ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.9, September 2009, pp.2376−2386
doi: 10.3724/SP.J.1001.2009.03369

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

基于相对流形的局部线性嵌入*

文贵华[1+], 陆庭辉[1], 江丽君[2], 文 军[3]

[1](华南理工大学 计算机科学与工程学院,广东 广州 510641)

[2](华南理工大学 电子材料科学与工程系,广东 广州 510641)

[3](湖北民族学院 理学院,湖北 恩施 445000)

Locally Linear Embedding Based on Relative Manifold

WEN Gui-Hua[1+], LU Ting-Hui[1], JIANG Li-Jun[2], WEN Jun[3]

[1](School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China)

[2](Department of Electronic Material Science and Engineering, South China University of Technology, Guangzhou 510641, China)

[3](School of Mathematical Science, Hubei Insitute for Nationalities, Enshi 445000, China)

+ Corresponding author: E-mail: crghwen@scut.edu.cn

Wen GH, Lu TH, Jiang LJ, Wen J. Locally linear embedding based on relative manifold. *Journal of Software*, 2009,20(9):2376−2386. http://www.jos.org.cn/1000-9825/3369.htm

Abstract: Locally linear embedding greatly depends on whether the neighborhood graph can realistically reflect the underlying geometry structure of the data manifolds. The topological structure of constructed neighborhood with the existing approaches is unstable. It is sensitive to the noisy and sparse data sets. Based on the relative cognitive law, the relative transformation is presented, by which the relative space and the relative manifold are further constructed. The relative transformation can improve the distinguishing ability between data points and reduce the impact of noise and sparsity of data. To determine the neighborhood in the relative space and the relative manifold can more truly reflect the manifold structure, based on which the enhanced local linear embedding algorithms are developed with significantly improved performance. Besides, the speed is also enhanced with this approach. The experiments on challenging benchmark data sets validate the proposed approach.

Key words: locally linear embedding; relative transformation; relative manifold; neighborhood graph

摘 要: 局部线性嵌入算法极大地依赖于邻域是否真实地反映了流形的内在结构,现有方法构造的邻域结构是拓扑不稳定的,对噪音和稀疏数据敏感.根据认知的相对性规律提出了相对变换,并用其构造了相对空间和相对流形.相对变换可以提高数据之间的可区分性,并能抑制噪音和数据稀疏的影响.在构造的相对空间和相对流形上确定数据点的邻域能够更真实地反映流形的内在结构,由此提出了增强的局部线性嵌入算法,明显地提高了性能,特别是基于流形的方法还同时提高了速度.标准数据集上的实验结果验证了该方法的有效性.

关键词: 局部线性嵌入;相对变换;相对流形;邻域图

* Supported by the Key Science-Technology Project of Hubei Province of China under Grant No.2005AA101C17 (湖北省科技攻关项目); the Key Science-Technology Project of Guangdong of China under Grant No.2007B030803006 (广东省科技攻关项目); the Project of Scientific Research Foundation for the Returned Overseas Chinese Scholars (国家教育部留学回国人员科研启动基金)

Received 2007-11-06; Revised 2008-02-01; Accepted 2008-03-14

中图法分类号: TP181 文献标识码: A

很多高维数据常常分布于较低维的流形,ISOMAP 和 LLE 是寻找描述这样低维流形参数空间的最有代表性的方法[1,2].ISOMAP 在降维过程中通过计算点对之间的测地距离,并采用 MDS 方法来获取全局最优的几何结构,获得了较好的效果,目前已经发展了很多改进算法.LLE 在降维过程中保持局部的几何结构不变,并能够避免局部极小,从而可获得全局的低维嵌入,目前发展的改进算法包括利用拉普拉斯(Laplacian)和赫森(Hessian)变换改进的算法 LE 和 HLLE[3,4];利用数据分类信息改进的监督 LLE[5]、增量式 LLE[6]、鲁棒性 LLE[7];利用连续 1-维拉普拉斯特征图改进的 LLE[8];利用神经网络改进的 LLE[9]、多方法的集成框架[10]等.性能上 HLLE 是对 LLE 的较大改进,它在某些情况下超越了 ISOMAP 的能力.ISOMAP 的基本假设是全局等距映射和凸的参数空间,这在很多情况下难以满足.而 HLLE 只要求局部等距映射和开的连通参数空间,因而应用范围更宽.但是,HLLE 需要保持邻域的线性化,当数据流形比较弯曲时,这难以满足,特别是它对噪音和孤立点(outlier)非常敏感,是拓扑不稳定的.噪音和孤立点容易在邻域图中产生大量的短路边,进而导致剧烈的嵌入偏差[11,12],因而需要邻域优化.目前有 4 类与邻域相关的工作,第 1 类构造连通的邻域图[13,14].第 2 类采用新的测度来选择邻域点,例如监督 LLE 利用分类信息来构造新的测度[5],对于无分类信息的数据,则可采用自动聚类来确定分类信息[15].同时,图代数、测地距离或局部估计的测地距离也被用来确定更好的邻域[16−18].第 3 类是利用“短路边”的一些启发式判别准则删除邻域图中的“短路边”[19,20].第 4 类是研究邻域大小的选择问题[21],例如,适用于非均匀流形的邻域大小自适应确定算法[22,23].

