

文章编号: 1000-5013(2008)04-0522-05

# 基于离群点检测的鲁棒局部切空间排列方法

王 靖

( 华侨大学 信息科学与工程学院, 福建 泉州 362021)

摘要: 研究局部切空间排列方法(LTSA)对离群点的敏感性,提出一种基于离群点检测的鲁棒局部切空间排列方法(RLTSA).该方法用样本点到切空间的投影距离检测离群点.在构造样本点局部邻域时,RLTSA 尽可能排除离群点,以构造稳定的局部邻域,而对离群点,RLTSA 把它们投影到更高维的切空间,以减少离群点的投影距离.模拟实验和实际例子说明,新方法能提高局部切空间排列方法处理离群样本点的能力.

关键词: 鲁棒; 离群点; 流形学习; 局部切空间排列

中图分类号: TP 181

文献标识码: A

数据降维问题,是从高维空间中找出隐藏的低维结构.传统的数据降维方法,如主分量分析法(PCA),能反映出数据的线性结构,但它并不能学习数据间内在的非线性几何结构.实际应用中,不少高维数据都是嵌入于一个非线性的低维流形.近年来,非线性降维在包括数据挖掘、机器学习、图像分析和计算机视觉等许多领域都得到广泛的关注,并已经发展出一些有效的算法,例如等距映射(Isomap)<sup>[1]</sup>、局部线性嵌入(LLE)<sup>[2]</sup>、多重权的局部线性嵌入(MLLE)<sup>[3]</sup>、拉普拉斯特征映射(LE)<sup>[4]</sup>,以及局部切空间排列(LTSA)<sup>[5]</sup>等.这些算法具有参数少、计算快、易求全局最优解等优点,在图像识别、文本聚类、谱重建、数据可视化等许多方面都得到广泛的应用<sup>[6]</sup>.尽管这些非线性降维方法能有效的学习出数据所在流形的低维结构,但它们对离群点缺乏鲁棒性.Zhang 等<sup>[7]</sup>提出了对离群点光滑化的方法,用加权 PCA 方法对流形进行移除离群点和减小噪音的预处理.这种方法对权的选取较为敏感,而且对流形进行预处理将难以体现离群点真正的低维结构.近来,Chang 等<sup>[8]</sup>提出了鲁棒局部线性嵌入方法(RLLE).RLLE 尽管对离群点体现出较好的鲁棒性,但它用迭代算法检测离群点,需要较高的计算代价,且仍然保留了 LLE 原有的缺点.目前,在研究流形学习的离群点问题的工作中,还没有关于 LTSA 的改进工作.本文提出了一种鲁棒的局部切空间排列方法(Robust Local Tangent Space Alignment, RLTSA),并通过模拟实验和实际例子进行验证.

## 1 局部切空间排列方法

### 1.1 局部切空间排列算法

局部切空间排列(LTSA)算法,是一种基于切空间的流形学习方法.其基本思想是,利用样本点邻域的切空间来表示局部的几何性质,然后将这些局部切空间排列起来构成流形的全局坐标<sup>[5]</sup>.

假设样本点集 $\{x^1, \dots, x^N\}$ ,  $x_i \in \mathbf{R}^m$  采自有噪声的低维流形.即  $x_i = f(\mathbf{t}_i) + \varepsilon$ ,  $i = 1, \dots, N$ , 其中  $f: \Omega \subset \mathbf{R}^d \rightarrow \mathbf{R}^m$ ,  $d < m$ ,  $\Omega$  是开连通集合,  $\varepsilon$  为样本点的噪声. LTSA 首先寻找每个样本点的邻域, 设  $X_i = [x_{i_1}, \dots, x_{i_k}]$  为样本点  $x_i$  包括自身在内的最近的  $k$  个邻域点所构成的矩阵. 然后, LTSA 利用主分量分析法(PCA) 计算一个  $d$  维的仿射子空间, 用于逼近  $X_i$  中的点. 即

$$\min_{x, \Theta, Q} \sum_{j=1}^k \|x_{i_j} - (x + Q\theta_j)\|^2 = \min_{x, \Theta, Q} \|X_i - (x\mathbf{e}^T + Q\Theta)\|_F^2. \quad (1)$$

收稿日期: 2008-05-27

作者简介: 王 靖(1981-), 男, 讲师, 博士, 主要从事数据减维与矩阵计算的研究. E-mail: wroaring@yahoo.com.cn.

