ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.4, April 2007, pp.869–877
DOI: 10.1360/jos180869

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

一种更具拓扑稳定性的 ISOMAP 算法*

邵 超[1,2+], 黄厚宽[1], 赵连伟[1]

[1](北京交通大学 计算机与信息技术学院,北京 100044)
[2](河南财经学院 计算机科学系,河南 郑州 450002)

A More Topologically Stable ISOMAP Algorithm

SHAO Chao[1,2+], HUANG Hou-Kuan[1], ZHAO Lian-Wei[1]

[1](School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)
[2](Department of Computer Science, He'nan University of Finance and Economics, Zhengzhou 450002, China)
+ Corresponding author: Phn: +86-10-51683602, Fax: +86-10-51840526, E-mail: shaochao051227@gmail.com

Shao C, Huang HK, Zhao LW. A more topologically stable ISOMAP algorithm. *Journal of Software*, 2007, 18(4):869–877. http://www.jos.org.cn/1000-9825/18/869.htm

Abstract: The success of ISOMAP depends greatly on being able to choose a suitable neighborhood size, however, it is still an open problem how to do this effectively. Based on the fact that "short circuit" edges pass the area with the relatively lower local densities, this paper presents a new variant of ISOMAP, i.e. P-ISOMAP (pruned-ISOMAP), which can prune effectively "short circuit" edges existed possibly in the neighborhood graph and thus is much less sensitive to the neighborhood size and more topologically stable than ISOMAP. Consequently, P-ISOMAP can be applied to data visualization more easily than ISOMAP because the open problem described above can be avoided to the utmost extent. The effectivity of P-ISOMAP is verified by the experimental results very well.

Key words: ISOMAP; P-ISOMAP (pruned-ISOMAP); neighborhood size; topological stableness; residual variance; kernel density estimation; local density

摘 要: ISOMAP 算法能否被成功运用,很大程度上依赖于邻域大小的选取是否合适.然而,如何有效地选取合适的邻域大小,目前还是一个尚未解决的难题.根据"短路"边会途经相对的低密度区域这一特点,能够有效删除邻域图中可能存在的"短路"边,提出了 P-ISOMAP(pruned-ISOMAP)算法,这极大地削弱了 ISOMAP 算法对邻域大小的依赖程度,从而使其更具拓扑稳定性.由于避免了邻域大小难以有效选取的问题,P-ISOMAP 算法能够更容易地对数据进行可视化.实验结果很好地验证了该算法的有效性.

关键词: ISOMAP;P-ISOMAP(pruned-ISOMAP);邻域大小;拓扑稳定性;残差;核密度估计;局部密度
中图法分类号: TP181 文献标识码: A

作为一种有效的非线性降维技术,ISOMAP 算法[1,2]采用能够有效表征数据全局几何结构(global geometric

* Supported by the National Natural Science Foundation of China under Grant No.60443003 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant Nos.2007CB307100, 2007CB307106 (国家重点基础研究发展规划(973)); the Foundation of Beijing Jiaotong University of China under Grant No.2003SZ003 (北京交通大学基金)
Received 2005-11-13; Accepted 2006-04-27

structure)的测地距离(geodesic distance)对古典 MDS(multidimensional scaling)算法进行了非线性扩展,从而能够很好地对具有良好抽样且内在扁平(intrinsically flat)的数据集进行可视化,如 swiss roll 数据集和 S 形数据集等.

ISOMAP 算法有很多优点,其中之一就是它只有一个参数——邻域大小(neighborhood size).与其他流形学习算法,如 LLE(locally linear embedding)算法[3]一样,ISOMAP 算法能否被成功运用很大程度上依赖于该参数的选取是否合适,这也造成了 ISOMAP 算法的所谓拓扑不稳定性(topological unstableness)[4].然而,如何有效地选取合适的邻域大小,目前还是一个尚未解决的难题.通常做法是根据最终映射"质量"的高低反过来判断对应邻域大小的选取是否合适[2,5],这是极其耗时且不实用的[6].我们知道,ISOMAP 算法是用邻域图中的最短路径长度来逼近测地距离的,为了得到一个能够正确表达数据邻域结构的邻域图,邻域大小应该足够大,但不能使邻域图中出现"短路(short circuit)"边[4].在本文中,我们根据"短路"边会途经低密度区域这一特点,对邻域图中可能存在的"短路"边进行了有效的删除,从而得到了对邻域大小不再敏感、也更具拓扑稳定性的 P-ISOMAP(pruned-ISOMAP)算法.由于避免了邻域大小难以有效选取的问题,该算法能够更容易地对数据进行可视化.

