一种基于 Biharmonic 样条插值的流形学习算法*

顾艳春[1,2]，马争鸣[2]，梁宇滔[2]
（1. 佛山科学技术学院电子与信息工程学院，广东 佛山 528000；
2. 中山大学信息科学与技术学院，广东 广州 510220）

摘 要：作为一种有效的非线性降维方法，流形学习在众多领域吸引了广泛的关注并取得了长足的发展。但当样本点较为稀疏时，样本点的局部邻域很难满足流形学习局部同胚的前提条件，此时流形学习算法往往效果变差甚至失效。一种有效的解决方法是增加一些新的插值点。但已有的插值方法选取的插值点与原样本点均存在线性关系。从线性代数的理论来说，由插值点和原有邻域点张成的线性子空间与原有邻域点张成的子空间是一样的，因此，不会改善线性逼近的误差。而且，插值点没有反应出流形的本质结构和特征，从理论上背离了数据降维的目的。为此，提出了一种基于 Biharmonic 非线性插值技术的流形学习算法 BbMLA。由于是从高维曲面逼近的角度非线性的选择插值点，插值出的样本点不会被原有邻域点线性表示，从而能更好的重构原样本点。将 BbMLA 应用到多个数据集后，图示说明了插值点能够有效的改善邻域内的样本点结构，同时插值后的流形学习算法具有较好的有效性和稳定性。

关键词：流形学习；数据降维；曲面拟合；插值
中图分类号：TP391.4 **文献标志码**：A **文章编号**：0529－6579（2013）05－0082－10

A Manifold Learning Algorithm Based on Biharmonic Spline Interpolation Technique

GU Yanchun[1,2], *MA Zhengming*[2], *LIANG Yutao*[2]
（1. School of Electronics and Information Engineering, Foshan University, Foshan 528000, China;
2. School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China）

Abstract: As an effective non-linear dimension reduction method, manifold learning has attracted widespread attention and made great progress. But when sample points are not dense, these algorithms often become worse or even failed just because the points in some neighborhoods do not meet the requirement of local homeomorphism. An effective solution to this question is to increase some new interpolation points. Unfortunately, the points selected by existing interpolation methods nowadays are all linear with the original sample points. From the theory of linear algebra, the subspace spanned by the interpolation points and the original neighbors is the same as the subspace spanned by the original ones; therefore, the interpolation points will not improve the linear approximation error either. Moreover, the interpolation points have no consideration to the native structure and characteristics of the manifold, which deviates from the purpose of data dimensionality reduction. To this end, a new manifold learning algorithm based on a non-linear interpolation method called Biharmonic is proposed. Experimental results demonstrate the improvement of the neighborhood structure. The effectiveness and stability of this algorithm are further confirmed by applying it to the classical manifold learning algorithms.

Key words: manifold learning; dimensionality reduction; surface fitting; interpolation

---

* **收稿日期**：2013－02－26
基金项目：广东省自然科学基金资助项目（8452800001001086）；佛山科学技术学院资助项目（2010X063）
作者简介：顾艳春（1981 年生），男；**研究方向**：机器学习、图像处理；E-mail：gufenger@fosu.edu.cn

流形学习是一种有效的非线性降维方法。近年来，流形学习方法在数据挖掘、机器学习、图像处理和计算机视觉等多个研究领域吸引了广泛的关注。典型的流形学习方法有 Isometric Feature Mapping（ISOMap）[1]、Locally Linear Embedding（LLE）[2]、Hessian Eigenmaps（HLLE）[3]、Local Tangent Space Alignment（LTSA）[4]、Laplacian Eigenmaps（LE）[5]等。这些算法具有一个共同的特征：找出每个数据点周围的局部性质，并将这些局部性质信息映射到一个低维空间中。显然，局部几何结构信息的保持和恢复程度决定了流形学习算法的优劣。在获取流形的局部信息时，流形学习算法假定流形在一个很小的范围内，局部同胚于一个欧式空间的一个连通开集，这就决定了流形学习算法在选择邻域时，要尽可能保证邻域内的点满足局部同胚条件。而当样本点较为稀疏时，邻域内的样本点很难保持局部同胚条件，从而导致上述流形学习算法在处理稀疏数据集时会造成较大的误差，甚至失效。

针对流形学习算法无法有效处理样本点稀疏的问题，目前主要有三种解决方法。一类是根据样本点的稀疏程度，自适应的改变邻域大小，从而尽可能的使邻域内的样本点满足同胚条件[6-8]。在样本点比较稀疏时，此种方法会使得邻域相对较小，这很容易造成在将局部坐标信息排列成全局坐标时由于交叠不够而使算法效果难以令人满意的现象。第二种方法是改变邻域内的局部信息选取方式，例如，Wu 等[9]求取邻域时，首先对样本点集做预处理，去除样本集中的“短路”边，然后利用最短路径算法迭代出样本点间的测地线距离来选取邻域；Song 等[10]通过最小化邻域内样本点间的梯度值来实现高维数据的局部线性逼近。此类方法计算复杂，受流形本身形状影响较大从而稳定性较差。另一类比较有效的做法是添加一些虚拟样本点，使得样本点相对稠密，从而改善降维效果。例如，Zhan 等[11]利用样本点到邻域内其他两个点组成连线的垂足来添加样本点，提出了基于邻域线的 LLE 算法。但该方法并没有考虑流形本身的性质和曲率等因素对降维的影响，添加的虚拟样本点与原样本点之间为线性关系，因此，效果有限，只能针对特定的流形。

为此，我们提出了一种新的基于 Biharmonic 样条插值的流形学习算法 BbMLA，通过非线性的获取插值点来有效改善邻域内样本点的稠密程度，同时插值点又能忠实的保持流形本身的结构和性质。在本文提到的算法中，我们利用 Biharmonic 样条插值算法[12]，首先在样本点的各邻域内做曲面插值，而后根据流形本身的特点和性质，从插值曲面中非线性的选取插值点；然后利用这些插值点与原样本点一起组成新的样本点集，并求取其低维坐标；最后，将原样本点的坐标抽离和表示出来，最终得到原样本点集的低维坐标值。通过对插值点的图示，我们说明了算法得到的插值点与流形的本质结构较为匹配，而且插值点考虑了流形的密度和曲率等因素。在将本文提到的插值算法应用到经典的流形学习算法如 LTSA、LLE 后，实验结果证实了我们的算法的有效性和稳定性。

