

小世界邻域优化的局部线性嵌入算法

张育林, 庄健, 李小虎, 王孙安

(西安交通大学机械工程学院, 710049, 西安)

摘要: 通过分析稀疏数据或噪声数据, 导出局部线性嵌入(LLE)算法出现失效的原因, 由此提出了一种基于小世界邻域优化的局部线性嵌入(SLLE)算法. 将复杂网络算法引入到流形学习中, 利用小世界算法对 LLE 算法进行数据优化, 并以最短路径和局部集群系数作为局部优化参数, 解决了数据点不规则时以欧氏空间作为邻域判别标准在构建局部超平面造成嵌入结果扭曲的难题. 通过 3 组标准测试数据集比较了 SLLE、LLE 算法, 结果表明 SLLE 算法的计算效果、鲁棒性、非理想数据的降维结果均优于 LLE 算法, 且计算正确率至少提高 10%.

关键词: 局部线性嵌入; 降维; 小世界邻域

中图分类号: TP181 **文献标志码:** A **文章编号:** 0253-987X(2008)12-1486-04

Small World Neighborhood Optimized Local Linear Embedding Algorithm

ZHANG Yulin, ZHUANG Jian, LI Xiaohu, WANG Sun'an

(School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: By analyzing the invalidity reason of the local linear embedding (LLE) algorithm in case of the sparse data or the high noise data, small world neighborhood optimization LLE algorithm (SLLE) is proposed based on the complex networks theory. The data in LLE are optimized using the small world algorithm, and the shortest path and the local neighbor set clustering coefficients are used as the local parameters. As a result, the problem of the embedding distortion using only local linear patch of the manifold to define neighborhood in Euclidean space is effectively solved. Three groups of standard data sets are selected to test and to compare the efficiency and robustness of SLLE and LLE. The experimental results show that the calculation results, robustness and dimension reduction of SLLE are all better than those of LLE, and accuracy rate of SLLE is at least 10 percent higher than that of LLE.

Keywords: local linear embedding; dimension reduction; small world neighborhood

现实世界的一些数据集, 如语音、文本、基因、图像等, 具有高维、非线性等特性, 分析这些数据很困难^[1], 而降维的目的就是要发现高维数据集分布的内在规律. 2000 年《Science》上提出了 2 种最具代表性的方法: 局部线性嵌入(LLE)^[2] 和等距特征映射(ISOMAP)^[3], 使得流形学习迅速发展. LLE 算法采用局部线性化方法, 通过构建局部线性超平面, 将高维、非线性数据映射到全局的低维空间中, 以保持邻域点集合的结构不变, 保证平移、旋转不变性^[4].

但是, LLE 要求采集足够量的自平滑嵌入流形数据, 每个采样点与构成的邻域数据之间具有较强的关联性. 当数据集噪声较大或样本点数据稀疏时, 流形上的邻域数据间为弱关联, 因此在选取邻域数据时会将噪声数据或相隔较远的点作为邻近点, 从而导致嵌入结果产生扭曲^[4-5].

为此, 本文提出了一种基于小世界邻域优化的局部线性嵌入(SLLE)算法, 以提高 LLE 的对于非理想采样数据的降维效果和稳定性.

1 LLE 算法及失效分析

LLE 算法可将高维空间中的数据向低维目标空间映射,从而降低数据维数,并在映射过程中保证高维数据和低维数据的邻域结构.假设输入数据为采样于流形空间的 n 个 D 维实向量 $\mathbf{x}_i, \mathbf{x}_i \in \mathbf{R}^D, i=1, 2, \dots, n$ (由向量 $\mathbf{x}_i$ 构成的数据集记作 $\mathbf{X}$) 及低维嵌入向量 $\mathbf{y}_i, \mathbf{y}_i \in \mathbf{R}^d, i=1, 2, \dots, n$ (由 n 个向量 $\mathbf{y}_i$ 构成的数据集记作 $\mathbf{Y}$), 其中 $d \ll D$, 那么 LLE 算法的主要步骤如下.

