文章编号:1001 - 9081(2006)04 - 0895 - 03

基于局域主方向重构的适应性非线性维数约减

侯越先[1],吴静怡[2],何丕廉[1]
(1. 天津大学 电子信息工程学院,天津 300072; 2. 天津大学 管理学院,天津 300072)
(yxhou@ tju. edu. cn)

摘 要:现有的主要非线性维数约减算法,如 SIE 和 Isomap 等,其邻域参数的设定是全局性的。仿真表明,对于局域流形结构差异较大的数据集,全局一致的邻域参数可能无法获得合理的嵌入结果。为此给出基于局域主方向重构的适应性邻域选择算法。算法首先为每个参考点选择一个邻域集,使各邻域集近似处于局域主线性子空间,并计算各邻域集的基向量集;再由基向量集对各邻域点的线性拟合误差判定该邻域点与主线性子空间的偏离程度,删除偏离较大的点。仿真表明,基于局域主方向重构的适应性邻域选择可有效处理局域流形结构差异较大的数据集;且相对于已有的适应性邻域选择算法,可以更好屏蔽靠近参考点的孤立噪声点及较大的空间曲率导致的虚假连通性。

关键词:非线性维数约减;适应性邻域选择;局域主方向;流形学习
中图分类号: TP181 **文献标识码**:A

Locally adaptive nonlinear dimensionality reduction

HOU Yue-xian[1], WU Jing-yi[2], HE Pi-lian[1]
(1. *School of Electronic Information Engineering, Tianjin University, Tianjin* 300072, *China*;
2. *School of Management, Tianjin University, Tianjin* 300072, *China*)

Abstract: Popular nonlinear dimensionality reduction algorithms, such as SIE and Isomap suffer a difficulty in common: global neighborhood parameters often fail in tackling data sets with high variation in local manifold. To improve the availability of nonlinear dimensionality reduction algorithms in the field of machine learning, an adaptive neighbors selection scheme based on locally principal direction reconstruction was proposed. The method involves two main computation steps. First, it selects an appropriate neighborhood set for each data points such that all neighbors in a neighborhood set form a d-dimensionality linear subspace approximatively and computes locally principal directions for each neighborhood set respectively. Secondly, it fits each neighbor by means of locally principal directions of corresponding neighborhood set and deletes the neighbors whose fitting error exceed a predefined threshold. The simulation show that the method can deal with data set with high variation in local manifold effectively. Moreover, comparing with other adaptive neighbors selection strategy, this method can circumvent false connectivity introduced by noise or high local curvature.

Key words: nonlinear dimensionality reduction; adaptive neighbors selection; locally principal direction; manifold learning

0 引言

自动维数约减和流形学习是机器学习研究的重要主题,其实质是利用多维数据的冗余性,抽取最小数目的独立变量描述数据的背景动力特征。更一般地,维数约减模型较中肯地把握了生物感知和思维活动中抽象过程的形式特征:基本的视觉感知过程被认为是冗余约减的过程,视觉神经元的输出需要尽可能地彼此独立[1];约减的表示形式被认为是心智活动中思想和意义形成的首要条件[2,3];日常智慧或科学活动中的抽象理解和理论的形成过程,是利用经验材料之间的统计相关性,寻求复杂关系的压缩表示和简洁解释的过程。

已有若干算法实现维数约减目的。主元分析(Principal Component Analysis, PCA)和 CMDS(Classic Multidimensional Scaling)[4]等线性算法实现较为简单,但无法揭示复杂的非线性流形结构[5]。因此非线性维数约减算法(Nonlinear Dimensionality Reduction, NDR)受到广泛的关注。现有主要的 NDR 算法,如 LLE[6]、Isomap 算法族[5,7]、Laplacian 特征映象[10]和 SIE[8]等,其邻域参数的设定是全局性的。仿真表明,对于局域流形结构差异较大的数据集,由全局一致的邻域参数可能无法获得合理的嵌入结果[9]。

适应性非线性维数约减的实现大致遵循两种不同的思路,其一是全局方法,即依照某些一般性判定原则,定义相应的统计量,对 NDR 算法在不同参数下的嵌入结果做评价和筛选。这类方法是后验、静态的,不改变基本 NDR 算法。另一思路是局域性的,在基本 NDR 算法中直接引入适应性的参数选择过程。例如,基于局域切空间的邻域选择算法(Locally Tagent Space Alignment, LTSA)[9]通过判定流形上某参考点附近的一阶泰勒展开对于其邻域集的近似程度,选择合适的邻域点数。LTSA 的一个问题是当邻域集中包括噪声点或邻域集附近的背景流形的空间曲率较大时,可能误选邻域点,破坏嵌入流形的全局拓扑。

本文给出一种局域性的适应性邻域选择算法:基于局域主方向重构的适应性邻域选择(Locally Principal Direction Reconstruction, LPDR)。LPDR 分为两步:首先利用 PCA 过

收稿日期:2005 - 10 - 08
作者简介:侯越先(1972 -),男,天津人,副教授,博士,主要研究方向:人工智能、算法理论; 吴静怡(1981 -),女,北京人,硕士研究生,主要研究方向:机器学习; 何丕廉(1941 -),男,天津人,教授,主要研究方向:人工智能.

