# 基于最优密度方向的等距映射降维算法

梁少军[1], 张世荣[2†], 孙澜琼[1]

(1. 陆军工程大学 军械士官学校, 湖北 武汉 430075; 2. 武汉大学 电气与自动化学院, 湖北 武汉 430072)

**摘要**: 等距映射算法(ISOMAP)是一种典型的非线性流形降维算法, 该算法可在尽量保持高维数据测地距离与低维数据空间距离对等关系的基础上实现降维. 但ISOMAP容易受噪声的影响, 导致数据降维后不能保持高维拓扑结构. 针对这一问题, 提出了一种基于最优密度方向的等距映射(ODD–ISOMAP)算法. 该算法通过筛选数据的自然邻居确定每个数据沿流形方向的最优密度方向, 之后基于与各近邻数据组成的向量相对最优密度方向投影的角度、方向和长度合理缩放局部邻域距离, 引导数据沿流形方向计算测地距离, 从而降低算法对噪声的敏感度. 为验证算法有效性, 选取了2类人工合成数据和5类实测数据作为测试数据集, 分别使用ISOMAP, LLE, HLLE, LTSA, LEIGS, PCA和ODD–ISOMAP算法对数据集降维, 并对降维数据进行K-mediods聚类分析. 通过比对聚类正确率以及不同幅度噪声对此正确率的影响程度评价各算法降维效果优劣. 结果表明, ODD–ISOMAP算法较其他6种常见算法降维效果提升显著, 且对噪声干扰有更强的抵抗能力.

**关键词**: 等距映射; 流形学习; 自然邻居; 最优密度

**引用格式**: 梁少军, 张世荣, 孙澜琼. 基于最优密度方向的等距映射降维算法. 控制理论与应用, 2021, 38(4): 467 – 478

DOI: 10.7641/CTA.2020.00454

# Optimal density direction based isometric mapping dimensionality reduction algorithm

LIANG Shao-jun[1], ZHANG Shi-rong[2†], SUN Lan-qiong[1]

(1. School of Ordnance Sergeant, Army Engineering University, Wuhan Hubei 430075, China;
2. School of Electrical Engineering and Automation, Wuhan University, Wuhan Hubei 430072, China)

**Abstract:** As a typical nonlinear dimensionality reduction algorithm, ISOMAP can realize the dimensionality reduction based on the corresponding relationship between the high-dimensional geodesic distance and the low-dimensional spatial distance. However, the classical ISOMAP algorithm is heavily affected by noise making it difficult to maintain the topology in low dimensionality. To address this problem, an optimal density direction based ISOMAP algorithm is proposed. Firstly, the algorithm screens the optimal density direction of each data along the manifold direction by filtering its natural neighbors, then composes a vector with the data and its neighbors. After that, the local neighborhood distance is reasonably scaled according to the angle, direction and length of the projection of the vector relative to the optimal density direction. In this way, the geodesic distance is guided to be calculated along the manifold direction so as to reduce the sensitivity to noise. In order to verify the effectiveness of the algorithm, 2 types of artificial data sets and 5 types of measured data sets are selected as the test cases. ISOMAP, LLE, HLLE, LTSA, LEIGS, PCA, and ODD–ISOMAP algorithms are applied on data sets respectively. Moreover, K-mediods clustering algorithm is performed after dimension reduction. The dimensionality reduction effect of each algorithm is evaluated by comparing the clustering accuracy and the influence degree of different amplitude noise on the accuracy. Experimental results show that the effect of ODD–ISOMAP algorithm has been significantly improved comparing with the other 6 common algorithms, and it has stronger resistance to noises as well.

**Key words:** ISOMAP; manifold learning; natural neighbor; optimal density

**Citation:** LIANG Shaojun, ZHANG Shirong, SUN Lanqiong. Optimal density direction based isometric mapping dimensionality reduction algorithm. *Control Theory & Applications*, 2021, 38(4): 467 – 478

收稿日期: 2020−07−14; 录用日期: 2020−12−23.

†通信作者. E-mail: srzhang@whu.edu.cn; Tel.: +86 18971289275.

本文责任编委: 王卓.

国家自然科学基金项目(51475337), 陆军军内科研项目(LJ20182B050054, LJ20191C040483, LJ20202C020416, LJ20202C050412)资助.