(1) 以欧氏距离为依据, 为每个样本点 $\mathbf{x}_i$ 寻找 k 个最近点邻域 $\mathbf{x}_{ij} (j=1, 2, \dots, k)$, 记作 $N_i = \{N(\mathbf{x}_i) | \mathbf{x}_i \in \mathbf{X}\}$.

(2) 用每个 $\mathbf{x}_i$ 的最近邻域线性组合近似地表示为 $\mathbf{x}_i$, 通过最小化重构误差 ϵ , 得到局部重构权值矩阵 $\mathbf{W}_{ij}$, 则有

$$\epsilon(\mathbf{x}_i) = \sum_{j=1}^n \|\mathbf{x}_i - \sum_{j=1}^n \mathbf{W}_{ij} \mathbf{x}_{ij}\|^2 \quad (1)$$

$\mathbf{W}_{ij}$ 矩阵遵循 2 个约束条件: ① $\sum_{j=1}^n \mathbf{W}_{ij} = 1$; ② 对于非 $\mathbf{x}_i$ 邻域的点, $\mathbf{W}_{ij} = 0$.

(3) 保持步骤(2)中的重构系数 $\mathbf{W}_{ij}$ 不变, 通过求解稀疏对称矩阵, 可最小化代价函数

$$E(\mathbf{y}_i) = \sum_{i=1}^n \|\mathbf{y}_i - \sum_{j=1}^n \mathbf{W}_{ij} \mathbf{y}_{ij}\|^2 \quad (2)$$

从而求出低维嵌入向量 $\mathbf{y}_i$. 式(2)遵循 2 个约束条件: ① $\sum_{i=1}^n \mathbf{y}_i = 0$, 以限定输出结果的平移; ②

$\frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^T = \mathbf{I}$, 以限定输出结果的旋转, 其中 $\mathbf{I}$ 为 $d \times d$ 单位矩阵.

由以上步骤可以看出, LLE 算法相当于把整个流形分成 n 个局部区域, 每个区域的 $\mathbf{x}_i$ 可以用邻域线性地表示. 只要邻域参与重构, 便可仅从邻域几何特征获取整个低维流形, 但这种局部性的方法依赖于邻域的选取, 所以通常假设采样于流形空间的数据是理想的, 比如数据点需足够多, 采样要均匀或没有噪声.

将流形 Peak 图均匀地分成若干个块 (见图 1a), 用灰度值 (仅限本文) 标识邻域关系, 那么图 1b 为 Peak 图均匀采集 450 个数据点的结果, 图 1c 为非理想随机采样 450 个数据点的结果, 图 1d 为理想采样数据下 LLE 降维结果 (邻域 $k=6$), 图 1e 为非理想数据下 LLE 降维结果 (邻域 $k=6$). 从图 1 可

以看出, 对于同样的采样数量, 当采样为非理想数据时, LLE 算法将源数据映射到了目标空间邻域; 当采样为非理想数据时, LLE 算法降维后的数据点混杂在一起. 其主要的原因是, 在构建局部超平面时, 如果数据点不规则, 单凭欧氏空间作为邻域判别的标准, 会将流形上相距较远的点划为邻域点, 使嵌入结果明显扭曲. 针对这一问题的改进算法是, 采用一个球形空间作为邻域选取参考, 但当选取邻域的欧氏距离超过这个空间时, 就舍弃该点^[4], 而对于非理想数据, 该方法的稳定性较差.