1　流形学习中的样本点稀疏问题

流形学习的方法可以分为两类：一类是全局方法（如 Isomap），另一类是局部方法（如 LLE、LE、HLLE、LTSA 等）。由于局部方法只需要考虑流形临近点之间的关系，无须要求流形所对应的低维空间为凸，且计算复杂度较低，因此局部方法有着更广泛的适用对象[13]。

局部保持的流形学习方法正是通过保持邻域内的局部近邻结构来构造全局低维表示，所以，邻域结构的表示和保持程度将直接影响最终的嵌入效果。在刻画流形的局部几何特性时，需要尽可能的保证局部邻域能够同胚于欧氏空间的一个连通开集。显然，邻域越小，邻域的低维结构越明显，近邻结构越容易忠实保持。另一方面，邻域之间需要有足够的交叠以保证全局排列时有足够的联系，这又使得邻域不能过小。这种矛盾一直伴随着流形学习算法，当样本点比较稀疏时，邻域内的局部同胚条件更加难以保持，这就造成了目前绝大多数流形学习算法在样本点较为稀疏时的失效。

图 1 标示了样本点稀疏程度不同时某一点的邻域结构，稀疏程度不同时，邻域内的线性程度也不同。其中，采样点数据来自于 Swiss Roll，星点为从 Swiss Roll 随机选择的某一个样本点，实心点为采样点为 800 个点时的邻域点，空心圆点为采样点为 100 时的邻域点（邻域值为 8，邻域包括自身点）。显然，当采样点比较密集时，我们可以认为其局部同胚于一个欧式空间，此时，样本点在由邻域点线性表出时的误差较小。而当采样点较为稀疏时，局部同胚条件较难保持，此时刻画和表示的邻域内的结构信息，便带有较大的误差，从而导致算法效果变差乃至失效。

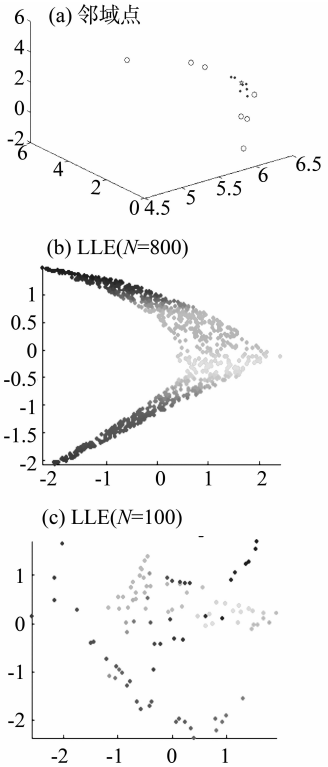


图 1 样本点稀疏程度不同时的邻域点集

Fig. 1 Selected neighborhood with different denseness of the sample points

对于流形学习算法不能有效处理稀疏样本点集的问题，目前常用的解决方法，是通过插值增加一些新的样本点以使样本点密集。具体来说，是利用样本点有限的邻域点插值出新的邻域点，然后再由这些原有的邻域点和插值出的新的邻域点张成一个线性子空间去逼近原样本点。例如，NL^3E 方法利用样本点到邻域内其他两个点组成连线的垂足来添加样本点。

这类插值方法一定程度上改善了样本点稀疏时的算法效果。但是这些方法都采用线性插值的方法去产生新的样本点，也就是说，新的邻域点都是原有邻域点的线性组合，从线性代数的理论来说，由插值点和原有邻域点张成的线性子空间与原有邻域点张成的子空间是一样的，因此，也不会改善线性逼近的误差。而且，插值点并没有反应出流形的本质结构和特征，从理论上背离了数据降维的目的。为此，我们利用 Biharmonic 样条插值法非线性的获取插值点。此时，插值出的样本点不会被原有邻域点线性表示，也就是说，新插值出的样本点不会落在原邻域点张成的线性子空间里，因此，由插值点和原有邻域点张成的线性子空间是原有邻域点张成子空间的真扩展。如图 2 所示，线性插值方法是从原邻域点张成的子空间内选取合适的样本点作为插值点，而非线性插值方法是从高维空间逼近的角度选取插值点，由这个子空间去逼近样本点会更有效的减少逼近误差。另外，由于是从邻域内曲面重建中非线性的获取插值点，插值出的点能够更好的反映流形的曲面性质而不是平面性质，从而更好的保持和揭示了流形的本质特征。

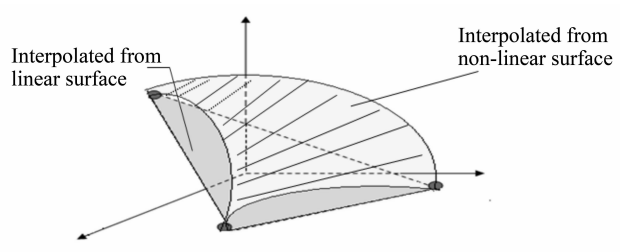


图 2 线性插值与非线性插值方法选取插值点的不同

Fig. 2 The difference of interpolation points chosen by linear and non-linear interpolation method

2 基于 Biharmonic 样条插值的流形学习算法

算法主要用于解决样本点稀疏问题，对于稀疏样本点，根据其本质结构特点，利用 Biharmonic 样条插值方法在样本点的邻域内构造插值曲面，并从插值曲面中选取一定数目的样本点作为插值点。而后，利用这些插值点与原样本点一起作为新的样本点集。待利用各种经典的流形学习算法求得样本点的全局低维坐标后，取出原样本点集的低维坐标。

2.1 Biharmonic 样条插值

解决样本点稀疏问题的有效方法之一，是根据流形特点，添加新的插值点。为了合理的构造插值点，我们首先需要用一个光滑的曲面来逼近这些无规则的散乱抽样数据点，即曲面拟合问题；然后从拟合的曲面上选取合适的点作为新样本点。流形上散乱数据的曲面拟合，其难点在于，如何得到邻近点间正确的拓扑连接关系，而正确的拓扑连接关系将有效的揭示散乱数据集所蕴涵的本质形状和拓扑结构。

