基于紧集子覆盖的流形学习算法

张绍群
（四川大学数学学院 成都 610024）

摘 要 2000 年以后新兴了一系列非线性降维的方法，流形学习中的 Isomap 就是其中的代表。该算法能够反映数据集的全局结构且简单高效，但是存在低维流形等距的欧氏子集必须是凸集和计算复杂度高等缺点。L-Isomap 成功降低了算法的计算复杂度，但是对于地标点（landmark points）的选取大多采用随机的方法，致使该算法不稳定。依据拓扑学和泛函分析中有限维空间有界闭集与紧集（compact set）等价、紧集的任一开覆盖存在有限子覆盖等经典定理，分析数据集所在区域的拓扑结构，确定了一系列能够反映数据结构的地标点。这样的方法计算复杂度低，比 L-Isomap 稳定，且将数据集是凸集的要求弱化到紧集（有界闭集），避免了传统 Isomap 算法放大不完整流形中的“空洞”误差等问题。

关键词 流形学习，等距映射，地标点，紧性
中图法分类号 TP311.12 **文献标识码** A

Manifold Learning Algorithm Based on Compact Setsub-coverage

ZHANG Shao-qun
(College of Mathematics, Sichuan University, Chengdu 610024, China)

Abstract Since 2000, a series of nonlinear dimensionality reduction methods have been emerging, and Isomap in manifold learning is one of the representatives. The algorithm can reflect the global structure of the data set and is simple and efficient. But there are some shortcomings that low-dimensional manifold must be convex set and the computational complexity is large. L-Isomap successfully reduces the computational complexity, but majority of the landmarks is selected by random method, which makes the algorithm unstable. In this paper, according to the classical theoremes that bounded closed set is equivalent to compact set in finite-dimensional space and there is finite sub-coverage covering the compact set, we analyzed the topology of the area of the data set and selected a series of landmarks. This method has low computational complexity and is more stable than L-Isomap. In addition, this method weakens the condition that the data set is a convex set to a compact set (bounded closed set), which avoids enlarging "the hollow" error in the incomplete manifold.

Keywords Manifold learning, Isomap, Landmark points, Compact

1 背景

数据降维的目的是从高维数据集中找到其隐藏的低维数据结构。随着大数据时代的来临，数据降维已经深入到计算机科学等学科的各个领域，比如图像分类系统、文本分类系统、基因序列的建模等。现有的数据降维方法大致可以分为两类：1）线性降维的方法，比如主成分分析法（PCA）、非负矩阵分解（NMF）和多维尺度变换算法（MDS）等；2）非线性降维方法，如等距映射算法（Isomap）、局部线性嵌入法（LLE）和局部切空间排列法（LTSA）等[1-3]。

目前，非线性降维的一个主流方向是流形学习（Manifold Learning）。流形学习假设数据集分布在一个高维欧氏空间中的低维流形上，其目的是从高维数据集中恢复这个低维流形结构，并求出相对应的嵌入映射，以实现降维。流形学习可以大致分为两类：一类是全局优化算法，以等距映射（Isomap）为代表；另一类是局部优化算法，以局部线性嵌入（LLE）、局部切空间排列法（LTSA）等为代表。在流形学习方面最有影响力的文章是 2000 年 J. B. Tenenbaum 等和 S. T. Roweis 等人在《Science》同一期上发表的两篇文章，他们各自提出了自己的流形学习算法：等距映射（Isomap Mapping，Isomap）和局部线性嵌入（Locally Linear Embedding）[2-3]。

Isomap 是一种高效的全局优化流形学习算法，能较好地保持数据集的全局结构特点，但是其计算复杂度高，且要求低维流形等距的欧氏子集必须是一个凸集。后来，Vin de Silva 和 J. B. Tenenbaum 提出了带地标点（Landmark Points）的快速算法 L-Isomap，降低了计算复杂度。但是对于如何选取地标点，L-Isomap 采取随机选取的方法，很难保证算法的稳定性，且要求低维流形等距的欧氏子集是凸集的缺点依然存在。本文依据拓扑学和泛函分析中有限维空间有界闭集与紧集（compact set）等价、紧集的任一开覆盖存在有限子覆盖等定理，分析数据集所在区域的拓扑结构，确定了一系列能够反映数据结构的地标点，从而提出一种基于紧集子覆盖地标点的快速流形学习算法（CL-Isomap）。该算法降低了计算复杂度，将低维流形等距的欧氏子集必须是凸集的要求弱化到紧

