

基于随机游走的流形学习与可视化

邵超 万春红 张啸剑

(河南财经政法大学计算机与信息工程学院, 郑州, 450046)

摘要: 现有的全局流形学习算法都敏感于邻域大小这一难以高效选取的参数, 它们都采用了基于欧氏距离的邻域图创建方法, 从而使邻域图容易产生“短路”边。本文提出了一种基于随机游走模型的全局流形学习算法(Random walk-based isometric mapping, RW-ISOMAP)。和欧氏距离相比, 由随机游走模型得到的通勤时间距离是由给定两点间的所有通路以概率为权组合而成的, 不但鲁棒性更高, 而且还能在一定程度上度量具有非线性几何结构的数据之间的相似性。因此采用通勤时间距离来创建邻域图的RW-ISOMAP算法将不再敏感于邻域大小参数, 从而可以更容易地选取邻域大小参数, 同时还具有更高的鲁棒性。最后的实验结果证实了该算法的有效性。

关键词: 全局流形学习; 等距映射; 邻域图; 随机游走; 通勤时间距离

中图分类号: TP181 **文献标志码:** A

Manifold Learning and Visualization Based on Random Walk

Shao Chao, Wan Chunhong, Zhang Xiaojian

(School of Computer and Information Engineering, Henan University of Economics and Law, Zhengzhou, 450046, China)

Abstract: The existing global manifold learning algorithms are relatively sensitive to the neighborhood size, which is difficult to select efficiently. The reason is mainly because the neighborhood graph is constructed based on Euclidean distance, by which shortcut edges tend to be introduced into the neighborhood graph. To overcome this problem, a global manifold learning algorithm is proposed based on random walk, called the random walk-based isometric mapping (RW-ISOMAP). Compared with Euclidean distance, the commute time distance, achieved by the random walk on the neighborhood graph, can measure the similarity between the given data within the nonlinear geometric structure to a certain extent, thus it can provide robust results and is more suitable to construct the neighborhood graph. Consequently, by constructing the neighborhood graph based on the commute time distance, RW-ISOMAP is less sensitive to the neighborhood size and more robust than the existing global manifold learning algorithms. Finally, the experiment verifies the effectiveness of RW-ISOMAP.

Key words: global manifold learning; isometric mapping; neighborhood graph; random walk; commute time distance

引言

随着大数据时代的到来,人们越来越重视通过维数约简方法来获得高维数据的低维简洁表示,这样可以极大地提高数据洞察力。近年来,人们提出了多种流形学习算法,可以发现很多高维非线性数据内在的低维流形,成为目前比较有效的非线性维数约简方法^[1],如典型的全局流形学习算法等距映射^[2-4](Isometric mapping, ISOMAP)和典型的局部流形学习算法局部线性嵌入^[5-6](Locally linear embedding, LLE)等。其中,以等距映射为代表的一类流形学习算法通过保持数据点之间的全局度量如测地距离来获得数据的低维嵌入,从而能更好地保持数据的全局结构、获得更好的可视化结果,称为全局流形学习算法。这些全局流形学习算法都敏感于邻域大小这一实际中难以高效选取的参数^[7-9]。究其原因,它们都采用了基于欧氏距离的邻域图创建方法,容易在邻域图中产生“短路”边,从而使这些全局流形学习算法都敏感于邻域大小参数和数据中的噪声,鲁棒性较低。目前人们大都采用欧氏距离来创建邻域图,主要有 K-最近邻(K-nearest neighbor, KNN)法和 ϵ -邻域(ϵ -neighborhood)法等。为使邻域图能正确反映数据的内在几何结构,从而成功运行流形学习算法,人们通常使用残差来选取合适的邻域大小参数^[2,10],但需要多次运行整个流形学习算法,从而使该参数在实际中难以高效选取。通过对多个运行结果进行集成^[8],可以减弱邻域大小参数对流形学习算法的影响,但时间复杂度会明显增大。通过剔除邻域图中的“短路”边^[11-12],也可以减弱邻域大小参数对流形学习算法的影响,但同时又会引入新的同样难以高效选取的额外参数。文献[13,14]为每一个数据点自适应地选取不同的邻域大小参数,但计算相对复杂。

近年来,人们利用随机游走(Random walk)理论,在邻域图上得到了通勤时间距离^[15-18]这一新型的距离度量,它由邻域图上给定两点间的所有通路以概率为权组合而成。因此,和 ISOMAP 算法采用的单一最短路径距离相比,通勤时间距离更加鲁棒;和创建邻域图常用的欧氏距离相比,通勤时间距离能在一定程度上度量具有非线性几何结构的数据之间的相似性^[15-18]。因此,采用通勤时间距离来创建邻域图不但能更好地避免“短路”边,减弱邻域大小参数对流形学习算法的影响,从而更容易地选取邻域大小参数,而且还能使流形学习算法获得更高的鲁棒性,这就是本文算法的出发点。

1 随机游走模型和通勤时间距离

数据分析首先需要计算任意两个数据点之间的相似度,通常采用

$$w_{ij} = \begin{cases} e^{-\frac{|x_i - x_j|^2}{t}} & i \in N_k(j) \text{ 或 } j \in N_k(i) \\ 0 & \text{其他} \end{cases} \quad (1)$$

式中: w_{ij} 为数据点 x_i 和 x_j 之间的相似度,它有两个参数: t 和 k , t 为尺度参数, k 为邻域大小参数; $|x_i - x_j|$ 为 x_i 和 x_j 之间的欧氏距离; $N_k(j)$ 为 x_j 的 k 最近邻点集合(采用欧氏距离), $i \in N_k(j)$ 表示 x_i 是 x_j 的一个 K-最近邻点。由此 K-最近邻法可以得到该数据集 $\mathbf{X} = (x_i), i = 1, \dots, n$ 上的一个邻域图 $G = (\mathbf{V}, E, \mathbf{W}), \mathbf{V} = \mathbf{X} = (x_i), i = 1, \dots, n, \langle i, j \rangle \in E$ 当且仅当 $i \in N_k(j)$ 或 $j \in N_k(i), \mathbf{W} = (w_{ij})_{n \times n}$ 为 G 的相似度