程,在参考点附近选择可近似实现线性嵌入的邻域集,并计算邻域集的 d 维主线性子空间的基向量集,这里 d 是嵌入维;再由基向量集对各邻域点的线性拟合误差判定该邻域点与主线性子空间的偏离程度,从邻域集中删除偏离较大的点。此方法可有效屏蔽靠近参考点的孤立噪声点及较大的空间曲率导致的虚假的连通性,且几何意义清晰,实现简单。

1 原理

设原数据集 $X_D \equiv \{x_1, x_2, \cdots, x_N\}$ 是 D 维空间中的点集,维数约减问题需在保持 X_D 的内在拓朴、几何特征的条件下,求解其在 $d(d < D)$ 维空间中的嵌入集 X_d,这里 X_D 和 X_d 所在的空间分别称为原空间和嵌入空间,d 称为嵌入维。

Isomap、SIE、LLE 等非线性维数约减算法的基本思想是以局域线性指标近似全局非线性指标:Isomap 和 SIE 假设 $x_i(i = 1,2,\cdots,N)$ 的 k 邻域集上的欧几里德距离近似对应于局域测地线距离,再以此为基础计算全局测地线距离;LLE 则假设 k 邻域集对 x_i 的线性拟合系数记录了此 k 邻域集的局域几何信息,以此强迫出保留全局几何特征的嵌入。上述假设成立的共同前提是 x_i 的 k 邻域集近似形成 D 维原空间中的 d 维线性子空间,k 值的选择须确保此条件的成立。

如果 x_i 的 k 邻域集 $\overline{N_i^k} \equiv \{x_i, x_{i1}, x_{i2}, \cdots, x_{ik}\}$ 近似形成 d 维线性子空间,则通过对其进行主元素分析(PCA),可获得方差残留较小的 d 维线性嵌入。反之,通过检验 k 邻域集的 d 维线性嵌入的方差残留,可初步判定邻域点数 k 是否合适(LTSA 利用动力系统中的局域泰勒线性化方法[9] 实现类似的思路)。此方法可较好地处理由于局域流形结构差异或不均匀的样本点空间密度所导致的邻域可变性。

但是仅以方差残留判定 k 值有较大的局限性。某些情况下 k 邻域集的选择并不合适,但其 PCA 嵌入的方差残留可能很小。例如,x_i 附近背景流形的空间曲率较大(图 1(a)),使得个别欧几里德距离较小的邻域点与 x_i 的实际流形距离较大,这样的邻域选择将导致嵌入流形的虚假贯通;又如 x_i 附近有小扰动的噪声点(图 1(b)),使得 k 邻域集中除了包括处于 d 维线性子空间的清洁数据点,还包括处于 d 维线性子空间之外的噪声点(对于使用对称测地线距离的 Isomap 算法,噪声点经常是难以有效屏蔽的。SIE 算法的自组织嵌入机制允许非对称的测地线距离[8],可以在一定程度上屏蔽噪声点对于嵌入流形的整体影响。因此辨识出噪声点对于 SIE 算法具有更多的实际意义)。图 1 中由点线形成的曲线表示流形的某个局部,O 表示参考点,以 O 为圆心的圆周表示其邻域,经过 O 的实线表示流形在该点的切空间。在这两个例子中,虽然邻域集不能被精确地 1 维嵌入,但由于处于切空间之外的邻域点较少,其 1 维 *PCA* 嵌入的方差残留仍较小。

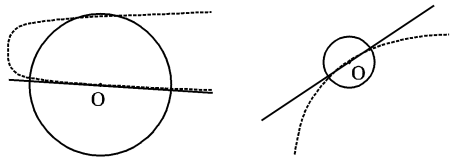


图 1 两个具有小的 PCA 嵌入的方差残留邻域集

LPDR 与 LTSA 的一个重要区别在于:通过考察各个邻域点与 $\overline{N_i^k}$ 中具有较大方差的方向所张成的线性子空间(非形式地,称其为 $\overline{N_i^k}$ 的主线性子空间)的接近程度,对邻域点做进一步筛选,而不是如 LTSA 那样简单地从 x_i 的 k 邻域集中删除若干与 x_i 的欧几里德距离最远的点以保证获得较好的局域泰勒近似[9]。