张绍群（1992－），男，硕士生，主要研究方向为流形学习和信息几何，E-mail：zsqhndn@126.com。

集(有界闭集),并通过实验验证了该算法的高效性、稳定性和优越性。

2 Isomap

Isomap 建立在多维尺度变换(MDS)的基础上,力求保持数据集的内在几何性质。Isomap 与 MDS 的最大区别在于:MDS 构造距离矩阵时采取的是样本之间的欧氏距离,而 Isomap 采取的是样本之间的测地距离。在 Isomap 中,测地距离的计算方法如下:样本点 x 和它的邻域点之间的测地距用它们之间的欧氏距离来代替,这里认为局部邻近的关系可以用线性来逼近;样本点 x 和它邻域外的点用流形上他们之间的最短路径来代替。

Isomap 的步骤如下。

输入:数据集 X(N 个 l 维的数据样本),近邻数 k 或者邻域距离 ε

输出:降维后的数据集 Y

Step1　选取邻域,构造无向图 G。计算每个样本点 x_i 与其余样本点之间的欧氏距离 d。ε-邻域:当 x_i 与 x_j 的欧氏距离 $d(x_i,x_j)$ 小于固定值 ε 时,认为 x_i 与 x_j 相邻,则图 G 有边 x_ix_j;k-邻域:当 x_j 是 x_i 最近的 k 个点之一时,认为 x_i 与 x_j 相邻,则图 G 有边 x_ix_j。这里采取哪一种计算邻域的方式都可以,设图 G 的边 x_ix_j 的权为 $d(x_i,x_j)$。

Step2　计算测地距离矩阵。当无向图 G 有边 x_ix_j 时,设测地距离为 $d_G(x_i,x_j)=d(x_i,x_j)$;否则设 $d_G(x_i,x_j)=\infty$。对于 $m=1,\cdots,N$,$d_G(x_i,x_j)=\min\{d(x_i,x_j),d(x_i,x_m)+d(x_m,x_j)\}$。这样即可得到测地距离矩阵 $D_G=\{d_G^2(x_i,x_j)\}_{i,j=1}^N$。

Step3　计算 d 维流形嵌入。将 MDS 的方法应用到距离矩阵 D_G 上。记 $H=-(I-1_N1_N^T)D_G(I-1_N1_N^T)/2$,H 的最大 d 个特征值 $\lambda_1,\cdots,\lambda_d$ 以及对应的特征列向量 $u_1,\cdots,u_d$ 所构成的矩阵为 $U=\{u_1,\cdots,u_d\}$,则 $T=\mathrm{diag}(\lambda_1^{\frac{1}{2}},\cdots,\lambda_d^{\frac{1}{2}})U^T$ 是 d 维流形嵌入的嵌入结果。

从算法可以看出,Isomap 是一种全局优化方法,它的嵌入结果能够反映出高维数据集中存在的低维流形结构,且这个低维流形结构与欧氏空间的一个子集是整体等距的。但需要注意的是,Isomap 要求与低维流形等距的欧氏子空间的子集是凸集。当这个欧氏子集非凸即数据集所在区域中间存在"空洞"时,计算样本间的测地距离时会出现较大的偏差,致使嵌入结果产生明显的变形。具体例子可以参见文献[7]。

Isomap 还存在计算复杂度高的问题:1)选取邻域时,构造无向图 G 的计算复杂度为 $O(lN^2)$;2)计算测地距离矩阵的计算复杂度为 $O(N^3)$;3)在计算 d 维流形嵌入、求解 D_G 的特征值方面,由于 D_G 的稠密性,其计算复杂度为 $O(N^3)$。显然,影响 Isomap 计算复杂度的因素主要有两个:计算测地距离矩阵 D_G 和求解 D_G 的特征值。采用 Dijkstra 算法或者 Floyd 算法可以将计算测地距离的计算复杂度降低到 $O(N^2\log N)$。

