软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(7):1571−1579 [doi: 10.3724/SP.J.1001.2011.03839]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于重构权的离群点检测方法*

王 靖+

(华侨大学 计算机科学与技术学院,福建 厦门 361021)

Outlier Detection Approach Based on Reconstruction Weights

WANG Jing+

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

+ Corresponding author: E-mail: wroaring@yahoo.com.cn

Wang J. Outlier detection approach based on reconstruction weights. *Journal of Software*, 2011,22(7): 1571−1579. http://www.jos.org.cn/1000-9825/3839.htm

Abstract: In past years, the problem with nonlinear dimensionality reduction has aroused a great deal of interest in many research fields, including pattern analysis, machine learning, and data mining. However, the general manifold learning methods are not robust on the outliers. In the paper, an outlier detection method, based on reconstruction weights, is proposed. The proposed algorithm constructs local 'strong' neighborhoods on each sample point, and computes the reliability score of each sample point using local reconstruction weights, and then detects the outliers using the reliability scores. The advantages of the algorithm are that it has fast computation, low parameter, and low parameter sensitivity. Based on the proposed outlier detection method, the robust Isomap algorithm is proposed in this paper. Experimental results illustrate that the proposed algorithm can detect the outliers efficiently and make the manifold learning methods more robust on the outliers.

Key words: manifold learning; reconstruction weight; outlier; robust

摘 要: 近几年来,流形学习在模式识别、机器学习和数据挖掘等许多领域都受到了广泛的关注.但是,通常的流形学习方法对离群点缺乏鲁棒性.对此,提出了一种基于重构权的流形离群点检测方法.该方法在每个样本点构造局部"强"邻域,再利用局部重构权来计算每个样本点的可靠值,最后利用可靠值检测出离群点.该算法具有计算快、参数少、参数敏感性小等优点.基于此离群点检测方法,提出了鲁棒的 Isomap 算法.实验结果表明,该方法能够有效检测离群点,从而提高流形学习方法对离群点的鲁棒性.

关键词: 流形学习;重构权;离群点;鲁棒

中图法分类号: TP181 **文献标识码**: A

近几年来,流形学习成为模式识别、机器学习和数据挖掘等领域的研究热点之一,目前已经发展出一些有效的流形学习方法,如等距映射(Isomap)[1]、局部线性嵌入(locally linear embedding,简称 LLE)[2]以及局部切空间排列(local tangent space alignment,简称 LTSA)[3]等.这些流形学习方法有着共同的框架:(1) 寻找每个样本点的

* 基金项目: 国家自然科学基金(10901062); 福建省自然科学科学基金(2010J01336); 华侨大学基本科研业务专项基金(JB-SJ1004)

收稿时间: 2010-01-11; 定稿时间: 2010-03-11

局部邻域并构造局部线性特征;(2) 利用构造的局部线性特征将样本点投影到一个低维空间.尽管这些流形学习方法已经在许多领域中得到应用,能够有效地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.这主要是因为无法准确构造离群点的局部线性特征,从而导致低维嵌入结果产生偏差.因此,在对数据降维之前需要先检测数据中的离群点,然后再将离群点光滑化或降低离群点在低维重构时的重要性,从而提高流形学习算法的鲁棒性[4,5].

目前已有一些有效的离群点检测方法,如基于统计分布、基于距离、基于密度和基于偏差等方法[6,7],但大多数并不适用于检测流形中的离群点,这主要是因为这些算法忽视了流形数据的局部线性和整体非线性的特点.近来,Chang 和 Yeung 提出了一种基于鲁棒 PCA 的离群点检测算法(robust PCA,简称 RPCA)[8].RPCA 在每个样本点的局部邻域用迭代加权最小二乘法估计流形的局部切空间,并利用样本点到局部切空间的投影距离计算样本点的可靠值,这些值的大小反映了样本点是离群点的可能性.但它也有一些不足:(1) 算法对邻域参数较为敏感;(2) 流形维数是输入参数,但在许多实际数据中流形本质维数无法得知;(3) 算法采用迭代方法估计流形局部切空间,计算量较大.

对此,本文从 4 个方面考虑设计流形离群点的检测方法:(1) 应该能够有效地检测流形中的离群点;(2) 对邻域参数应该具有较强的鲁棒性;(3) 应该避免将流形维数作为输入参数;(4) 计算量小,能够处理高维数据.基于这些考虑,本文提出一种流形离群点检测方法:对样本点的初始邻域,进一步选取“强”邻域点以降低对邻域参数的敏感性;计算邻域的局部重构权,并利用权值来计算每个样本点的可靠值,可靠值的大小反映了样本点是离群点的可能性.同时,本文还将对邻域的局部重构权进行分析,从理论上说明离群点和权值大小的联系.最后,本文将以 Isomap 为例,通过实验,表明本文方法可以提高流形学习方法对离群点的鲁棒性.