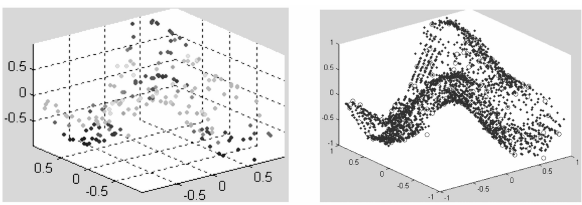
在众多的曲面拟合算法中，Biharmonic 样条插值方法[12]是一种效果较好的曲面构造方法。与其他曲面拟合算法如双三次样条插值和 B 样条插值算法相比，Biharmonic 样条插值方法拟合的曲面较为光滑，局部性能较好，能够根据散乱数据点发现和保持曲面的本质结构和特征，而且算法计算量较小，效率较高[14]。

Biharmonic 样条可以对散乱分布的数据进行曲面插值。插值产生的曲面是以各数据点为中心的 Green 函数的线性组合[12]。Biharmonic 方程在不同维空间中的解就是不同维的 Green 函数。对于 D 维空间中散乱分布的 K 个控制点 $x_k, k = 1,2,\cdots,K$，Biharmonic 样条 D 维插值问题转化为对公式（1）的求解

$$\nabla^4 W(X) = \sum_{k=1}^{K} \alpha_k \delta(X - x_k) \qquad (1)$$

其中，∇^4 为 Biharmonic 算子，δ 为单位冲击函数，$W(X)$ 为 X 位置处的值。

图 3 为在 Twin Peaks 样本集上做 Biharmonic 样条插值方法后从插值曲面上选取部分插值点的图示。



(a) Twin Peaks样本集　　(b)插值后的样本集

图 3　Biharmonic 样条插值方法选取的插值点

Fig. 3　Effect by Biharmonic spline interpolation algorithm

其中，图 3（b）中空心圆点为原样本点（原样本点数目为 200），实心点为从插值曲面上选取的部分插值点。由图 3 可以看出，Biharmonic 样条插值法得到的曲面，与原流形曲面较为匹配，比较忠实的体现了原流形的特征和结构，并且，插值函数本身动态的考虑了流形的曲率和密度变化等因素。

2.2　插值点的选取

插值点的选取是指从插值曲面上，取合适的点作为新的样本点，并放入样本集中。为了提高插值精度，我们要产生尽可能多的点来逼近原流形曲面。但是，过多的插值点参与到流形学习算法会很严重的影响算法的效率。而且，按照文献［11］的理论，为每一个样本点插入不少于其维数的插值点即可。从直观上考虑，样本点稀疏处，应选择较多的插值点，曲率较大处，应选择较多的插值点。通常，插值点的选取有两种方法，一种为从插值曲面上均匀采样，另一种是根据流形及样本集本身的特点（如样本稠密度和曲率的不同）来抽取样本点。由于 Biharmonic 样条插值法在插值时，已经考虑了流形局部的密度和曲率等因素，因此，我们只需要选取合适数目的样本点作为插值点。

选出的插值点，有两种利用方式。一种是让插值点和原样本点集组合起来，一起参与流形学习算法；另一种是只利用局部范围内的插值点，来修正每个样本点的局部坐标，但这种方法，不能有效的处理邻域间交叠不够的问题。本文中，我们选取第一种方法。

2.3　BbMLA 算法框架

为了解决流形学习算法不能有效处理稀疏样本点的问题，针对线性插值方法的不足，我们提出了基于 Biharmonic 样条插值的流形学习算法，即 BbMLA 算法。算法首先选取生成插值点的邻域，然后利用 Biharmonic 样条插值方法在样本点的邻域内构造插值曲面，并从中选取一定数目的样本点作为插值点。选取插值点后，将插值点并入原样本点集中并利用经典的流形学习算法获取新的样本点集的低维坐标；而后，将原样本点集分离出来从而得到最终的原样本点集得低维坐标。算法过程如表 1 所示：

算法中，X 为原始样本点集，V 为新插入点的样本集，L 为 Biharmonic 样条插值时的邻域选取参数，为了保证邻域内的点满足同胚条件，可根据样本点密度或曲率变化动态调整 L。λ 为从重建曲面中采样时选取的新样本点个数，可为每一个样本点选取不同个数的插值点。*MLA* 为调用流形学习算法得到低维坐标，可选择多种流形学习算法如 LLE、ISOMAP、LE、HLLE、LTSA 等。

表 1　BbMLA 算法过程

Table 1　Pseudo-code of BbMLA

T = BbMLA（X, K, L, d）
输入：
X：原始样本点集 $\{x_1, x_2, \cdots, x_N\} \in \mathbf{R}^D$
L：插值时的邻域选取参数
K：流形学习算法中的邻域选取参数
d：低维值
输出：
T：原样本点集的低维坐标
过程：
1 $V = \varnothing$
2FOR i = 1, 2, …, N DO
3 确定 x_i 的最近 L 个邻域点
4 根据 x_i 和 L 个邻域点做 Biharmonic 样条插值
5 选取插值点，新采样的点集合记为 V_i
6 $V = V \cup V_i$
7END FOR
8 V = unique（V）% 去除 V 中相同的点
9 $X' = X \cup V$
10 获取插值点个数，记为 λ
11 $T' = \mathrm{MLA}\left(X', d, \left(\frac{\lambda}{N} + 1\right) \times K\right)$
12 T = get$X$Cor(T', X)

3 实验及分析

为了更好的比较和分析插值前后算法的效果差异，我们设计了以下实验。实验中，CPU 频率为 1.86GHz，内存容量为 2GB，运行环境为 Matlab 7.0。

3.1 插值点效果对比

我们首先对线性插值和非线性插值方法得到的插值点的效果进行了对比。

图 4 标示了样本点数为 200，邻域值取 8 时的插值点效果对比图，其中(a),(a′),(a″),(a‴)为原始样本点集图，(b),(b′),(b″),(b‴)为线性插值（NL^3E 为例）后的样本点集图，(c),(c′),(c″),(c‴)为 Biharmonic 插值算法得到的样本点集图。(b),(b′),(b″),(b‴)、(c),(c′),(c″),(c‴)图中红色圈点为原始样本点，蓝色实点为选取的插值点（并非改变原采样点的颜色向量，在此只是为了区分原采样点和新插值点）。由图 4 可以看出，通过非线性插值方法插值后的样本点集，较好的保持了流形的本质特征。与线性插值方法相比，得到的插值点更加忠实于流形本身。