矩阵。令 $\mathbf{D} = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix}$ (或记作 $\mathbf{D} = \text{diag}(d_i)$), 其中, $d_i = \sum_j w_{ij}$ 为 x_i 的度(degree)。则 $\mathbf{P} =$

$(p_{ij})_{n \times n} = \mathbf{D}^{-1} \mathbf{W}$ 可视为邻域图 G 上的一个转移概率矩阵,从 x_i 转移到 x_j 的概率为 $P(x_j | x_i) = w_{ij} / d_i = p_{ij}$ (存在 $\sum_j p_{ij} = 1$), 由此便得到了 G 上的一个随机游走模型。在该模型中,从 x_i 出发,以概率 p_k 转移到 x_k , 最终首次到达 x_j 的期望步数 $h(j | i)$ 被定义为平均首次命中时间^[15-16] (Average first passage/hitting time), 则

$$\begin{cases} h(j|j) = 0 \\ h(j|i) = 1 + \sum_{k=1}^n p_{ik} h(j|k) \quad i \neq j \end{cases} \quad (2)$$

该度量将邻域图上从 x_i 到 x_j 的所有通路以概率为权进行组合,不但能在一定程度上度量具有非线性几何结构的数据之间的相似性^[15-18](据此创建的邻域图不容易产生“短路”边,因此对邻域大小参数并不敏感),而且比单一的最短路径距离更加鲁棒,但它并不对称,经对称化后为通勤时间距离^[15-16],则

$$c(i, j) = h(i|j) + h(j|i) \quad (3)$$

经证明,通勤时间距离可以根据式(4)计算得到^[15,16],则

$$c^2(i, j) = V_G(L_{ii}^+ + L_{jj}^+ - 2L_{ij}^+) \quad (4)$$

式中: $c(i, j)$ 为 x_i 和 x_j 之间的通勤时间距离, $V_G = \sum_i d_i$ 为 G 的总度, $L^+ = \{L_{ij}^+\}_{n \times n}$ 为 Laplacian 矩阵 $L = D - W$ 的广义逆(Moore-Penrose pseudoinverse)矩阵。和主成分分析(Principal components analysis, PCA)算法类似,文献[15-18]给出了基于通勤时间距离的维数约简方法-通勤时间嵌入(Commute time embedding, CTE)算法,则

$$Y = \Phi \Lambda^{-\frac{1}{2}} \quad (5)$$

式中: Λ 为包含 L^+ 最大几个(为了数据可视化,通常为 2 或 3)特征值的对角矩阵, Φ 为对应的特征向量矩阵。给定两点间的通路越多,或某些通路越短,其通勤时间距离就会越小^[15-18](如图 1 所示)。因此,CTE 算法会扭曲数据的全局几何结构(见随后的实验结果)。另外,基于通勤时间距离来创建的邻域图可能会漏掉其中某些最近邻点(如图 1 所示,如果采用通勤时间距离,数据点 B 就不再是 A 的 k -最近邻点($k=2$)),从而难以确保最短路径距离对测地距离这一全局度量的良好逼近^[2]。在图 1 中,数据集 $\{A, B, C, D\}$ 的邻域图($k=2$),如细实线表示,其中,数据点 A 的两个 k -最近邻点为 B 和 C 。然而,如果采用通勤时间距离,则数据点 A 的两个 k -最近邻点不再是 B 和 C ,而是 C 和 D ,如粗虚线表示,这是因为 D 和 A 之间存在多条相对较短的通路,如 (A, D) 和 (A, C, D) 。

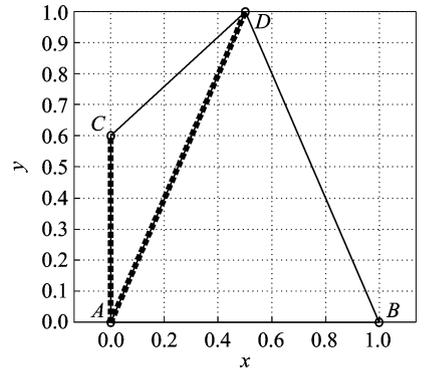


图 1 数据集 $\{A, B, C, D\}$ 的邻域图($k=2$)

Fig. 1 Neighborhood graph of data set $\{A, B, C, D\}$ ($k=2$)

2 RW-ISOMAP 算法

为了更好地展现数据的全局几何结构,本文依然采用 ISOMAP 算法的做法,即在低维嵌入空间中保持由最短路径距离逼近得到的流形上唯一有意义的全局度量—测地距离;但不同之处在于本文采用随机游走模型得到的通勤时间距离来创建邻域图,以此来减弱邻域大小参数对流形学习算法的影响,并提高流形学习算法的鲁棒性。为了确保对测地距离的良好逼近,首先应为每个数据点都保留若干个最近邻点,同时为了能在最大程度上避免“短路”边,本文采用使邻域图能够连通的最小 k 值(记作 k_m)来创建邻域图(称为最小连通邻域图,记作 G_m)。然后在最小连通邻域图的基础上使用通勤时间距离来为每个数据点增加更多的最近邻点。由于通勤时间距离考虑了数据内在的非线性几何结构,能在避免“短路”边的情况下使邻域图足够稠密,从而使测地距离能够得到更加精确的逼近^[19]。因此本文算法基于随机游走模型的全局流形学习算法(Random walk-based ISOMAP, RW-ISOMAP)可简要描述为

(1) 给定数据集 $\mathbf{X} = (x_i), i = 1, \dots, n$, 采用欧氏距离来创建最小连通邻域图 G_m ;

- (a) $k=0$;
 (b) Do
 (i) $k=k+1$;
 (ii) 采用 k -最近邻法创建邻域图 G ;
 (iii) 在 G 上运行广度优先搜索算法判断其是否连通;
 (c) Until G 连通;
 (d) $G_m=G; k_m=k$ 。

(2) 计算任一数据点 x_i 基于通勤时间距离的 k -最近邻点集合 $N_k(i)$:

- (a) 给定邻域大小参数 k_1 , 根据式(1)计算相似矩阵 $W=(w_{ij})_{n \times n}$;
 (b) 计算 Laplacian 矩阵 $L=D-W$, 其中, $D=\text{diag}(\sum_j w_{ij})$;
 (c) 计算 Laplacian 矩阵的广义逆矩阵 $L^+=(L-\frac{e \cdot e^T}{n})^{-1}+\frac{e \cdot e^T}{n}$, 其中, $e=(1, \dots, 1)^T$;
 (d) 根据式(4)计算所有数据点之间的通勤时间距离 $C=(c(i, j))_{n \times n}$;

(e) 计算 x_i 基于通勤时间距离的 k -最近邻点集合 $N_k(i)=\{j|j \neq i \wedge \text{IND}(c(i, j)) \leq k\}$ (和基于欧氏距离的 k -最近邻点集合 $N_k(i)$ 不同), 其中, k 为邻域大小参数, $\text{IND}(c(i, j))$ 为 $c(i, j)$ 在 $c(i, \cdot)$ (除 $c(i, i)$ 外) 升序排列中的次序。

(3) 给定邻域大小参数 k_2 , 对于任一 x_i , 如果 $j \in N_{k_2}(i)$, 则将连接边 $\langle i, j \rangle$ 加入到邻域图 G_m 中, 最终得到邻域图 G 。

(4) 在 G 上计算所有最短路径距离 (同 ISOMAP 算法)。

(5) 运行古典多维尺度分析 (Classical multi-dimensional scaling, CMDS) 算法, 得到所有数据点的低维嵌入 (同 ISOMAP 算法)。如上所述, RW-ISOMAP 算法主要有 3 个参数: k_1, k_2 和 t 。其中, 邻域大小参数 k_1 和尺度参数 t 用于在式(1)中计算通勤时间距离。由于通勤时间距离以概率为权对给定两点间的所有通路进行加权, 转移概率相对较小且稀疏得多的“短路”边对通勤时间距离的影响会比较小 (如图 2 所示, 当 k_1 取不同值 (如 $k_1=8, 16, 24$) 时, 基于通勤时间距离的 CTE 算法其结果几乎相同), 从而使 RW-ISOMAP 算法对 k_1 并不敏感。图 2 所用算法为 CTE, 数据集为 Swiss roll (如图 3(a) 所示)。

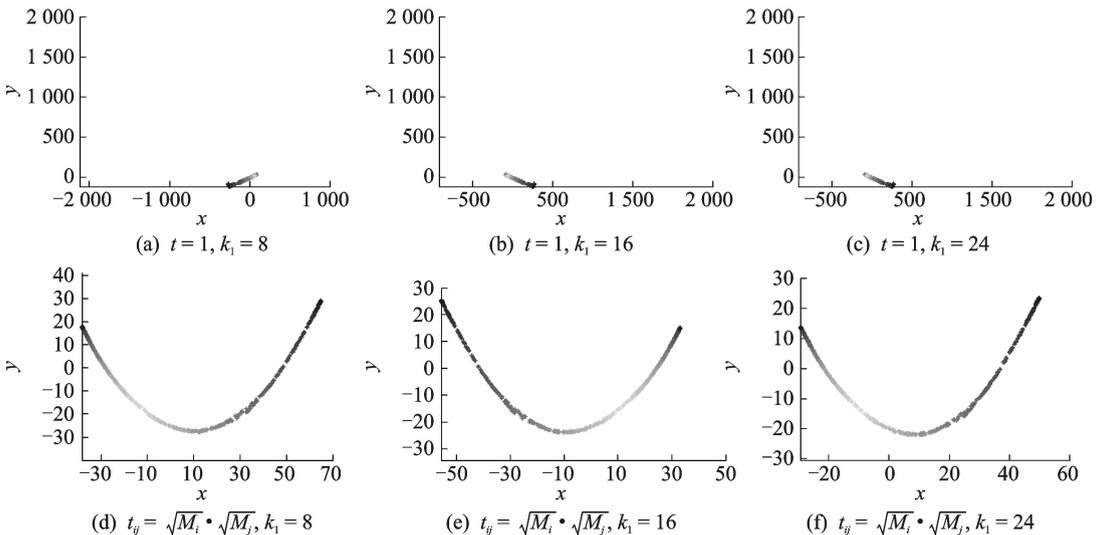


图 2 参数 t 对可视化结果的影响

Fig. 2 Influence of parameter t on visualization results

稀疏得多的“短路”边通常都是长边,为了给长边的两个顶点赋予更小的相似度(使其具有更小的转移概率,这样可以进一步降低“短路”边对通勤时间距离的不良影响),尺度参数 t 通常取比较小的值,如 $t=1$ 。为了更好地对分布不均匀的数据进行可视化,RW-ISOMAP 算法采用保角等距映射^[20](Conformal ISOMAP,C-ISOMAP)算法的作法,即

$$w_{ij} = \begin{cases} e^{-\frac{|x_i - x_j|^2}{t_{ij}}} & i \in N_{k_1}(j) \text{ 或者 } j \in N_{k_1}(i) \\ 0 & \text{其他} \end{cases} \quad (6)$$

$$t_{ij} = \sqrt{M_i} \cdot \sqrt{M_j} = \sqrt{\sum_{p=1}^{k_1} |x_i - x_{i_p}| / k_1} \cdot \sqrt{\sum_{p=1}^{k_1} |x_j - x_{j_p}| / k_1}$$

式中: x_{i_p} 和 x_{j_p} 分别为 x_i 和 x_j 的第 p 个 k_1 -最近邻点。该方法得到的可视化结果如图 2(d-f)所示。和图 2(a-c)相比,给不同的连接边 $\langle i, j \rangle$ 赋予不同的参数 t_{ij} 能够得到更好的可视化结果。因此随后实验中的 CTE 算法也将采用这一方法。邻域大小参数 k_2 用于在最小连通邻域图 G_m 上加入更多的连接边,由于通勤时间距离能够在一定程度上度量具有非线性几何结构的数据之间的相似性(如图 4(d-f)所示,在流形上相距较远的数据点,其通勤时间距离也比较大),采用通勤时间距离来创建邻域图不容易产生“短路”边,从而使 RW-ISOMAP 算法对 k_2 也不敏感。

