

一种邻域线性竞争的排列降维方法^{*}

闫德勤¹, 吕志超¹, 刘胜蓝²

(1. 辽宁师范大学 计算机与信息技术学院, 辽宁 大连 116081; 2. 大连理工大学 电子信息与电气工程学部 计算机学院, 辽宁 大连 116024)

摘要: 局部线性嵌入算法以及局部切空间排列算法是目前对降维研究有着重要影响的算法, 但对于稀疏数据及噪声数据, 在使用这些经典算法降维时效果欠佳。一个重要问题就是这些算法在处理局部邻域时存在信息涵盖量不足。对经典算法中全局信息和局部信息的提取机制进行分析后, 提出一种邻域线性竞争的排列方法(neighborhood linear rival alignment algorithm, NLRA)。通过对数据点的近邻作局部结构提取, 有效挖掘稀疏数据内部信息, 使得数据整体降维效果更加稳定。通过手工流形和真实数据集的实验, 验证了算法的有效性和稳定性。

关键词: 流行学习; 线性化; 局部线性嵌入; 降维; 稀疏数据

中图分类号: TP181

文献标志码: A

文章编号: 1001-3695(2014)01-0099-03

doi:10.3969/j.issn.1001-3695.2014.01.023

Neighborhood linear rival alignment dimensionality reduction algorithm

YAN De-qin¹, LV Zhi-chao¹, LIU Sheng-lan²

(1. College of Computer & Information Technology, Liaoning Normal University, Dalian Liaoning 116081, China; 2. College of Computer, Faculty of Electronic Information & Electrical Engineering, Dalian University of Technology, Dalian Liaoning 116024, China)

Abstract: LLE and LTSA algorithm are famous canonical dimensionality reduction algorithms. However, the algorithms do not perform well on sparse and noise data. The main reason is insufficient information acquisition in the algorithms. Based on the analysis to the canonical algorithms for the mechanism of global and local information acquisition, this paper presented a new algorithm called a NLRA algorithm. The NLRA used neighborhood linear rival alignment algorithm to extract the local structure of the close neighbors of the data points, and effectively mined the internal information of sparse data. The principle of the algorithm enabled stable global dimensionality reduction effect. The experimental results on the manual manifold and sparse real-world datasets show the efficiency and stability of the algorithm.

Key words: manifold learning; linearization; local linear embedding; dimensionality reduction; sparse data

0 引言

流形学习是把高维数据嵌入到低维流形结构中, 达到降低数据维数的目的。它是机器学习中的重要方法, 在计算机应用的很多领域有着广泛的应用, 成为近年来的研究重点和热点。早期提出的线性降维方法有主成分分析(principal component analysis, PCA)^[1]、多维尺度变换(MDS)^[2]等。PCA是广泛使用的线性降维方法, 以保持降维前后样本集方差最大为主要思想, 尽可能地保证样本间的相互关系。MDS在降维过程中, 为了保持数据点间的相似性, 利用欧氏距离构造优化函数。推动降维研究走向高潮的是近年非线性降维方法的提出。这些方法有等距映射(Isometric feature mapping, Isomap)^[3]、局部线性嵌入(locally linear embedding, LLE)^[4]和局部切空间排列(local tangent space alignment, LTSA)^[5]等。等距映射 Isomap 在保持 MDS 样本最大相似性基础上提出, 用近似样本间的测地距离作为最短路径, 实现了局部与全局的统一。Local MDS^[6]在 MDS 的基础上描述了样本集的局部相似性。LLE 利用局部近似线性思想构建权值矩阵, 使每个样本点所在邻域尽可能地线性化, 并希望得到潜在的低维嵌入流形能保持原样本点间的权

值关系。LLC(locally linear coordination)^[7]在 LLE 的基础上对计算出的每个邻域的局部坐标进行排列, 进而得到了全局的低维坐标。LLE 和 LLC 算法在连续的数据集上都能够得到较好的低维流形, 但在处理信息缺失严重的稀疏数据集的时候, 局部信息量不足的问题非常突出, 影响高维数据映射到低维空间后数据间的距离关系。

局部切空间排列(LTSA)是利用局部切空间来反映流形的局部几何特征, 进而能够比较好地描述细节的非线性。LTSA 继承了 PCA 的优点, 对每个样本局部实施 PCA 降维, 并对局部仿射坐标进行全局排列, 进而得到全局的低维流形。但是 LTSA 也面临着一些问题: 数据集稀疏时, 它的局部邻域到局部切空间的投影距离往往很大, 影响降维效果。此外, LTSA 对于高曲率的流形学习效果也不好, 原因是当曲率大的时候, 拟合坐标之间的关系同样本点之间的弧长关系会有较大的偏差, 如果用这样的拟合坐标来反映局部邻域结构不可避免地会对最终的嵌入结果带来影响。