基金项目: 福建省青年科技人才创新基金(2007F3067); 华侨大学高层次人才科研启动项目(06BS304)

式(1)中,  $e_k$  为所有分量为 1 的  $k$  维列向量,  $\Theta = [\theta_1, \dots, \theta_k]$ , 且  $Q$  的列数为  $d$ . 记  $\bar{x}_i = X_i e_k$  为邻域矩阵  $X_i$  的中心点,  $Q_i \Sigma_i V_i^T$  为中心化邻域矩阵  $X_i - \bar{x}_i e_k^T = [x_{i_1} - \bar{x}_i, \dots, x_{i_k} - \bar{x}_i]$  的奇异值分解, 即  $Q_i, V_i$  分别为对应于最大的  $d$  个奇异值  $\sigma_1^{(i)}, \dots, \sigma_d^{(i)}$  的左右奇异向量所构成的矩阵. 这样容易求出式(1)的最优解为  $x = \bar{x}_i, Q = Q_i, \Theta = Q_i^T (X_i - \bar{x}_i e_k^T)$ , 从而可以得到局部坐标系为

$$\Theta_i = [\theta_1^{(i)}, \dots, \theta_k^{(i)}] = [Q_i^T (x_{i_1} - \bar{x}_i), \dots, Q_i^T (x_{i_k} - \bar{x}_i)].$$

LTSA 将所有这些有交叠的局部坐标系  $\Theta_i$  排列起来, 得到一个全局坐标系  $T = [\tau_1, \dots, \tau_w]$ . 即认为全局坐标  $\tau_j$  应该能反映由局部坐标  $x_j^{(i)}$  所决定的局部几何结构. 记  $T_i = [\tau_1, \dots, \tau_k]$ . 为了尽可能保持局部的低维特征, LTSA 极小化的重建误差为

$$E(T) = \sum_i \|E_i\|^2 = \sum_i \min_{L_i} \|T_i(I - e_k e_k^T/k) - L_i \Theta_i\|^2. \tag{2}$$

式(2)中,  $L_i$  是一个待定的局部仿射变换矩阵. 为了得到唯一解, LTSA 给全局坐标  $T$  加上中心化和标准化约束. 由于重建误差(2)又可以写成

$$E(T) = \sum_i \|T_i(I - e_k e_k^T/k)(I - \Theta_i^* \Theta_i)\|^2 = \text{trace}(T \Phi T), \tag{3}$$

式(3)中,  $\Phi = \sum_{i=1}^N S_i W_i W_i^T S_i^T$  为排列矩阵,  $S_i \in \mathbf{R}^{N \times k}$  是满足  $X S_i = X_i$  的选择矩阵, 且  $W_i = I - [e_k/\sqrt{k}, V_i] \cdot [e_k/\sqrt{k}, V_i]^T$ . 这样极小化重建误差(2)的最优解, 能通过计算矩阵  $\Phi$  从第 2 到第  $d+1$  小的特征值所对应的特征向量  $u_2, \dots, u_{d+1}$  来获得, 即  $T = [u_2, \dots, u_{d+1}]^T$ .

1.2 LTSA 对离群点的敏感性

例 1 首先生成一个具有  $N = 1\,500$  个样本点的 Swiss-Roll, 数据点生成(用 MATLAB 记号)为  $t = (3 \times \pi/2) \times (1 + 2 \times \text{rand}(1, N))$ ,  $s = 21 \times \text{rand}(1, N)$ ,  $X = [t \times \cos t, s, t \times \sin t]$ . 数据点集和其理想的嵌入结果(即同流形等距的二维生成坐标), 以及 LTSA 的嵌入结果, 如图 1(a)~(c) 所示. 显然, LTSA 能得到理想的嵌入结果.