目前的这些方法在计算数据点之间的距离时,均没有考虑其他数据点的影响,这使得噪音或孤立点与正常点的待遇相同,同时也没有考虑数据的稀疏性,因此我们提出新的处理方法,根据认知的相对性规律提出一种相对变换[21],将原始数据空间转换到相对空间,之后在相对空间中虽然仍采用原来的距离公式,但计算出的值却考虑了所有数据点的影响,因此在相对空间中测量数据的相似性或距离能够更符合我们的直觉,从而提高数据分析的准确性,之后我们还发展了局部相对变换、核相对变换等,并用其改进了 ISOMAP[25]和 HLLE[26,27],效果非常明显.但它们采用殴氏距离构造相对变换,不能变换非线性的弯曲空间,为此,本文利用流形上的测地距离构造相对变换,并用其确定 HLLE 的邻域,进而提出了一种新的 HLLE 增强算法.

1 相对变换与相对流形

ISOMAP 和 LLE 可认为是基于认知的方法,因为研究表明人类的感知是流形[28],而流形概念是流形机器学习方法的核心,正因为如此,这两种方法才具有原始的创新性,引起目前广泛的研究.但是它们仍然面临数据噪音等困难,需要模型化更多的认知规律,因此我们考察认知的相对性规律.经验表明,人类的感知具有相对性,例如,在观察图 1 中的两个圆 x 和 y,通常都会认为圆 x 比圆 y 大,而实际上它们一样大[29].发生的原因是在观察圆 x 时,与其周围相比,显得很大,而在观察圆 y 时,与其周围相比,显得很小,因此这是一种相对性比较的结果.为模型化这种认知规律,我们以原始数据空间 $X=\{x_1,\ldots,x_i,\ldots,x_N\}$中的每个数据点作为基向量来构造新的空间,这样任意点 x_i 到所有点的距离就构成该点在新空间中的坐标,这个过程称为相对变换.

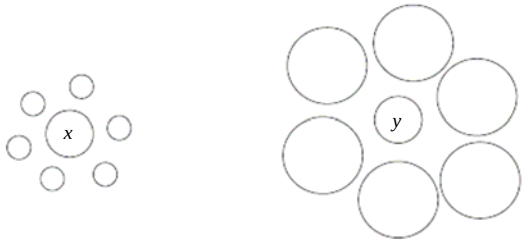


Fig.1 Vision perception is relative

图 1 视觉感知的相对性

$\Gamma_x: X \to Y \subset R^{|X|}$.

$y_i=\Gamma_x(x_i)=(d_{i1},\ldots,d_{ij},\ldots,d_{i|X|}) \in Y, d_{ij}=\|x_i-x_j\|$为距离.

其中$|X|=N$为集合X的元素个数,通过相对变换构造的空间称为相对空间.

定理 1[21]. 对任意 $x_i,x_j \in X$ 有 $d(x_i,x_j) \le d(y_i,y_j)$.

因此相对变换不是等距变换,而是具有放大作用,这有利于我们观察数据之间的拓扑结构的细节.

定理 2. 相对变换可以提高数据之间的可区分性.

证明:

$$\begin{aligned} & d(y_i,y_j)^2 - d(y_i,y_m)^2 \\ &= \sum_{k=1,k\neq j}^{|X|}(d(x_i,x_k)-d(x_j,x_k))^2 + d(x_i,x_j)^2 - \sum_{k=1,k\neq m}^{|X|}(d(x_i,x_k)-d(x_m,x_k))^2 - d(x_i,x_m)^2 \\ &= \sum_{k=1,k\neq j}^{|X|}(d(x_i,x_k)-d(x_j,x_k))^2 - \sum_{k=1,k\neq m}^{|X|}(d(x_i,x_k)-d(x_m,x_k))^2 + d(x_i,x_j)^2 - d(x_i,x_m)^2. \end{aligned}$$

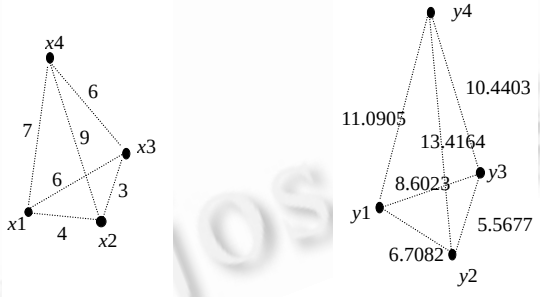
令

$$\nabla = \sum_{k=1,k\neq j}^{|X|}(d(x_i,x_k)-d(x_j,x_k))^2 - \sum_{k=1,k\neq j}^{|X|}(d(x_i,x_k)-d(x_m,x_k))^2,$$

则 $d(y_i,y_j)^2 - d(y_i,y_m)^2 = \nabla + d(x_i,x_j)^2 - d(x_i,x_m)^2$.

此时分两种情况,若 $d(x_i,x_j)=d(x_i,x_m)$,则在原始空间中 x_i 到 x_j 和 x_m 的距离是一样的,难以区分 x_j 和 x_m 谁是 x_i 的最近邻.此时若$\nabla \neq 0$,那么 $d(y_i,y_j) \neq d(y_i,y_m)$,这表示在相对空间中,相对于 x_i 来说,x_j 和 x_m 变得可区分了.反之的情况也可能成立,即在原始空间中,$d(x_i,x_j) \neq d(x_i,x_m)$,但 $d(y_i,y_j)=d(y_i,y_m)$,即将原来可分的数据变得不可分了,此时 $\nabla=-(d(x_i,x_j)^2-d(x_i,x_m)^2)$,这种情况发生的概率要远远小于第 1 种情况发生的概率,因为第 1 种情况的概率 $P(\nabla\neq 0)=P(\nabla>C)+P(\nabla=C)+P(0<\nabla<C)+P(\nabla<0) >> P(\nabla=C)$,这里 $C=-(d(x_i,x_j)^2-d(x_i,x_m)^2)$,且假定 $C>0$.对 $C<0$ 的情况推导类似. □