Supported by the National Natural Science Foundation of China (51475337) and the Army Research Project (LJ20182B050054, LJ20191C040483, LJ20202C020416, LJ20202C050412).

## 1 引言

随着信息技术的飞速发展, 数据产生与获取方式的渠道增多, 数据样式呈现多样化, 如音视频数据、图像数据、文本数据等. 众多研究领域的数据维度也越来越高, 如人脸识别[1]、信息检索[2]、气候变化、恒星光谱、基因分布[3]等. 高维数据会导致数据计算的高度复杂化, 给后期数据处理工作带来巨大挑战, 这通常被称为“维数灾难”[4]. 高维数据包含大量的冗余信息, 体现为数据的稀疏性, 数据降维是应对高维稀疏数据的有效手段. 常用的线性降维方法有主成分分析(principal components analysis, PCA)算法[5]、线性判别(linear discriminant analysis, LDA)算法[6]、多维尺度变换(multi-dimensional scaling, MDS)算法[7]等. 在实际应用中大多数数据各维度间呈现非线性相关特征, 因此出现了一些非线性降维方法, 如拉普拉斯映射(Laplacian eigenmaps, LEIGS)算法[8]、局部线性嵌入 (locally linear embedding, LLE) 算法[9]、局部切空间排列(local target space alignment, LTSA)算法、等距映射 (isometric mapping, ISOMAP) 算法[10] 等. 其中, ISOMAP算法是一种使用广泛的典型非线性降维算法, 该算法基于局部邻域距离计算各数据点之间的全局测地距离, 在尽量保持高维数据测地距离与低维数据空间距离对等关系的基础上实现降维. 但ISOMAP算法本身缺乏噪声应对能力, 噪声会破坏算法的拓扑稳定性影响降维效果[11]; 局部邻域距离的计算没有方向性, 当高维流形存在较大曲率时将导致“短路现象”[12], 降维后数据不能很好保持高维拓扑结构.

为解决以上问题, 提出了一些改进方法, 例如, 监督的 ISOMAP (supervised isometric mapping, S–ISOMAP)算法[13]、多流形判别 ISOMAP (multi-manifold dimensional isometric mapping, MMD–ISOMAP) 算法[14]等. 但此类方法无法适用于无标签数据集. 基于密度缩放因子的ISOMAP(density scaling factor based isometric mapping, D–ISOMAP)算法[15]虽然降低了对噪声的敏感性, 但仍无法应对较大曲率引起的“短路现象”. 本文针对现有算法存在的上述问题, 提出了一种基于最优密度方向的流形降维方法(optimal density direction based ISOMAP, ODD–ISOMAP), 该算法用向量表征高维数据与近邻关系, 通过寻找高维数据最优密度方向构建距离缩放矩阵, 旨在增强算法应对强的噪声抗干扰的能力并避免高维流形较大曲率导致的“短路现象”. 文章将对比ODD–ISOMAP算法与ISOMAP, HLLE, LTSA, LEIGS, LLE和PCA算法的降维性能, 验证了算法的有效性.

## 2 经典ISOMAP算法

设原始高维数据为 $\boldsymbol{X}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_m]^{\mathrm{T}}\in\mathbb{R}^{m\times n}$, 降维后的数据矩阵为$\boldsymbol{Y}=[\boldsymbol{y}_1\ \boldsymbol{y}_2\ \cdots\ \boldsymbol{y}_m]^{\mathrm{T}}\in\mathbb{R}^{m\times d}$; 其中: $m$为采样数, $n$为原始变量维度, $d$为降维后变量维度. 经典的ISOMAP算法包括以下几个核心环节: 1) 构建无向邻域距离矩阵; 2) 获取测地距离矩阵; 3) 获取低维嵌入表示矩阵.