针对经典流行算法中的不足, 近些年一些学者提出了一些改进方法, 如正交的邻域保持投影(orthogonal neighborhood preserving projection, ONPP)^[8]、邻域保持嵌入(neighbor-hood pre-

收稿日期: 2013-05-11; 修回日期: 2013-06-23 基金项目: 国家自然科学基金资助项目(61105085)

作者简介: 闫德勤(1962-), 男, 教授, 博士, 主要研究方向为模式识别、数据挖掘等(yandeqin@163.com); 吕志超(1988-), 男, 硕士研究生, 主要研究方向为模式识别; 刘胜蓝(1984-), 男, 博士研究生, 主要研究方向为模式识别。

serving embedding, NPE)^[9]、局部线性切空间排列 (linear local tangent space alignment, LLTSA)^[10]、随机样本重构算法 (random sample reconstruction, RSR)^[11] 等能够提取局部信息的线性流形学习算法。还有改进的局部线性嵌入算法 (improved locally linear embedding, ILLE)^[12], 提出自适应选取的权值矩阵的方法保持降维的稳定性。稀疏局部嵌入算法 (sparse-NE-based, SpNE)^[13], 它是一种利用 k 均值聚类确定邻域数据点个数 k 的算法, 来保持嵌入的稳定性。Du 等人提出的一种自适应的邻域选取方法^[14], 用于 Isomap 的局部邻域处理。Zou 等人提出了 natural nearest neighbor (3N) 邻域选择方法^[15], 用于 LLE 的降维对权值矩阵的分析。后来又把 3N 用在改进 LTSA 中, 提出一个最大 k (Sup), 能更好地找到局部切空间^[16]。这些算法都是将高维数据映射到低维空间的非线性隐式映射明确为一种线性映射而得到的, 能很好地把握局部信息, 但是没有提出对稀疏矩阵有效提取局部信息的办法。

数据的稀疏性是自然数据的属性之一^[17,18], 用降维方法处理稀疏数据至关重要。若采用一般的降维方法, 不能充分地挖掘高维空间中的局部信息, 很难保证降维的稳定性。针对稀疏数据的特性, 本文提出了一种邻域线性竞争的排列方法 (NLRA)。该方法采用了一种新的排列算法, 通过对数据点的近邻作局部结构提取排列, 充分挖掘局部信息, 使低维空间中全局信息更加完整、详细。使用本文提出的 NLRA 与 LTSA、LLE 算法在手工流形、Frey 人脸和图像检索中进行了对比实验, 实验结果显示出新算法具有较好的稳定性和有效性。

1 局部线性竞争嵌入方法

1.1 竞争模型

考虑高维观测样本 $X = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{D \times N}$, 通过流形学习后的低维数据为 $Y = [y_1, y_2, \dots, y_N] \in \mathbb{R}^{d \times N}$ 的过程, 若利用 LLE 和 LTSA 学习样本稀疏的高维数据集, 样本间的信息量非常少, 容易出现欠学习的现象。本文考虑通过增加对目标点邻域的局部结构提取, 充分挖掘局部的信息, 提出了局部线性竞争排列方法。图 1 为二维空间竞争模型中目标点邻域内的结构。

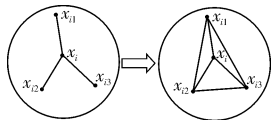


图 1 x_i 的局部近邻结构

现令 x_i 和其 $k-1$ 个近邻点构成一个邻域, 下标集合为 $N(i)$ 。在低维流形局部嵌入的时候, 与 LLE 计算权值不同 (图 1 左), 对邻域 $N(i)$ 中的每个点作局部结构提取, 让邻域 $N(i)$ 中的 x_{i1} 也跟邻域中其他点的线性组合作差再平方, 使得线性表达误差最小 (图 1 右), 从而解得权值 w_{ij} 。对应的优化函数如下:

$$\min \varepsilon(Y_i) = \sum_{j \in N(i)} \|y_i - \sum_j w_{ij} y_j\|^2 r_{i,t} \quad (1)$$

其中: y_i 对应高维空间的 x_i , w_{ij} 为 x_i 与邻域 $N(i)$ 内的 x_j 之间的权值, 该权值由原始样本点局部结构决定; $r_{i,t}$ 为 x_i 与 x_t 之间的权值, 该权值由样本整体分布结构决定, $r_{i,t} = \exp(x_i - x_t)^2$ 。

