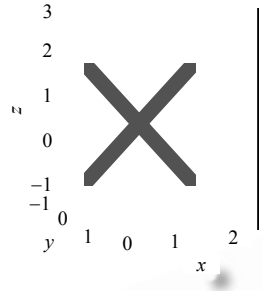
图 4 过渡曲线算法与分解-整合算法在 4 类非连通流形数据集上的对比表现

从图 4 可以观察到,分解-整合算法在此非连通数据集上的降维结果发生了异常:位于不同类上的数据发生了混叠,从而无法从降维结果中分辨出 4 个类.这说明分解-整合算法未能正确保持数据所在类间的关系,未能体

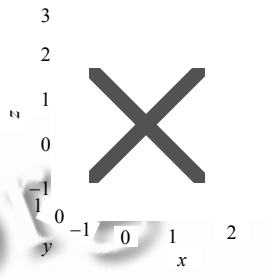
现数据准确的本质维形态.相比之下,过渡曲线算法的表现具有明显的改进:它较为准确地找到了各数据子类的近似类边缘表示集,所获得的降维结果成功地将不同类区分开来,较为准确地保持了原非连通数据集的类间关系,同时也良好地恢复了各个非连通类的圆形的本质维形态.这些结果均验证了过渡曲线算法的有效性.

4.2 应用到具有不同分布密度的3类非连通流形数据集

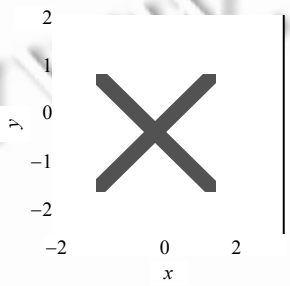
第 2 组实验旨在测试新算法在具有不同类分布密度的非连通流形数据集上的效能.我们采用的测试数据集分布在如图 5(a)所示的 3 类非连通流形上,其中:各流形类均为形状相同的半球面流形;3 个非连通类(如图 5(b)所示)的数据个数分别为 200,700 和 1 200,它们由相应流形类上的均匀分布产生.图 5(c)、图 5(e)分别展示了采用分解-整合算法与新的过渡曲线算法(邻域尺寸 k=5)所得的降维结果.图 5(d)则展示了过渡曲线算法获得的近似类边缘点集.



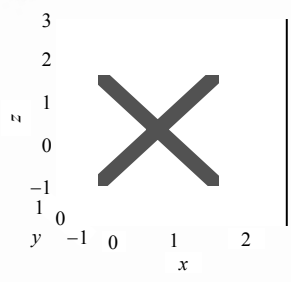
(a) 3-Cluster disconnected manifold
(a) 3 类非连通流形



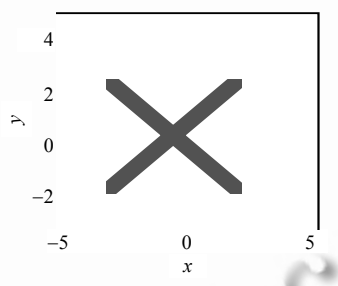
(b) 2 100 data points randomly distributed on the manifold
(b) 随机分布在此流形上的 2 100 个数据点



(c) Performance of D-C algorithm
(c) 分解-整合算法的表现



(d) Edge points found by transition curve method
(d) 过渡曲线方法寻找到的边缘点



(e) Performance of the transition curve method
(e) 过渡曲线方法的表现

Fig.5 Comparative performance of transition curve method and D-C algorithm for data sets lying on 3-cluster disconnected manifold with different densities on 3-cluster disconnected manifold with different densities

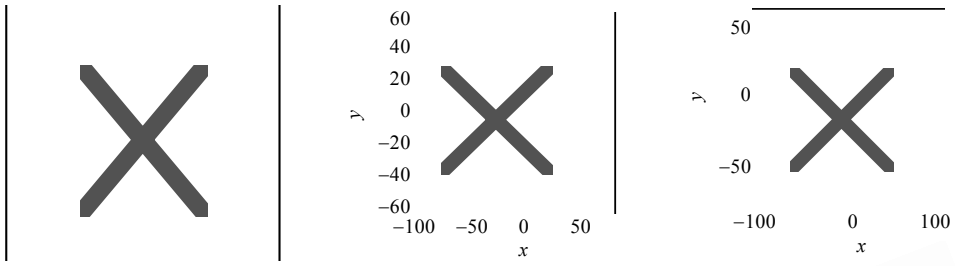
图 5 过渡曲线算法与分解-整合算法在 3 类不同密度的非连通流形数据集上的对比表现