无向邻域距离的计算方式有两种: 一种基于距离, 将$\boldsymbol{x}_i$与其邻域半径$r$范围内的数据间距离作为邻域距离; 另一种基于数量, 将$\boldsymbol{x}_i$与其$k$个最近邻集合$k(\boldsymbol{x}_i)$数据间的距离作为邻域距离. 本文选择后者构建邻域距离矩阵, 将$\boldsymbol{x}_i$与$\boldsymbol{X}$中数据间的无向邻域距离定义为

$$d(\boldsymbol{x}_{i,j})=\begin{cases}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2, & \boldsymbol{x}_j\in k(\boldsymbol{x}_i),\\ 0, & \boldsymbol{x}_j=\boldsymbol{x}_i,\\ \infty, & \boldsymbol{x}_j\notin k(\boldsymbol{x}_i),\end{cases}\tag{1}$$

则$\boldsymbol{X}$数据间的无向邻域距离矩阵$\boldsymbol{D}$可表示为

$$\boldsymbol{D}=\begin{bmatrix}d(\boldsymbol{x}_{1,1}) & d(\boldsymbol{x}_{1,2}) & \cdots & d(\boldsymbol{x}_{1,m})\\ d(\boldsymbol{x}_{2,1}) & d(\boldsymbol{x}_{2,2}) & \cdots & d(\boldsymbol{x}_{2,m})\\ \vdots & \vdots & & \vdots\\ d(\boldsymbol{x}_{m,1}) & d(\boldsymbol{x}_{m,2}) & \cdots & d(\boldsymbol{x}_{m,m})\end{bmatrix}_{m\times m}.\tag{2}$$

再使用Dijkstra或Floyd最短路径算法即可得到任意两个数据间的最短路径距离, 即测地距离$d_G(\boldsymbol{x}_{i,j})$.

获取低维嵌入表示可通过求解以下最优化问题来实现:

$$\min_{\boldsymbol{Y}}=\sum_{i,j}\left(d(\boldsymbol{x}_{i,j})-d_G(\boldsymbol{x}_{i,j})\right)^2.\tag{3}$$

使用MDS算法求解该问题, 即得到高维数据$\boldsymbol{X}$的低维嵌入表示矩阵$\boldsymbol{Y}$.

## 3 ODD–ISOMAP算法

ODD–ISOMAP算法的基本思想是首先确定每个高维数据$\boldsymbol{x}_i$的最优密度方向, 然后获取$\boldsymbol{x}_i$的$k$个近邻在最优密度方向上的投影. 再以投影角度、方向和相对大小修订$\boldsymbol{x}_i$与$k$个近邻的局部邻域距离, 即得到有向邻域距离, 从而引导数据沿着流形方向计算测地距离, 降低短路风险, 增强算法抵制噪声干扰的能力.

### 3.1 基于自然邻居的最优密度方向

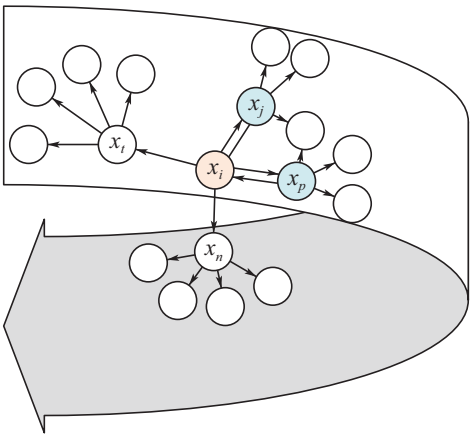
高维流形上的任一数据$\boldsymbol{x}_i$都有一个指向局部最高数据密度且处于流形表面的方向, 将该方向定义为最优密度方向, 并记为$\bar{\boldsymbol{x}}_i$. 为了确保最优密度方向沿流形方向, 同时具有较强的抗噪声干扰能力, 本文引入自然邻居(natural neighbor, NN)概念.

先按照经典ISOMAP算法计算高维数据$\boldsymbol{X}$中任一采样数据$\boldsymbol{x}_i$的$k$近邻集合$k(\boldsymbol{x}_i)$. 在$k(\boldsymbol{x}_i)$基础上$\boldsymbol{x}_i$的自然邻居集合定义为

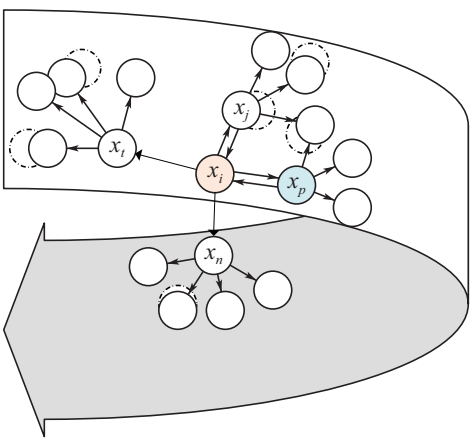
$$\psi(\boldsymbol{x}_i)=\{\boldsymbol{x}_j|\boldsymbol{x}_i\in k(\boldsymbol{x}_j),\boldsymbol{x}_j\in k(\boldsymbol{x}_i)\},\tag{4}$$