1 构造“强”局部邻域

假设样本点集$\{x_1,\ldots,x_N\}$,$x_i\in R^D$ 采自一个低维流形,对每个样本点 x_i,可以选择离它最近的 k 个点作为邻域点,而这 k 个邻域点构造了 x_i 的 k-近邻.在流形学习算法中,邻域大小是影响嵌入结果的一个重要因素:过大的邻域可能不具有局部线性关系;邻域过小则可能导致流形被分割成多个不连通区域[9].在流形离群点的检测中也面临同样的问题,即如何减小算法对邻域大小的敏感性.对此,本文参考文献[10]的思想,先在每个样本点构造一个 k-近邻,然后从这个 k-近邻中选取“强”邻域点构造“强”局部邻域.记 x_i 的 k-近邻为 $N_i=\{x_{i_j}\mid j=1,\ldots,k\}$,则 x_i 的“强”邻域点为

$$SN(x_i)=\{y\in N_i\mid (x_i-x_{i_j})^T(y-x_{i_j})\geqslant 0, j=1,\ldots,k\} \tag{1}$$

图 1 给出了一个“强”邻域的例子.在这个例子中,当 k 大于 5 后,所寻找的“强”邻域点并不会随着局部邻域大小的变化而发生改变.这也说明了“强”邻域点对邻域大小具有很好的鲁棒性.需要注意的是,“强”邻域点虽然剔除了局部邻域中一些较远的点,但仍可能包含了离群点.因此,接下来需要考虑的一个关键问题是如何判断“强”邻域点中的离群点.

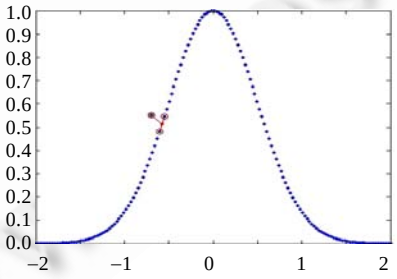


Fig.1 An example for illustrating the ‘strong’

图 1 一个曲线的“强”邻域点例子

2 局部重构权

在 LLE 中,每个样本点 x_i 与其邻域点之间构造了一个重构权向量,并用这个局部重构权来刻画流形的局部几何性质[11].本文将基于 LLE 中的局部重构权设计流形离群点的检测方法,在此之前,首先回顾一下 LLE 中的局部重构权.

对于样本点 x_i 及其局部邻域 $N_i=\{x_{i_j}\mid j=1,...,k\}$,LLE 通过极小化下面的重构误差来计算重构权:

$$\min_{w_{ji}}\left\|x_i-\sum_{j=1}^{k}w_{ji}x_{i_j}\right\|^2,\ \text{s.t.}\sum_{j=1}^{k}w_{ji}=1 \tag{2}$$

记 w_i 为由 w_{ji},j=1,…,k 构成的 k 维列向量,1_k 为所有分量均为 1 的 k 维列向量,$G_i=[x_i-x_{i_1},...,x_i-x_{i_k}]$,则最小化问题(2)可以等价地变换为

$$\min_{w_i}\|G_iw_i\|^2,\ \text{s.t.}\ 1_k^Tw_i=1 \tag{3}$$

采用拉格朗日乘子法,易知最小化问题(3)的最优解满足:

$$G_i^TG_iw_i^*-\lambda\mathbf{1}_k=0,\ \mathbf{1}_k^Tw_i^*=1,$$

其中,λ是拉格朗日乘子.因此,当 G_i 列满秩时,可以通过下列方法来计算最优解:

$$G_i^TG_im_i=\mathbf{1}_k,\ w_i^*=\frac{m_i}{\mathbf{1}_k^Tm_i} \tag{4}$$

当 G_i 不是列满秩或近似秩亏损时,公式(4)中的线性方程组可能无解.此时,通过给矩阵 $G_i^TG_i$ 加上一个小的对角矩阵以保持线性方程组求解的稳定性,即求解

$$(G_i^TG_i+\gamma\|G_i\|^2\ I_k)m_i=\mathbf{1}_k,\ w_i^*=\frac{m_i}{\mathbf{1}_k^Tm_i} \tag{5}$$

其中,γ是一个给定的小参数,I_k 是一个 $k\times k$ 的单位矩阵.

从上面的求解过程可以看出,LLE 的局部重构权求解包含两部分:(1) 求解线性方程组获得 m_i;(2) 对 m_i 进行正则化处理获得最优解 w_i^*.显然,对 m_i 的正则化处理不会改变它分量间的大小关系.但是对不同样本点的 m_i,正则化处理可能会改变权向量之间的大小关系.因此,本文将基于 m_i 而不是 w_i^* 设计流形离群点的检测方法.对于 m_i,本文仍称其为 x_i 的重构权向量.

重构权 m_{ji} 反映了邻域点 x_{i_j} 对样本点 x_i 在重构时的贡献大小.从直观上看,如果 x_{i_j} 是离群点,则它在 x_i 重构时的贡献就很小,这就导致 m_{ji} 很小.接下来,本文用一个简单的例子说明这种现象.在如图 2(a)所示的 5 个点中,星号点 x_3 可以视为离群点.对每个点 x_i,i=1,…,5,其余的 4 个点都可以被当作是其邻域点.