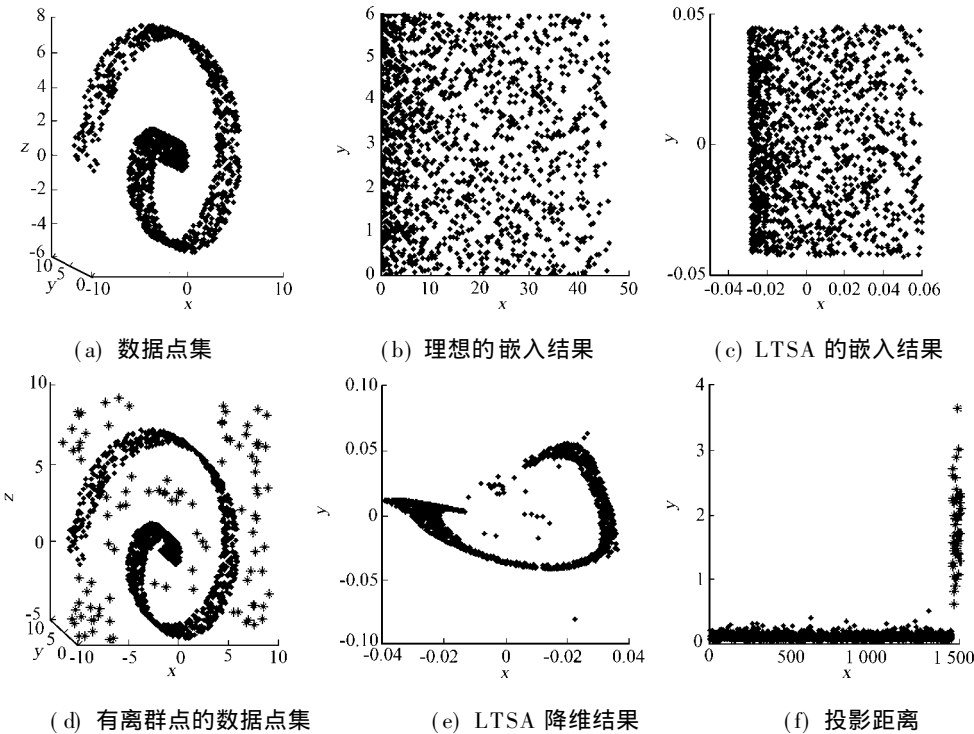


图 1 LTSA 在数据点集的嵌入结果

Fig.1 Embedding results of LTSA on sample data points

其次, 往数据集中加入 50 个一致分布的随机噪声点作为离群点, 这些噪声点被限制与样本点保持一定的距离. 图 1(d) 为有离群点的 Swiss-Roll 数据点集, 其中“\*”点表示离群点. 采用 LTSA(邻域大小  $k = 15$ ) 降维得到的二维嵌入结果, 如图 1(e) 所示. 从图 1(e) 中可看出, LTSA 的降维结果有明显的扭曲变形, 并不能保持流形的局部几何性质.

最后, 对于所有样本点  $x_i(i=1, \cdots, 1550)$ , 采用 PCA 对  $x_i$  和最近的 14 个样本点构成的邻域  $X_i$  进行降维, 可以得到切空间  $Q_i \in \mathbf{R}^{15 \times 2}$  和局部坐标  $\Theta_i \in \mathbf{R}^{2 \times 15}$ , 从而计算出邻域内样本点到切空间的投影距离  $p_i = \|X_i - X_i e_k e_k^T / k - Q_i \Theta_i\|, i=1, \cdots, 1550$ . 样本点  $x_i$  到它们局部切空间上的投影距离  $p_i$ , 如图 1(f) 所示. 从图 1(f) 可知, 所有离群点的投影距离都比较大. 说明离群点的邻域不具有低维的线性结构, 采用 PCA 降维无法得到它们准确的切空间和局部坐标. 值得注意的是, 离群点还会影响非离群点的切空间和局部坐标的计算. 在图 1(f) 中, 有几个非离群点的投影距离比较大. 主要是因为非离群点的局部邻域包含了离群点, 故无法准确地计算出切空间和局部坐标.