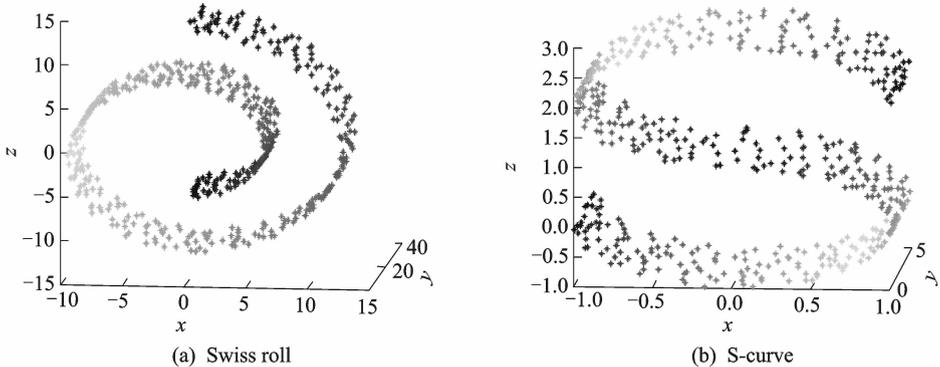


图 3 实验中的两个数据集

Fig. 3 Two data sets in experiments

3 实验结果

3.1 完整流形的实验结果

为了验证本文提出的 RW-ISOMAP 算法的有效性,本文在 Swiss roll^[2] 和 S-curve^[5] 数据集(具有 500 个数据点,如图 3 所示)上分别运行 ISOMAP,LLE,CTE 和 RW-ISOMAP 算法,实验结果分别如图 4,5 所示。如图 4(a,b,d,e)和图 5(a,b,d,e)所示,当邻域大小参数合适时,如 $k=8$ (Swiss roll)和 $k=16$ (S-curve),ISOMAP 和 LLE 算法都能较好地展现数据的内在几何结构,然而,当邻域大小参数取更大值时,如 $k=16$ (Swiss roll)和 $k=24$ (S-curve),ISOMAP 算法和 LLE 算法会扭曲数据的内在几何结构,相比之下,CTE 算法能较好地展现数据的内在几何结构,但会扭曲数据的全局几何结构,如图 4(c,f)和图 5(c,f)所示。如图 4(g-j)和图 5(g-j)所示,和 ISOMAP,LLE 算法相比,RW-ISOMAP 算法并不敏感于它的两个邻域大小参数 k_1 和 k_2 。和 LLE,CTE 算法相比,RW-ISOMAP 算法能更好地展现数据的全局几何结构,这是因为 RW-ISOMAP 算法保持的也是测地距离这一全局度量,而不是和通路条数相关的通

勤时间距离。为了验证 RW-ISOMAP 算法的鲁棒性,本文在这两个数据集上分别加入了一定程度的噪声^[21],并再次运行 ISOMAP 算法、LLE 算法、CTE 算法和 RW-ISOMAP 算法,实验结果分别如图 6,7 所示。如图 6(a-c)和图 7(a-c)所示,在加入了噪声的这两个数据集上,ISOMAP 和 LLE 算法已经不能正确地展现数据的内在几何结构,而 CTE 算法却依然能较好地展现数据的内在几何结构,从而具有一定

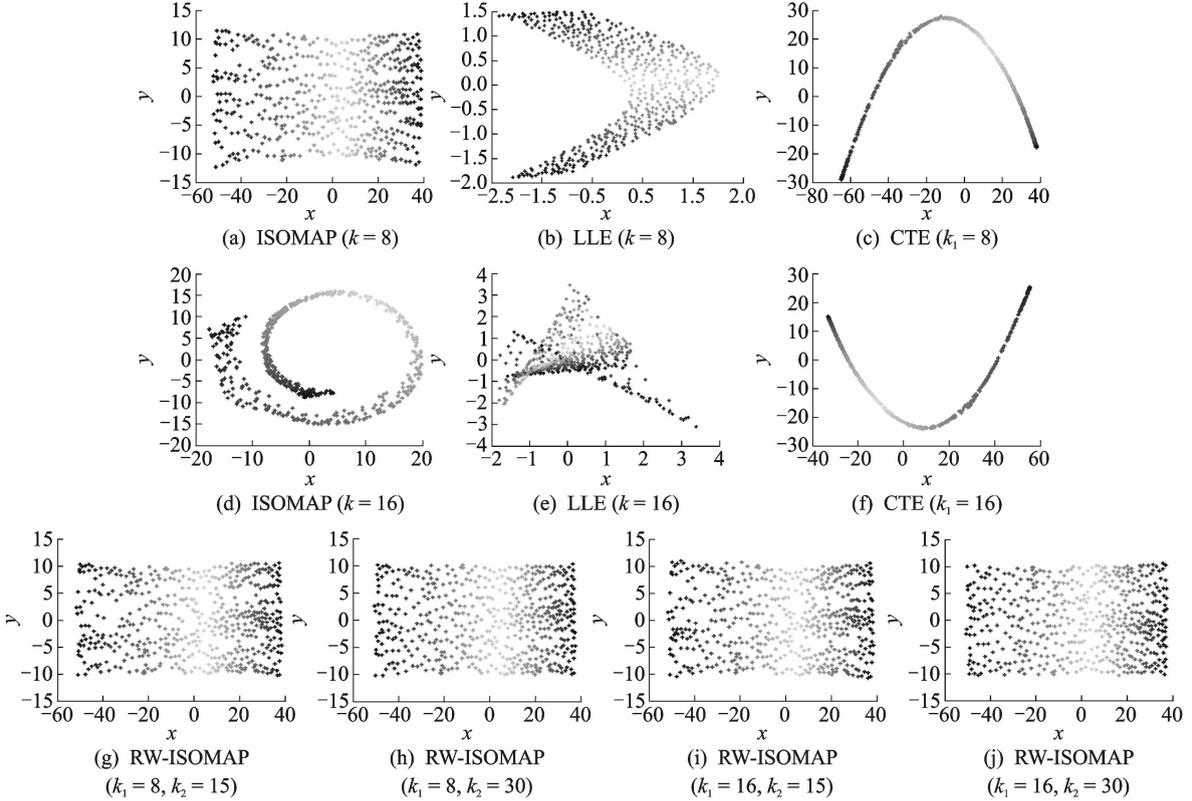
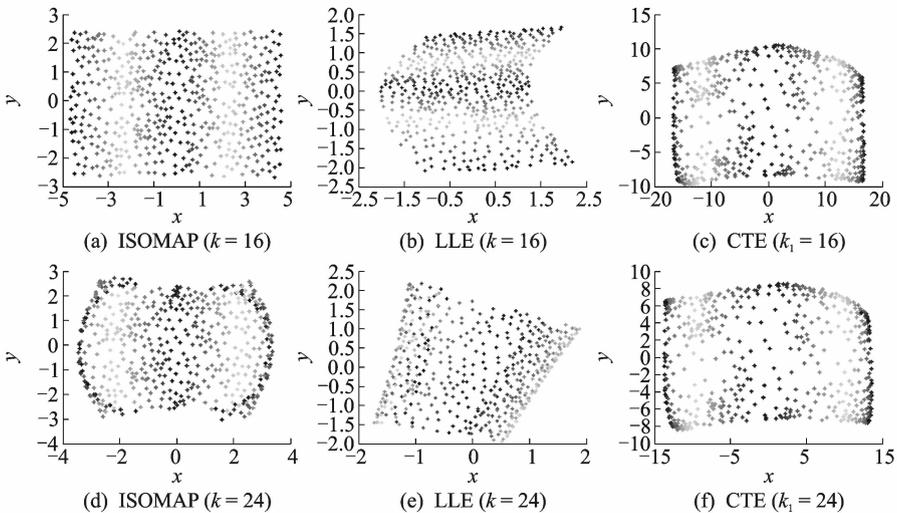