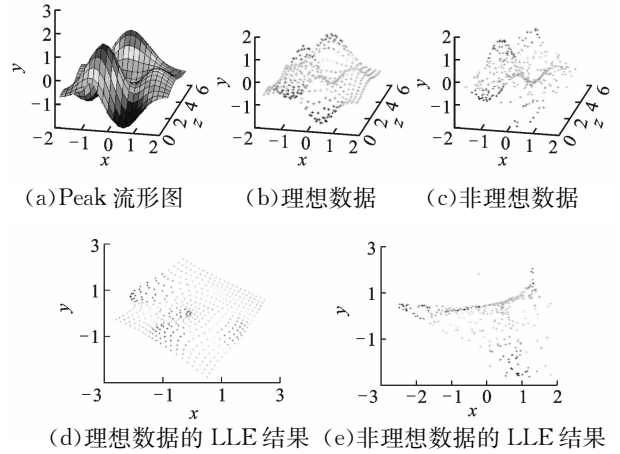


图 1 Peak 流形采样数据降维示意图

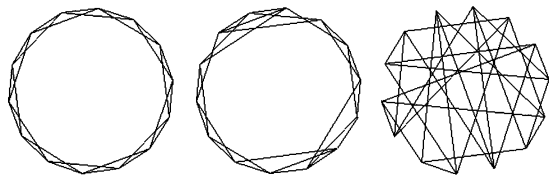
2 小世界网络的邻域优化及应用

小世界网络源于社会心理学家 Milgram 在研究社会网络时提出的“六度分离”理论^[6], 而 1998 年 Watts 和 Strogatz 在《Nature》上发表的论文^[7], 使得小世界现象逐渐被计算机科学家所重视, 并被迅速上升为网络论, 成为人工智能、复杂性理论探讨的热点^[8]. 小世界网络已在工程中得到应用, 如函数优化、复杂系统诊断等^[8-9].

2.1 小世界原理定义

小世界网络是从规则网络向随机网络过渡的中间网络. 在一个有 n 个节点规则的环状网络中, 采用欧氏距离为每个结点 $\mathbf{x}_i (i=1, 2, \dots, n)$ 寻找 k 个邻域, $\mathbf{x}_i$ 与最近邻的 k 个结点连接出 k 条边 (见图 2a), 这样就会产生规则网络. 以概率 p 对节点进行重新随机连接, 当 p 较小时, 形成小世界网络 (见图 2b), 当 p 较大时 ($p \approx 1$), 形成随机网络 (如图 2c).

在社会网络中, 如果 A 与 B 熟悉, 而 B 又与 C 熟悉, 则很有可能 A 与 C 也相互认识. 网络能够显示这种良好、局部的重叠性, 这种特性称之为网络的集群性. 集群性可体现在集群系数量化上, 集群系数



(a) 规则网络 (b) 小世界网络 (c) 随机网络

图 2 小世界网络的演化过程

分为网络整体集群系数和节点局部集群系数,本文采用节点局部集群系数^[10].

设点 x_{ij} 为 x_i 的邻域 $N_i = \{N(x_i) | x_i \in \mathbf{X}\}$ 中第 j 个点, $j=1,2,\dots,k$, 则 x_{ij} 的局部集群系数

$$C_{ij} = \frac{2E_{ij}}{k(k-1)} \quad (3)$$

式中: k 为邻域大小; E_{ij} 为 x_{ij} 与 k 个邻域点形成的边数 (如果 x_{ij} 同时也是 $N(x_i)$ 的点邻域, 就认为在 x_{ij} 与这个邻域点之间会形成一条连接边).

小世界网络是集群系数大的局部短连接和较长连接构成的网络, 其中最短路径和集群系数是 2 个重要的参数.

2.2 LLE 小世界邻域优化算法

自流形的采样数据同样具有小世界聚类特征. 在对近邻数据点进行连接时, 那些性质相近的点之间联系紧密, 节点的局部聚集系数较大. 如果采集到的自平滑流形的数据点很规则 (如图 1b), 对数据点进行邻域选择和连接时就会形成类似图 2a 的规则拓扑结构. 当数据集噪声较大或数据稀疏时, 流形上的邻域数据会产生弱关联, 近类数据在欧氏空间的分布就会出现不规则的现象. 对同类点进行连接, 会产生类似图 2b 的小世界网络结构.

在对数据进行邻域选取时, 如果仅用欧氏距离来评判, 很可能将噪声数据或相隔较远的点作为邻近点, 从而扭曲了数据的邻域结构, 并产生错误的低维嵌入结果 (如图 1e). 为了解决这一问题, 本文将小世界网络中的 2 个重要指标——最短路径和集群系数, 作为选取 k 邻域的依据对邻域进行优化. 算法描述如下.

初始数据: n 个 D 维实向量 x_i 构成的高维数据集 $\mathbf{X} = \{x_i | x_i \in \mathbf{R}^D\}$.

参数: 嵌入空间维数 d 、邻域大小 k 、邻域优化范围 m , 选择概率 p .