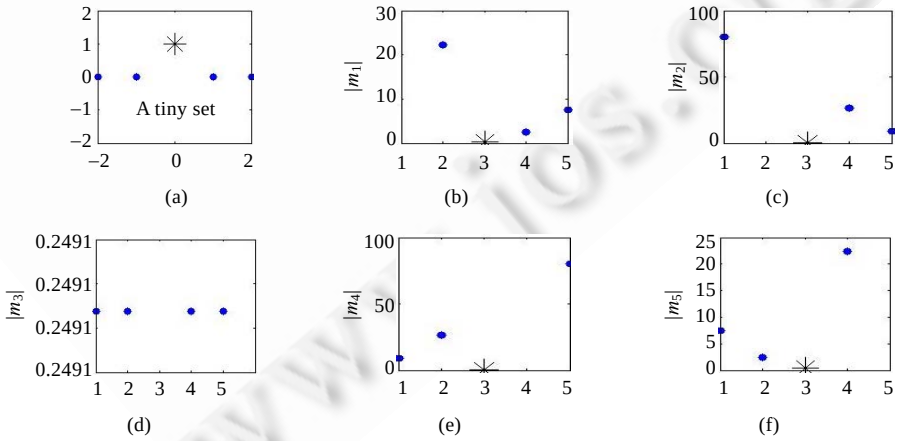


Fig.2 A simulation example for illustrating the neighbors of a curve property of the reconstruction weights

图 2 一个说明重构权性质的模拟例子

利用公式(4)可以计算出每个 x_i 对应的重构权向量 m_i,所有权值构成了如下的矩阵:

$$M=\begin{pmatrix} 0 & 80.3869 & 0.2491 & 8.8393 & -7.5112 \\ 22.3157 & 0 & 0.2491 & 26.7262 & 2.4311 \\ 0.3720 & 0.7026 & 0 & 0.7026 & 0.3720 \\ 2.4311 & 26.7262 & 0.2491 & 0 & 22.3157 \\ -7.5112 & 8.8392 & 0.2941 & 80.3869 & 0 \end{pmatrix}.$$

在图 2(b)~图 2(f)中,画出了每个 x_i 对应的重构权向量的绝对值$|m_i|$.从图和矩阵中可以看出,离群点 x_3 对每个 $x_i,i=1,2,4,5$ 重构时的贡献都很小,因此所有的值$|m_{3i}|$都很小.另一方面,非离群点 $x_i,i=1,2,4,5$ 对离群点 x_3 在重构时的贡献也很小,它们也导致了小的权值$|m_{i3}|$.本文将从理论上对这些现象进行进一步的分析.

首先,对流形中离群点的重构权向量进行分析.假设样本点集$\{x_1,\ldots,x_N\}$,$x_i\in R^D$ 采自 d 维流形 $M=f(\Omega)$,其中,$f:\Omega\subseteq R^d\to R^D$ 是一个定义在开连通集Ω上的光滑映射.对于流形上的某个离群点 $x_i=f(\tau_i)+\varepsilon_i$,假设$\varepsilon_i$ 为离群点的噪声且正交于流形在 $f(\tau_i)$处的切空间.对函数 $f(x)$在 $f(\tau_i)$作一阶泰勒展开,则 x_i 的无噪邻域点 $x_{i_j}=f(\tau_{i_j}),j=1,\ldots,k$ 可以表示成

$$x_{i_j}=f(\tau_i)+J_{\tau_i}\cdot(\tau_{i_j}-\tau_i)+o(\|\tau_{i_j}-\tau_i\|^2)=x_i+J_{\tau_i}\cdot(\tau_{i_j}-\tau_i)-\varepsilon_i+o(\|\tau_{i_j}-\tau_i\|^2),j=1,\ldots,k,$$

其中,$J_{\tau_i}\in R^{D\times d}$ 是函数 $f(x)$相对于点τ_i 的雅克比矩阵而且张成 $f(\tau_i)$处的切空间,ε_i 与 J_{τ_i} 正交.因此,局部矩阵 G_i 可以表示成

$$\begin{aligned} G_i&=[x_{i_1}-x_i,\ldots,x_{i_{k-1}}-x_i,x_{i_k}-x_i]\\ &=[J_{\tau_i}\cdot(\tau_{i_1}-\tau_i),\ldots,J_{\tau_i}\cdot(\tau_{i_{k-1}}-\tau_i),J_{\tau_i}\cdot(\tau_{i_k}-\tau_i)]-\varepsilon_i 1_k^T+E_i\\ &=J_{\tau_i}\cdot T_i-\varepsilon_i 1_k^T+E_i \end{aligned} \tag{6}$$

其中,$T_i=[\tau_{i_1}-\tau_i,\ldots,\tau_{i_k}-\tau_i],E_i=[o(\|\tau_{i_1}-\tau_i\|^2,\ldots,o(\|\tau_{i_k}-\tau_i\|^2)]$.注意,重构权 m_i 由公式(4)或公式(5)计算,而这两个公式也可以写成

$$[G_i^TG_i+\gamma\|G_i\|^2\,I_k\quad -1_k]\begin{pmatrix}m_i\\1\end{pmatrix}=0 \tag{7}$$

其中,γ为 0 或者是一个小数.