我们举一个实例来说明.从图 2 中可以看出在原始数据空间中 $d(x3,x1)=d(x3,x4)$,此时,$x3$ 无法决定 $x1$ 和 $x4$ 谁离自己更近,这对基于最近邻选择的机器学习方法产生了不利影响.但是在转换后的相对空间中,$d(y3,y1)<d(y3,y4)$,很容易决定 $y1$ 与 $y3$ 更近,特别是这种情形也更符合人类的直觉,因此相对变换不是线性变换,它能够将原来在原始数据空间中不能区分的数据在相对空间中区分开来,从而提高了数据之间的可区分性.同时,相对变换对抑制噪音或识别孤立点都非常有用,从而提高机器学习的鲁棒性.例如,图 2 中的 $x4$ 可能是孤立点,但是在原始数据空间中,$d(x3,x1)=d(x3,x4)$,这使 $x1$ 和 $x4$ 有相同的机会成为点 $x3$ 的近邻,这与我们的直觉不一致.而在相对空间中,$d(y3,y1)<d(y3,y4)$,这意味着孤立点 $y4$ 更加远离正常数据点,从而更加易于区分.



(a) Original data space (a) 原始空间 (b) Relative data space (b) 构造的相对空间

Fig.2 Relative transformation can weaken the influence of noise on machine learning

图 2 相对变换能抑制噪音影响

相对变换中的距离通常是 Euclidean 距离,但也可以是任意距离,如测地距离.计算数据集 X_i 中任意两点之间的测地距离可以采用 ISOMAP 的方法,主要包括两步:① 根据 X_i 和 k 确定每个点的 k 邻域,然后构造权重图 $G=(V,E)$.V 对应于 X_i 中的数据,E 为连接 V 中两点的边集合,$(x_i,x_j) \in E$,若 x_i 是 x_j 的 k 最近邻,x_i 与 x_j 之间的距离为

欧氏距离 $d_e(x_i,x_j)$.② 通过求 G 上任意两点之间的最短距离来估计 X_i 所形成的局部流形上的所有点对之间的测地距离 $d_g(x_i,x_j)$.首先对所有 $(x_i,x_j)\in E$ 令 $d_g(x_i,x_j)=d_e(x_i,x_j)$,否则令 $d_g(x_i,x_j)=\infty$.然后利用所有 t,迭代计算所有的 $d_g(x_i,x_j)=\min\{d_g(x_i,x_j),d_g(x_i,x_t)+d_g(x_t,x_j)\}$.我们利用计算出的测地距离构造相对变换,进而构造相对空间,并称此相对空间为相对流形,然后用其改进 HLLE.

2 增强的 HLLE 算法

假定有一个参数空间 $\Theta\subset R^d$ 和一个光滑映射 $\varphi:\Theta\to R^n$,其中,嵌入空间 R^n 满足 $n>d$,则称 $M=\varphi(\Theta)$ 为流形,流形学习的目的是根据观察数据确定参数空间 Θ.ISOMAP 采用等距嵌入来实现流形学习,而 HLLE 则采用局部线性方法实现流形学习,其理论依据来源于流形切空间上的 Hessian 变换.框架上 HLLE 与 LE 和 LLE 一致,不同的是,用 Hessian 变换取代了 LE 的 Laplacian 算子,而 LLE 是 LE 理论框架下的一种经验实现,但是 HLLE 要求每个点邻域是线性的,当邻域高度弯曲时,极易面临短路威胁,我们通过构造相对空间和相对流形,并用之确定邻域的方法来解决这个问题.虽然采用 Euclidean 距离作为测度只能发现球形邻域,但在相对空间中或相对流形上发现的球形邻域在原始数据空间中却可能是任意形状的,同时还抑制了噪音和数据稀疏的影响,因此构造的邻域图能够明显地减少短路现象.图 3 的原始数据是从 Swiss roll surface 上采样 400 个点形成的数据,并添加了均值为 0、方差为 0.4 的随机高斯噪音,然后绘制的几个邻域图.不难看出,在原始空间中构造的邻域图随着邻域参数 k 的增大,短路的边数也快速增加.相反,在相对空间中构造的邻域图中,短路的边数明显减少,特别是在邻域参数取 5 和 10 时没有短路边出现,因此在相对空间中构造 HLLE 的邻域图能够获得更好的低维嵌入,记此改进的算法为 R-HLLE.

(a) k=5 (b) k=10 (c) k=15 (e) k=5 (f) k=10 (g) k=15

Fig.3 (a)~(c) shows the neighbourhood graph constructed in the original data space, while (e)~(g) shows the neighbourhood graph constructed in the relative data space

图 3 (a)~(c)是在原始数据空间构造的邻域图,(e)~(g)是在相对数据空间中构造的邻域图

算法 1. R-HLLE 算法($\boldsymbol{X}$,k,d).