1 ISOMAP 算法

人们通常用欧氏距离来表达数据间的相异度,但这是建立在全局线性结构假设基础之上的[7].在数据的全局几何结构未知的情况下,欧氏距离还能否用来表达数据间的相异度,就不那么确定了;但所幸的是,在一个很小的邻域内,我们仍然有理由这么做.因此,我们可以用已知的这些局部欧氏距离来逼近数据未知的全局几何结构,ISOMAP 算法就是这样做的.如果数据位于单一流形之上,并且具有良好抽样,那么,能够有效表征数据全局几何结构的测地距离,就可以用邻域图中的最短路径长度来进行逼近了[8].为了能对测地距离进行良好逼近,该邻域图应能正确表达数据的邻域结构,因此,必须具有合适的邻域大小.ISOMAP 算法可简要描述如下:

(1) 对于大数据集,为了降低计算量,从中选取 n 个代表点以执行下面的操作.选取的方法很多,本文采用的是矢量量化方法,因为它能得到更具代表性的数据点,可视化效果会更好一些[9].

(2) 用 K 近邻法(K nearest neighbors)创建能够正确表达数据邻域结构的邻域图(为了更好地对数据进行可视化,该邻域图至少应是连通的),这需要一个合适的邻域大小 K.

(3) 在该邻域图上运行最短路径算法得到所有数据点间的最短路径长度,用来逼近相应的测地距离.

(4) 将这些最短路径长度作为输入运行古典 MDS 算法,将数据重建在一个低维可视空间中.

在给定了一个具有良好抽样且内在扁平的数据集之后,ISOMAP 算法能否被成功运用的关键,就在于邻域大小的选取是否合适了,因为只有合适的邻域大小才能保证对测地距离的良好逼近[4].通常的做法是根据最终映射"质量"的高低反过来判断对应邻域大小的选取是否合适,常用来衡量映射"质量"高低的标准是残差(residualvariance)[2,4,5]:$1-\rho^2_{\hat{D}_X(K)D_Y}$,其中:$\hat{D}_X(K)$ 和 D_Y 分别表示数据点在原数据空间中的测地距离矩阵(由邻域图中的最短路径长度来进行逼近,在给定数据集的情况下为邻域大小 K 的函数)和在低维可视空间中的欧氏距离矩阵而 $\rho_{\hat{D}_X(K)D_Y}$ 则表示它们之间的线性相关系数(linear correlation coefficient).残差越小,表示映射的"质量"越高,对应的邻域大小也就越合适.因此,最优邻域大小可定义如下[5]:

$$K_{opt}=\arg\min_K(1-\rho^2_{\hat{D}_X(K)D_Y}) \tag{1}$$

由于残差用到了 ISOMAP 算法的运行结果 Y,因此,计算残差需要运行整个 ISOMAP 算法.另外,残差只能衡量两个邻域大小的相对合适程度,不能用来判断某一个邻域大小合适与否;同时,残差还具有多峰性(multimodality),因此,该方法需要就每一个可能的邻域大小分别计算相应的残差,这是极其耗时且不实用的,从而使邻域大小在实际中难以有效选取.

2 P-ISOMAP 算法

显而易见,为了更好地对测地距离进行逼近,邻域大小应在邻域图不存在"短路"边的前提下尽可能地大.如

果能对邻域图中可能存在的“短路”边进行鉴别和删除,我们就可以极大地削弱 ISOMAP 算法对邻域大小的依赖程度,并使其更具拓扑稳定性.

所谓“短路”边是指那些会途经非数据流形区域(偏离了数据流形区域),从而将流形上本不相邻的两个数据点连接起来的边.也就是说,“短路”边会途经某些低密度区域.与之相对应的是,非“短路”边因为都位于数据流形区域之上,从而不会像“短路”边那样途经某些低密度区域.因此,我们可以根据“短路”边会途经低密度区域这一特点,来对邻域图中可能存在的“短路”边进行鉴别和删除.

估计数据密度的方法有很多,比较常用的是核密度估计法(kernel density estimation)[10,11],如果采用高斯核函数,则有,

$$f(X)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}\cdot h^{D}}\mathrm{e}^{-\frac{d(X,X_i)^2}{2h^2}} \tag{2}$$