从图 5 中可以观察到,新算法的表现优于分解-整合算法.特别地,该算法成功地恢复了非连通流形类的类间关系,从而完全克服了分解-整合算法中存在的边缘问题.这不仅进一步验证了所提出算法的有效性,而且证明了其降维结果对于类密度差异的不敏感性.

4.3 应用到3类图像数据集

该组实验的测试对象是 3 类图像数据集,每类分别包含 100 张小熊、螃蟹和蜜蜂不同姿势的图像(如图 6(a)所示).我们将其转换为黑白图像,并将图像像素值归一化.这些图像均为 50 像素×60 像素,因此可表示为一个 3 000 维空间中的数据集.对其分别应用分解-整合算法与新的过渡曲线算法(邻域尺寸 k=4),可得到如图 6(b)、

图 6(c)所示的结果.可以看出,新的过渡曲线算法得到的降维结果并没有出现分解-整合算法中的异常混叠现象.也就是说,新算法成功地恢复了各非连通流形类的类间关系.本组实验的结果进一步验证了所提出算法对分解-整合算法的效果改善.



(a) Representational images from 3-cluster image data sets
(a) 3 类图像数据集的代表图像

(b) Performance of D-C algorithm
(b) 分解-整合算法的表现

(c) Performance of the transition curve method
(c) 过渡曲线方法的表现

Fig.6 Comparative performance of transition curve method and D-C algorithm on 3-cluster image data set

图 6 过渡曲线算法与分解-整合算法在 3 类图像数据集上的对比表现

4.4 应用到带噪声的3类非连通流形数据集

最后一组实验旨在检验过渡曲线算法对数据集噪声干扰的敏感性.采用的数据集是将第 4.2 节中的实验数据集中每一元素加入[−0.1,0.1]的随机噪声,从而产生一组带噪声的 3 类非连通流形数据集,如图 7(b)所示(分布于 3 个半球面的数据个数分别为 200,700 和 1 200).对其分别应用分解-整合算法与过渡曲线算法(邻域尺寸 k=5),可得到如图 7(c)、图 7(e)所示的结果.从图 7 中可以观察到,加入噪声后,算法得出的结果与未加噪声时不存在明显差异.这进一步证明了所提出的过渡曲线算法对噪声干扰具有不敏感性.

上述 4 个数值实验说明,本文所提出的过渡曲线算法能够建立类间的有效连接关系,解决了分解-整合算法中出现的边缘问题,对于非连通数据有着良好的低维特征提取效果,且对噪声干扰具有不敏感性.