1.2 局部结构的刻画 (权值确定)

x_i 和其 k 个近邻点构成一个邻域为 $N(i)$ 。权值 w_{ij} 由下式 (2) 确定:

$$\min \varepsilon(w) = \sum_{i=1}^N \sum_{j \in N(i)} \|x_i - \sum_j w_{ij} x_j\|^2 \quad (2)$$

其中: $x_j (j=1, 2, \dots, k)$ 为 x_i 的 k 个近邻点, w_{ij} 是 x_i 与 x_j 之间的权值, 且要满足条件:

$$\sum_{j=1}^N \sum_{i \in N(j)} w_{ij} = 1 \Leftrightarrow \mathbf{1}_N^T \mathbf{1}_N = 1 \quad (3)$$

通过式 (2) 和 (3), 利用 Lagrange 乘子法求解出 w_{ij} 。权值 $r_{i,t}$ 由下面确定:

$$r_{i,t} = \exp(x_i - x_t)^2 \quad (4)$$

2 NLRA 理论分析及算法

2.1 NLRA 理论分析

就局部来看, 邻域线性竞争排列思想形式上可以用降维优化函数表示:

$$\min \varepsilon(Y_i) = \sum_{t \in N(i)} \|y_i - \sum_j w_{ij} y_j\|^2 r_{i,t} \quad (5)$$

用式 (5) 求出的 y_i 为局部降维后的坐标。为了使整体坐标有局部约束, 本文用下面的优化函数 (广义的排列约束):

$$f = \min \varepsilon(Y) = \sum_{i=1}^N \sum_{t \in N(i)} \|y_i - \sum_j w_{ij} y_j\|^2 r_{i,t} \quad (6)$$

设 $w'_i = (w_{i1}, w_{i2}, \dots, w_{ik})^T$, I'_t 为第 t 个分量为 1, 其他为 0 的 k 维列向量。由式 (6) 得

$$f = \min \varepsilon(Y) = \sum_{i=1}^N \sum_{t \in N(i)} \|Y_i(I'_t - w'_i)\|^2 r_{i,t} \quad (7)$$

其中: $Y_i = (y_{i1}, y_{i2}, \dots, y_{ik})$ 为 y_i 的 k 个近邻向量构成的矩阵, 也可写为 $Y_i = YS_i$, S_i 为选择矩阵。令 $D_i = \text{diag}(r_{i,i_1}, r_{i,i_2}, \dots, r_{i,i_k})$ 为对角阵, 对角线上的元素为 x_i 和其 k 个近邻点的距离权值, 即 $(r_{i,i_1}, r_{i,i_2}, \dots, r_{i,i_k}) = r_i S_i$ 。这里 r_i 为距离权值矩阵 R 的第 i 行, S_i 为选择矩阵。得到

$$f = \min \varepsilon(Y) = \sum_{i=1}^N \text{Tr}(Y_i(I - W_i)D_i(I - W_i)^T Y_i^T) =$$

$$\sum_{i=1}^N \text{Tr}(Y_i M_i Y_i^T) = \text{Tr}(\sum_{i=1}^N (Y_i M_i Y_i^T))$$

其中: $M_i = (I - W_i)D_i(I - W_i)^T$, I 为 $k \times k$ 阶单位矩阵, $W_i = (w'_{i1}, \dots, w'_{ik})$ 为 k 个列向量 w'_i 组成的 $k \times k$ 阶矩阵。

令 $M = \text{diag}(M_1, M_2, \dots, M_N)$, 有

$$f = \min \varepsilon(Y) = \text{Tr}(\sum_{i=1}^N Y_i M_i Y_i^T) = \text{Tr}(\sum_{i=1}^N YS_i M_i S_i^T Y^T) =$$

$$\text{Tr}(YSMS^T Y^T) = \text{Tr}(YBY^T)$$

其中: S_i 是满足 $Y_i = YS_i$ 的选择矩阵, $B = SMS^T$, $S = [S_1, S_2, \dots, S_N]$, 约束条件为 $Y^T Y = I_N$ (这里 I_N 为 $N \times N$ 阶单位矩阵)。其中: $Y = (y_1, y_2, \dots, y_N)$, $y_i \in \mathbb{R}^d$, $d \ll D$ 。

最终的优化模型为: $f = \min \varepsilon(Y) = \text{Tr}(YBY^T)$ 。约束条件 $Y^T Y = I_N$ 是为了防止出现退化解。