其中:$d(X,X_i)$表示 X 和数据点 X_i 之间的距离,通常采用欧氏距离$||X-X_i||$;h 表示带宽(bandwidth)参数;n 和 D 分别表示数据点的个数及其维数.如果数据不服从全局线性结构假设[7],那么,式(2)中的 $d(X,X_i)$就应该采用能够有效表征数据全局几何结构的测地距离 $\hat{d}(X,X_i)$ [11].然而,在 ISOMAP 算法中,未知的测地距离是用邻域图中的最短路径长度来进行逼近的,如果邻域图中出现了“短路”边,那么,对测地距离的逼近效果就会很差,用这样的测地距离估计出来的数据密度也就会很差.为了克服这一问题,根据流形的局部欧氏特性,类似于 ISOMAP 算法对欧氏距离的处理方式,我们仅考虑邻域内的核函数平均,那么,式(2)中的 $d(X,X_i)$就可以采用欧氏距离$||X-X_i||$了,这样得到的数据密度我们称为局部密度(local density):

$$\underline{f}(X)=\frac{1}{|NE(X)|}\sum_{X_i\in NE(X)}\frac{1}{\sqrt{2\pi}\cdot h^{D}}\mathrm{e}^{-\frac{\|X-X_i\|^2}{2h^2}} \tag{3}$$

其中,$NE(X)$和$|NE(X)|$分别表示 X 的邻域集合及其包含的数据点个数.借用由 ISOMAP 算法得到的邻域结构,数据点 X_i 处的局部密度可表示为

$$\underline{f}(X_i)=\frac{1}{|NE(X_i)|}\sum_{X_j\in NE(X_i)}\frac{1}{\sqrt{2\pi}\cdot h^{D}}\mathrm{e}^{-\frac{\|X_i-X_j\|^2}{2h^2}} \tag{4}$$

其中:X_i 的邻域集合为 $NE(X_i)=\{X_i,X_{i_1},...,X_{i_K}\}$ (X_{i_k} 为 X_i 的第 k 近邻数据点,k=1,2,…,K),其中包含的数据点个数为$|NE(X_i)|=K+1$.

给定了一条边,它所途经区域的局部密度可以用该边中点处的局部密度来近似表示.更鲁棒的做法是用该边上 3 个四分位点处的平均局部密度来近似表示.形式化地,给定了连接数据点 X_i 和 X_j 的边(X_i,X_j),该边上的 3 个四分位点可分别表示为

$$X_{ij_1}=\frac{3\cdot X_i+X_j}{4},X_{ij_2}=\frac{X_i+X_j}{2},X_{ij_3}=\frac{X_i+3\cdot X_j}{4} \tag{5}$$

它们的局部密度可分别表示为

$$\underline{f}(X_{ij_k})=\frac{1}{|NE(X_{ij_k})|}\sum_{X_v\in NE(X_{ij_k})}\frac{1}{\sqrt{2\pi}\cdot h^{D}}\mathrm{e}^{-\frac{\|X_{ij_k}-X_v\|^2}{2h^2}},k=1,2,3 \tag{6}$$

其中, $NE(X_{ij_k})$ 和$|NE(X_{ij_k})|$分别表示 X_{ij_k} 的邻域集合及其包含的数据点个数.由于数据集的邻域结构在 ISOMAP 算法中已被求出,即 $NE(X_i)$和 $NE(X_j)$已知,则 $NE(X_{ij_k})$ 就可以简单地按照下式进行处理(如图 1 所示):

$$NE(X_{ij_k})=NE(X_i)\cup NE(X_j),k=1,2,3 \tag{7}$$

这样,我们就可以如下来表示边(X_i,X_j)途经区域的局部密度:

$$\underline{f}(X_i,X_j)=\frac{1}{3}\sum_{k=1}^{3}\underline{f}(X_{ij_k}) \tag{8}$$

然后,按照下式计算边(X_i,X_j)途经区域的相对局部密度:

$$r(X_i,X_j)=\frac{\underline{f}(X_i,X_j)}{\max(\underline{f}(X_i),\underline{f}(X_j))} \tag{9}$$

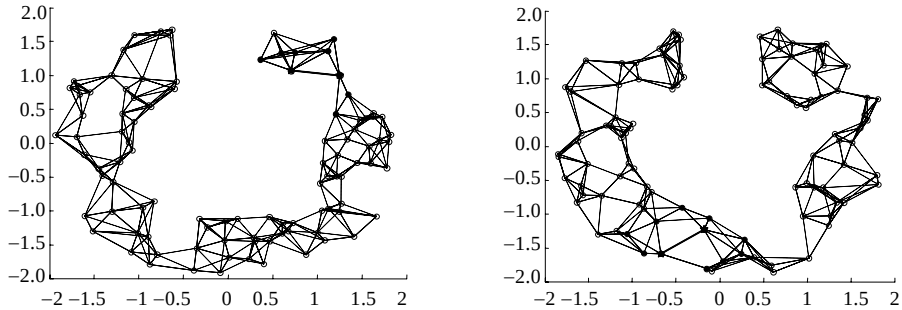


Fig.1 The same neighborhood sets (represented by solid circles) of three quantiles of the given edges (represented by thick solid lines) in the neighborhood graphs