图 4 不同算法对 Swiss roll 的可视化结果

Fig. 4 Visualization results of different algorithms on Swiss roll



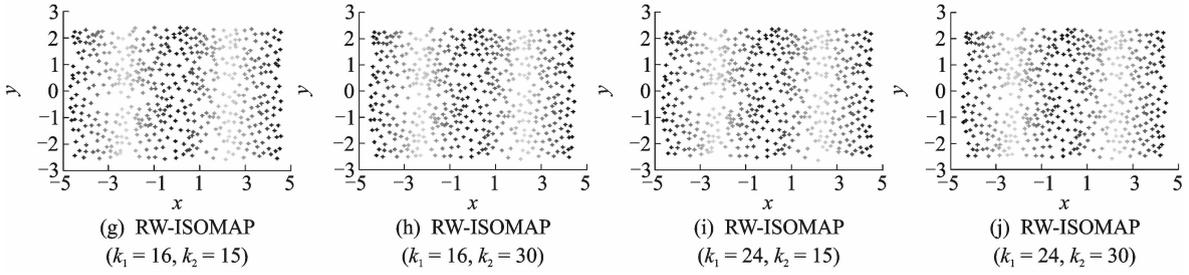


图5 不同算法对 S-curve 的可视化结果

Fig. 5 Visualization results of different algorithms on S-curve

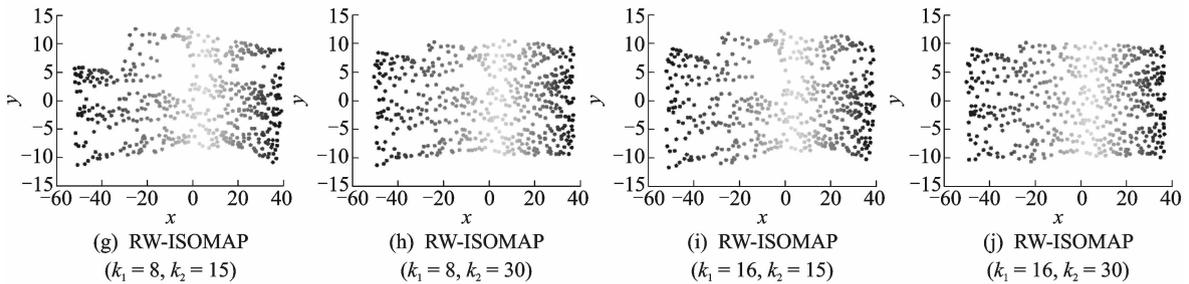
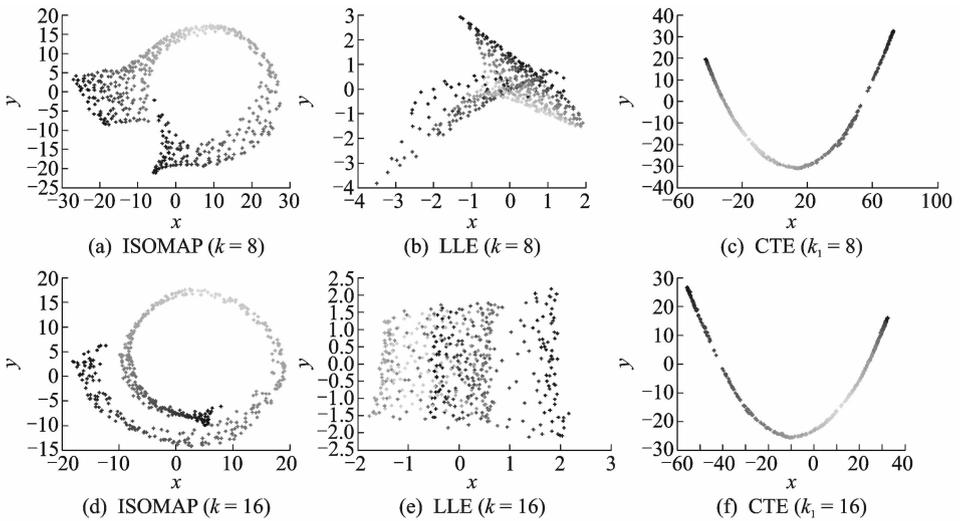
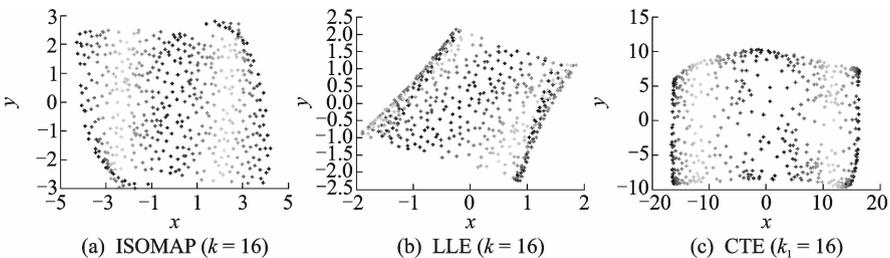