3 L-Isomap

为了弥补 Isomap 计算复杂度大的缺点,Vin de Silva 和 J. B. Tenenbaum 提出了带地标点(Landmark Points)的快速算法 L-Isomap,即在 N 个样本点中随机挑选 n 个点作为地标点($n\leqslant N$),在构计算测地距离矩阵时,并不是计算所有样本间的测地距离,而是计算 n 个地标点和 N 个样本之间的测地距离,得到的测地距离矩阵是一个 $n\times N$ 的矩阵 $D_{n\times N}$,然后将 L-MDS 的方法运用到测地距离矩阵 $D_{n\times N}$ 上得到嵌入结果。

L-Isomap 的步骤如下。

输入:数据集 X(N 个 l 维的数据样本),近邻数 k 或者邻域距离 ε,地标点个数 n

输出:降维后的数据集 Y

Step1　选取邻域,构造无向图 G。计算每个样本点同其余样本点之间的欧氏距离 d。ε-邻域:当 x_i 与 x_j 的欧氏距离 $d(x_i,x_j)$ 小于固定值 ε 时,认为 x_i 与 x_j 相邻,则图 G 有边 x_ix_j;k-邻域:当 x_j 是 x_j 最近的 k 个点之一时,认为 x_i 与 x_j 相邻,则图 G 有边 x_ix_j。这里采取哪一种计算邻域的方式都可以,设图 G 的边的权为 $d(x_i,x_j)$。

Step2　从 N 个样本点中随机挑选 n 个样本作为地标点。

Step3　计算基于地标点的测地距离矩阵 $D_{n\times N}$。

Step4　应用 L-MDS 求解低维嵌入流[4]。

由上可知,L-Isomap 算法主要分为两部分:从样本中随机选取 $n(n\ll N)$ 个地标点,计算地标点和样本之间测地距离矩阵 $D_{n\times N}$,然后使用 L-MDS 的方法求低维流形嵌入。这两步的计算复杂度分别为 $O(n^2N)$ 和 $O(nN\log N)$。L-Isomap 虽然降低了计算复杂度,但是地标点选取的随机性影响到了其嵌入结果的稳定性,而且要求与低维流形等距的欧氏子空间的子集是凸集的弊端并没有得到改善,相反还可能扩大其"空洞"误差。

本文提出了一种基于紧集子覆盖地标点的快速流形学习算法(CL-Isomap)。该算法基于拓扑学和泛函分析有限维空间紧集及紧性定理的方法选取地标点,在计算复杂度增幅不大的情况下取得了稳定解,并成功规避了流形结构中的"空洞"放大问题。

4 CL-Isomap

首先引入泛函分析中几个重要的定义和定理(参见文献[5-6])。

定义 1　距离空间 Z 中的集合 X 称为紧的,如果 X 的任一开覆盖都存在有限的子覆盖,集合 X 也称为紧集(Compact set)。

紧性本身是一种拓扑结构性质,一个集合 X 是紧集,则说明这个集合的结构是紧的,即存在一个有限的覆盖结构。以距离空间 Z 取定为欧氏空间为例,如果 X 是 Z 上的紧集,则存在一个有限的序列 $\{\alpha_i\}_{i=1}^n\subset X$ 及及其邻域集 $\{B(\alpha_i,\varepsilon_i)\}_{i=1}^n$,满足 $X\subset\coprod_{i=1}^n B(\alpha_i,\varepsilon_i)$。

定理 1　有限维欧氏空间中的紧集等同于有界闭集。

定理 2　紧集 X_1 经过等距映射得到的集合 X_2 依然是紧的。

假设所采用的原始数据集 X 为 N 个 l 维的样本,由以上的定理及定义不难得到以下结论:若高维数据集 X 是欧氏空间 R^l 中的一个有界闭集,即紧集,则存在一个有限的序列 $\{\alpha_i\}_{i=1}^n\subset X$ 及其邻域集 $\{B(\alpha_i,\varepsilon_i)\}_{i=1}^n$ 可以覆盖 X。其中,序列 $\{\alpha_i\}_{i=1}^n$ 即是将要选取的地标点。若 Isomap 算法整体等距地将 X 映射为欧氏空间 R^d 中的数据集 M,则 M 也是一个紧集,即 R^d 中的有界闭集,而不是 R^d 中的凸集。从而 X 在 R^l