图 1 邻域图中给定边(用粗实线表示)上的 3 个四分位点所共同的邻域集合(用实心园表示)

如上所述,当边(X_i,X_j)途经区域的相对局部密度 $r(X_i,X_j)$很小(小于某个阈值ε)时,我们有理由相信该边即为“短路”边,如图 2 所示(对应的邻域图是在 swiss roll 数据集[1,2](如图 3(a)所示)上用 K 近邻法得到的,图 2(a)中 K=10;图 2(b)中 K=12),其中最左边的那一小部分就是由“短路”边途经区域的相对局部密度组成的,它们明显要小得多,能够很容易地进行区分.在删除了这些“短路”边的邻域图上运行最短路径算法和古典 MDS 算法,就得到了 P-ISOMAP 算法.该算法可简要描述如下:

(1) 对于大数据集,为降低计算量,用矢量量化的方法从中选取 n 个代表点以执行下面的操作.

(2) 给定一个较大的 K 值(至少应能创建一个连通的邻域图),用 K 近邻法创建相应的邻域图.

(3) 鉴别和删除邻域图中可能存在的“短路”边:

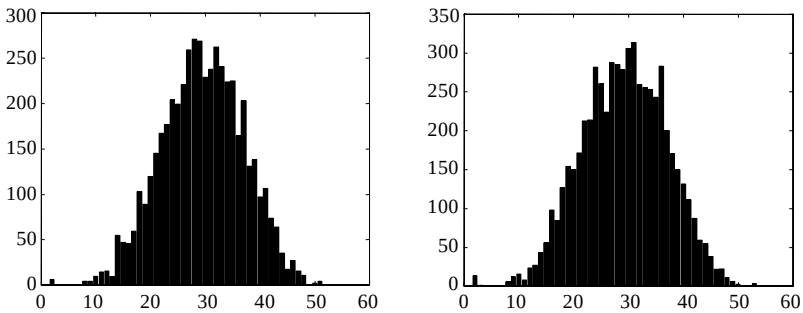
a) 根据式(9)计算邻域图中每一条边所途经区域的相对局部密度.

b) 选取一个合适的阈值ε(选取方法见下一节).

c) 对邻域图中的每一条边(X_i,X_j),如果 $r(X_i,X_j)<\varepsilon$,则使其为“短路”边,从邻域图中删除.

(4) 在删除了“短路”边的邻域图上运行最短路径算法,得到所有数据点间的最短路径长度,用来逼近相应的测地距离.

(5) 将这些最短路径长度作为输入运行古典 MDS 算法,将数据重建在一个低维可视空间中.



(a) (b)

Fig.2 The distributions of relative local densities of passing areas of all the edges in the neighborhood graphs

图 2 邻域图中所有边途经区域的相对局部密度的分布图

邻域大小 K 越大,邻域图中边的条数以及其中的“短路”边就越多,P-ISOMAP 算法附加的第 3 步操作计算量也就越大.很显然,这一步附加操作的时间复杂度可表示为 $O(nK^2)$,由于 K 比 n 明显要小得多,因此,与随后 ISOMAP 算法所固有的两步操作(最短路径算法和古典 MDS 算法)相比,这一步附加操作的计算量相对较小.

3 P-ISOMAP 算法中的参数选取问题

P-ISOMAP 算法通过删除邻域图中可能存在的“短路”边,极大地削弱了邻域大小的影响,从而可以比较容易地对邻域大小进行指定:我们可以不必考虑过大的邻域大小会在邻域图中引入“短路”边这一问题而选用更大的邻域大小,从而可在一定程度上避免文献[2,4]采用过于耗时的残差方法(见第 1 节)来为 ISOMAP 算法选取邻域大小这一问题.除邻域大小以外,P-ISOMAP 算法又引入了另外两个参数:带宽 h 和阈值 ε.

在核密度估计中,带宽 h 越大,得到的密度函数就越平坦.由于我们采用的是局部密度,为了区分“短路”边和非“短路”边,带宽 h 应稍小一些.然而,在由式(9)得到的相对局部密度中,带宽 h 的影响得到了一定程度的抑制;另外,“短路”边和非“短路”边途经区域的相对局部密度相差也比较大(如图 2 所示),从而使 P-ISOMAP 算法对带宽 h 不很敏感,实验结果也很好地对此进行了验证.因此,我们可以简单地设定 h=1,将其从 P-ISOMAP 算法中除去,其合理性可以通过下一节的实验结果得以验证.