/* $\boldsymbol{X}$ 是高维数据,k 是局部线性邻域大小,d 是低维参数空间的维数,$\min(k,n)>d$.输出是低维参数空间 $\boldsymbol{W}$ */

1) 将 X 转换到相对空间,然后在相对空间计算每个点 $x_i\in\boldsymbol{X}$ 的 k-NN 邻域 $N(x_i)$.
2) 将每个点 x_i 的邻域表达为中心化的行向量 $k_i\times n$ 矩阵 M^i.
3) 采用奇异值分解每个邻域矩阵 M^i,将其正交向量 V 的前 d 个分量作为其切空间.
4) 求切空间的 Hessian 矩阵.当 d=2 时,根据切空间中的点形成如下矩阵 X^i=[1 $V_{\cdot,1}$ $V_{\cdot,2}$ $(V_{\cdot,1})^2$ $(V_{\cdot,2})^2$ $(V_{\cdot,1}\times V_{\cdot,2})$],其中 $V_{\cdot,1}$ 表示切空间中所有点的第 1 个维的值,对 d>2 采用相同的方法创建 $1+d+d(d+1)/2$ 列的矩阵,然后用 Gram-Schmidt 正交化 X^i 产生新的正交矩阵,并将其转置后取最后的 $d(d+1)/2$ 列构成 Hessian 矩阵 H^i.
5) 构造二次型 $H_{ij}=\sum_l\sum_r((H^l)_{r,i}(H^l)_{r,j})$,对 H 进行特征分析,获取其$(d+1)$个最小特征值对应的$(d+1)$维子空间,第 1 个特征值 0 对应于常函数,接下来的 d 个特征向量就构成 d 维空间,对其选择一个正交基,变换就可以获得要恢复的参数空间 W.

与 HLLE 相比,算法只增加了第 1 步和修改了第 2 步,其增加的时间复杂度是 $O(|X|^2)$,其余步骤与 HLLE 相同,数据的嵌入仍然在原始数据空间中完成.但是 R-HLLE 构造相对空间时采用的是 Euclidean 距离.考虑到大多数高维数据是流形,采用测地距离更合适,而计算全局的测地距离需要很大的时间代价,此时可构造局部相对流

形并用其确定任意点的邻域.具体方法是为任意点 x_i 根据最近邻方法定义一个局部区域,然后构造局部相对流形,进而在此局部相对流形上确定邻域,提出基于局部相对流形的局部线性嵌入算法 RM-HLLE.

算法 2. RM-HLEE 算法(X,L,k_g,k,d).

/* $\boldsymbol{X}$ 是高维数据,L 定义局部区域大小($L>k$ 且 $L>k_g$),k_g 是估计测地距离的邻域参数,k 是局部线性邻域大小,d 是低维参数空间的维数,$\min(k,n)>d$.输出是低维参数空间 $\boldsymbol{W}$ */

1) 计算每个点 $x_i\in\boldsymbol{X}$ 的 k-NN 邻域 $N(x_i)$.
   a) 根据 L-NN 方法和欧氏距离确定每个点 x_i 的局部区域 $X_i=\{x_j|x_j$ 是 x_i 的 L 个最近邻成员$\}$.
   b) 以 k_g 为邻域参数,计算每个点 x_i 到其局部区域 X_i 中任意点之间的局部测地距离.
   c) 根据每个局部区域 $\boldsymbol{X_i}$ 和测地距离构造其相对流形 $X_{im}=\{y_j|y_j=\Gamma_{xi}(x_j\in X_i)$.
   d) 对任意点 $x_i\in X$,在其局部相对流形 X_{im} 中用欧氏距离计算其 k-NN 邻域 $N(x_i)$.
2) 其余步骤与 R-HLLE 相同.

RM-HLLE 用 Floyd 算法估计局部区域中的测地距离所需要的时间复杂度为 $O(|X|L^2)$,因为 L 通常很小,可认为是常数,因此增加的时间复杂度是线性的.特别是其确定的邻域具有良好的结构,这将加速后继的低维嵌入过程,因此整体时间相对 HLLE 并不增加,对大规模数据还有减少趋势,这将从后面的实验得到验证.

3 实验分析

实验比较 LLE,ISOMAP,HLLE,R-HLLE,RM-HLLE 的性能和时间复杂度,评价准则采用定性的可视化嵌入效果和定量标准 Spearman's rho 和 procrustes 值[6],Spearman's rho 越大越好,procrustes 则越小越好.

3.1 实验参数设置

实验中,LLE 和 HLLE 的邻域参数设置为 12,ISOMAP 中的邻域参数选择 7,它们是 HLLE 和很多流形学习算法在实验中采用的参数值.为了在相同的参数下比较,R-HLLE 和 RM-HLLE 的线性邻域参数也选择 12,RM-HLLE 中估计测地距离的邻域参数也选择 7.RM-HLLE 新增加的局部区域大小参数 L 则通过实验选择.方法是以 Spearman's rho 和 procrustes 作为评价标准,在数据上抽样实验.例如,从 Swiss roll surface 上随机采样规模分别为 800 个点和 400 点的 10 个样本,L 为 30,40,…,120.RM-HLLE 以 L 的每个值运行每种规模的 10 个样本,计算 Spearman's rho 和 procrustes 的平均值,结果如图 4 所示,不难看出,对两种规模,RM-HLLE 都在 L=40 时取得较好的 Spearman's rho 和 procrustes 平均值,因此下面的实验均取此值.除非另有说明,否则以上参数对以下所有的实验适用.

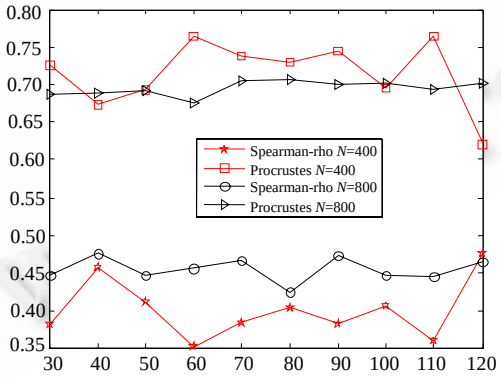


Fig.4 Average value of Spearman's rho and procrustes vary with the local region sizes

图 4 Spearman's rho 和 procrustes 的平均值随局部区域大小的变化情况

3.2 Swiss roll surface数据

Swiss roll surface 是标准数据集,我们采用 HLLE 的采样方法[4],从其表面随机采样 N 个点的长方形,但同时

从其中心移去一个小的长方形以使得数据不再是凸的,这是一个很有挑战性的数据集,很多算法都得不到理想的结果,在此数据集上我们做几个实验.