中的地标点集$\{\alpha_i\}_{i=1}^n$和覆盖$\{B(\alpha_i,\varepsilon_i)\}_{i=1}^n$通过 Isomap 的等距映射后成为 M 在 R^d 中的点集和覆盖。进一步，通过紧性和覆盖可以将 Isomap 方法中所有点的降维过程压缩为地标点集$\{\alpha_i\}_{i=1}^n$的降维，从而大大减少了计算量。

以上的推理基于全局非线性降维方法在几何上的理论基础。不难发现，只要选取合适的地标点$\{\alpha_i\}_{i=1}^n$和覆盖$\{B(\alpha_i,\varepsilon_i)\}_{i=1}^n$，并将其与 Isomap 方法相结合，就可以保证高维原始数据集 X 和低维嵌入数据集 M 具有一样的拓扑结构。当然，这也从侧面说明了地标点不能随机选取。

下面具体介绍如何选取地标点$\{\alpha_i\}_{i=1}^n$。为了选取有代表性的地标点并保证选取地标点的方法稳定，从数据集 X 的欧氏距离矩阵 d 中距离最大的两个点开始搜索，即

$$d(x_i,x_j)=\max d(x_t,x_s)$$

其中，$t,s=1,\cdots,N$。

令 x_i（或者 x_j）为第一个地标点 α_1，并计算其邻域 $B(\alpha_1,\varepsilon_1)$，然后从剩余样本集 $X\backslash B(\alpha_1,\varepsilon_1)$ 中以同样的方法挑选第二个地标点 α_2，并计算其邻域 $B(\alpha_2,\varepsilon_2)$。以此类推，可求得全部地标点$\{\alpha_i\}_{i=1}^n$及其邻域$\{B(\alpha_i,\varepsilon_i)\}_{i=1}^n$。这样的选取方法既可以保证地标点分散均匀，也可以使邻域间部分重叠。为了计算方便，通常令邻域半径 $\varepsilon_1=\varepsilon_2=\cdots=\varepsilon_n$，如果邻域半径不等且数值相差不大，则在本质上不会对算法产生退化等影响。这里选取地标点的方法称为 Compact SelectLandmark (CS)。显然 CS 的计算复杂度为 $O(N)$。

挑选好地标点之后，将 CS 部分嵌入到 L-Isoamp 之中，得到新算法 CL-Isomap。

CL-Isomap 的步骤如下。

输入：数据集 X(N 个 l 维的数据样本)，近邻数 k 或者邻域距离 ε，地标点个数 n

输出：降维后的数据集 Y

Step1 选取邻域，构造无向图 G。计算每个样本点 x_i 同其余样本点之间的欧氏距离 d，如果 x_i 与 x_j 相邻，则图 G 有边 x_ix_j，边权为 $d(x_i,x_j)$。

Step2 应用 CS 方法选取地标点集$\{\alpha_i\}_{i=1}^n$。

Step3 计算地标点和样本点之间的测地距离矩阵 $D_{n\times N}$。

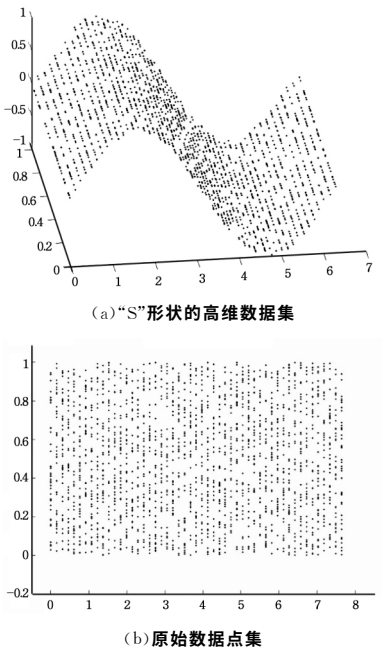
Step4 应用 L-MDS 求解低维嵌入流[4]。

由上述算法可知，相对于 L-Isomap 算法，CL-Isomap 只在挑选 landmark points 方面多了计算复杂度为 $O(N)$ 的 Step2。因此在计算复杂度方面，其与 L-Isomap 算法并无太多差别。