显然,$\begin{pmatrix}m_i\\1\end{pmatrix}$属于矩阵$[G_i^TG_i+\gamma\|G_i\|^2\,I_k\quad -1_k]$的零空间,且正交于矩阵$[G_i^TG_i+\gamma\|G_i\|^2\,I_k\quad -1_k]^T$的值域.因此,对任意 $x\in R^k$,有

$$[m_i^T\quad 1]\begin{pmatrix}G_i^TG_i+\gamma\|G_i\|_F^2\\-1_k^T\end{pmatrix}x=0 \tag{8}$$

取向量 x 满足 $T_ix=0,1_k^Tx=1$,并将公式(6)中的 G_i 代入公式(8),则有

$$\begin{aligned} 0&=m_i^T(G_i^TG_i+\gamma\|G_i\|_F^2)x-1_k^Tx\\ &=m_i^TG_i^T(J_{\tau_i}\cdot T_ix-\varepsilon_i1_k^Tx+E_ix)+\gamma\|G_i\|_F^2\,m_i^Tx-1_k^Tx\\ &\approx -m_i^TG_i^T\varepsilon_i+\gamma\|G_i\|_F^2\,m_i^Tx-1\\ &=-m_i^TT_i^T\cdot J_{\tau_i}^T\varepsilon_i+m_i^T1_k\varepsilon_i^T\varepsilon_i-m_i^TE_i\varepsilon_i+\gamma\|G_i\|_F^2\,m_i^Tx-1\\ &\approx(1_k^Tm_i)\|\varepsilon_i\|_F{}^2+\gamma\|G_i\|_F^2\,m_i^Tx-1. \end{aligned}$$

由此可以得出 $1_k^Tm_i\approx\dfrac{1-\gamma\|G_i\|_F^2\,m_i^Tx}{\|\varepsilon_i\|^2}$.由于$\varepsilon_i$ 反映了离群点的误差,$\|\varepsilon_i\|$并不小,而γ为 0 或者是一个小数,因此 $1_k^Tm_i$ $1_k^Tm_i$ 将比较小,即离群点的邻域点重构权之和比较小.

接下来分析无噪数据的局部邻域内,离群点的权值大小.

假设无噪数据 $x_i=f(\tau_i)$和它的局部邻域为 $N_i=\{x_{i_j}\mid j=1,\ldots,k\}$.不失一般性,假设前 $k-1$ 个邻域点为无噪数据

点,第 k 个邻域点为离群点,即 $x_{i_j}=f(\tau_{i_j}), j=1,...,k-1$ 和 $x_{i_k}=f(\tau_{i_k})+\varepsilon_k$. 这里,$\varepsilon_k$ 表示离群点 x_{i_k} 的噪声.对函数 $f(x)$在 $f(\tau_i)$作一阶泰勒展开,则 x_i 的无噪邻域点 $x_{i_j}, j=1,...,k-1$ 可以表示成

$$x_{i_j}=x_i+J_{\tau_i}\cdot(\tau_{i_j}-\tau_i)+o(\|\tau_{i_j}-\tau\|^2).$$

而 x_i 的含噪邻域点 x_{i_k} 可以表示成

$$x_{i_k}=x_i+J_{\tau_i}\cdot(\tau_{i_k}-\tau_i)+\varepsilon_k+o(\|\tau_{i_k}-\tau_i\|^2).$$

这里,$J_{\tau_i}\in R^{D\times d}$ 为函数 $f(x)$相对于点 τ_i 的雅克比矩阵.因此,局部矩阵 G_i 可以表示成

$$G_i=[x_{i_1}-x_i,...,x_{i_{k-1}}-x_i,x_{i_k}-x_i]=J_{\tau_i}\cdot T_i+\varepsilon_k e_k^T+E_i,$$

其中,e_k 表示单位矩阵 I_k 的第 k 列,$T_i=[\tau_{i_1}-\tau_i,...,\tau_{i_k}-\tau_i], E_i=[o(\|\tau_{i_1}-\tau_i\|^2,...,o(\|\tau_{i_k}-\tau_i\|^2)]$.将此 G_i 代入最小化问题(3),则有

$$\begin{aligned}\|G_i w_i^*\|&=\min_{\mathbf{1}_k^T w_i=1}\|G_i w_i\|=\min_{\mathbf{1}_k^T w_i=1}\|(J_{\tau_i}\cdot T_i)w_i+\varepsilon_k e_k^T w_i+E_i w_i\|\\&\leqslant\min_{\mathbf{1}_k^T w_i=1}(\|(J_{\tau_i}\cdot T_i)w_i\|+\|\varepsilon_k e_k^T w_i\|+\|E_i w_i\|)\\&\approx\min_{\mathbf{1}_k^T w_i=1}(\|(J_{\tau_i}\cdot T_i)w_i\|+\|w_{ki}\varepsilon_k\|)\\&=\min_{\mathbf{1}_k^T w_i=1}(\|(J_{\tau_i}\cdot T_i)w_i\|+\|(\|\varepsilon_k\|e_k^T)w_i\|)\end{aligned}\tag{9}$$