实验 1:噪音数据.

真实数据一般都有噪音,具有拓扑稳定性的算法受噪音的影响较少.我们从 Swiss roll surface 上随机采样 800 个点,然后叠加均值为 0 和方差为 0.4 的高斯噪音.按此方法采样多次并实验.分析发现,HLLE 在少部分情况下能够将数据嵌入在二维空间.ISOMAP 总是将去除的区域强烈膨胀,并扭曲其余的数据点.LLE 在绝大多数情况下都得不到正确结果.而 *R*-HLLE 和 RM-HLLE 也受噪音的影响,在部分情况下也不能正确嵌入,原因是噪音影响了测地距离的估计,导致最终的嵌入偏差.但相对稳定,特别是 RM-HLLE 表现最好,在较多情况下都能够较完美地将数据嵌入在二维空间,其中心移去的一个小长方形也能在嵌入的二维空间中正确体现,图 5 是其中的一个结果,可以看出,R-HLLE 和 RM-HLLE 表现最好,这能够从其获得的 Spearman's rho 和 procrustes 值得到支持.

实验 2:稀疏数据.

大量真实数据是稀疏的,很多算法难以处理.我们从 Swiss roll surface 上随机采样数据规模为 400 点的多个稀疏数据集,实验发现在很多情况下,HLLE 和 LLE 获得的结果是混乱的.*R*-HLLE 和 RM-HLLE 在部分情况下也不能正确嵌入,但相对而言,RM-HLLE 表现最好,图 6 是其中的一个结果,可以看出,RM-HLLE 表现最好,这证实了在相对流形上,原始数据空间中的稀疏数据变得相对密集.

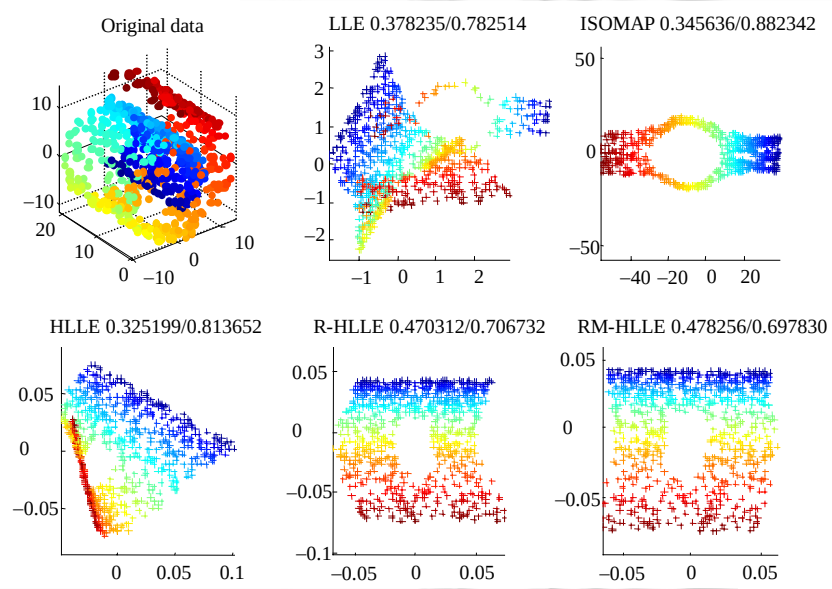


Fig.5 Embedding results of compared approaches on noisy Swiss roll surface data set

图 5 几种方法在含噪音密集的 Swiss roll surface 数据上的降维结果

实验 3:大规模数据.

RM-HLLE 只对局部数据作相对变换,与整体的规模没有关系,因而与 HLLE 一样能够处理大规模数据.我们从 Swiss roll surface 上采样 2 500 个样本点的较大规模数据,并叠加均值为 0 和方差为 0.1 的高斯噪音.结果如图 7 所示,可以看出,HLLE,R-HLLE 和 RM-HLLE 在此数据集上都取得了较好的嵌入效果,但 RM-HLLE 表现最好,这也可以从其 Spearman's rho 和 procrustes 值得到一致支持.

3.3 S-Curve数据

S-Curve 是另一个常用的标准数据集[6].我们随机采样 800 个点,然后叠加均值为 0 和方差为 0.1 的高斯噪

音,按此方法多次采样实验发现,HLLE 在少部分情况下能够将数据嵌入在二维空间.LLE 在性能上是最差的,绝大多数情况下都得不到正确结果.而 ISOMAP,R-HLLE 和 RM-HLLE 也受噪音的影响,在部分情况下也不能正确嵌入,但相对稳定,特别是 RM-HLLE 表现最好,图 8 是其中的一个结果,可以看出,RM-HLLE 表现最好,这也可以从其 Spearman's rho 和 procrustes 值得到一致的支持.

