两种流形学习算法的对比研究

王 博，刘美玲，张学敏
(西安建筑科技大学，陕西 西安 710055)

摘 要：介绍了局部线性嵌套和等距映射两种最基本的非线性降维方法，对比测试了两种降维方法在不同参数下的执行效果与效率，总结了两种降维方法所适合的数据特点，并应用于图像识别中，比较了两者在图像识别中的识别率。

关键词：非线性降维；流形学习；局部线性嵌套；等距映射；人脸识别
中图分类号：TP391.41 **文献标识码**：A **文章编号**：1674-7720(2013)08-0042-03

A comparative study of two kinds of manifold learning algorithms

Wang Bo, Liu Meiling, Zhang Xuemin
(Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: Data dimensionality reduction ingeneral can be divided into linear dimensionality reduction and nonlinear dimensionality reduction,locally linear embedding and isometric mapping is one of the two most recently the nonlinear dimensionality reduction method,for different sets of data dimensionality reduction have their advantages and disadvantages.The contrast test of weo dimension reduction method under different parameters of effectiveness and efficiency,summarizes itsdimensionality reduction fit for the characteristics of data, is aoolied to image recognition,compared to the respective image recognition in recognition rate.

Key words: nonlinear dimensionality reduction; manifold learning; locally linear embedding; isometric mapping; face recognition

流形的概念最早是由德国数学家黎曼在1854年提出的，它是微分几何学的基础[1]。流形本质上是局部可坐标化的拓扑空间，可以看作是欧式空间的非线性推广。

1 局部线性嵌入算法

局部线性嵌入算法LLE(Locally Linear Embedding)是ROWEIS S T和SAUL L K于2000年提出的一种非线性降维方法[2]，该方法主要认为在局部意义下，数据结构是线性的，或者说局部意义下的点是在一个超平面上，故可以使用任意一点的邻近点的线性组合来表示该点。对于一组具有嵌套流形的数据集，在嵌套空间与内在低维空间局部邻域间的点的关系应该保持不变。即在嵌套空间，每个采样点可以用它的近邻点线性表示，在低维空间中保持每个邻域中的权值不变，重构原数据使重构误差最小。

通过最小化这种线性表示的误差，可以建立如下数学模型：

$$J_{\mathrm{LLE}}=\arg\ \min \sum_{i=1}^{n} \| y_i-\sum_{i=1}^{k} w_{ij}y_j \|^2 \tag{1}$$

该方法的输入样本 $x=\{x_1,x_2,\cdots,x_n\in R^N\}$，样本的维数为 d，邻域参数为 k(ε 邻域)；输出为低维嵌入 $Y=\{y_1,y_2,\cdots,y_n\in R^d\}$。算法的基本步骤如下。

(1)计算每个样本点的 x_i 的 k 个邻近点，并且把相对于所求样本点距离最近的 k 个样本点规定为所求样本点的 k 个邻近点。其中，k 是一个预先给定的数值。

(2)通过 x_i 的近邻域计算，得到权值矩阵 $\boldsymbol{W}$，其中，若 x_i 和 x_j 不是近邻点，则 $W_{ij}=0$ 且 $\sum W_{ij}=1$。式(1)中 W_i 可以通过如下方法获得。首先计算 x_i 的近邻协方差矩阵 $\boldsymbol{C}$：

$$C_{jk}=(x_i-x_{ij})(x_i-x_{ik}) \tag{2}$$

再通过求解线性方程 $\sum \boldsymbol{C}\boldsymbol{W}=1$ 来求解 W_{ij}。

(3)最小化重构成本函数：

$$g(W)=\sum_{i=1} |x_i-\sum W_{ij}x_j|^2 \tag{3}$$

(4)通过稀疏矩阵 $\boldsymbol{M}(M_{ij}=\delta_{ij}-w_{ij}-w_{ji}+\sum_k w_{ki}w_{kj})$ 来求解式(2)中的特征向量，从而计算由 W_{ij} 最优重构的低维嵌入向量 $\boldsymbol{y}_i$。

该算法有两个待定的参数 k 和 d，由于重构成本函数同时最小化得到的最优权值应该遵循对称性，因此每个点的邻近权值在进行平移、伸缩和旋转变换时保持不变[3]。

2 等距映射

等距映射算法是由 TENENBAUM J B 等人于 2000 年提出的一种非线性降维方法[4]。该方法试图保持数据内部几何特征，从而获得流形上数据之间的测地距离。与传统的非线性降维方法所不同的是，利用等距映射方法可以求得高维数据的本征维数，将本征维数较低的高维数据投影到低维空间中去[5]，使得高维数据可以直接观察。等距映射有两个假设：(1)高维数据所在的低维流形与欧式空间的一个子集是整体等距的；(2)与数据所在的流形等距的欧式空间的子集是一个凸集。

假设 $X_i(i=1,2,\cdots,n)$ 是一组 D 维观测数据，等距映射的步骤如下[6]。

(1)对任意的 X_i、X_j，计算它们之间的距离 $d_{ij}^2=(X_i-X_j)^{\mathrm{T}}(X_i-X_j)$，得到欧式距离阵 $\boldsymbol{D}=(d_{ij})$。