当 w_i 逼近正交于矩阵 $\begin{pmatrix}J_{\tau_i}\cdot T_i\\\|\varepsilon_k\|e_k^T\end{pmatrix}$ 时,由于$\|\varepsilon_k\|$并不是一个小数,此时必定有 $e_k^T w_i=w_{ki}$ 趋近于 0.同时,公式(9)的上界趋近于 0,此时 w_i 逼近最优解 w_i^*,显然 $w_{ki}^*\approx w_{ki}$ 趋近于 0.需要注意的是,权向量 $m_i=(\mathbf{1}_k^T m_i)w_i^*$,这也说明了离群点的权值$|m_{ki}|$会远小于无噪邻域点的权值$|m_{ji}|$,$j$=1,…,$k$–1.

3 离群点检测算法

上节的实验和分析表明:样本点邻域中的离群点的权值远小于无噪邻域点的权值.另一方面,离群点的邻域点的权值之和也比较小.基于这些结论,本文将提出一种基于重构权的流形离群点的检测方法.对于样本点 x_i, i=1,…,N和它的局部邻域 $N_i=\{x_{i_j}\mid j=1,...,k\}$,为了提高算法对邻域大小 k 的鲁棒性,首先生成 x_i 的局部“强”邻域 $SN(x_i)=\{y\in N_i\mid(x_i-x_{i_j})^T(y-x_{i_j})\geqslant 0, j=1,...,k\}$.记 J_i 为 x_i 的“强”邻域的下标集,k_i=$|J_i|$为“强”邻域点个数,则 x_i 的重构权向量 m_i 为 k_i 维列向量.将 m_i 嵌入到一个 N 维空间中,即 $M_i(J_i)$=m_i,$M_i(J)$=0,$j\notin J_i$,则所有嵌入后的权向量构成了 $N\times N$ 的矩阵 M.这样,可以计算样本点 x_i 的可靠值:

$$r_i=\sum_{j=1}^{N}|M_{ij}|+\sum_{j=1}^{N}|M_{ji}|\tag{10}$$

公式(10)中,$\sum_{j=1}^{N}|M_{ij}|$表示所有样本点对 x_i 重构的贡献之和,$\sum_{j=1}^{N}|M_{ji}|$表示 x_i 对所有样本点重构的贡献之和.可靠值 r_i 越小,则 x_i 越可能是一个离群点.

综上,本文提出基于重构权的流形离群点的检测方法(outlier detection based on reconstruction weight,简称ODBRW),算法步骤如下:

输入:样本集 X={x_1,…,x_N},$x_i\in R^D$,邻域大小 k,参数α;

输出:离群点集 X_o 和非离群点集 X_c.

Step 1. 寻找每个样本点 x_i 的最近 k 个点构成的邻域 N_i.在局部邻域 N_i 中按照公式(1)构造局部“强”邻域:

$$SN_i=\{y_1,...,y_{k_i}\}.$$

Step 2. 构造每个样本点的局部矩阵 $G_i=[y_1-x_i,...,y_{k_i}-x_i]$,并按照公式(5)计算重构权向量 m_i.

Step 3. 利用公式(10)计算样本点 x_i 的可靠值 r_i.

Step 4. 将数据集 X 分成离群点集 $X_o=\{x_i|r_i\leqslant a,i=1,\ldots,N\}$和非离群点集 $X_c=\{x_i|r_i>a,i=1,\ldots,N\}$.

本算法需要的计算量较小.在算法步骤中,寻找最近邻域点是计算代价最大的步骤,它的计算复杂度为 $O(DN^2)$.构造局部“强”邻域的计算复杂度为 $O(k^3N)$,而计算重构权向量涉及到规模不超过 $k\times k$ 的线性方程组求解,其计算复杂度小于 $O(k^3N)$.对于高维数据而言,邻域大小 $k<<N,k<<D$,因此,构造局部“强”邻域和计算重构权的计算量远小于寻找最近邻域点的计算量.此外,与寻找最近邻域相比,计算可靠值 r_i 需要的计算量可以忽略.综上,本算法所需要的总计算量只是略大于寻找最近邻域点的计算量.需要注意的是,寻找最近邻域点是流形学习算法中首要的步骤,因此,将本算法应用于流形学习算法时并不需要太多额外的计算代价.

最后,基于此离群点检测算法,本文提出一种鲁棒的 Isomap 算法(robust Isomap,简称 Risomap)以提高 Isomap 算法处理离群点的能力.其他鲁棒的流形学习算法也可以类似地构造.具体算法步骤如下:

输入:样本集$\{x_1,\ldots,x_N\}$,$x_i\in R^m$,低维空间维数 d,邻域大小 k;

输出:低维嵌入坐标 $T=\{t_1,\ldots,t_N\}$.

Step 1. 用流形离群点检测方法检测样本集 X,得到离群点集 X_o 和非离群点集 X_c.

Step 2. 对非离群点集 X_c 运用 Isomap 算法,得到 $x_i\in X_c$ 的低维坐标 t_i.