输出数据: 小世界算法优化后的邻域数据 $N = \{N(x_i) | x_i \in \mathbf{X}, i=1,2,\dots,k\}$.

步骤 1: 为每个样本点 x_i 寻找 m 个最近邻域 $x_{ij} (j=1,2,\dots,m), m > k$.

步骤 2: 将 x_{ij} 按照欧氏距离升序排序. 选取前 $k(1-p)$ 个点作为邻域, 记作 $l = \{x_{il} | x_{il} \in \mathbf{X}, l=1,2,\dots,k(1-p)\}$, 它相当于小世界网络中的局域短连接.

步骤 3: 在剩余的 $m-k(1-p)$ 点中, 计算每个点的局部聚集系数, 并降序排列. 选取前 $k \times p$ 个点作为邻域, 记作 $C = \{x_{iq} \in \mathbf{X} | q=k(1-p)+1,\dots,k\}$, 并作为小世界网络中的较长连接.

步骤 4: 将步骤 2 和步骤 3 中的数据集进行合并, 得到优化后的邻域数据集 $N = \{N(x_i) | x_i \in \mathbf{X}, i=1,2,\dots,k\}$.

可见, 优化后的数据具有明显的小世界特征, 既保证数据之间具有最短路径, 也使数据集的集群系数较大. 优化后的数据, 可按照第 1 节基本 LLE 算法中步骤 2 和步骤 3 进行局部重构和低维嵌入计算.

3 实验结果与分析

为了验证算法的有效性, 采用 SLLE 算法与 LLE 算法, 对 Peak、S-Curve、Swiss-Roll 流形稀疏随机采样数据进行降维对比. 相关实验在 Matlab 语言、主频为 2.4 GHz、内存为 512 MB 的 PIV 计算机上完成.

非理想 Peak 数据的 SLLE 降维结果如图 3 所示, 其中 $m=10, p=0.09$. 数据点在 2 个波峰之间形成均匀过渡, 三维数据在二维空间正确展开, 可见 SLLE 算法能较好地体现流形结构.

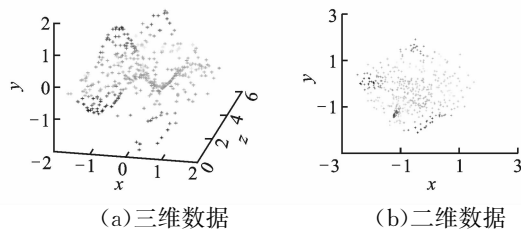


图 3 非理想 Peak 数据的 SLLE 降维结果

LLE 算法对噪声比较敏感, 而 SLLE 算法由于增加了数据聚类判断, 所以降低了噪声的敏感性. 图 4 是 SLLE 对含有噪声的 S-Curve 数据进行降维的结果, 其中 $n=800$, 噪声数据点数为 50, $k=10, m=16, p=0.1$. 用 LLE 和 SLLE 算法对此类数据进行了 100 次实验, 参数和结果对比见表 1. 实验表明, 在一定范围的噪声下, SLLE 的稳定性较好.

当数据稀疏时, 流形上的邻域数据会产生弱关联, 这会将相隔较远的点作为邻近点处理, 结果扭曲了数据的邻域结构. 用 LLE 和 SLLE 算法对 Swiss-Roll ($m=9, p=0.12$) 流形进行降维, 结果如图 5 所

示. 为了验证算法的稳定性, 分别进行了 100 次实验, 实验参数和对比结果同见表 1.