图6 不同算法对加入了噪声的 Swiss roll 的可视化结果

Fig. 6 Visualization results of different algorithms on Swiss roll with noise



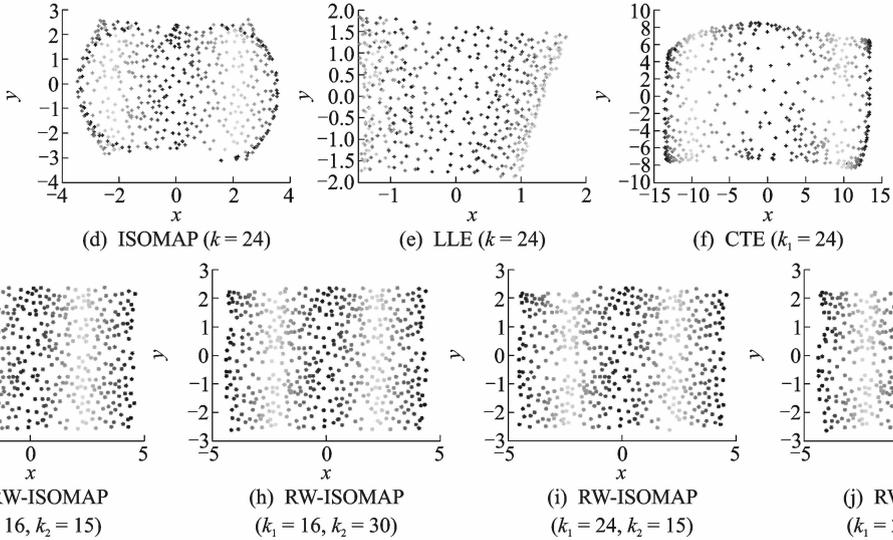


图7 不同算法对加入了噪声的 S-curve 的可视化结果

Fig. 7 Visualization results of different algorithms on S-curve with noise

的鲁棒性。相比之下,如图 6(g-j)和图 7(g-j)所示,在加入了噪声的这两个数据集上,RW-ISOMAP 算法依然能较好地展现数据的内在几何结构,不但具有和 CTE 算法(如图 6(c,f)和图 7(c,f)所示)相当的鲁棒性,而且具有更好的全局保持能力。众所周知,残差可用来衡量低维嵌入映射的质量,进而用来判断流形学习算法对邻域大小参数的敏感程度。如图 8 所示,RW-ISOMAP 和 CTE 算法对邻域大小参数的敏感程度远低于 ISOMAP 和 LLE 算法。此外 RW-ISOMAP 算法得到的映射质量要远好于 CTE 算法(表现为更小的残差)。

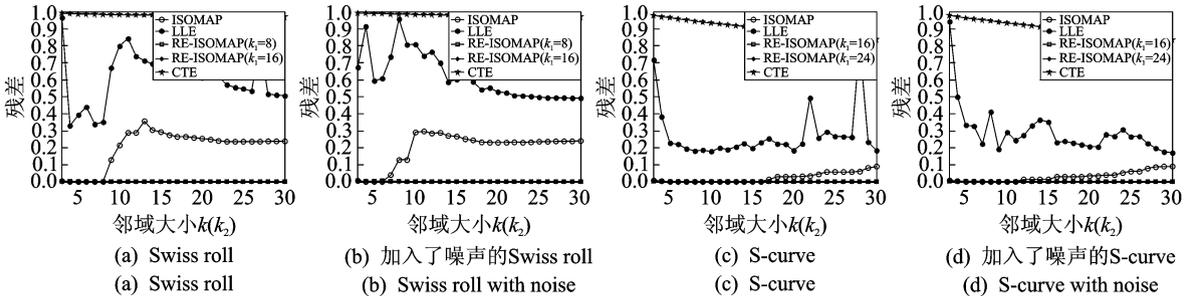


图8 不同算法得到的残差

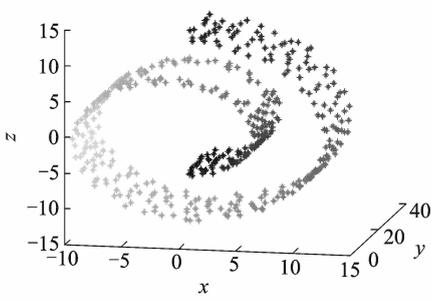
Fig. 8 Residual variances obtained by different algorithms on different data sets

3.2 不完整流形的实验结果

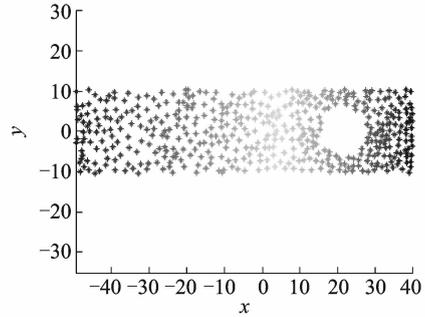
为了验证 RW-ISOMAP 算法在不完整流形上的有效性,本文在 Swiss roll 数据集上去掉一些数据点形成一个空洞(如图 9 所示),再次运行 ISOMAP 算法、LLE 算法、CTE 算法和 RW-ISOMAP 算法,实验结果如图 10 所示,RW-ISOMAP 算法不但能较好地展现该数据的内在几何结构(ISOMAP 算法只有在邻域大小参数合适时才能做到这一点),而且对邻域大小参数也不敏感。

3.3 实际数据集的实验结果

为了验证 RW-ISOMAP 算法在实际数据集上的有效性,本文在人脸数据集^[2]上再次运行 ISOMAP 算法、LLE 算法、CTE 算法和 RW-ISOMAP 算法。如图 11 所示,RW-ISOMAP 和 CTE 算法对邻域大小