Step 3. 对离群点 $x_i\in X_o$

Step 3.1. 寻找它在 X_o 上最近的 k 个点为邻域点.记 J_i 为 x_i 的邻域点的下标集.

Step 3.2. 令 $G_i=[\ldots,x_i-x_j,\ldots]_{j\in J_i}$, 按照公式(5)计算重构权 w_i.

Step 3.3. 计算 $t_i=\sum_{j\in J_i}w_{ji}t_j$, 则 t_i 就是离群点 x_i 的低维坐标.

4 数值实验

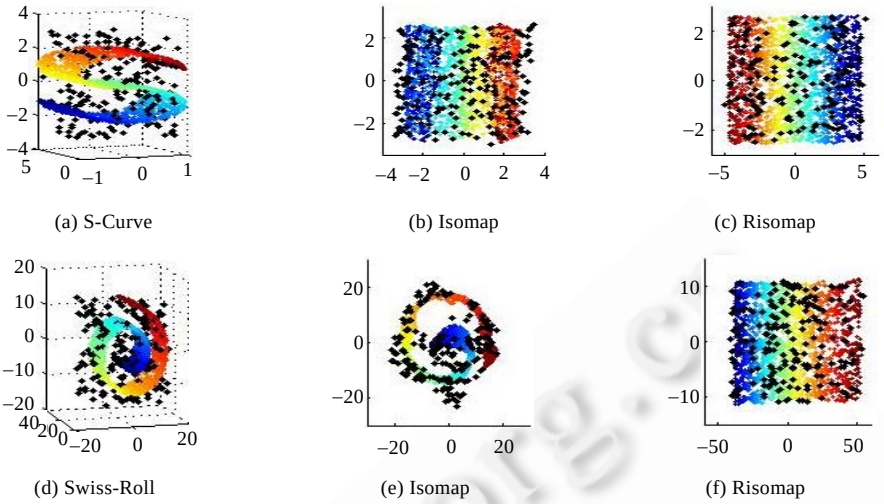
本节将通过模拟例子和实际例子对本文提出的算法进行实验.一方面,将本文提出的离群点算法与其他离群点算法进行比较,如基于鲁棒 PCA 的离群点检测算法和基于局部邻域距离的离群点算法,以说明算法检测离群点的效率.另一方面,将鲁棒的 Isomap 算法和 Isomap 算法、鲁棒的 LLE 算法(robust locally linear embedding,简称 RLLE)[8]进行比较,以说明算法处理离群点的能力.

4.1 模拟实验

本文以含离群点的 S-曲面和 Swiss-roll 曲面为例,说明算法检测和处理离群点的能力.每个数据集都包含 2 000 个无噪数据点和 200 个离群点,离群点和无噪数据点的距离至少为 0.1.图 3(a)和图 3(d)中画出了这两个数据集.在此实验中,将本文的算法(ODBRW)与基于鲁棒 PCA 的离群点检测算法(RPCA)、基于局部邻域距离(NB-distance)[8]的离群点检测算法应用于这两个数据集进行比较实验,并采用 ROC 曲线分析 3 种算法的检测结果.ROC 曲线绘制从左下角开始,检查每个数据点的检测情况.当离群点被正确检测出时,则它是一个真的正例(TP),曲线向上移动并绘制一个点.当无噪数据点被错误检测为离群点时,则它是一个假的正例(FP),曲线向右移动并绘制一个点.显然,ROC 曲线下区域的面积越大,则算法检测离群点越准确.图 4(a)和图 4(b)画出了 3 种算法应用于这两个数据集的 ROC 曲线.这里,3 种算法选取的邻域大小参数均是最优参数,分别为 15,20,3.从图中可以看出,本文的算法和 RPCA 算法均能很好地检测出这两个数据集中的离群点,而基于邻域距离的离群点检测算法则难以有效地检测流形中的离群点.为了比较算法对邻域大小的敏感性,3 种算法分别取 k=5 和 k=10 用于含噪 S-曲面.图 4(c)和图 4(d)分别画出了当邻域大小为 5,10 时,3 种算法应用于含噪 S-曲面的 ROC 曲线.比较图 4(a)、图 4(c)和图 4(d)可以看出,对于不同的邻域大小参数,本文算法都体现了很好的离群点检测效果,而另两种算法则对邻域大小较为敏感.本文算法的另一个优势在于计算效率.当邻域大小为 10 时,3 种算法应用于含噪 S-曲面所需的 CPU 时间分别为 3.65s,6.17s 和 1.97s,本文算法所需的计算时间要远少于 RPCA 算法.

为了体现鲁棒 Isomap 算法处理离群点的能力,本文还将鲁棒的 Isomap 算法和 Isomap 算法应用于这两个模拟数据集.在此实验中,两种算法的邻域大小均为 15.图 3(b)和图 3(e)给出了 Isomap 算法的两维嵌入结果.显然,嵌入结果有着明显的形变,Isomap 算法无法挖掘出这两个数据集本质的低维关系.而从图 3(c)和图 3(f)中可

以看出,鲁棒 Isomap 算法的嵌入结果能够很好地保持数据点之间的测地距离关系.