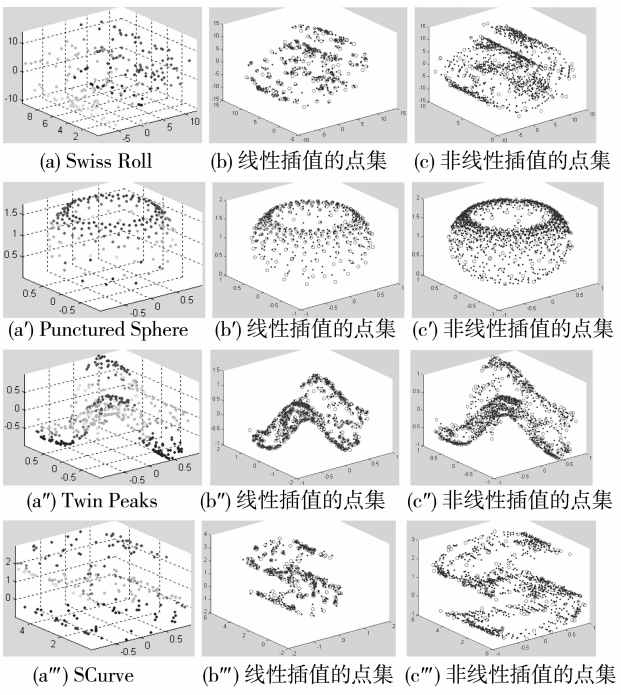


图 4 插值点效果对比（$N=200$，$L=8$）

Fig. 4 Interpolation points by linear and nonlinear methods ($N=200$, $L=8$)

3.2 插值前后流形学习算法效果对比

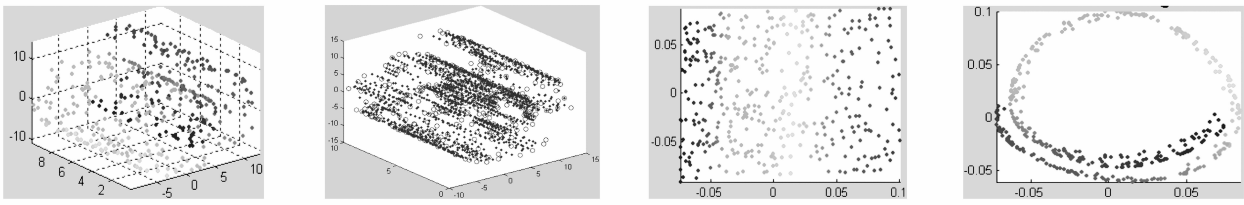
插值算法可以应用到数据集。我们首先 Mani 程序中的数据集（Swiss Roll、Punctured Sphere 和 Twin Peaks），Mani 数据集是一种在流形学习中广泛使用的数据集，可以方便的从 http://www.math.ucla.edu/~wittman/mani/index.html 处免费下载。

图 5 标示了在原样本点数目为 400，邻域取 8 时，原 LTSA 算法的效果图以及相应的在插入插值点后的算法效果图。其中(a),(a′),(a″)为原始流形采样图；(b),(b′),(b″)为插值后的采样图，其中红色圈点为原始样本点，蓝色实点为选取的插值点；(c),(c′),(c″)为原 LTSA 算法效果图；(d),(d′),(d″)为插值后的 LTSA 算法效果图。由图 5 可以看出，插值后的算法效果跟原始算法效果相比基本相同，这主要是因为原始采样点比较密集，邻域内基本满足局部同胚关系，故虽然插入的样本点基本保持了流形本身的形状且使得样本点集更为稠密，但对整体效果的影响有限。

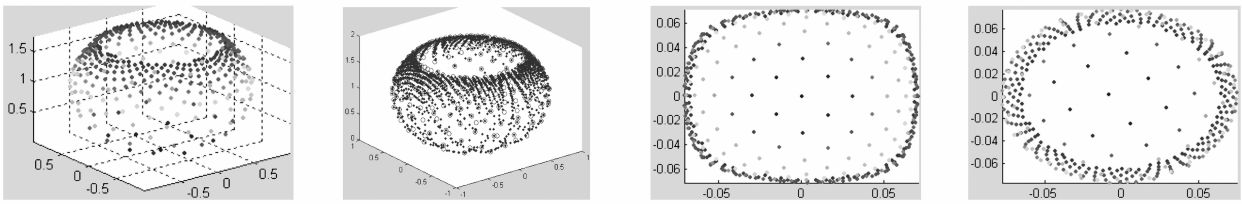
图 6 标示了在原样本点数目为 200，邻域取 8 时，原 LTSA 算法的效果图以及相应的在插入插值点后的算法效果图。由图 6 可以看出，原始的 LTSA算法得到的降维图，效果已显著下降，这主要是因为原始采样点比较稀疏，邻域值取 8 时，邻域内的样本点已难以满足局部同胚关系，故得到的降维效果欠佳。插值后，新插入的样本点较好的保持了原流形的本质结构，邻域内的样本点重新较好的满足了局部同胚关系，故插值后的算法取得了较好的效果。

图 7 标示了在原样本点数目为 100，邻域取 8 时，原 LTSA 算法的效果图以及相应的在插入插值点后的算法效果图。由图 7 可以看出，原始的算法已基本失效，而插值后的算法仍保持了较好的效果。这主要是由于插值前的样本非常稀疏，局部很难保持同胚条件，而插值后的新的样本点集有效的克服了这一现象。

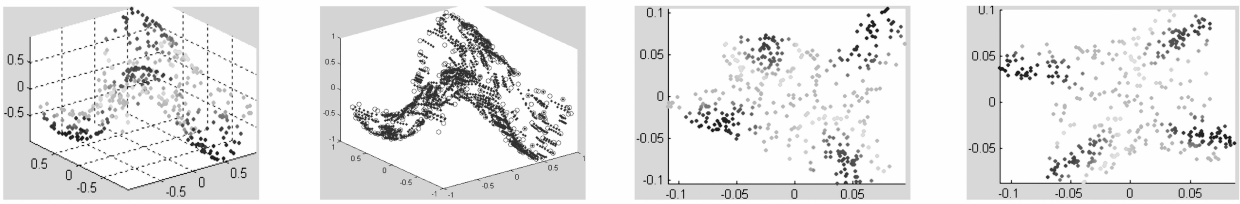
当样本点较为稀疏时，为了保持局部同胚关系，我们可适当的降低邻域值。但太小的邻域值会使得邻域间缺乏足够的交叠，从而使得全局排列受到较大影响，甚至导致算法失效。图 8 标示了在原样本点数目为 100，邻域取 4 时，原 LTSA 算法的效果图以及相应的在插入插值点后的算法效果图。由图 8 可以看出，原 LTSA 算法由于邻域间缺乏足够的交叠，导致算法失效，而插值后的算法，由于添加了样本点，使得邻域间的同胚关系得到较好保持的同时，也增强了邻域间的交叠关系，从而使得算法效果有了较为明显的改善。