记 $v_{i1}, v_{i2}, \cdots, v_{id}$ 为 $\overline{N_i^k}$ 的协方差矩阵的最大 d 个特征值对应的特征向量,$v_{i1}, v_{i2}, \cdots, v_{id}$ 张成了 $\overline{N_i^k}$ 的主线性子空间。几何上,处于 $\overline{N_i^k}$ 的主线性子空间中的点可由 $v_{i1}, v_{i2}, \cdots, v_{id}$ 的线性组合表示。构造向量 $y_{i1}, y_{i2}, \cdots, y_{ik}$,其中 $y_{ij} \equiv x_{ij} - x_i, j = 1,2,\cdots,k$,则由 $v_{i1}, v_{i2}, \cdots, v_{id}$ 对 y_{ij} 的线性拟合误差,可判定 y_{ij} 是否近似处于 $\overline{N_i^k}$ 的主线性子空间。形式地,有:

$$y_{ij} = V_i w_{ij} + \varepsilon_{ij} \qquad j = 1,2,\cdots,k \tag{1}$$

其中 $V_i = [v_{i1}, v_{i2}, \cdots, v_{id}]$,$w_{ij} = [w_{ij}^{(1)}, w_{ij}^{(2)}, \cdots, w_{ij}^{(d)}]^T$ 是线性拟合系数向量,ε_{ij} 是拟合误差向量。对 $y_{ij}, j = 1,2,\cdots,k$,需要求出优化的线性拟合系数 w_{ij} 所对应的拟合误差 ε_{ij},这是典型的二次优化问题。由多元回归理论[4],优化的线性拟合系数向量:

$$w_{ij} = (V_i^T V_i)^{-1} V_i^T y_{ij} \tag{2}$$

2 算法

LPDR 过程为:1)为每个点初步选择一个的邻域集,使各个邻域集近似处于 d 线性子空间;2)计算各邻域集的主线性子空间,从邻域集中删除偏离主线性子空间的邻域点。

算法 1:基于局域主方向重构的适应性邻域选择(LPDR)算法

输入:D 维原空间中的点集 $\{x_1, x_2, \cdots, x_N\}$

输出:$\{x_1, x_2, \cdots, x_N\}$ 中各点的适应性邻域

参数:嵌入维 d,最大邻域点数 $k_{\max}$,最小邻域点数 $k_{\min}$,邻域点数的变化量 Δk,方差残留阈值 th_{cov},拟合误差阈值 Th_{err}

算法过程:对 $x_i, i = 1,2,\cdots,N$,分别执行以下过程

1) $k := k_{\max}$;

2) 生成 x_i 的 k 邻域集 $\overline{N_i^k} \equiv \{x_i, x_{i1}, x_{i2}, \cdots, x_{ik}\}$,计算其 PCA 的方差残留 R_{cov};

3) if $R\,cov < Th_{cov}$ or $k = k_{\min}$

then 保存 $\overline{N_i^k}$ 的协方差矩阵的最大 d 个特征值对应的特征向量 $v_{i1}, v_{i2}, \cdots, v_{id}$

else $k := \max\{k - \Delta k\}, k_{\min}$,返回 2)

4) 由 $\overline{N_i^k}$ 构造向量 $y_{i1}, y_{i2}, \cdots, y_{ik}$,其中 $y_{ij} = x_{ij} - x_i, j = 1,2,\cdots,k$;

5) 根据式(2)求解分别 $y_{i1}, y_{i2}, \cdots, y_{ik}$ 优化线性拟合系数向量 $w_{i1}, w_{i2}, \cdots, w_{ik}$;

6) 根据式(1)分别计算 $y_{i1}, y_{i2}, \cdots, y_{ik}$ 的线性拟合误差 $\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{ik}$;

7) 计算归一化线性拟合误差 $\varepsilon_{i1}^n, \varepsilon_{i2}^n, \cdots, \varepsilon_{ik}^n$,其中 $\varepsilon_{ij}^n \equiv \varepsilon_{ij} / \| y_{ij} \|_2, i = 1,2,\cdots,k$;