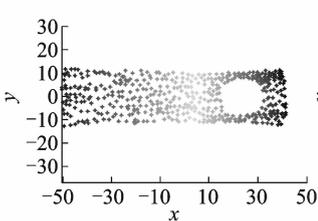
(a) 带有空洞的Swiss roll
(a) Swiss roll with hole



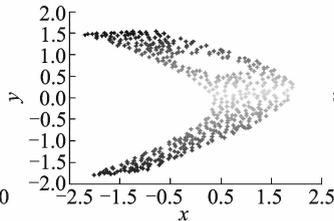
(b) 其内在低维流形
(b) Its intrinsic low-dimensional manifold

图9 带有空洞的Swiss roll及其内在低维流形

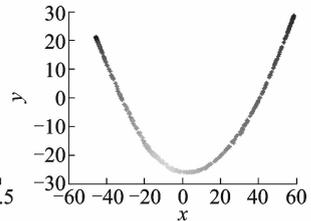
Fig. 9 Swiss roll data set with hole and its intrinsic low-dimensional manifold



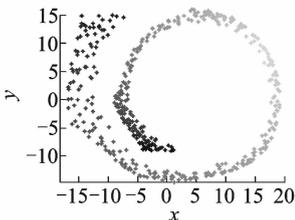
(a) ISOMAP ($k=8$)



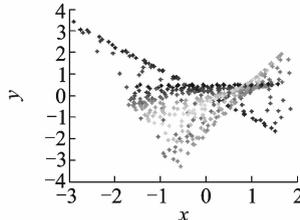
(b) LLE ($k=8$)



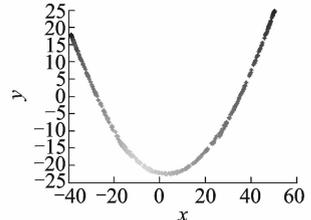
(c) CTE ($k_1=8$)



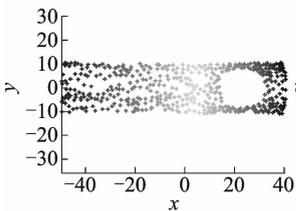
(d) ISOMAP ($k=16$)



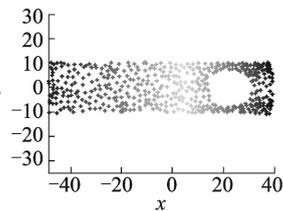
(e) LLE ($k=16$)



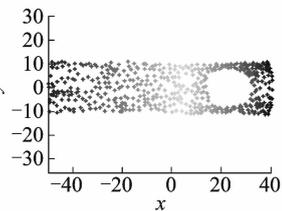
(f) CTE ($k_1=16$)



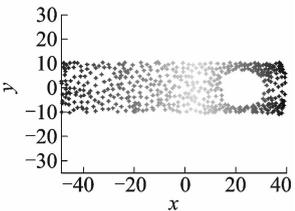
(g) RW-ISOMAP
($k_1=8, k_2=15$)



(h) RW-ISOMAP
($k_1=8, k_2=30$)



(i) RW-ISOMAP
($k_1=16, k_2=15$)



(j) RW-ISOMAP
($k_1=16, k_2=30$)

图10 不同算法对带有空洞的Swiss roll的可视化结果

Fig. 10 Visualization results of different algorithms on Swiss roll with hole

参数的敏感程度依然低于 ISOMAP 和 LLE 算法。给定邻域大小参数 $k=k_1=k_2=10$ (比最大合适邻域大小 $k=8$ (如图 11 所示) 稍大), 图 11 由该邻域大小参数在 ISOMAP, LLE 和 CTE 这 3 个算法中为 k , 而在 RW-ISOMAP 算法中为 k_2 。RW-ISOMAP 算法依然能获得人脸图像的有意义排列 (如图 12 所示)。