(a) S-Curve (b) Isomap (c) Risomap

(d) Swiss-Roll (e) Isomap (f) Risomap

Fig.3 Noisy S-curve and Swiss-roll, and the results of Isomap and robust Isomap on these two data sets

图 3 含噪 S-曲面和 Swiss-roll 曲面以及 Isomap 和鲁棒 Isomap 在这两个数据集上的嵌入结果

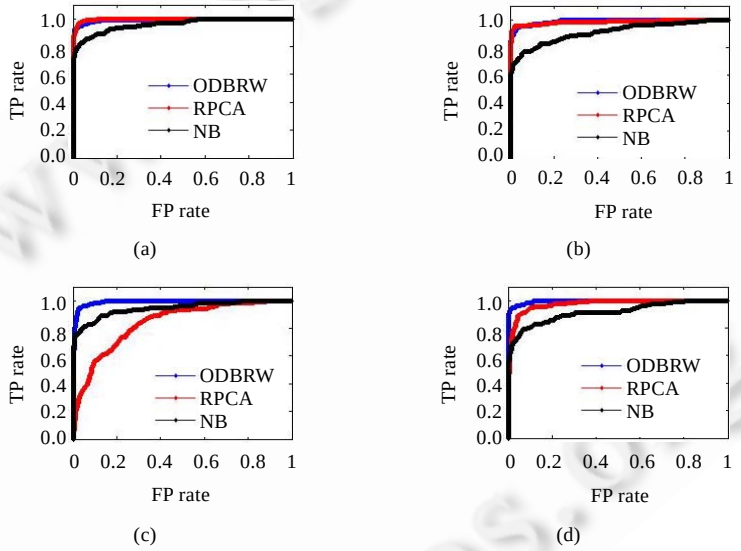


Fig.4 ROC curve of three outlier detection methods on noisy S-curve and noisy Swiss-roll

图 4 3 种离群点检测算法用于 S-曲面和 Swiss-roll 曲面的 ROC 曲线

4.2 人脸图像数据实验

本实验中,将鲁棒的 Isomap 算法和 Isomap 算法、鲁棒的 LLE 算法应用于现实中的人脸数据集[1].这个人脸数据集由 698 幅 64×64 的人脸图像组成,每幅人脸图像由 3 个隐藏的参数所决定:人脸左右方向上的姿态、人脸上下方向上的姿态以及拍摄时的光线亮度.图 5 给出了部分人脸图像.每幅图像可以转化成 4 096 维的高维向量,而且所有图像本质上近似位于一个三维流形.为了比较几种流形学习算法处理离群点的能力,从人脸数据集中随机选出 150 幅图像,并在这些图像中随机添加 30%的黑白噪声.显然,这 150 副图像可以视为离群点.实验中,分别用鲁棒 Isomap 算法、Isomap 算法和 RLLE 算法对含噪人脸数据集进行降维.所有方法采用的邻域大小

均为 10,嵌入空间维数为 3.需要注意的是,降维后得到的嵌入结果 T 和真实的参数 P 之间会相差平移、旋转等变换,因此,本文定义相对重构误差如下:

$$error = \min_{c\in R^d, L\in R^{d\times d}} \frac{\| P-(c1^T+LT)\|_F}{\| P\|_F}.$$

易知,最优解 c 为 P 的均值,$L=(P-c1^T)T^+$.在本实验中,鲁棒 Isomap 算法、Isomap 算法和 RLLE 算法在含噪人脸数据集上的相对重构误差分别为 0.075 8,0.238 8 和 0.223 5.显然,鲁棒 Isomap 算法的嵌入结果的相对重构误差远小于 Isomap 算法和 RLLE 算法.这也表明,鲁棒 Isomap 算法能够很好地恢复出含噪人脸数据的姿态参数和光线亮度参数.

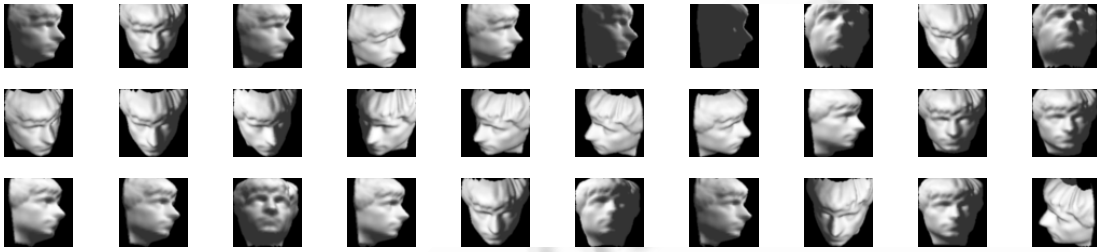
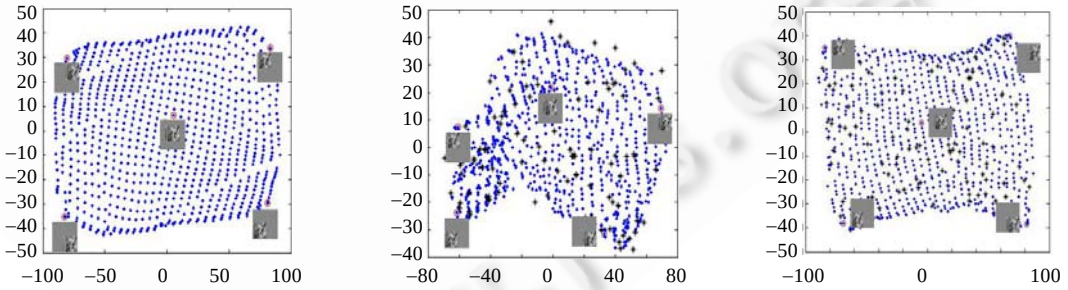