LTSA 对离群点的敏感性主要体现在两个方面. (1) 离群点的局部邻域不具有低维的线性结构; (2) 非离群点的邻域可能包含离群点, 因而无法准确地计算出这些非离群点切空间和局部坐标.

## 2 鲁棒局部切空间排列方法

### 2.1 离群点的检测

假设样本点  $x_i \in \mathcal{N}(i=1, \cdots, N)$  采自  $d$  维流形, 由于离群点的局部邻域不具有  $d$  维的线性结构, 因此, 离群点到它的  $d$  维切空间上的投影距离通常要大于非离群点的投影距离. 这样, 可以用样本点到它们局部  $d$  维切空间上的投影距离来检测离群点, 具体有如下 3 个步骤. (1) 寻找每个样本点  $x_i$  的邻域点. 记  $X_i = [x_{i_1}, \cdots, x_{i_k}]$  为样本点  $x_i$  包括自身在内的最近的  $k$  个邻域点所构成的矩阵. (2) 对每个样本点的邻域  $X_i$ , 采用 PCA 降到  $d$  维, 并计算邻域内样本点到  $d$  维切空间的投影距离  $p_i$ . 记  $q^{(i)}_1, \cdots, q^{(i)}_k$  为中心化矩阵  $X_i - \bar{x}_i e_k^T$  的  $k$  个奇异值, 则有  $p_i = \sqrt{\sum_{j=1}^d q^{(i)j2}}$ . (3) 将数据集  $\mathcal{N}$  分成  $\mathcal{N}_0 = \{x_i | p_i \leq a^*, i=1, \cdots, N\}$  和  $\mathcal{N}_1 = \{x_i | p_i > a^*, i=1, \cdots, N\}$ .

在步骤(3)中, 参数  $a^*$  用来区分离群点和非离群点, 理想的值应该是介于离群点和非离群点的投影距离之间. 注意到离群点和非离群点的投影距离有较大的间隔, 因此, 参数  $a^*$  的选取并不敏感. 采用简单的选取方式就可以较准确的区分离群点和非离群点. 一种简单的选取方式为取  $a^*$  为投影距离  $p_i$  的平均值, 即  $a^* = \sum_{i=1}^N p_i / N$ .

经过检测步骤后, 可以把数据集  $\mathcal{N}$  分成点集  $\mathcal{N}_0$  和  $\mathcal{N}_1$ . 通常, 点集  $\mathcal{N}_0$  为非离群点的集合, 而离群点只包含在  $\mathcal{N}_1$  中. 但需要指出的是, 由于一些非离群点的局部邻域里可能包含离群点, 这会使得这些非离群点的投影距离偏大, 从而被划分至点集  $\mathcal{N}_1$ . 也就是说, 点集  $\mathcal{N}_1$  中可能会包含一些非离群点.