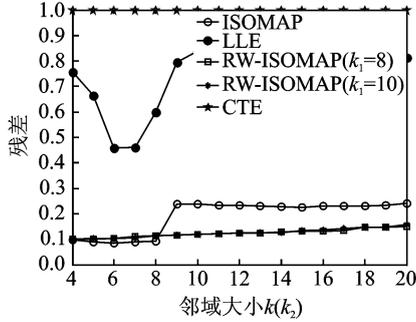


图 11 不同算法在人脸数据集上得到的残差

Fig. 11 Residual variances obtained by different algorithms on face data set

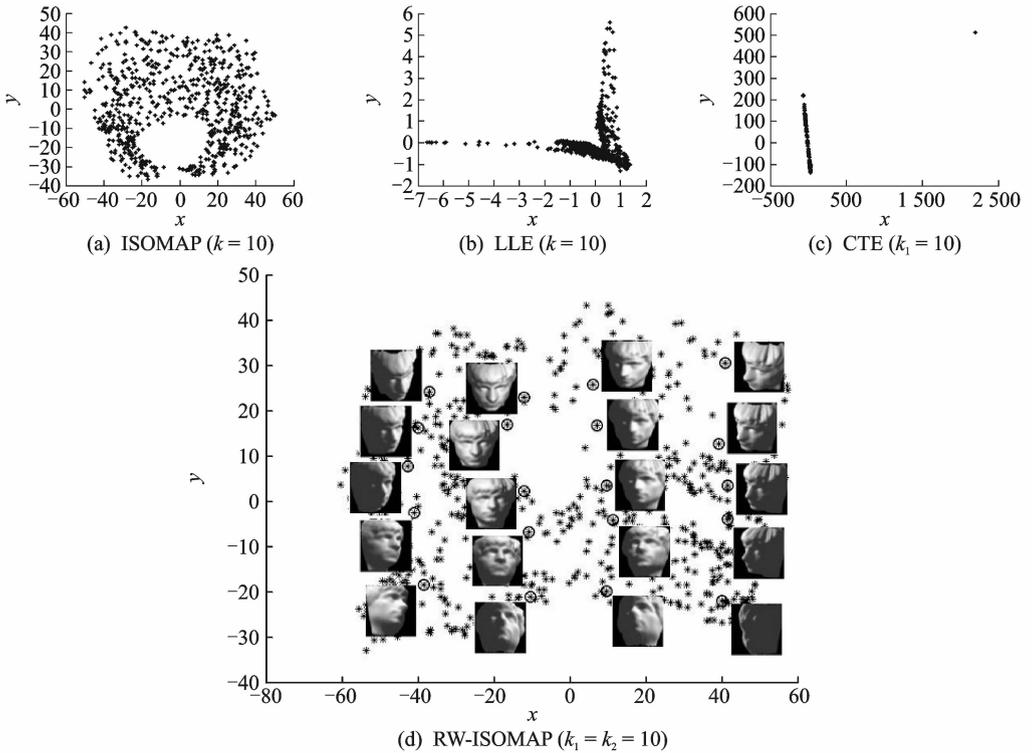


图 12 不同算法对人脸数据集的可视化结果

Fig. 12 Visualization results of different algorithms on face data set

4 结束语

由随机游走模型得到的通勤时间距离具有良好的鲁棒性,且考虑了数据内在的非线性几何结构,因此本文提出了基于随机游走的全局流形学习算法 RW-ISOMAP,以克服传统的基于欧氏距离的邻域图创建方法鲁棒性差、拓扑不稳定等问题。该算法在最小连通邻域图的基础上加入更多的基于通勤时间距离的连接边,从而能在避免“短路”边的前提下使邻域图足够稠密,以实现测地距离的良好逼近。因此该算法不但能真实地展现数据的全局几何结构,而且具有较高的鲁棒性。如何将该算法扩展至多流形数据将是下一步的研究内容。

参考文献:

- [1] 杨剑,李伏欣,王珏.一种改进的局部切空间排列算法[J].软件学报,2005,16(9):1584-1590.
Yang Jian, Li Fuxin, Wang Jue. A better scaled local tangent space alignment algorithm[J]. Journal of Software,2005,16(9):1584-1590.
- [2] Tenenbaum J B, de Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000,290(5500):2319-2323.
- [3] 王耀南,张莹,李春生.基于核矩阵的 ISOMAP 增量学习算法研究[J].计算机研究与发展,2009,46(9):1515-1522.
Wang Yaonan, Zhang Ying, Li Chunsheng. Kernel matrix based incremental learning isomap algorithm[J]. Journal of Computer Research and Development,2009,46(9):1515-1522.
- [4] 吴文通,李元祥,韦邦合,等.局部测地距离估计的增量等距特征映射算法[J].上海交通大学学报,2013,47(7):1082-1086.
Wu Wentong, Li Yuanxiang, Wei Banghe, et al. Incremental ISOMAP method based on locally estimated geodesic distance [J]. Journal of Shanghai Jiaotong University,2013,47(7):1082-1086.
- [5] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science,2000,290(5500):2323-2326.
- [6] Zhang S. Enhanced supervised locally linear embedding[J]. Pattern Recognition Letters,2009,30(13):1208-1218.
- [7] Saul L K, Roweis S T. Think globally, fit locally: Unsupervised learning of low dimensional manifolds[J]. Journal of Machine Learning Research,2003,4(1):119-155.
- [8] 詹德川,周志华.基于集成的流形学习可视化[J].计算机研究与发展,2005,42(9):1533-1537.
Zhan Dechuan, Zhou Zhihua. Ensemble based manifold learning for visualization[J]. Journal of Computer Research and Development,2005,42(9):1533-1537.
- [9] 邵超,黄厚宽,赵连伟.一种更具拓扑稳定性的 ISOMAP 算法[J].软件学报,2007,18(4):869-877.
Shao Chao, Huang Houkuan, Zhao Lianwei. A more topologically stable ISOMAP algorithm[J]. Journal of Software,2007,18(4):869-877.
- [10] Samko O, Marshall A D, Rosin P L. Selection of the optimal parameter value for the ISOMAP algorithm[J]. Pattern Recognition Letters,2006,27(1):968-979.
- [11] Wen G, Jiang L, Shadbolt N R. Using graph algebra to optimize neighborhood for isometric mapping[C]//Velo M M. Proceedings of the 20th International Joint Conference on Artificial Intelligence. Menlo Park,CA: AAAI Press,2007:2398-2403.
- [12] Yang L. Building k edge-disjoint spanning trees of minimum total length for isometric data embedding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2005,27(10):1680-1683.
- [13] 文贵华,江丽君,文军.邻域参数动态变化的局部线性嵌入[J].软件学报,2008,19(7):1666-1673.
Wen Guihua, Jiang Lijun, Wen Jun. Dynamically determining neighborhood parameter for locally linear embedding[J]. Journal of Software,2008,19(7):1666-1673.
- [14] Zhang Z, Wang J, Zha H. Adaptive manifold learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012,34(2):1473-1480.
- [15] Qiu H, Hancock E R. Clustering and embedding using commute times[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2007,29(11):1873-1890.
- [16] Fouss F, Pirotte A, Renders J M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation[J]. IEEE Transactions on Knowledge and Data Engineering,2007,19(3):355-369.
- [17] Albano J A, Messinger D W. Euclidean commute time distance embedding and its application to spectral anomaly detection [C]//Proceedings of SPIE 8390. Bellingham,WA:SPIE,2012:1-15.
- [18] Kim K H, Choi S. Walking on minimax paths for k-NN search[C]//Proceedings of the 27th AAAI Conference on Artificial Intelligence. Bellevue, Washington, USA: AAAI Press,2013:518-525.
- [19] Jing L, Shao C. Selection of the suitable parameter value for ISOMAP[J]. Journal of Software,2011,6(6):1034-1041.
- [20] de Silva V, Tenenbaum J B. Global versus local methods in nonlinear dimensionality reduction[J]. Advances in Neural Information Processing Systems,2003,15:721-728.
- [21] Balasubramanian M, Shwartz E L, Tenenbaum J B, et al. The ISOMAP algorithm and topological stability[J]. Science, 2002,295(5552):7.

作者简介:



邵超(1977-),男,博士,教授,硕士生导师,研究方向:机器学习、数据挖掘与可视化,E-mail:sc_flying@163.com。



万春红(1982-),女,讲师,研究方向:机器翻译与自然语言处理。



张啸剑(1980-),男,博士,副教授,研究方向:数据挖掘、图数据管理与隐私保护。