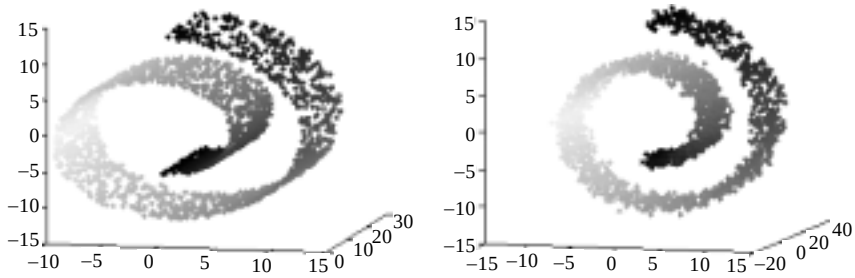
如上节所述,与非“短路”边相比,“短路”边途经区域的相对局部密度要小得多,它们能够很好地进行区分(如图 2 所示).因此,我们可以将邻域图中每一条边所途经区域的相对局部密度从小到大进行排序,在前半部分序列中找到增量最大(也就是区分最开)的那个相对局部密度作为阈值 ε.形式化地,dr_i 为邻域图中第 i 条边所途经区域的相对局部密度,p 为邻域图中边的条数.我们可以按照以下两步对阈值 ε 进行自适应选取:

(1) 对 $dr_1,dr_2,\ldots,dr_p$ 从小到大进行排序,取其前半部分序列,有 $dr'_1 \le dr'_2 \le dr'_{\lfloor \frac{p}{2} \rfloor}$.

(2) $\varepsilon = dr'_v$,其中, $v = \arg\max_i (dr'_i - dr'_{i-1})$.

4 实验结果

在这一节,我们将以不同的邻域大小 K 分别运行 ISOMAP 算法和 P-ISOMAP 算法,以比较它们对邻域大小的依赖程度.我们采用的数据集是具有 2 000 个数据点的 swiss roll 数据集[1,2](如图 3(a)所示).在实验中,我们设定 n=500,矢量量化方法采用 Matlab v6.5 工具箱中的 k-均值算法.如上节所述,我们也设定带宽 h=1.实验结果分别如图 4~图 6 所示.



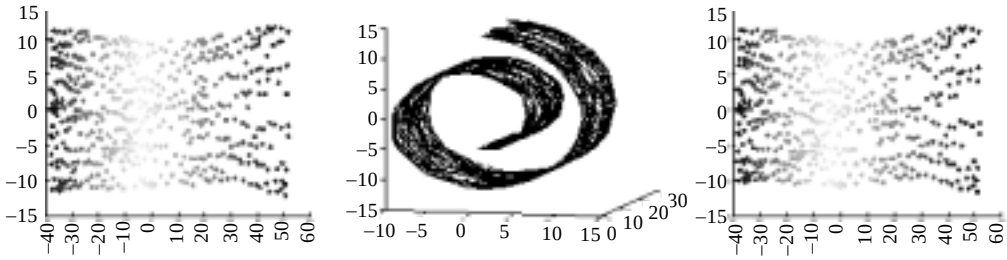
(a) (b)

Fig.3 (a) The swiss roll data set with 2 000 points[1,2], and (b) the data set shown in Fig.3(a), with zero-mean normally distributed noise, the standard deviation of which is 2% of smallest dimension of the bounding box enclosing the data[4]

图 3 (a) 具有 2 000 个数据点的 swiss roll 数据集[1,2],(b) 在图 3(a)中的每一个数据点上都加入一个服从正态分布的噪音,该正态分布均值为 0,标准差为该数据集在各维上跨度最小值的 2%[4]

从实验结果可以看出:P-ISOMAP 算法能够有效删除邻域图中可能存在的“短路”边(分别如图 5(b)和图 6(b)

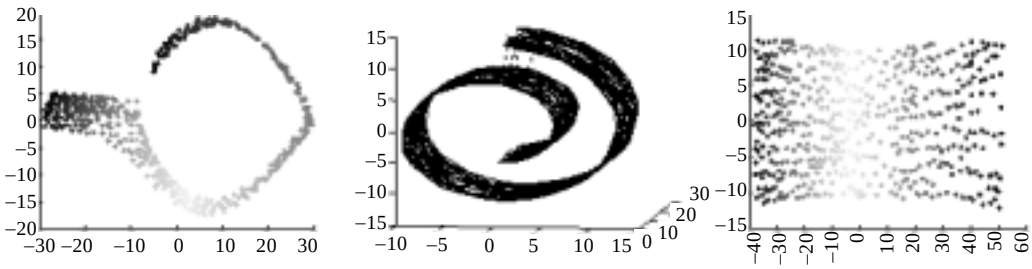
中的虚线所示),从而极大地削弱了邻域大小的影响(如图 5、图 6 所示,P-ISOMAP 算法在 K=10 和 K=12 时依然能正确发现数据的全局几何结构并对其进行可视化;而 ISOMAP 算法则因为邻域图中出现了“短路”边而无法被成功运用),也得到了更好的可视化效果.