8) if $\varepsilon_{i1}^n, \varepsilon_{i2}^n, \cdots, \varepsilon_{ik}^n$ 中小于 Th_{err} 的归一化线性拟合误差的数目大于 $k_{\min}$

then 选取小于 Th_{err} 的 ε_{ij}^n 所对应的 x_{ij} 构成 x_i 的适应性邻域 $\overline{N_i^a}$

else 选取 $k_{\min}$ 个最小的 ε_{ij}^n 所对应的 x_{ij} 作为 x_i 的适应性邻域 $\overline{N_i^a}$

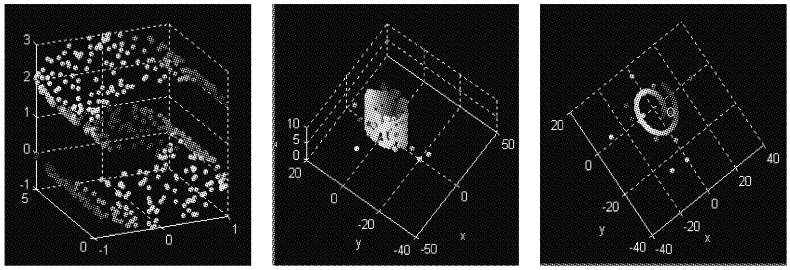
算法 1 可直接作为 Isomap、SIE、LLE 和 Laplacian 特征映象等基础 NDR 算法的附加子过程,实现适应性的邻域选择。

3 实验结果

使用清洁的 S 形数据集[6] 和含贯通型噪声的 Swiss roll 数据集[5] 进行实验(如图 2)。两个数据集各包含 500 个样本点,噪声点由随机扰动无噪数据集中的少量数据点获得。

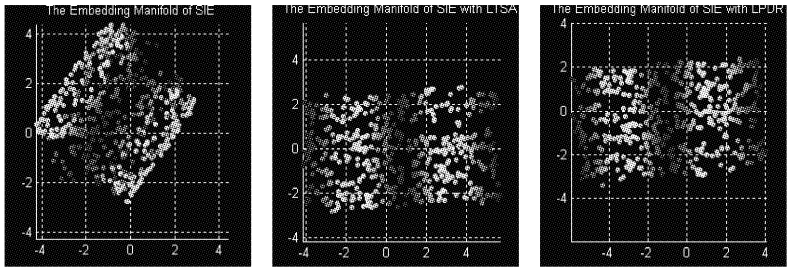
分别以基本 SIE 算法[8]、基于 LTSA 邻域选择的 SIE 算法和基于 LPDR 邻域选择的 SIE 对两个数据集进行嵌入。为两

种邻域选择算法选择相对优化的参数设置,LTSA 的参数[9]设定如下 $k_{\max}=60, k_{\min}=10, \Delta k=1, \eta=0.1$;LPDR 参数设定如下:$k_{\max}=60, k_{\min}=10, \Delta k=1, th_{cov}=0.005, Th_{err}=0.1$;基本 SIE 算法的邻域参数设定为$\lceil (k_{\max}+k_{\min})/2 \rceil$。图 3 和图 4 分别给出了各算法对清洁 S 形数据集和含噪 Swiss roll 数据集的嵌入结果。由图 3 可见,对于清洁 S 型数据,LTSA 和 LPDR 均可获得合理的嵌入结果,优于基本 SIE。



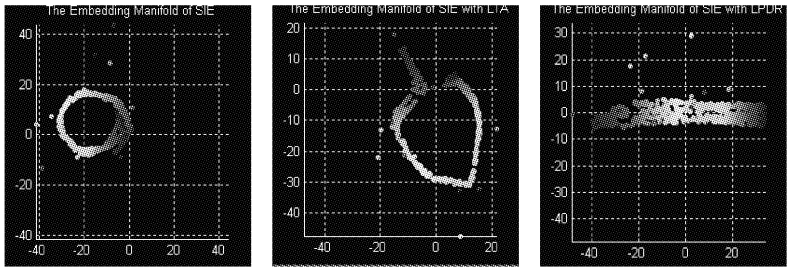
(a) 清洁S形数据集的3维原流形　(b) 含贯通型噪声的Swiss roll数据集的3维原流形　(c) 含贯通型噪声的Swiss roll数据集的2维投影