(2)计算每个点的邻近点。

(3)在样本集 X 上定义一个赋权无向图 G，当 X_j 为 X_i 的近邻点或者 X_i 为 X_j 的近邻点时，边的权值为 d_{ij}，即 $d_G(x_i,x_j)=d(x_i,x_j)$；当 X_j 不为 X_i 的近邻点且 X_i 不为 X_j 的近邻点时，权值为 ∞，即 $d_G=\infty$。

(4)用 Dijkstra 算法计算图中任意两点的最短距离，组成测地距离矩阵 $\boldsymbol{D}_G=\{d_G(\mathrm{i},\mathrm{j})\}$。

(5)用多维尺度方法(MDS)对测地距离矩阵 $\boldsymbol{D}_G$ 求得 $d(d<<D)$ 维嵌入向量 $\boldsymbol{Y}_i(i=1,2,\cdots,n)$。

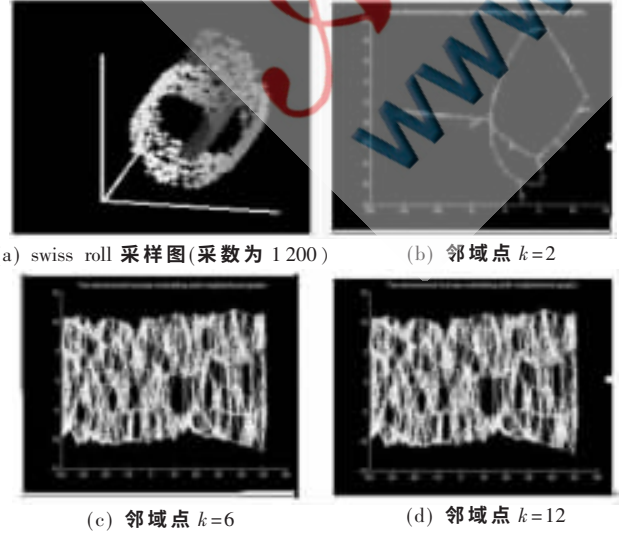
3 两种经典算法的仿真研究

实验 1

使用 MATLAB 软件用 siomap 方法对 swiss roll 数据集进行降维，分别选取数据点个数为 800、1 200，降维以后的维数为 2，在构造邻域图时选取 $k=2$、6、12。降低维数后的仿真结果如图 1 所示，数据降维的用时对比如表 1 所示。

实验 2

使用 MATLAB 软件用 LLE 方法对 swiss roll 数据集进行降维，分别选取数据点个数为 800、1 200。降维以后的维数为 2，在构造邻域图时选取 $k=6$、8、12。降低维数后的仿真结果如图 2 所示，数据降维用时对比如表 2 所示。



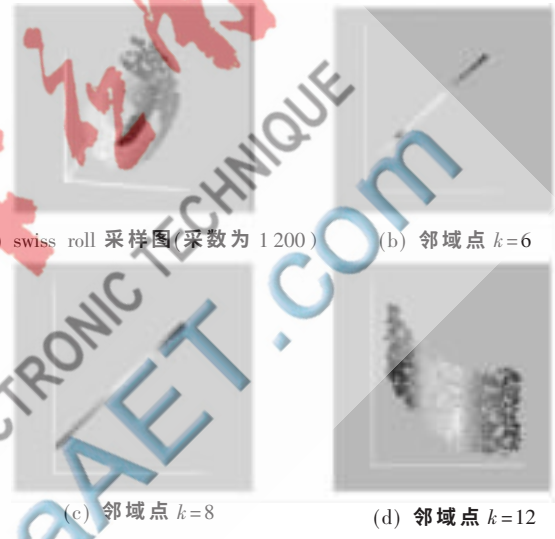
(a) swiss roll 采样图(采数为 1 200)　(b) 邻域点 $k=2$

(c) 邻域点 $k=6$　(d) 邻域点 $k=12$

图 1 siomap 方法对 swiss roll 数据集进行降维

表 1 数据集大小不同、邻域图不同时降维用时(s)

数据点个数	k			
	2	6	12	总计
800	31.23	43.21	48.51	122.95
1 200	50.12	54.31	63.87	168.30
总计	81.35	97.52	112.38	



(a) swiss roll 采样图(采数为 1 200)　(b) 邻域点 $k=6$

(c) 邻域点 $k=8$　(d) 邻域点 $k=12$

图 2 LLE 方法对 swiss roll 数据集进行降维

表 2 数据集大小不同、邻域图不同时的时间表(s)

数据点个数	k			
	2	6	12	总计
800	0.90	0.92	0.92	2.74
1 200	1.21	1.32	1.93	4.46
总计	2.11	2.24	2.85	