(a) The output of ISOMAP (b) The corresponding neighborhood graph (c) The output of P-ISOMAP
(a) ISOMAP 算法的运行结果 (b) 相应的邻域图 (c) P-ISOMAP 算法的运行结果

Fig.4 The results of ISOMAP and P-ISOMAP over the swiss roll data set, where K=8

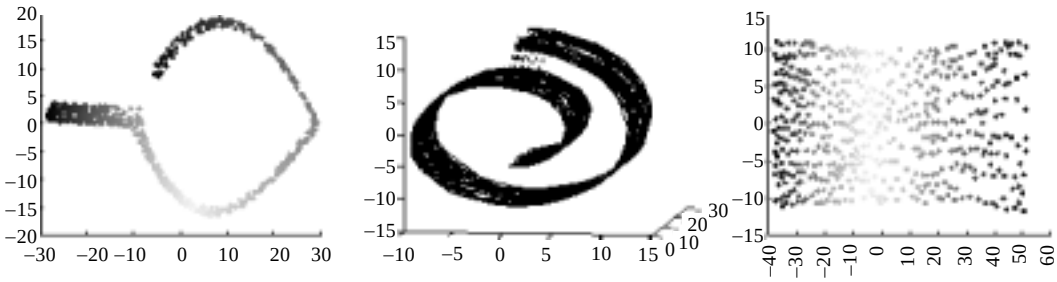
图 4 ISOMAP 算法和 P-ISOMAP 算法在 swiss roll 数据集上的运行结果,其中,K=8



(a) The output of ISOMAP (b) The corresponding neighborhood graph (c) The output of P-ISOMAP
(a) ISOMAP 算法的运行结果 (b) 相应的邻域图 (c) P-ISOMAP 算法的运行结果

Fig.5 The results of ISOMAP and P-ISOMAP over the swiss roll data set, where K=10

图 5 ISOMAP 算法和 P-ISOMAP 算法在 swiss roll 数据集上的运行结果,其中,K=10

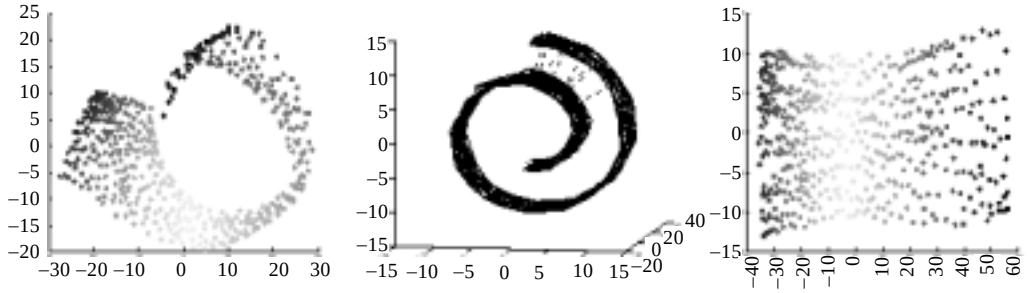


(a) The output of ISOMAP (b) The corresponding neighborhood graph (c) The output of P-ISOMAP
(a) ISOMAP算法的运行结果 (b) 相应的邻域图 (c) P-ISOMAP算法的运行结果

Fig.6 The results of ISOMAP and P-ISOMAP over the swiss roll data set, where K=12

图 6 ISOMAP 算法和 P-ISOMAP 算法在 swiss roll 数据集上的运行结果,其中,K=12

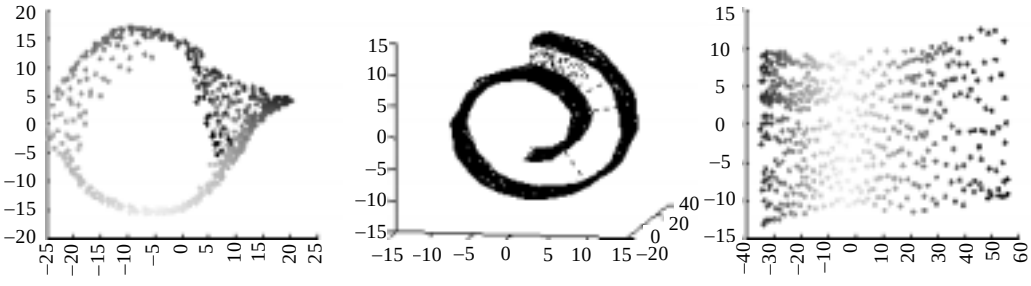
众所周知,ISOMAP 算法对噪音是比较敏感的,这也是因为邻域图中出现了“短路”边的缘故.为了验证 P-ISOMAP 算法能否通过删除“短路”边来克服这一问题,我们在加入了噪音的 swiss roll 数据集[4](如图 3(b)所示)上以同样的参数再次运行 ISOMAP 算法和 P-ISOMAP 算法.实验结果分别如图 7~图 9 所示.