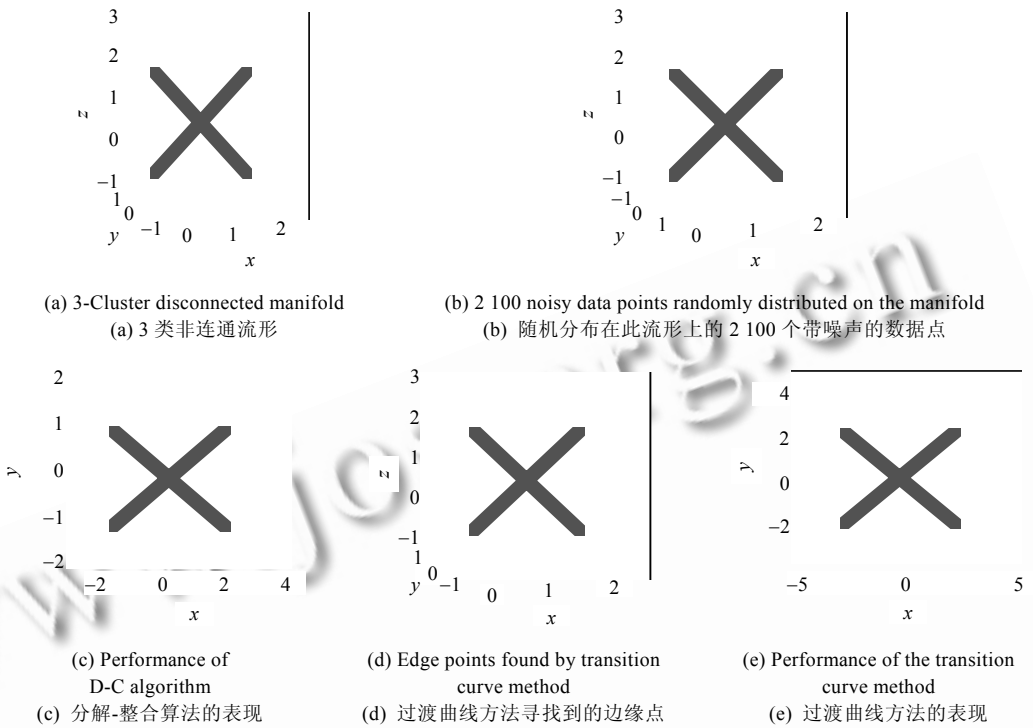
5 总 结

本文针对分解-整合算法处理非连通数据时出现的边缘问题,即当不同数据子类间最短距数据对位于类的内部而不是边缘时分解-整合算法失效的情况,提出了一种能够更准确地建立类间有效连接关系的方法——过渡曲线算法.该算法通过构建不同类边缘最近点之间的平滑过渡曲线,实现了对相应类中心间测地距离的准确估计,进而建立了类间的有效连接关系,保证了整体数据集良好的全局降维形态.一系列人工与图像非连通流形数据集上的数值实验结果验证了该算法的有效性.特别地,该算法完全解决了分解-整合算法中出现的边缘问题.

过渡曲线算法虽然能够解决分解-整合算法中出现的边缘问题,具有较好的特征提取能力,但它也存在一些问题.其中一个问题是,当数据中存在较大的噪声且将非连通数据变为连通数据时,所提方法会出现与分解-整合方法类似的异常表现[17].这一问题的出现是由于在通过数据邻域图的连通支寻找非连通类时,这些噪音点的存在会导致本不连通的两类变为连通状态,从而导致非连通类的识别错误.当遇到这种问题时,实际可采用更为有效的聚类方法(如 Mean-shift 等方法)实现有效的非连通类识别,然后进行之后的步骤,这样可以使该算法进一步适应于噪声干扰的情况.然而,这样的策略无疑会增加算法的计算效率与复杂性.我们将在继续的研究中致力于设计能够很好地嵌入于该算法框架的快速非连通类识别方法,从而使算法具有更好的鲁棒性与更广泛的有效应用范畴.

另一个需要特别强调的问题是计算效率问题.在本文涉及的 4 个实验中,所提出算法与分解-整合算法所需

的计算时间分别为 982.4074s 与 660s,6759s 与 739s,5.0210s 与 2.6938s,5938s 与 712s.显然,所提出算法尽管显著地改善了分解-整合算法的降维表现,却同时增加了计算时间.根据第 3.3 节的计算复杂度分析可知,这主要是由于过渡曲线算法计算最短路径树时采用的 Floyd 算法的较高复杂度所导致的.目前已存在一些针对最短路径树的效率改进算法[20].我们将在未来的工作中继续尝试更为先进的计算方法,从而对所提出算法的计算效率作进一步的改进与提高.



(a) 3-Cluster disconnected manifold
(a) 3 类非连通流形

(b) 2 100 noisy data points randomly distributed on the manifold
(b) 随机分布在此流形上的 2 100 个带噪声的数据点

(c) Performance of D-C algorithm
(c) 分解-整合算法的表现

(d) Edge points found by transition curve method
(d) 过渡曲线方法寻找到的边缘点

(e) Performance of the transition curve method
(e) 过渡曲线方法的表现

Fig.7 Comparative performance of transition curve method and D-C algorithm for data sets lying on 3-cluster disconnected manifold with noise

图 7 过渡曲线算法与分解-整合算法在带噪声的 3 类非连通流形数据集上的对比表现

References:

[1] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500):2319−2323. [doi: 10.1126/science.290.5500.2319]

[2] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290(5500):2323−2326. [doi: 10.1126/science.290.5500.2323]

[3] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 2003,15(6): 1373−1396. [doi: 10.1162/089976603321780317]

[4] Zhang ZY, Zha HY. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. Technical Report, CSE-02-019, Pennsylvania: Pennsylvania State University, 2002.