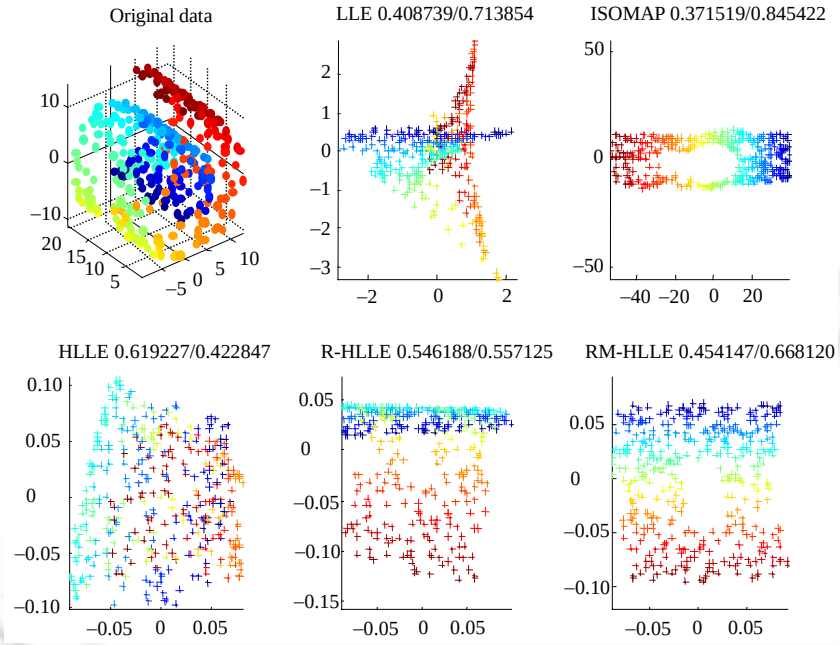


Fig.6 Embedding results of compared approaches on sparse Swiss roll surface data set

图 6 几种方法在稀疏的 Swiss roll surface 数据上的降维结果

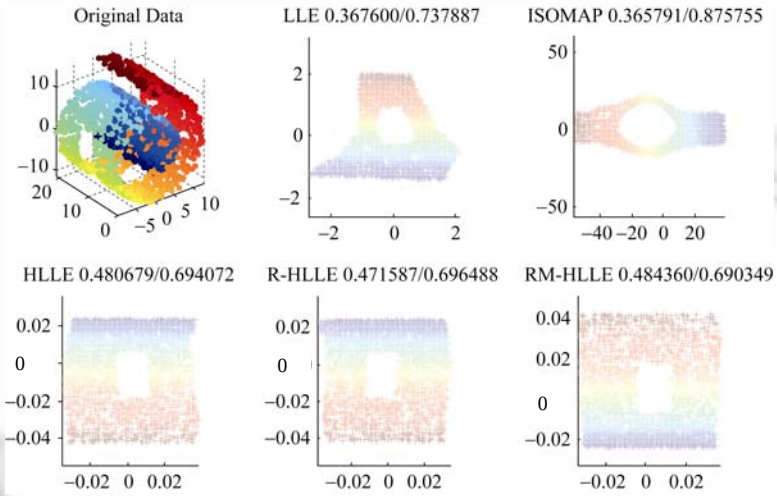


Fig.7 Embedding results of compared approaches on large Swiss roll surface data set