(a) Swiss Roll　(b) 插值点　(c) 原LTSA　(d) 插值后的LTSA



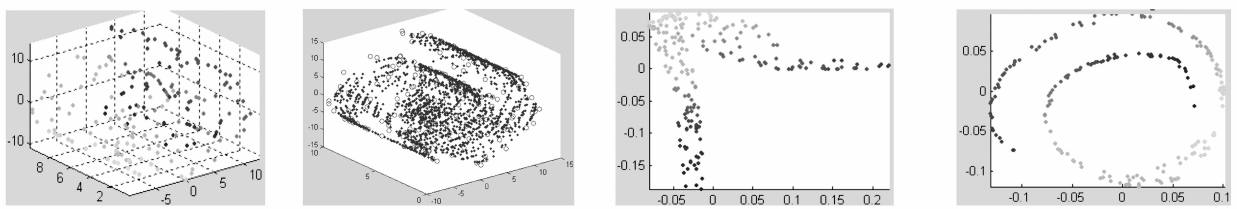
(a′) Punctured Sphere　(b′) 插值点　(c′) 原LTSA　(d′) 插值后的LTSA



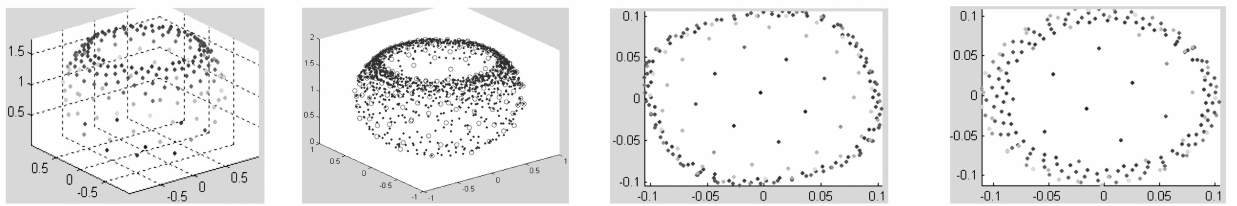
(a″) Twin Peaks　(b″) 插值点　(c″) 原LTSA　(d″) 插值后的LTSA

图 5　Mani 数据集插值前后 LTSA 算法效果对比图（$N=400$，$K=8$）

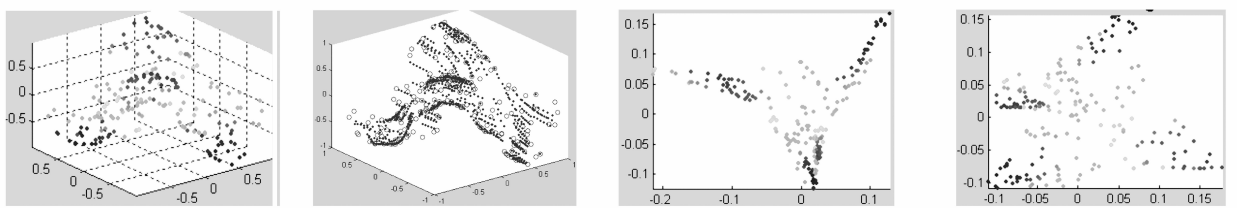
Fig. 5　Processed results by LTSA with the interpolation algorithm（$N=400$，$K=8$）