[5] Bachmann CM, Ainsworth TL, Fusina RA. Exploiting manifold geometry in hyperspectral imagery. IEEE Trans. on Geoscience and Remote Sensing, 2005,43(3):441−454. [doi: 10.1109/TGRS.2004.842292]

[6] Lee JG, Zhang CS. Classification of gene-expression data: The manifold-based metric learning way. Pattern Recognition, 2006, 39(1):2450−2463. [doi: 10.1016/j.patcog.2006.05.026]

[7] Venna J, Kaski S. Local multidimensional scaling. Neural Networks, 2006,19(6-7):889−899. [doi: 10.1016/j.neunet.2006.05.014]

[8] Yang L. Sammon's nonlinear mapping using geodesic distances. In: Kittler J, Petrou M, Nixon M, eds. Proc. of the 17th Int'l Conf. on Pattern Recognition (ICPR 2004). Washington: IEEE Computer Society Press, 2004. 303−306.

[9] Yang L. Building *k*-edge-connected neighborhood graph for distance-based data projection. Pattern Recognition Letters, 2005, 26(13):2015−2021. [doi: 10.1016/j.patrec.2005.03.021]

[10] Yang L. Building *k* edge-disjoint spanning trees of minimum total length for isometric data embedding. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(10):1680−1683. [doi: 10.1109/TPAMI.2005.192]

[11] Yang L. Building *k*-connected neighborhood graphs for isometric data embedding. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006,28(5):827−831. [doi: 10.1109/TPAMI.2006.89]

[12] Yang L. Tetrahedron mapping of points from *N*-space to three-space. In: Kasturi R, Laurendeau D, Suen C, eds. Proc. of the 16th Int'l Conf. on Pattern Recognition (ICPR 2002). Washington: IEEE Computer Society Press, 2002. 343−346.

[13] Yang L. *k*-edge connected neighborhood graph for geodesic distance estimation and nonlinear data projection. In: Kittler J, Petrou M, Nixon M, eds. Proc. of the 17th Int'l Conf. on Pattern Recognition (ICPR 2004). Washington: IEEE Computer Society Press, 2004. 196−199.

[14] Yang L. Distance-Preserving mapping of patterns to 3-space. Pattern Recognition Letters, 2004,25(1):119−128. [doi: 10.1016/j.patrec.2003.09.009]

[15] Yang L. Distance-Preserving projection of high dimensional data. Pattern Recognition Letters, 2004,25(2):259−266. [doi: 10.1016/j.patrec.2003.10.010]

[16] Yang L. Distance-Preserving projection of high dimensional data for nonlinear dimensionality reduction. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004,26(9):1243−1246. [doi: 10.1109/TPAMI.2004.66]

[17] Meng DY, Leung Y, Fung T, Xu ZB. Nonlinear dimensionality reduction of data lying on the multicluster manifold. IEEE Trans. on Systems, Man and Cybernetics—Part B: Cybernetics, 2008,38(4):1111−1122. [doi: 10.1109/TSMCB.2008.925663]

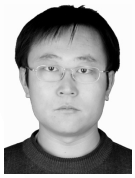
[18] Vardi Y, Zhang CH. The multivariate L_1-median and associated data depth. Proc. of the National Academy of Sciences, 2000,97(4): 1423−1426.

[19] Knuth DE. The Art of Computer Programming, Vol.3: Sorting and Searching. 2nd ed., Reading: Addison-Wesley, 1973. 144−148.

[20] Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms. 2nd ed., Cambridge: MIT Press, 2002. 614−618.

古楠楠(1985－),女,河南南阳人,博士生,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.



孟德宇(1978－),男,博士,讲师,主要研究领域为非线性降维,模式识别.