实验 3

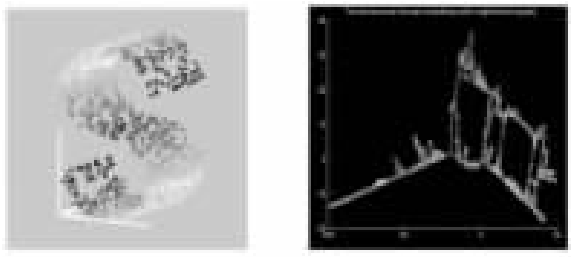
使用 MATLAB 软件用 siomap 方法对 scurve 数据集进行数据降维，分别选择数据点个数为 800、1 200，降维以后的维数为 2，在构造邻域图时选取 $k=2$、6、12。降低维数后的仿真结果如图 3 所示，数据降维用时对比如表 3 所示。

实验 4

使用 MATLAB 软件用 LLE 方法对 scurve 数据集进行降维，分别选择数据点个数为 800、1 200，降维后的维数为 2，在构造邻域图时选取 $k=6$、8、12。降低维数后的仿真结果如图 4 所示，数据降维用时对比如表 4 所示。

4 结果分析

实验 1 中，从图 1 可以看出样本点的分布及其邻域



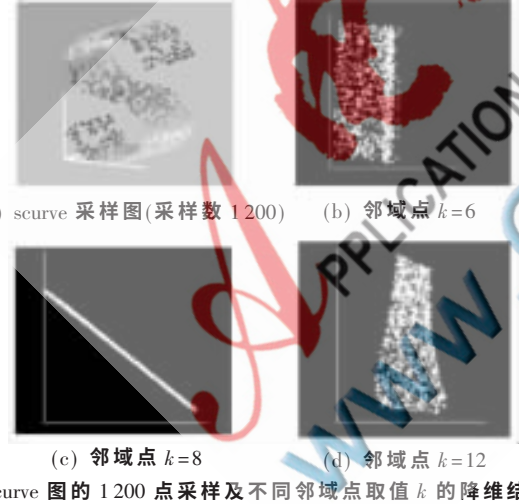
(a) scurve 采样图(采样数 1 200) (b) 邻域点 $k=2$

(c) 邻域点 $k=6$ (d) 邻域点 $k=12$

图 3 scurve 图的1 200 点采样及不同邻域点取值 k 的降维结果图

表 3 数据集大小不同、邻域图不同时的时间表(s)

数据点个数	k			
	2	6	12	总计
800	29.12	31.52	42.42	103.06
1 200	43.21	49.96	58.63	151.80
总计	72.33	81.48	101.05	



(a) scurve 采样图(采样数 1 200) (b) 邻域点 $k=6$

(c) 邻域点 $k=8$ (d) 邻域点 $k=12$

图 4 scurve 图的 1 200 点采样及不同邻域点取值 k 的降维结果图

表 4 数据集大小不同、邻域图不同时的时间表(s)

数据点个数	k			
	6	8	12	总计
800	2.21	2.52	2.63	7.36
1 200	3.73	3.92	3.94	11.59
总计	5.94	6.44	6.57	

点的取值对 isomap 的降维结果会产生比较大的影响[7]。实验 2 中,随着邻域点 k 取值的增加,图 2 有着明显的变化,说明随着邻域 k 的增加,LLE 所得的结果明显增强。在样本点稀疏的情况下,邻域 k 的取值对于 LLE 降维效果有比较明显的影响,因而选取合适的邻域取值对于 LLE 降维有非常重要的作用。对比实验 2 和实验 4 可知,邻域 k 的选择对于不同数据集的选取是不同的。LLE 算法中的待定参数很少(k 和 d),从图 3 可以看出,随着样本邻域选取的增加,会把其他较远点一起纳入,从而造成结果的误差,说明邻域的选取对于实验有着直接的影响。

通过对比实验运行的时间会发现,isomap 所用时间远远大于 LLE。其中主要原因是计算欧式距离矩阵花费时间比较长,计算赋权无向图运算量比较庞大,用多维尺度方法(MDS)时会用到大量的矩阵运算,对于每一个不同的数据集,需要重新计算距离矩阵等,算法复杂度比较高,而 LLE 运算量相对较少。

isomap 算法计算图上两点间的最短距离,执行起来比较慢,该方法适用于学习内部平坦的低维流形,不适于学习有较大内在曲率的流形。LLE 算法可以学习任意维数的低维流形,每个点的近邻权值在平移、旋转和伸缩变换下是保持不变的。在计算耗时上,isomap 远远大于 LLE。

参考文献

[1] 王泽杰.两类非线性降维流形学习算法的比较分析[J].上海工程技术大学学报,2008,22(1):54-59.

[2] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 26(8): 2323-2326.

[3] 赵连伟,罗四维,赵艳敞.高维数据的低维嵌入及嵌入维数研究[J].软件学报,2005,12(8):1423-1430.

[4] REINHARD K, NIRANJAN M. Subspace models for speech transitions using principal curves[J]. Proceedings of Institute of Acoustics, 1998:53-60

[5] 王靖.流形学习的理论与方法研究[D].杭州:浙江大学,2006.

[6] 孙明明.流形学习理论与算法研究[D].南京:南京理工大学,2007.

[7] 刘小明.数据降维及分类中的流形学习研究[D].杭州:浙江大学,2007.

(收稿日期:2012-11-16)

作者简介:

王博,男,1987 年生,硕士研究生,主要研究方向:图像处理。