(a) Swiss Roll　(b) 插值点　(c) 原LTSA　(d) 插值后的LTSA



(a′) Punctured Sphere　(b′) 插值点　(c′) 原LTSA　(d′) 插值后的LTSA



(a″) Twin Peaks　(b″) 插值点　(c″) 原LTSA　(d″) 插值后的LTSA

图 6　Mani 数据集插值前后 LTSA 算法效果对比图（$N=200$，$K=8$）

Fig. 6　Processed results by LTSA with the interpolation algorithm（$N=200$，$K=8$）

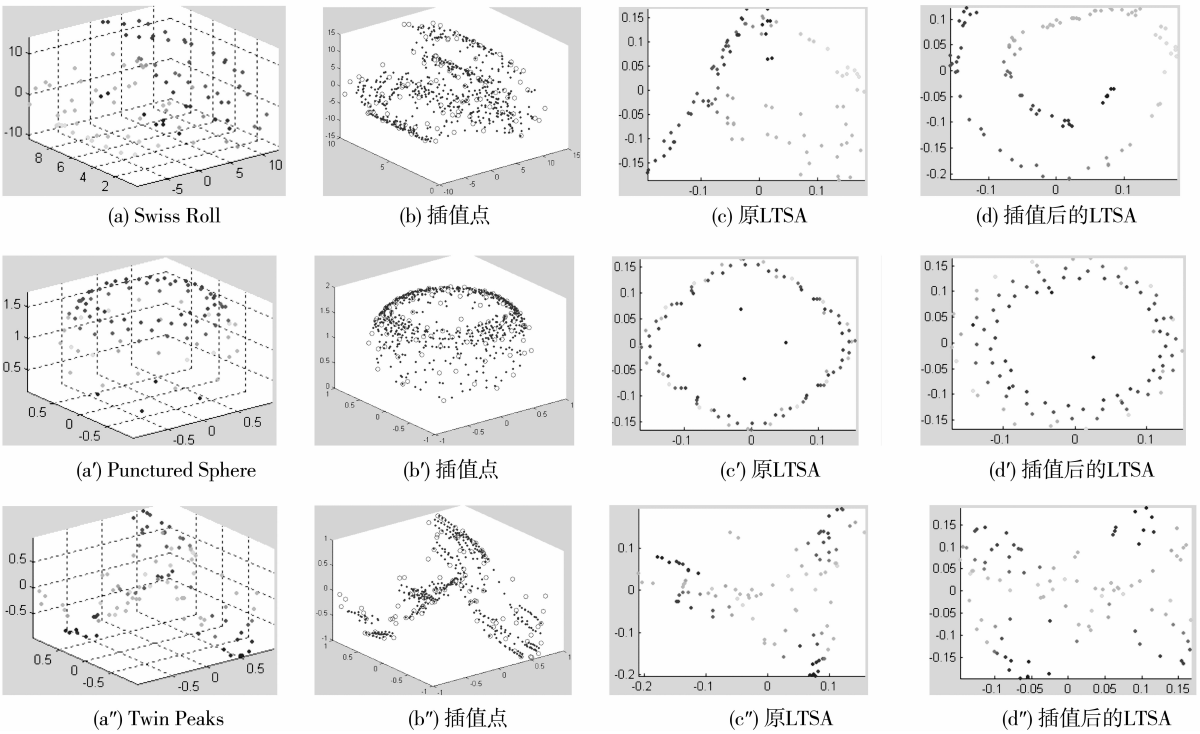


图 7　Mani 数据集插值前后 LTSA 算法效果对比图（$N=100$，$K=8$）

Fig. 7　Processed results by LTSA with the interpolation algorithm（$N=100$，$K=8$）

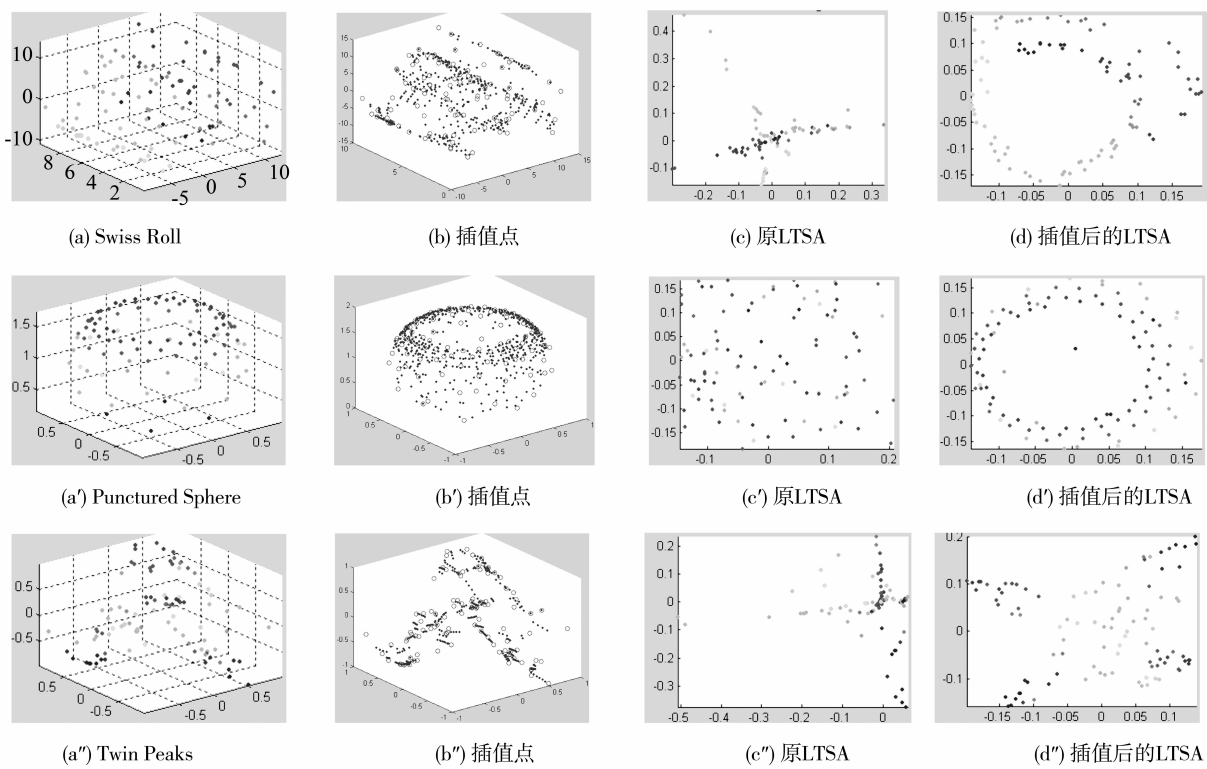


图 8　Mani 数据集插值前后 LTSA 算法效果对比图（$N=100$，$K=4$）

Fig. 8　Processed results by LTSA with the interpolation algorithm（$N=100$，$K=4$）

图 9 标示了插值前后在 SCurve 数据集上 LTSA 算法效果对比图。与在 Mani 数据集上基本类似，当样本点较为稀疏时，插值算法取得了较好的效果。多个数据集上的效果，说明了我们的算法的健壮性和鲁棒性。

我们的插值算法也适用于其他经典流形学习算法如 LLE、HLLE、Diffusion Maps 等。图 10 标示了插值前后 LLE 算法效果对比图。由图 10 可以看出，我们的插值算法在 LLE 等其他流形学习算法中也取得了较好的效果。