图 7 几种方法在较大规模的 Swiss roll surface 数据上的降维结果

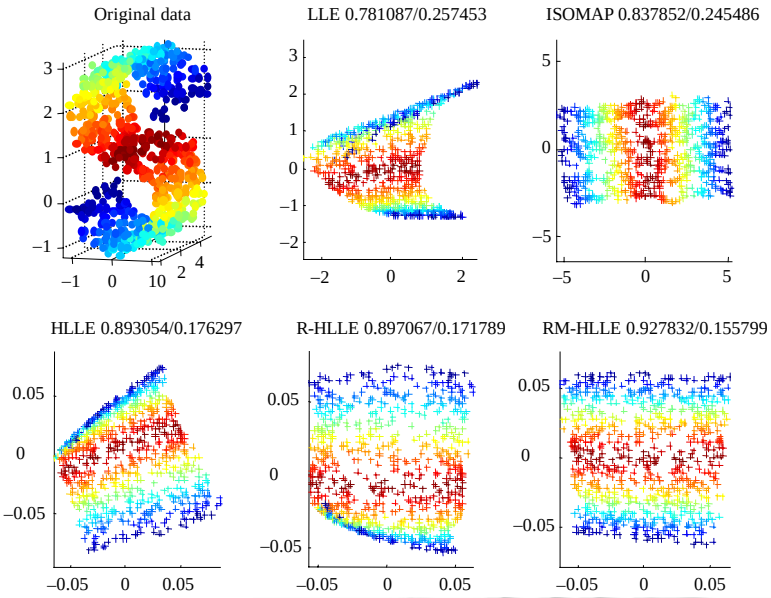


Fig.8 Embedding results of compared approaches on S-Curve data set

图 8 几种方法在含噪音且密集的 S-Curve 数据上的降维结果

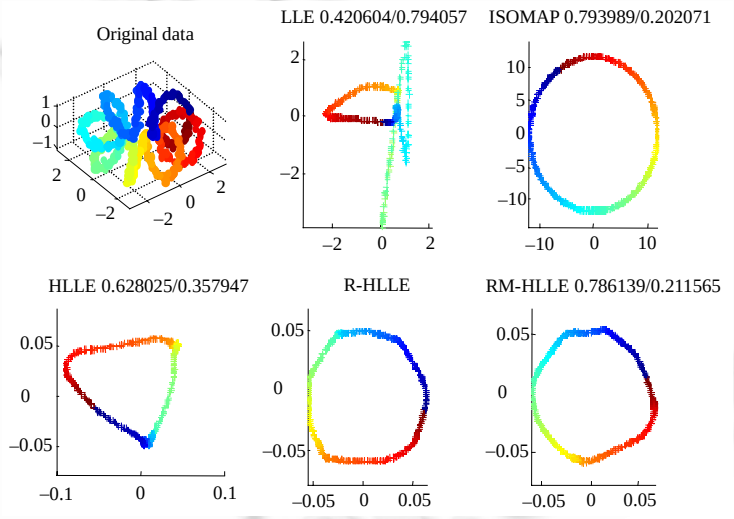


Fig.9 Embedding results of compared approaches on Toroidal Helix data set

图 9 几种方法在含噪音的 Toroidal Helix 数据上的降维结果

3.4 Toroidal Helix数据

Toroidal Helix 是常用的另一标准数据(http://www.math. umn.edu/~wittman/mani/index.html).好的流形学习方法应能够将此盘绕的曲线还原成一个圆.我们采样 600 个点,并叠加均值为 0 和方差为 0.05 的高斯噪音.根据 Spearman’s rho 和 procrustes 值,以及图 9 的嵌入效果可知,ISOMAP 表现最好,它属于全局性方法.RM-HLLE 则是局部线性嵌入方法中表现最好的,说明了相对变换和相对流形的有效性.

3.5 Sculpture人脸数据

Sculpture 人脸数据是 ISOMAP 使用的数据,包含 698 幅 4 096 维的灰度人脸图像.由于 Spearman’s rho 和

procrustes 值对非常高维的数据集计算复杂,内存要求太高,不适用.因此我们采用残差作为定量评估的标准,残差越小越好.根据图 10 所示的嵌入效果和残差可以发现,ISOMAP(k_g=6)表现最好,RM-HLLE(k=12,L=26,k_g=L)比 HLLE(k=12)的嵌入效果要好,说明了相对变换和相对流形的有效性.

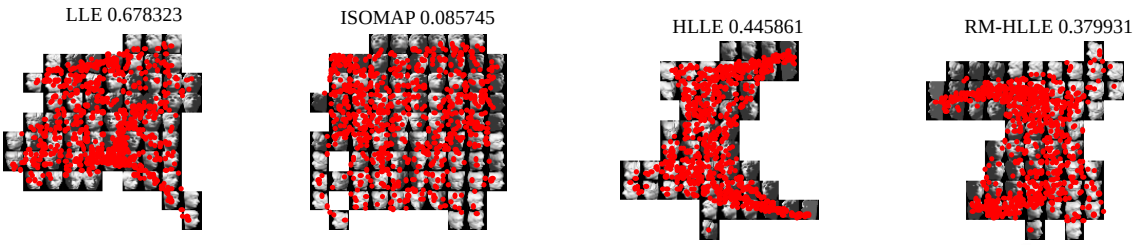


Fig.10 Embedding results of compared approaches on Sculpture data set

图 10 几种方法在 Sculpture 人脸数据上的降维结果

3.6 时间复杂度比较

我们从 Swiss roll surface 上采样 5 类规模的数据样本,每类规模的样本随机采样 5 次,记录每种方法分别运行这 5 个样本的平均时间作为该类规模的时间,则 5 种方法在 5 类规模数据上的平均时间见表 1 所示,不难发现:① RM-HLLE 不仅提高了性能,而且速度比 HLLE 更快,规模越大,差距越大,原因是优化的邻域加速了后继的嵌入过程,这特别适合于大规模数据.② R-HLLE 的运行时间略高于 HLLE,因为数据点变成高维,高维计算需要更多的时间.③ LLE 运算速度最快,是我们未来拟改造的主要方法,而 ISOMAP 对数据规模十分敏感,增长率最快,对大规模数据需要采用其有效的改进算法.

Table 1 Comparison among average running time of five approaches on five data sizes (s)

表 1 5 种方法在 5 种样本规模上的平均运行时间比较 (秒)

	Data size				
	500	1 000	1 500	2 000	2 500
LLE	0.501 4	1.643 8	3.626 9	6.473 3	10.375 0
RM-HLLE	**2.464 2**	**18.179 6**	**59.620 5**	**138.740 7**	**276.609 0**
HLLE	2.521 9	18.365 5	60.056 0	139.692 1	279.609 0
R-HLLE	2.728 4	19.659 3	63.526 6	147.585 9	305.406 0
ISOMAP	7.228 2	57.534 4	185.231 4	444.890 6	865.195 0

4 结论及未来的工作

根据认知的相对性规律提出了相对变换,并构造了相对空间和相对流形,它们可以提高数据之间的可区分性,并能够抑制噪音和数据稀疏的影响,将其用于增强 HLLE 算法,明显提高了性能和速度,适合于大规模数据处理,实验验证了相对流形的有效性.未来的工作是,首先将相对变换用于增强 LLE,以处理大规模很高维的数据,例如图象和文本数据.第二,继续研究相对变换,特别是理解与核函数的内在联系.第三,目前对高维数据与内在低维流形之间的联系已有很好的研究成果可以借鉴[30],以此为基础,研究降维过程中的拓扑不变性规律,进而提出新的流形学习方法.

References:

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

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

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

[4] Donoho DL, Grimes C. Hessian eigenmaps: Locally linear embedding,techniques for high-dimensional data. Proc. of the National Academy of Sciences, 2003,100(10):591−596.

[5] de Ridder D, Kouropteva O, Okun O, Pietikainen M, Duin RPW. Supervised locally linear embedding. LNAI 2714, 2003. 333−341.