2.2 NLRA

通过上述的理论分析, 给出 NLRA 的步骤:

a) 初始化数据集 X , 及适应选取 k, d , 利用 k 邻域方法构建邻域集。

b) 对 $i = 1 : N$, 计算 $w_i = \frac{Z_i^{-1} \cdot \mathbf{1}_N}{\mathbf{1}_N^T Z_i^{-1} \mathbf{1}_N}$, 得到权值阵 W , 同时计算权值阵 R , r_i 作为于点和点距离的权值, $r_{i,t} = \exp(x_i - x_t)^2$ 。

c) $D_i = \text{diag}(r_{i,i_1}, r_{i,i_2}, \dots, r_{i,i_k})$ 为对角阵, 得到 $M_i = (I - W_i)D_i(I - W_i)^T$, S_i 是 $Y_i = YS_i$ 的选择矩阵, 得到 $B = SMS^T$ 。

d) 对矩阵 B 进行 EVD, 对其特征值进行排序 $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$, 则 B 的第 2 ~ 第 $d+1$ 个最小特征值对应的特征向量为低维嵌入 $Y = [q_2, \dots, q_{d+1}]^T$ 。

3 实验结果分析

3.1 手工流形实验

本节采用 S-curve 的生成函数为: $t = \pi(1.5\text{rand}(1, N/2) - 1)$; $\text{height} = 15\text{rand}(1, n)$; $X = [\cos(t), -\cos(t)]$; height ; $[\sin(t), 2 - \sin(t)]$ 。Swiss-roll 的生成函数为: $t = (3\pi/2) \times (1 + 2\text{rand}(1, N))$; $\text{height} = 21\text{rand}(1, N)$; $X = [t \cdot \cos(t); \text{height}; t \cdot \sin(t)]$ 。实验结果如图 2 和 3 所示。

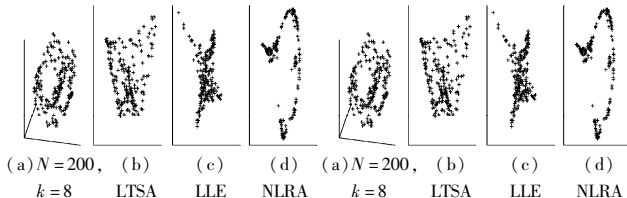


图2 利用 LTSA、LLE 及 NLRA 对 Swiss-roll 降至二维
图3 利用 LTSA、LLE 及 NLRA 对 S-curve 降至二维

以上实验说明,在样本数据稀疏(数据点稀少)的条件下, LLE 及 LTSA 处理效果不理想,低维拓扑流行数据交叠严重,颜色变化混乱,如图 2(b)(c)、图 3(b)(c)所示。主要原因在于样本局部信息有限,原数据空间中每个数据点和其近邻点组成的邻域就很难保证是线性的。但是 NLRA 却能在原数据空间中信息量少的情况下,深度挖掘数据信息,能保持邻域线性重构的有效性,降至二维后基本保持原始流形数据点的颜色变化结构,数据点平铺效果良好,如图 2(d)、图 3(d)所示。

3.2 Frey 人脸实验

Frey 人脸库中共 1 960 幅人脸,每幅图像为 20×28 灰度图片,该表情库是由视频设备连续拍摄得到的,包括一个人的各种表情如:中性(自然)、高兴和不高兴等。本实验随机选取 300 幅人脸图像,邻域参数 $k = 6$,通过 LLE、LTSA 及 NLRA 降至二维后的效果如图 4 所示。

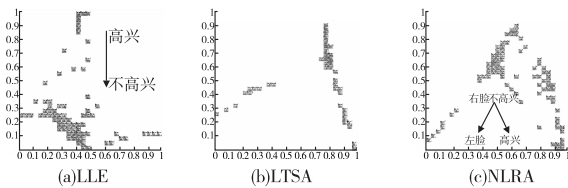


图4 Frey 人脸表情库二维可视化效果

在 Frey 人脸实验中,三种不同算法表现好坏在图 5 中是显而易见的。LLE 处理效果不够理想,图片交叠严重,不能很好地分清各种表情。主要原因是在处理稀疏数据时 LLE 算法不能很好地挖掘局部信息,导致降至低维空间后表情混杂。LTSA 也没达到理想的降维效果,奇异值现象明显。主要原因是在处理稀疏数据时,LTSA 很容易遇到构建切空间信息量不够的情况,使得分解矩阵出现较严重奇异的状况。相比较而言, NLRA 的处理效果理想,能清楚地分开高兴表情和不高兴表情,很好地解决了稀疏数据中信息量不足的问题。