5 实验验证

为了验证算法的稳定性和有效性，设计了两组实验，分别在完整的“S”形状的数据集和不完整的流形（含有“空洞”）的“S”形状数据集上比较 Isomap，L-Isomap 和 CL-Isomap 3 种算法。

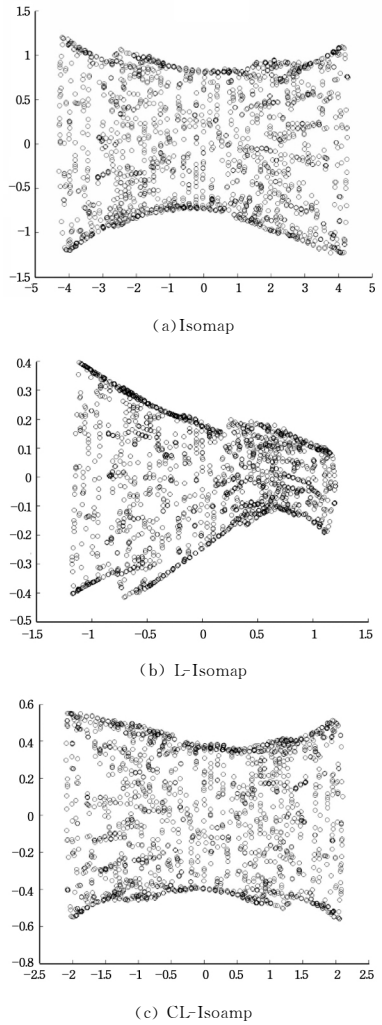
图 1 和图 2 为第一次实验的结果，该次实验采用完整的“S”形状的数据集，比较 Isomap，L-Isomap 和 CL-Isomap 3 种算法，实验参数保持一致：样本数 $N=2000$，近邻数 $K=20$，地标点个数 $n=100$。图 1 示出完整的“S”形状数据集和其二维平面上的原始数据点集，图 2(a)－图 2(c)分别是 Isomap，L-Isomap 和 CL-Isomap 3 种算法的实验结果。



(a)“S”形状的高维数据集

(b)原始数据点集

图 1 完整的“S”形状数据集和其二维平面上的原始数据点集



(a)Isomap

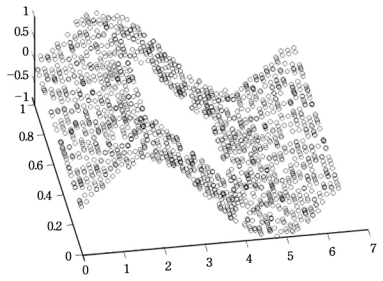
(b) L-Isomap

(c) CL-Isoamp

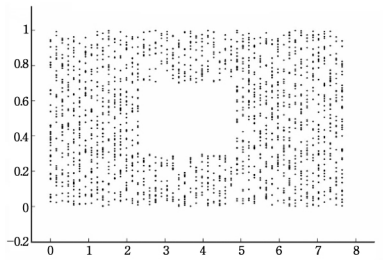
图 2 完整的“S”形状数据集上的 3 种嵌入结果比较

L-Isomap 算法中的地标点个数 n 只要满足 $n>d$（低维流形维数）即可，但是 CL-Isomap 的地标点需要多一些，为保证对比效果，取 $n=100$。从图 2 可以观察到：图 2(b)与图 2(a)的结果出现了较大的偏差，这是由 L-Isomap 算法选择地标点

的随机性造成的，由此可设想，如果在地标点个数更少，则结果会更加不够稳定；而图 2(c)和图 2(a)的结果没有较大差别，说明了 CL-Isomap 算法能够对 Isomap 算法保持足够忠实，多次实验结果比较稳定。