Fig.5 Part face images of the face images data set

图 5 人脸图像数据集中的部分人脸图像

4.3 可视化实验

在这个实验中,将一张 32×32 的灰度 lenna 图像嵌入到一个 61×61 的灰度背景中,每次嵌入可以生成一个高维(D=3721)的向量.本实验在背景中将 30×30 个不同的坐标嵌入 lenna 图像,由此生成 900 个 3 721 维的高维数据.显然,这些高维数据构成了一个二维流形,而流形的生成参数为 lenna 图像的嵌入坐标.本文先将 Isomap 算法应用于此数据集,采用的邻域大小为 12.图 6(a)画出了 Isomap 的二维嵌入结果.对于无噪数据,Isomap 算法的结果很好地体现了流形的本质结构.为了验证本文算法处理离群数据的能力,在数据集中随机选取 100 个高维数据,并在这些高维数据的 lenna 图像中加上 50%的噪声,从而生成 100 个离群点.图 6(b)和图 6(c)画出了 Isomap 和鲁棒 Isomap 应用于此含噪数据集的二维嵌入结果,两种算法的邻域大小均为 12.显然,鲁棒 Isomap 比 Isomap 更好地恢复了流形的生成参数,体现了对离群点的鲁棒性.



(a) Isomap results on clean data set (a) 无噪数据集上 Isomap 的结果

(b) Isomap results on noisy data set (b) 噪声数据集上 Isomap 的结果

(c) Risomap results on noisy data set (c) 噪声数据集上 Risomap 的结果

Fig.6 Embedding results of Isomap and robust Isomap on the clean data set and noisy data set

图 6 无噪数据集和噪声数据集上 Isomap 和鲁棒 Isomap 的嵌入结果

5 结束语

本文提出一种基于重构权的离群点检测方法,并以 Isomap 为例提出鲁棒的 Isomap 方法.实际上,此离群点检测方法可以适用于其他流形学习方法.在离群点检测中,参数 a 用以划分数据集中的离群点和非离群点,参数 a 的选取会影响最终的检测结果.但在实验中,离群点和非离群点的可靠值通常有较大的间隔,因此,对参数 a 的选取并不敏感.此外,也可以采用分类算法对可靠值进行分类以划分离群点和非离群点集.如何选取合适的分类算法以减少参数 a,这个问题值得进一步地研究和展开工作.

References:

[1] Tenenbaum J, De Silva V, Langford JC. A global geometric framework for nonlinear dimension reduction. Science, 2000,290(5500):2319−2323. [doi: 10.1126/science.290.5500.2319]

[2] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290(5500):2323−2326. [doi: 10.1126/science.290.5500.2323]

[3] Zhang ZY, Zha HY. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM Journal of Scientific Computing, 2005,26(1):313−338.

[4] Chen HF, Jiang GF, Yoshihira K. Robust nonlinear dimensionality reduction for manifold learning. In: Proc. of the 18th Int'l Conf. on Pattern Recognition, Vol.2. 2006. 447−450. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1699240 [doi: 10.1109/ICPR.2006.1011]

[5] Park J, Zhang ZY, Zha HY, Kasturi R. Local smoothing for manifold learning. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Vol.2. 2004. 452−459. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1315199 [doi: 10.1109/CVPR.2004.1315199]

[6] Kriegel HP, Schubert M, Zimek A. Angle-Based outlier detection in high-dimensional data. In: Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. 2008. 444−452. http://portal.acm.org/citation.cfm?id=1401946

[7] Maimon O, Rokach L. Data Mining and Knowledge Discovery Handbook. Kluwer Academic Publishers, 2005.

[8] Chang H, Yeung DY. Robust locally linear embedding. Pattern Recognition, 2006,39(6):1053−1065. [doi: 10.1016/j.patcog.2005.07.011]

[9] Wang J, Zhang ZY, Zha HY. Adaptive manifold learning. In: Advances in Neural Information Processing Systems 17. Cambridge: MIT Press, 2005. 1473−1480. http://books.nips.cc/nips17.html

[10] Lin T, Zha HB. Riemannian manifold learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2008,30(5):796−809. [doi: 10.1109/TPAMI.2007.70735]

[11] Saul LK, Roweis ST. Think globally, fit locally: Unsupervised learning of nonlinear manifolds. Journal of Machine Learning Research, 2003,4:119−155.

王靖(1981－),男,福建泉州人,博士,副教授,主要研究领域为流形学习,矩阵计算.