(a) The output of ISOMAP (b) The corresponding neighborhood graph (c) The output of P-ISOMAP
(a) ISOMAP算法的运行结果 (b) 相应的邻域图 (c) P-ISOMAP算法的运行结果

Fig.7 The results of ISOMAP and P-ISOMAP over the noisy swiss roll data set, where K=8

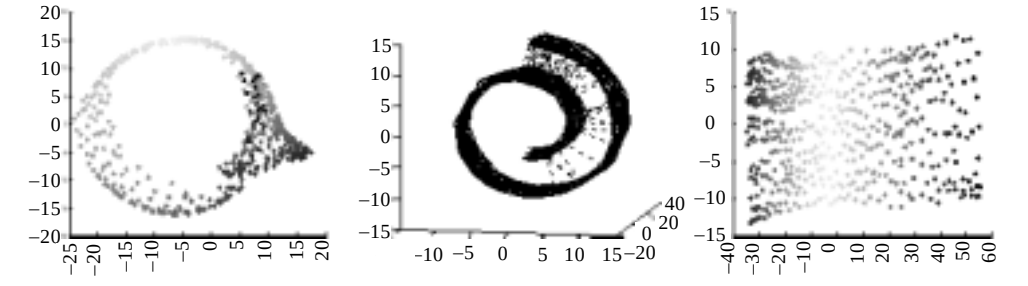
图 7 ISOMAP 算法和 P-ISOMAP 算法在加入了噪音的 swiss roll 数据集上的运行结果,其中,K=8



(a) The output of ISOMAP (b) The corresponding neighborhood graph (c) The output of P-ISOMAP
(a) ISOMAP 算法的运行结果 (b) 相应的邻域图 (c) P-ISOMAP 算法的运行结果

Fig.8 The results of ISOMAP and P-ISOMAP over the noisy swiss roll data set, where K=10

图 8 ISOMAP 算法和 P-ISOMAP 算法在加入了噪音的 swiss roll 数据集上的运行结果,其中,K=10



(a) The output of ISOMAP (b) the corresponding neighborhood graph (c) the output of P-ISOMAP
(a) ISOMAP 算法的运行结果 (b) 相应的邻域图 (c) P-ISOMAP 算法的运行结果

Fig.9 The results of ISOMAP and P-ISOMAP over the noisy swiss roll data set, where K=12

图 9 ISOMAP 算法和 P-ISOMAP 算法在加入了噪音的 swiss roll 数据集上的运行结果,其中,K=12

从实验结果可以看出:P-ISOMAP 算法在加入了噪音的 swiss roll 数据集上,同样能够有效删除邻域图中可能存在的“短路”边(分别如图 7(b)、图 8(b)和图 9(b)中的虚线所示),同样极大地削弱了邻域大小的影响(如图 7~图 9 所示:P-ISOMAP 算法在 K=8,K=10 和 K=12 时,依然能正确发现数据的全局几何结构并对其进行可视化;而 ISOMAP 算法则因为邻域图中出现了“短路”边而无法被成功运用),也得到了更好的可视化效果,进而说明了 P-ISOMAP 算法比 ISOMAP 算法更具鲁棒性.除此之外,从以上实验结果还可以看出,P-ISOMAP 算法比 ISOMAP 算法更具拓扑稳定性.

由于残差可用来衡量映射“质量”的高低以及邻域大小的相对合适程度,因此,我们在以上两个数据集上又分别计算了 ISOMAP 算法和 P-ISOMAP 算法相对于不同邻域大小 K 的残差(如图 10 所示),从中我们还可以进一步验证以上结论:P-ISOMAP 算法对邻域大小的依赖程度要比 ISOMAP 算法弱得多,这在一定程度上避免了邻域大小难以有效选取的问题,同时还能得到更好的可视化效果(具有更小的残差),进而也说明了 P-ISOMAP 算法更具鲁棒性和拓扑稳定性.