同时，我们也做了其他一些高维数据集的实验，如 Frey Faces 和 Handwritten Digits 等。算法同样能取得较好的效果。

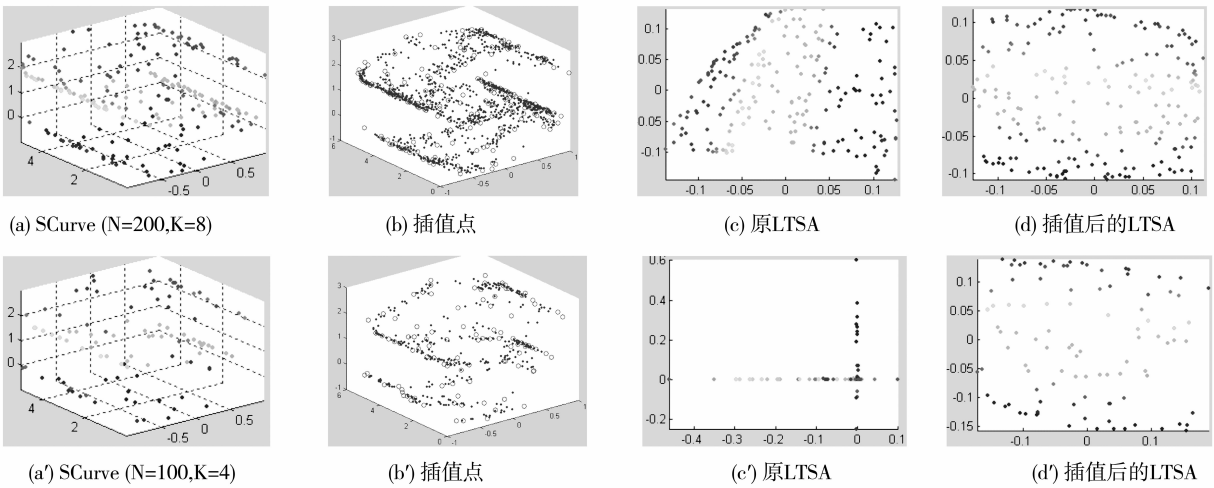


图 9　SCurve 数据集插值前后 LTSA 算法效果对比图

Fig. 9　Processed results by LTSA to SCurve with the interpolation algorithm

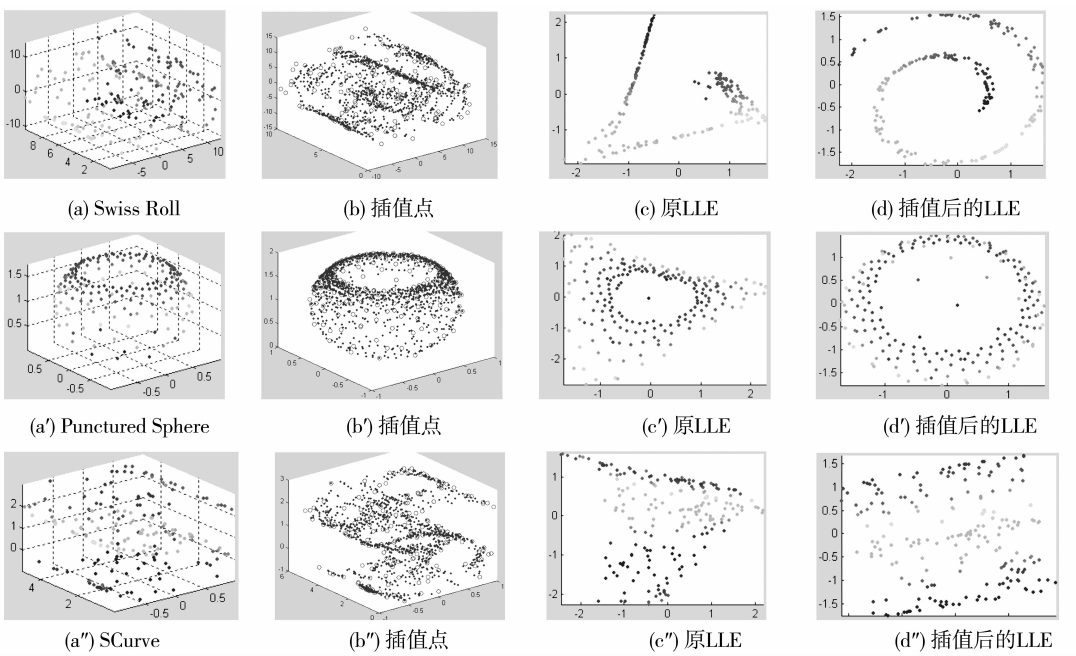


图 10　插值前后 LLE 算法效果对比图（$N=200$，$K=8$）

Fig. 10　Processed results by LLE with the interpolation algorithm（$N=200$，$K=8$）

3.3　参数调整及时间复杂度分析

将 BbMLA 算法应用到实际问题时，可以根据不同流形的特点，调整参数来获得更好的算法效果。在 BbMLA 算法中，主要有如下几个参数：

L，插值时的邻域选取参数。与流形学习算法用到的邻域选取参数 K 不同，L 用于设定 Biharmonic 样条插值时的邻域选取参数。对于稀疏样本点集，K 邻域方法取得的邻域点集很难满足局部同胚的条件，因此，通常有 $L < K$。另外，如同 K 可以为一个向量，即每个样本点可以有不同的邻域大小一样，我们也可以为每一个样本点构造局部插值曲面时选择不同的邻域，此时 L 为一向量。本文实验中，L 和 K 均设定为一常数值，其中：$L = \lfloor \frac{3}{4}K \rfloor$。

λ，插值点个数。从每个插值曲面选取的插值点数目可以不同，插值点数目越多，插值点便越能忠贞的体现流形本身的结构，但过多的插值点会大大增加算法运行的时间。而且，按照文献［11］的理论，为每一个样本点插入不少于其维数的插值点即可。

$\tilde{K}$，插值后流形学习算法的邻域选取参数，从直观上考虑，选取的插值点越多，满足局部同胚条件的邻域便越大，从而，可选择更大的邻域参数。本文算法中 $\tilde{K}$ 取做：$\tilde{K} = \frac{\lambda + N}{N}K$，即邻域大小随着插值点个数的多少动态调整。其他一些自适应邻域的选取方式同样适用于本算法中邻域的选取。

我们提出的算法中，由于需要对每个样本点做曲面插值和插值点的选择，并最终扩展了样本点集来参与流形学习算法，这导致算法的运行时间较长。表2中，我们比较了几种流形学习算法插值前后的运行时间（s），其中数据集取自 Swiss Roll 流形，采样点为200，邻域为8。由表2可以看出，插值算法和增加的插值点大大增加了算法的运行时间。可行的解决办法，一是选择合适的标志点而不是所有数据点的邻域来做曲面插值，二是选择插值点时在保持较好降维效果的同时尽可能选择较少的点；三是插值后的流形学习算法设定合适的邻域值，适当的减小邻域会降低算法的运行时间。

表2 不同插值点时几种流形学习算法运行时间对比

Table 2 Comparison of running time with different interpolation points

算法	插值前	$\lambda=500$	$\lambda=1\,000$	$\lambda=1\,500$	$\lambda=2\,000$
ISOMap	0.405 4	4.552 6	33.717 2	149.533 3	329.667 4
LLE	0.088 3	0.953 4	2.672 2	11.259 3	22.253 2
LTSA	0.084 8	0.478 8	0.981 2	2.464 6	5.310 7
HLLE	0.265 0	1.825 9	11.841 7	48.846 1	140.115 5