[6] Kouropteva O, Okun O, Pietikainen M. Incremental locally linear embedding. Pattern Recognition, 2005,38:1764−1767.

[7] Chang H, Yeung DY. Robust locally linear embedding, Pattern Recognition, 2006,39:1053−1065.

[8] Gerber S, Tasdizen T, Whitaker R. Robust non-linear dimensionality reduction using successive 1-dimensional laplacian eigenmaps. In: Oregon C, ed. Proc. of the 24th Int'l Conf. on Machine Learning. ACM, 2007. 281−288.

[9] Yin JS, Hu DW, Zhou ZT. Manifold learning using growing locally linear embedding. In: Duch W, ed. Proc. of the 2007 IEEE Symp. on Computational Intelligence and Data Mining. IEEE Press, 2007. 73−80.

[10] Yan SC, Xu D, Zhang BY, Zhang HJ, Yang Q, Lin S. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29:40−50.

[11] Balasubramanian M, Schwartz EL. The ISOMAP algorithm and topological stability. Science, 2002,295:7.

[12] Silva VD, Tenenbaum JB. Global versus local methods in nonlinear dimensionality reduction. Neural Information Processing Systems, 2003,15:705−712.

[13] Yang L. Building *k*-Connected neighborhood graphs for isometric data embedding. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006,28(5):827−831.

[14] Yang L. Building connected neighborhood graphs for locally linear embedding. In: Tang YY, ed. Proc. of the 18th Int'l Conf. on Pattern Recognition. IEEE Computer Society, 2006. 1680−1683.

[15] Wen GH, Jiang LJ, Wen J, Shadbolt NR. Clustering-Based nonlinear dimensionality reduction on manifold. LNAI 4099, 2006. 444−453.

[16] Wen GH, Jiang LJ, Shadbolt NR. Using graph algebra to optimize neighborhood for isometric mapping. In: Veloso MM, ed. Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2007). 2007. 2398−2403.

[17] Wen GH, Jiang LJ, Wen J. Using locally estimated geodesic distance to optimize neighborhood graph for isometric data embedding. Pattern Recognition, 2008,41(7):2226−2236.

[18] Varini C, Degenhard A, Nattkemper TW. ISOLLE: LLE with geodesic distance. Neurocomputing, 2006,69:1768−1771.

[19] Shao C, Huang HK, Zhao LW. A more topologically stable ISOMAP algorithm. Journal of Software, 2007,18(4):869−877 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18.869.htm

[20] Choi H, Choi S, Kernel R. Isomap. Pattern Recognition, 2007,40:853−862.

[21] Samko O, Marshall AD, Rosin PL. Selection of the optimal parameter value for the Isomap algorithm. Pattern Recognition Letters, 2006,27(9):968−979.

[22] Wen GH, Jiang LJ, Wen J, Shadbolt NR. Performing locally linear embedding with adaptive neighborhood size on manifold. LNAI 4099, 2006. 985−989.

[23] Wen GH, Jiang LJ, Wen J. Dynamically determining neighborhood parameter for locally linear embedding. Journal of Software, 2008,19(7):1666−1673 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/1666.htm

[24] Wen GH. Relative transformation for machine learning. Journal of Computer Research and Development, 2008,45(4):612−618 (in Chinese with English abstract).

[25] Wen GH. Relative transformation-based neighborhood optimization for isometric embedding. Neurocomputing, 2009,72(4−6): 1205−1213.

[26] Wen GH, Jiang LJ, Wen J. Improved locally linear embedding by cognitive geometry. LNAI, 2007. 317−325.

[27] Wen GH, Jiang LJ, Wen J. Kernel relative transformation with applications to enhancing locally linear embedding. In: Wang J, ed. Proc. of the Int'l Joint Conf. on Neural Networks (IJCNN 2008). IEEE, 2008. 3401−3406.

[28] Sung HS, Lee DD. The manifold ways of perception. Science, 2000,290:2268.

[29] Li DY, Liu CY, Du Y, Han X. Artificial intelligence with uncertainty. Journal of Software, 2004,15(11):1583−1594 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/1583.htm

[30] He L, Zhang JP, Zhou ZH. Investigating manifold learning algorithms based on magnification factors and principal spread directions. Chinese Journal of Computers, 2005,28(12):2000−2009 (in Chinese with English abstract).

附中文参考文献:

[19] 邵超,黄厚宽,赵连伟.一种更具拓扑稳定性的 ISOMAP 算法.软件学报,2007,18(4):869−877. http://www.jos.org.cn/1000-9825/18.869.htm

[23] 文贵华,江丽君,文军.邻域参数动态变化的局部线性嵌入.软件学报,2008,19(7):1666−1673. http://www.jos.org.cn/1000-9825/19/1666.htm

[24] 文贵华.面向机器学习的相对变换.计算机研究与发展,2008,45(4):612−618.

[29] 李德毅,刘常昱,杜鹢,韩旭.不确定性人工智能.软件学报,2004,15(11):1583−1594. http://www.jos.org.cn/1000-9825/15/1583.htm

[30] 何力,张军平,周志华.基于放大因子和延伸方向研究流形学习算法.计算机学报,2005,28(12):2000−2008.

文贵华(1968－),男,湖北利川人,博士,副研究员,主要研究领域为机器认知与创新,数据挖掘与知识发现,机器学习.

江丽君(1971－),女,讲师,主要研究领域为创新教育,智能 CAD.

陆庭辉(1984－),男,硕士生,主要研究领域为数据挖掘与知识发现.

文军(1964－),男,副教授,主要研究领域为创新计算,机器学习,智能软件.