### 2.2 算法

RLTSA 从两方面提高 LTSA 对离群点的鲁棒性. 一方面, 对每个样本点  $x_i \in \mathcal{N}$ , 在构造  $x_i$  的邻域时尽可能去除离群点, 从而构造其稳定的局部邻域. 首先, 检测  $x_i$  的  $k$  个最近邻域点组成的邻域  $N_i$  是否包含点集  $\mathcal{N}_1$  中的点. 如果不包含, 则  $N_i$  只包含了  $\mathcal{N}_0$  中的非离群点,  $N_i$  是  $x_i$  的稳定的局部邻域. 如果包含, 考虑到点集  $\mathcal{N}_1$  中除了离群点外, 还包含了少量的非离群点, 需要进一步明确  $x_i$  在  $\mathcal{N}_1$  中的邻域点是否是离群点. 在非离群点集  $\mathcal{N}_0$  中, 寻找  $x_i$  的邻域点并构造邻域  $N_i^0$ . 对于  $d$  维流形,  $x_i$  和其邻域应该逼近于一个  $d$  维切空间, 可以通过比较邻域  $N_i^0$  和  $N_i$  到  $d$  维切空间的投影距离  $p_i^0$  和  $p_i$  来选择稳定的邻域. 若  $p_i \leq p_i^0$ , 则  $N_i$  是  $x_i$  的一个稳定的邻域; 若  $p_i > p_i^0$ , 则  $N_i^0$  为  $x_i$  的一个更为稳定的邻域.

另一方面, 离群点的局部邻域不具有  $d$  维而是具有更高维的线性结构, 计算这些离群点更高维的切空间和局部坐标. 对样本点  $x_i \in \mathcal{N}_1$ , 通过将它们的邻域点投影到  $s_i(s_i \geq d)$  维切空间上, 以减小离群点的投影距离. 维数  $s_i$  的选取要满足 3 个原则. (1) 离群点到  $s_i$  维切空间上的投影距离要小于非离群点集  $\mathcal{N}_0$  中最大的投影距离. (2) 维数  $s_i$  应该不小于流形的维数  $d$ . (3) 维数  $s_i$  应该尽量小, 以避免丢失流形的局部低维特征. 选取  $s_i$  的方案为

$$s_i = \min \left\{ r \geq d, \sqrt{\sum_{j=1}^r q^{(i)j2}} \leq \max \{ p_j, x_j \in \mathcal{N}_0 \} \right\}. \tag{4}$$

式(4)中,  $q^{(i)j}$  为离群点中心化矩阵  $X_i - \bar{x}_i e_k^T$  的奇异值.

在计算出所有样本点的局部坐标  $\Theta_i$  后, 可以按(3)构造排列矩阵  $\Phi$ , 而  $\Phi$  的从第 2 到第  $d+1$  小的

特征值所对应的特征向量所构成的矩阵就是低维嵌入坐标  $T$ .

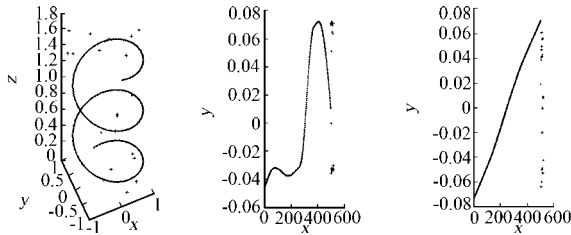
经过以上推导, RL TSA 的算法有如下 5 个步骤(输入为样本集  $\{x_1, \dots, x_N\}$ ,  $x_i \in \mathbf{R}^m$  低维嵌入空间维数  $d$ , 邻域大小  $k$ ; 输出为低维嵌入坐标  $T$ ).

- (1) 用离群点检测方法检测样本集  $N$ , 得到点集  $N_0$  和  $N_1$ .
- (2) 对  $x_i, i=1, \dots, N$ , 在寻找邻域点时尽可能去除离群点, 从而构造稳定的邻域. 记  $X_i = [x_{i_1}, \dots, x_{i_k}]$  为样本点  $x_i$  和其  $k-1$  个邻域点构成的矩阵.
- (3) 对  $x_i \in N_0$ , 采用 PCA 将其局部邻域降到  $d$  维, 从而得到  $d$  维局部坐标系  $\Theta_i$ , 并计算邻域内样本点到  $d$  维切空间的投影距离  $p_i$ .
- (4) 对  $x_i \in N_1$ , 按式(4) 确定局部切空间维数  $s_i$ . 采用 PCA 降维后得到  $s_i$  维局部坐标系  $\Theta_i$ .
- (5) 按式(3) 构造排列矩阵  $\Phi$ , 计算  $\Phi$  对应的特征向量  $u_2, \dots, u_{d+1}$ , 则  $T = [u_2, \dots, u_{d+1}]^T$  为计算的嵌入结果.