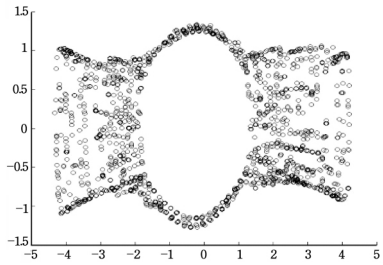


(a)带“空洞”的“S”形状的高维数据集

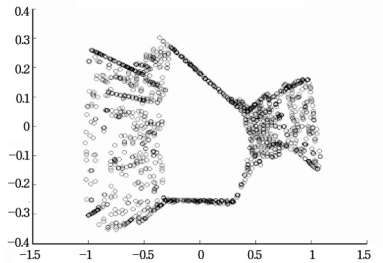


(b)原始数据点集

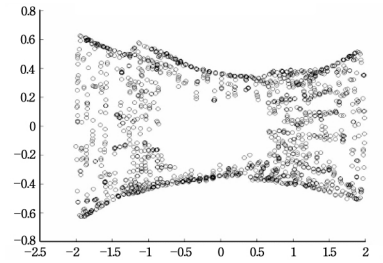
图 3　带“空洞”的“S”形状数据集和其二维平面上的原始数据点集



(a)Isomap



(b) L-Isomap



(c) CL-Isoamp

图 4　带“空洞”的“S”形状数据集上的 3 种嵌入结果

图 3 和图 4 为第二次实验的结果，本次实验采用不流形(含有“空洞”)的“S”形状的数据集来比较 Isomap，L-Isomap 和 CL-Isomap 3 种算法，实验参数保持一致：样本数 $N=1400$，近邻数 $K=20$，地标点个数 $n=70$。图 3 示出了带“空洞”的“S”形状数据集和其二维平面上的原始数据点集，图 4(a)—图 4(c)分别是 Isomap，L-Isomap 和 CL-Isomap 3 种算法的实验结果。

从图 4 可以观察到：图 4(a)放大了中间的“空洞”，图 4(b)出现了严重的失真，而图 4(c)相对较好地还原了“空洞”结构。

结束语　针对 Isomap 算法计算复杂度高、要求与低维流形等距的欧氏子集必须为凸集和 L-Isomap 算法随机选取地标点造成的不稳定性等缺点，本文提出的 CL-Isomap 充分考虑了高维数据集所在区域的拓扑结构性质，将每一个子覆盖视为等价类，选取等价元或者代表元作为地标点。这种选取方式即保证了原始数据集的拓扑结构，避免了“空洞”误差放大问题，又保证了地标点的足够分散和稳定，从而使嵌入结果更加稳定。实验表明，CL-Isomap 算法忠于原始数据集的数据结构，能够快速稳定地实现全局等距降维，且对带有“空洞”的不完整流形也有很好的嵌入结果。

另外，第二次实验建立在“空洞”较小、高维数据集所在区域还能够保持良好的连通性的基础上，如果“空洞”足够大，几乎可以使高维数据集所在的区域断片时，CL-Isomap 算法则不会表现出较强的优越性，这是因为数据集的流形拓扑结构已被破坏。这个问题也是流形学习需要解决的难题。

参 考 文 献

[1] RABINOWITZ, GEORGE B. An Introduction to Nonmetric[J]. American Journal of Political Science, 1975, 19(2): 343-390.

[2] TENENBAUM J B, DE SILVA V. Aglobel geometric frAmewok for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.

[3] ROWEIS S, SAUL L. Nonlineardimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.

[4] TENENBAUM J B, SILVA V De. Global versus local methods in nonlinear dimensionality reduction[C]// Proceeding of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2003.

[5] LAX P D. Functional Analysis[M]. Beijing: Higher Education Press, 2007: 43-51, 233-244.

[6] 江泽坚，孙善利. 泛函分析[M]. 北京：高等教育出版社，2005：19-25.

[7] 王靖. 流形学习的理论与方法研究[D]. 杭州：浙江大学理学院，2006.

[8] CARLOTTA O. An improved set covering problem for Isomap supervised landmark selecion[J]. Pattern Recognition Letters, 2014, 49: 131-137.