3.3 实验分析

从以上几个实验可以看出, NLRA 虽然与 LLE、LTSA 一样都是基于线性的非线性降维方法,但 NLRA 在处理稀疏数据时降维效果却远优于 LLE 及 LTSA,其主要原因为: a) NLRA 是一种邻域线性竞争的排列方法,能够更好地挖掘数据集在高维空间中的局部信息; b) NLRA 在把数据集从高维空间映射到低维

空间的过程中加上了点和点距离的权值, $r_{ii} = \exp(x_i - x_i)^2$, 可以根据不同的数据集取不同的值,类似于核函数。

以上两点大大减小了降维过程中局部空间的误差,进而防止了低维拓扑空间的变形及扭曲,增加了稳定性。时间复杂度分析: NLRA 与 LLE、LTSA 的时间复杂度都是 $O(pn^2)$, p 是稀疏矩阵非零元的个数, n 是样本点个数。

4 结束语

本文通过研究经典的流形学习算法 LLE 及 LTSA, 分析处理稀疏数据集时局部信息量不足的问题, 提出一种非线性降维算法 NLRA。该算法在数据集稀疏的情况下, 提取局部信息更加全面, 使得数据整体降维效果更加稳定。对于邻域大小的判断和选取仍然是以后工作的重点, 以使得这种排列的思想更好地应用到文本分类、人脸识别等实际领域。

参考文献:

- [1] WOLD S, ESBENSEN K, GELADI P. Principal component analysis [J]. *Chemometrics and Intelligent Laboratory Systems*, 1987, 2(1): 37-52.
- [2] KRUSKAL J B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis [J]. *Psychometrika*, 1964, 29(1): 1-27.
- [3] TENENBAUM J B, De SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. *Science*, 2000, 290(5500): 2319-2323.
- [4] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. *Science*, 2000, 290(5500): 2323-2326.
- [5] ZHANG Zhen-yue, ZHA Hong-yuan. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment [J]. *Journal of Shanghai University: English Edition*, 2004, 8(4): 406-424.
- [6] YANG Li. Alignment of overlapping locally scaled patches for multidimensional scaling and dimensionality reduction [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2008, 30(3): 438-450.
- [7] TEH Y W, ROWEIS S. Automatic alignment of local representations [M]// *Advances in Neural Information Processing Systems*. [S. l.]: MIT Press, 2002: 841-848.
- [8] KOKIOPOULOU E, SAAD Y. Orthogonal neighborhood preserving projections: a projection-based dimensionality reduction technique [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2007, 29(12): 2143-2156.
- [9] HE Xiao-fei, CAI Deng, YAN Shui-cheng, et al. Neighborhood preserving embedding [C]// *Proc of the 10th IEEE International Conference on Computer Vision*. [S. l.]: IEEE Press, 2005: 1208-1213.
- [10] ZHANG Tian-hao, YANG Jie, ZHAO De-li, et al. Linear local tangent space alignment and application to face recognition [J]. *Neurocomputing*, 2007, 70(7): 1547-1553.
- [11] GUHA T, WARD R K. Learning sparse representations for human action recognition [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2012, 34(8): 1576-1588.
- [12] WEN Ying, HE Liang-hua. An improved locally linear embedding for sparse data sets [J]. *International Journal of Pattern Recognition and Artificial Intelligence*, 2011, 25(5): 763-775.
- [13] GAO Xin-bo, ZHANG Kai-bing, TAO Da-cheng, et al. Image super-resolution with sparse neighbor embedding [J]. *IEEE Trans on Image Processing*, 2012, 21(7): 3194-3205.
- [14] DU Chun, ZHOU Shi-lin, SUN Ji-xiang, et al. Robust ISOMAP based on neighbor ranking metric [J]. *Intelligent Computing Technology*, 2012, 7389: 221-229.
- [15] ZOU Xian-lin, ZHU Qing-sheng, JIN Yi-fu. An adaptive neighborhood graph for LLE algorithm without free-parameter [J]. *International Journal of Computer Applications*, 2011, 16(2): 20-23.
- [16] ZOU Xian-lin, ZHU Qing-sheng. Adaptive neighborhood graph for LTSA learning algorithm without free-parameter [J]. *International Journal of Computer Applications*, 2011, 19(4): 28-33.
- [17] BALASUBRAMANIAN M, SCHWARTZ E L. The Isomap algorithm and topological stability [J]. *Science*, 2002, 295(5552): 7.
- [18] SILVA J, MARQUES J, LEMOS J. Selecting landmark points for sparse manifold learning [C]// *Advances in Neural Information Processing Systems*. 2006: 1241.