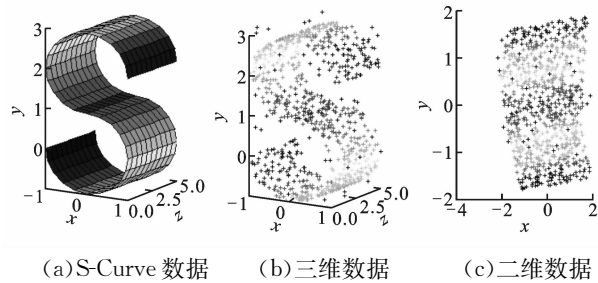


图 4 SLLE 对含噪声数据的降维结果

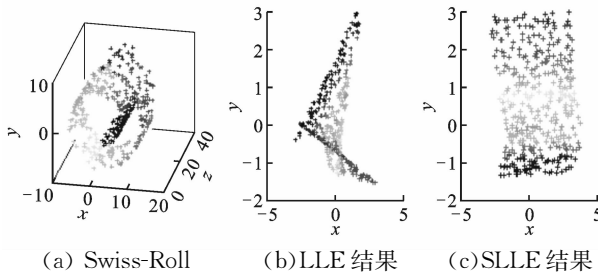


图 5 LLE 和 SLLE 对 Swiss-Roll 降维结果

表 1 降维实验参数及结果

实验参数与结果	实验流形			
	Peak	Swiss-Roll	S-Curve	S-Curve
D	3	3	3	3
d	2	2	2	2
k	7	8	10	8
n	450	400	800	300
噪声数据数	0	0	50	0
实验次数	100	100	100	100
错误率	LLE	0.39	0.45	0.20
	SLLE	0.21	0.28	0.09

对于理想采样数据, 如均匀采样或者采样点足够多, LLE 算法能够找出非线性数据之间的拓扑关系, 并成功地将高维数据映射到低维空间, 而不改变原有数据的相对位置. 但在对典型随机采样稀疏数据和含有噪声的非理想高维数据进行低维映射时, 它的错误映射率较高, 不同的样本点会混杂在一起. 对于典型非理想数据, SLLE 算法利用了数据的小世界特性, 从路径长度、集群系数优化和数据空间排列可以看出, 它将高维数据空间中相近数据成功地映射到了低维空间中同一邻域之内, 体现出了数据的聚类特性和相互关系. 多次实验结果也验证了 SLLE 算法的稳定性.

4 结束语

本文针对 LLE 算法对于稀疏、噪声数据稳定性差的问题, 采用了小世界算法对 LLE 算法进行局部邻域数据优化. 标准数据的实验结果表明, SLLE 算法可提高 LLE 算法的非理想数据降维的效果和稳定性.

参考文献:

[1] HASTIE T, TIBSHIRANI R, FREEDMAN J. The Element of statistical learning: data mining, inference, and prediction [M]. Berlin, Germany: Springer, 2001.

[2] ROWEIS S S, SAUL L K. Nonlinear dimensionality reduction by local linear embedding [J]. Science, 2000, 290(550): 2323-2326.

[3] TENENBAUM J B, SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(550): 2319-2323.

[4] LAWRENCE K S, SAM T R. Think globally, fit locally: unsupervised learning of low dimensional manifolds [J]. Journal of Machine Learning Research, 2003, 4: 119-155.

[5] SHA F, SAUL L K. Analysis and extension of spectral methods for nonlinear dimensionality reduction [C]//Proceedings of the 22nd International Conference on Machine Learning. New York, USA: ACM, 2005: 784-791.

[6] MILGRAM S. The small world problem [J]. Psychology Today, 1967, 1(1): 61-67.

[7] WATTS D J, STROGATZ S H. Collective dynamics of small-world networks [J]. Nature, 1998, 393 (4): 440-442.

[8] 杜海峰, 庄健, 张进华. 用于函数优化的小世界优化算法 [J]. 西安交通大学学报, 2005, 39(9): 1011-1015.

DU Haifeng, ZHUANG Jian, ZHANG Jinhua. Small-world phenomenon for function optimization [J]. Journal of Xi'an Jiaotong University, 2005, 39(9): 1011-1015.

[9] 李果, 高建民. 基于小世界网络的复杂系统故障传播模型 [J]. 西安交通大学学报, 2007, 41(3): 334-338.

LI Guo, GAO Jianmin. Failure propagation model of complex system based on small world net [J]. Journal of Xi'an Jiaotong University, 2007, 41(3): 334-338.

[10] 余根坚, 许力. 复杂动态网络模型研究进展 [J]. 福州大学学报, 2006, 34(5): 637-642.

YU Genjian, XU Li. The survey of complex dynamical networks models [J]. Journal of Fuzhou University, 2006, 34(5): 637-642.

(编辑 苗凌)