2.3 模拟例子

将 RL TSA 和 LT SA 应用到采自圆柱螺线的样本点上, 样本点中增加了 20 个一致分布的随机噪声点作为离群点. 图 2(a) 为圆柱螺线, 其中“\*”点表示离群点. 测试中采用的邻域大小  $k=15$ , LT SA 和 RL TSA 的嵌入结果, 如图 2(b), (c) 所示. 从图 2 中可以看出, LT SA 的嵌入结果不能保持曲线的邻域关系, 也不能保持样本点的弧长关系; RL TSA 的嵌入结果是平滑的直线, 显示了嵌入坐标和样本点坐标间的直接关系.

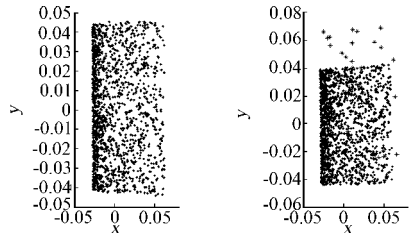
对例 1 中的有离群点和无离群点的 Swiss-Roll 数据集进行实验. 实验中, 采用邻域大小  $k=15$  的 RL TSA 算法, 分别对这两个数据集进行降维, 如图 3 所示. 从图 3 中可以看出, RL TSA 能很明显的增强对离群点的鲁棒性, 它的嵌入结果能很好的恢复出与 Swiss-Roll 数据点等距的二维嵌入结果.



(a) 有离群点的数据集 (b) LTSA (c) RL TSA

图 2 LTSA 和 RL TSA 在圆柱螺线上的嵌入结果

Fig. 2 LTSA and RL TSA applied to the helix data set



(a) 无离群点 (b) 有离群点

图 3 RL TSA 在数据集上的结果

Fig. 3 RL TSA applied to the data sets

2.4 木材纹理图像实验

采用的木材纹理图像来自于 USC-SIPI 图像数据库, 初使图像大小为  $512 \text{ px} \times 512 \text{ px}$ . 在实验中, 首先将图像顺时针旋转  $90^\circ$ , 并将每张图像(初始图像和旋转  $90^\circ$  后的图像) 分割成 784 个相互交叠的图像块, 每个图像块的大小为  $26 \text{ px} \times 26 \text{ px}$ . 这样, 木材纹理数据集共包含有 1 568 个 676 维( $26 \text{ px} \times 26 \text{ px}$ ) 的图像数据. 从这个数据集中随机选出 50 张图像, 并往图像中随机加入 30 个黑白噪点, 这些有噪音的图像数据可以看成是数据集中的离群点. 初始图像和部分噪音图像, 如图 4 所示.

将 LT SA 应用到无噪音的数据集, 再分别用 LT SA 和 RL TSA 对有噪音的数据集进行降维. 在这组实验中, 采用的邻域大小  $k=11$ , 降维后的维数为二维. LT SA 在无噪音数据集上的嵌入结果, 如图 5(a) 所示. 从图 5(a) 中可以看出, 对无噪音的数据集, LT SA 将两个方向的纹理图像很好的分离开. LT SA 和 RL TSA 在有噪音的数据集上的嵌入