图 2　实验所用仿真数据集



(a) 基本SIE嵌入结果　(b) 基于LTSA邻域选择的SIE嵌入结果　(c) 基于LPDR邻域选择的SIE的入结果

图 3　清洁的 S 形数据集的 2 维嵌入



(a) 基本SIE嵌入结果　(b) 基于LTSA邻域选择的SIE嵌入结果　(c) 基于LPDR邻域选择的SIE的入结果

图 4　含贯通型噪声的 Swiss roll 数据集的 2 维嵌入

表 1　LSTA 在不同参数设置下对 O 点邻域集的计算结果

$\eta=$	0.02	0.04	0.06	0.08	0.1
LSTA	不正确	不正确	不正确	不正确	不正确

表 2　LPDR 在不同参数设置下对 O 点邻域集的计算结果

Th	<0.005,0.05>	<0.005,0.1>	<0.01,0.005>	<0.01,0.1>
LPDR	正确	正确	正确	不正确

对于含噪 Swiss roll 数据集,三个嵌入结果中只有基于 LPDR 的 SIE 所获得的结果定性正确。造成此结果的原因是数据集含有导致了流形的虚假贯通的小扰动噪声点(图 2(c)中加白框的点)。LTSA 利用局域泰勒近似的误差判定邻域选择的合理性,若近似误差超过阈值,则依次删除到参考点的欧氏最远的邻域点,直到近似误差小于阈值。这里近似误差阈值正比于用户参数[9]。由于选择很小 η 经常会导致邻域点数过少,破坏流形的连通性,所以这种机制使距参考点较近的虚假贯通点或噪声点很难被删除。而 LPDR,在由 PCA 嵌入的方差残留初步确定邻域集之后,由局域流形的主方向进一步对邻域点做筛选,可有效地删除距参考点较近的噪声点。

直观起见,以图 5 所示数据集例示两种邻域选择的区别。表 1 和表 2 分别总结了在典型误差控制参数下,两种邻域选择算法对参考点 O 的邻域集的计算结果的定性正确性,其中对 LTSA 考察 η 值,对 LPDR 考察误差控制参数的组合,以序偶 $<Th_{cov}, Th_{err}>$ 表示。显然,正确的邻域选择不应包括噪声点。

上述实验说明,通过选择合理的误差控制参数,基于 LPDR 的 SIE 在很大程度上克服了 SIE 算法对于邻域参数的敏感性。可有效处理局域流形结构差异较大的数据集,并屏蔽噪声点的影响。我们的其他实验表明,此结论亦适用于基于 LPDR 的 Isomap 算法,这里不再附图。

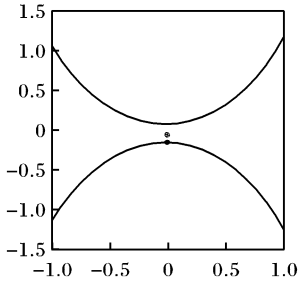


图 5　含有导致了流形的虚假贯通的小扰动噪声点的数据集(孤立点为噪声点)

4　结语

本文给出一种适应性领域选择算法:基于局域主方向重构的邻域选择。仿真表明,LPDR 可有效实现非线性维数约减算法的适应性邻域选择。相对于已有的 LTSA 邻域选择[9],新算法可以更好地屏蔽靠近参考点的孤立噪声点及较大的空间曲率导致的虚假的贯通。

进一步的工作方向是局域适应性与全局适应性的综合。局域适应性邻域机制在提供邻域适应性的同时,引入了额外的阈值参数。全局适应性机制所提供的统计判据可望在一定程度上解决自动参数辨识的问题。

参考文献:

[1] BARLOW HB. Unsupervised learning[J]. Neural Computation, 1989, 1(3): 295 – 311.

[2] MARCUS G. Programs of the Mind[J]. Science, 2004, 304(5676): 1450 – 1451.

[3] BAUM E. What Is Thought?[M]. Cambridge, MA: MIT Press, 2004.

[4] MARDIA KV, KENT JT, BIBBY JM. Multivariate Analysis[M]. Academic Press, London, 1979.

[5] TENENBAUM JB, *et al*. A Global Geometric Framework for Nonlinear Dimensionality Reduction[J]. Science, 2000, 290(12): 2319 – 2323.

[6] ROWEIS ST, *et al*. Nonlinear Dimensionality Reduction by Locally Linear Embedding[J]. Science, 2000, 290(12): 2323 – 2326.

[7] DE SILVA V, TENENBAUM J. Global versus local methods in nonlinear dimensionality reduction[A]. Neural Information Processing Systems 15(NIPS'2002)[C]. 2002.

[8] 侯越先, 丁铮, 何丕廉. 基于自组织的鲁棒非线性维数约减算法[M]. 计算机研究与发展, 2005, 42(2).

[9] WANG J, ZHANG Z, ZHA H. Adaptive Manifold Learning[A]. NIPS 2004[C]. 2004.

[10] BELKIN M, NIYOGI P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation[J]. Neural Computation, 2003, 15(6): 1373 – 1396.