4 结 论

近年来，流形学习方法在数据挖掘、机器学习、图像处理和计算机视觉等多个研究领域吸引了广泛的关注并取得了长足的发展。但当样本点较为稀疏时，这些流形学习算法往往效果变差甚至失效。解决此问题的有效方法，是根据流形特点增加一些插值点。但已有的算法均采用线性插值的方法获取插值点。从线性代数的理论来说，由插值点和原有邻域点张成的线性子空间与原有邻域点张成的子空间是一样的，新的插值点不会改善线性逼近的误差。而且，插值点并没有反应出流形的本质结构和特征，从理论上背离了数据降维的目的。本文利用 Biharmonic 样条插值法非线性的获取插值点，新的插值点能有效的改善稀疏样本集的局部结构，并且插值点能较好的体现流形本身的结构和性质。在将本文提到的插值算法应用到经典的流形学习算法如 LTSA、LLE 后，实验结果证实了我们的算法的有效性和稳定性。

值得注意的是，我们提出的算法中，由于需要对每个样本点做曲面插值和插值点的选择，并最终扩展了样本点集来参与流形学习算法，这导致算法的运行时间较长，尤其是对于较高维数的样本集，算法的运行时间更加难以接受。由此，如何有效的提高算法的执行效率将是本文未来的研究内容。

参考文献：

［1］ TENENBAUM J B, SILVA V DE, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction［J］. Science, 2000, 290(5000): 2219 – 2323.

［2］ ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding［J］. Science, 2000, 290(5000): 2323 – 2326.

［3］ DONOHO D, GRIMES C. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data［J］. Proceedings of the National Academy of Sciences, 2003, 100(10): 5591 – 5599.

［4］ ZHANG Z Y, ZHA H Y. Principal manifolds and nonlinear dimension reduction via local tangent space alignment［J］. SLAM Journal of Scientific Computing, 2004, 26(1): 313 – 338.

［5］ BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation［J］. Neural Computation, 2002, 15: 1373 – 1396.

（下转第96页）

Applied and Environmental Microbiology, 2011, 77(21): 7787 – 7796.

[16] BLACKBURNE R, YUAN Z, KELLER J. Partial nitrification to nitrite using low dissolved oxygen concentration as the main selection factor [J]. Biodegradation, 2008, 19(2):303 – 312.

[17] 傅金祥,韩晋英,齐建华,等. 常温下低DO和高pH短程硝化过程研究[J]. 水处理技术, 2008, 34(12):19 – 23.

[18] LAANBROEK H J, GERARDS S. Competition for limiting amounts of oxygen between *Nitrosomonas europaea* and *Nitrobacter winogradskyi* grown in mixed continuous cultures [J]. Archives of Microbiology, 1993, 159(5):453 – 459.

[19] SLIEKERS A O, HAAIJER S C M, STAFSNES M H, et al. Competition and coexistence of aerobic ammonium and nitrite-oxidizing bacteria at low oxygen concentrations [J]. Applied Microbiology and Biotechnology, 2005, 68(6):808 – 817.

[20] ITOKAWA H, HANAKI K, MATSUO T. Nitrous oxide production in high-loading biological nitrogen removal process under low COD/N ratio condition [J]. Water Research, 2001, 35(3): 657 – 664.

[21] BOLLMANN A, LAANBROEK H J. Influence of oxygen partial pressure and salinity on the community composition of ammonia-oxidizing bacteria in the Schelde estuary [J]. Aquatic Microbial Ecology, 2002, 28(3): 239 – 247.

[22] KIM J H, GUO X, PARK H S. Comparison study of the effects of temperature and free ammonia concentration on nitrification and nitrite accumulation [J]. Process Biochemistry, 2008, 43(2): 154 – 160.

[23] STRAUSS E A, MITCHELL N L, LAMBERTI G A. Factors regulating nitrification in aquatic sediments: effects of organic carbon, nitrogen availability, and pH [J]. Canadian Journal of Fisheries and Aquatic Sciences, 2002, 59(3): 554 – 563.

[24] SAMPEI Y, MATSUMOTO E. C/N ratios in a sediment core from Nakaumi Lagoon, southwest Japan-usefulness as an organic indicator [J]. Geochemical Journal, 2001, 35(3): 189 – 205.

[25] ISNANSETYO A, THIEN N D, SEGUCHI M, et al. Nitrification potential of mud sediment of the Ariake Sea tidal flat and the individual effect of temperature, pH, salinity and ammonium concentration on its nitrification rate [J]. Research Journal of Environmental and Earth Sciences, 2011, 3(5): 587 – 599.

（上接第90页）

[6] KARBAUSKAITË R, KURASOVA O, DZEMYDA G. Selection of the number of neighbors of each data point for the locally linear embedding algorithm [J]. Information Technology and Control, 2007, 36: 359 – 364.

[7] VALENCIA-AGUIRRE J, ÁLVAREZ-MESA A, DAZA-SANTACOLOMA G. Automatic choice of the number of nearest neighbors in locally linear embedding [C]// CIARP, 2009:77 – 84.

[8] WEN G, JIANG L, WEN J, et al. Performing locally linear embedding with adaptable neighborhood size on manifold [C]// 9th Pacific Rim International Conference on Artificial Intelligence, Springer Verlag, 2006: 985 – 989.

[9] WU S, QUAN X W, CHEN X C. CN-isomap algorithm for nonlinear dimensionality reduction of sparse data [J]. Mathematics in Practice and Theory, 2010, 17(40): 182 – 188.

[10] SONG X, YE S W. Data dimensionality reduction algorithm when source data is spare [J]. Computer Engineering and Application, 2007, 43(28): 181 – 183.

[11] ZHAN D C, ZHOU Z H. Neighbor line-based locally linear embedding [J]. PAKDD, Springer Verlag, 2006: 806 – 815.

[12] SANDWELL D T. Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data [J]. Geophysical Research Letters, 1987, 2: 139 – 142.

[13] ZHANG T H, TAO D C, LI X L. A unifying framework for spectral analysis based dimentionality reduction [C]//International Joint Conference Neural Networks, 2008: 1670 – 1677.

[14] WANG Y T, DONG L F, NI K. Image morphing algorithm based on Biharmonic spline interpolation and its implementation [J]. Journal of Image and Graphics, 2007, 12(12): 2189 – 2194.