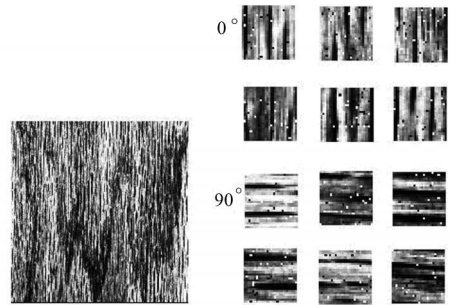


图 4 木材纹理图像

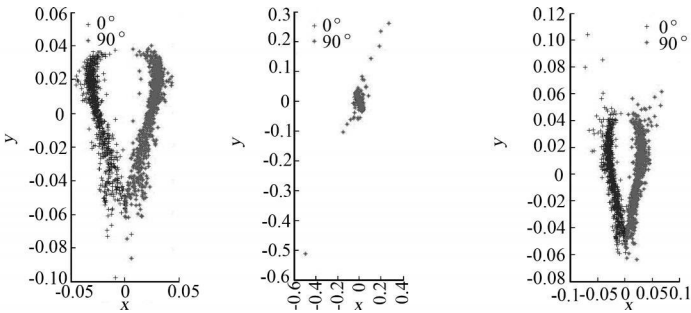
Fig. 4 Wood texture images

结果, 如图 5(b), (c) 所示. 显然, 由于离群点的存在, LT SA 的嵌入结果无法区分出两类纹理图像, 两类纹理图像完全被混在一起; 而 RL TSA 则体现出对离群点的鲁棒性, 对有噪音的数据集, RL TSA 可以得

到和无噪音数据集相类似的结果, 在二维嵌入空间中有噪音数据集仍能够很好的保持两类纹理图像的分离度.

3 结束语

RLTSA 通过计算样本点到局部切空间的投影距离, 来区分非离群点和离群点. 然后, 将离群点的邻域投影到较高维的切空间, 以减小投影距离. 数值实验验证了 RLTSA 的有效性. 但是, 当流形的局部低维性质不明显时, 将难以判定这些数据点是否是离群点, 此时, RLTSA 的效果将被减弱. 这些问题需要得到进一步的研究.



(a) 无噪音的 LTSA      (b) 有噪音的 LTSA      (c) 有噪音的 RLTSA

图 5 木材纹理图像数据集上的嵌入结果

Fig. 5 Embedding results of the data sets of the wood texture images

参考文献:

[ 1 ] TENENBAUM J, SILVA V D, LANGFORD J. A global geometric framework for nonlinear dimension reduction [ J ] . Science, 2000, 290: 2319-2323.

[ 2 ] ROWEIS S, SAUL L. Nonlinear dimensionality reduction by locally linear embedding[ J]. Science, 2000, 290: 2323-2326.

[ 3 ] ZHANG Z, WANG J. MLLE: Modified locally linear embedding using multiple weights[ C] // Advances in Neural Information Processing Systems, 2007.

[ 4 ] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[ M ] // Advances in Neural Information Processing Systems Cambridge: MIT Press, 2001.

[ 5 ] ZHANG Z, ZHA H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment[ J]. SIAM J Scientific Computing, 2005, 26( 1 ) : 313-338.

[ 6 ] SILVA V, TENENBAUM J B. Global versus local methods in nonlinear dimensionality reduction[ C] // Advances in Neural Information Processing Systems Cambridge: MIT Press, 2003: 705-712.

[ 7 ] ZHANG Z, ZHA H. Local linear smoothing for nonlinear manifold learning[ R ] . PA: Pennsylvania Statue University, 2003.

[ 8 ] CHANG H, YEUNG D Y. Robust locally linear embedding[ J] . Pattern Recognition, 2006, 39(6): 1053-1065.

Robust Local Tangent Space Alignment Based on Outlier Detection

WANG Jing

( College of Information Science and Engineering, Huaqiao University, Quanzhou 362021, China)

**Abstract:** The paper focuses on the sensitivity of local tangent space alignment (LTSA) to outliers, and presents a robust local tangent space alignment (RLTSA) based on outlier detection. RLTSA detects the outliers by the projection distances of the sample points onto their tangent spaces. When constructing the neighborhoods of the sample points, RLTSA removes the outliers to construct stably local neighborhood. For the outliers, RLTSA projects them onto higher dimensional tangent spaces to reduce their projection distances. Simulation and real examples show that the new approach can improve the ability of LTSA to deal with outliers.

**Keywords:** robust; outlier; manifold learning; local tangent space alignment

(责任编辑: 黄仲一      英文审校: 吴逢铁)