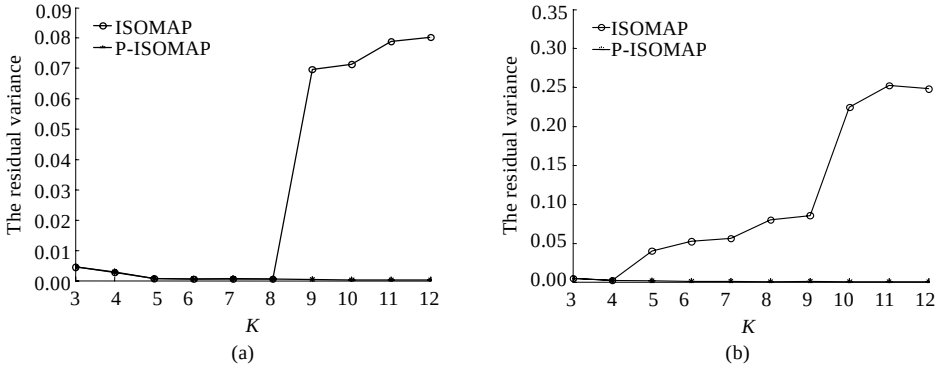


Fig.10 The residual variances obtained by ISOMAP and P-ISOMAP with different K over (a) the swiss roll data set, and (b) the noisy swiss roll data set

图 10 ISOMAP 算法和 P-ISOMAP 算法以不同邻域大小 K 在(a) swiss roll 数据集和(b) 加入了噪音的 swiss roll 数据集上得到的残差

5 结束语

针对 ISOMAP 算法很大程度上依赖于邻域大小,而邻域大小又难以有效选取这一问题,本文根据“短路”边会途经低密度区域的这一特点,提出了 ISOMAP 算法的一个变种——P-ISOMAP 算法.该算法通过有效删除邻域图中可能存在的“短路”边,极大地削弱了邻域大小的影响,从而更具鲁棒性和拓扑稳定性.由于该算法在一定程度上避免了邻域大小难以有效选取的问题,从而能够更容易地对数据进行可视化.

References:

[1] Tenenbaum JB. Mapping a manifold of perceptual observations. In: Jordan MI, Kearns MJ, Solla SA, eds. Advances in Neural Information Processing Systems 10. Denver: MIT Press, 1997. 682−689.

[2] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(22):2319−2323.

[3] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290(22):2323−2326.

[4] Balasubramanian M, Schwartz EL, Tenenbaum JB, de Silva V, Langford JC. The ISOMAP algorithm and topological stability. Science, 2002,295(5552):7. http://www.sciencemag.org/cgi/content/full/295/5552/7a

[5] Kouropteva O, Okun O, PietikÄainen M. Selection of the optimal parameter value for the locally linear embedding algorithm. In: Wang L, Halgamuge SK, Yao X, eds. Proc. of the 1st Int'l Conf. on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age. Singapore, 2002. 359−363. http://www.ee.oulu.fi/research/imag/document/publications/FSKD02.pdf

[6] Shao C, Huang HK. Selection of the optimal parameter value for the isomap algorithm. In: Gelbukh A, de Albornoz A, Terashima H, eds. Proc. of the 4th Mexican Int'l Conf. on Artificial Intelligence. Monterrey: Springer-Verlag, 2005. 396−404.

[7] Zhou ZH, Cao CG. Neural Networks and Its Applications. Beijing: Tsinghua University Press, 2004. 172−207 (in Chinese).

[8] Bernstein M, de Silva V, Langford JC, Tenenbaum JB. Graph approximations to geodesics on embedded manifolds. Technical Report, Stanford University, 2000. http://isomap.stanford.edu/BdSLT.pdf

[9] Lee JA, Lendasse A, Verleysen M. Nonlinear projection with curvilinear distances: ISOMAP versus curvilinear distance analysis. Neurocomputing, 2004,57(1):49–76.

[10] Carreira-Perpiñán MÁ. A review of dimension reduction techniques. Technical Report, CS-96-09, Sheffield: University of Sheffield, 1996. 46–53.

[11] Pelletier B. Kernel density estimation on riemannian manifolds. Statistics & Probability Letters, 2005,73(3):297–304.

附中文参考文献:

[7] 周志华,曹存根.神经网络及其应用.北京:清华大学出版社,2004.172–207.

邵超(1977－),男,河南渑池人,博士,讲师,主要研究领域为人工神经网络,机器学习,数据可视化,数据挖掘.

赵连伟(1976－),男,博士,主要研究领域为流形学习,统计学习理论,人工神经网络.

黄厚宽(1940－),男,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能,机器学习,数据仓库,数据挖掘,多智能体系统.

****