上式中$\psi(\boldsymbol{x}_i)$表示$\boldsymbol{x}_i$的自然邻居集合. 由该式可知, 数据$\boldsymbol{x}_j$属于$\psi(\boldsymbol{x}_i)$表示$\boldsymbol{x}_i$与$\boldsymbol{x}_j$都在对方的$k$近邻集合中.

图1展示了三维流形中$\boldsymbol{x}_i$自然邻居选取过程及自然邻居的抗干扰能力, 此例中近邻数$k$ =4. $k(\boldsymbol{x}_i)$包含4个数据$\boldsymbol{x}_j$, $\boldsymbol{x}_p$, $\boldsymbol{x}_t$与$\boldsymbol{x}_n$, $\boldsymbol{x}_i$与$\boldsymbol{x}_n$由于流形紧密折叠出现了短路现象. 如图1(a)所示, 在噪声干扰前$\boldsymbol{x}_j$与$\boldsymbol{x}_p$被选为$\boldsymbol{x}_i$的自然邻居, $\boldsymbol{x}_n$和$\boldsymbol{x}_t$虽然属于$\boldsymbol{x}_i$的最近邻但$\boldsymbol{x}_t$局部范围内有更相近的数据, $\boldsymbol{x}_n$在流形的另一端其近邻数据大概率也不包含$\boldsymbol{x}_i$, 故$\boldsymbol{x}_n$和$\boldsymbol{x}_t$均未入选$\boldsymbol{x}_i$的自然邻居. 可见自然邻居的选取标准比较严格, 这一特征能确保$\boldsymbol{x}_i$筛选出的自然邻居沿局部最大密度方向, 由于数据$\boldsymbol{x}_i$的局部最大密度方向一般沿着局部流形的方向分布, 故可以基于自然邻居得到的最优密度方向也会沿着局部流形方向. 另一方面, 如图1(b)所示, 当部分数据受到噪声干扰发生位移时, 自然邻居的筛选结果不容易改变. 这是由于自然邻居选择过程需要检查数据双方相互的近邻关系, 这一近邻关系是基于近邻位置顺序而非直接距离的, 因此具有一定稳定性. 所以基于自然邻居得到的最优密度方向也将是稳定的.




(a) 受噪声干扰前




(b) 噪声干扰后

图 1　自然邻居的抗干扰能力

Fig. 1　Resistance of noise of natural neighbors

将$\vec{x}_i$视为$\boldsymbol{x}_i$在高维空间对应的向量, 同理$\vec{x}_j$为$\boldsymbol{x}_j$的对应向量, $k(\vec{x}_i)$为$k(\boldsymbol{x}_i)$中所有数据在高维空间对应向量的集合, $\psi(\vec{x}_i)$为$\psi(\boldsymbol{x}_i)$中所有数据在高维空间对应向量的集合. 则所有属于自然邻居集合$\psi(\vec{x}_i)$的向量$\vec{x}_j$与$\vec{x}_i$的差向量$\vec{x}_{i,j}^N$可表示为

$$\vec{x}_{i,j}^N = \vec{x}_j - \vec{x}_i,\ \text{s.t.}\ \vec{x}_j \in \psi(\vec{x}_i). \tag{5}$$

在示意图2中, 差向量如红色箭头所示.

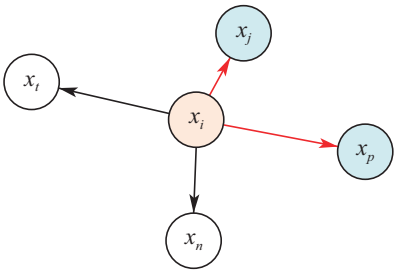



图 2　差向量

Fig. 2　Difference vector

获取差向量后, 可进一步获得$\boldsymbol{x}_i$对应的最优密度方向向量$\hat{x}_i$

$$\hat{x}_i = \sum \frac{\vec{x}_{i,j}^N}{|\vec{x}_{i,j}^N|^2} / \left|\sum \frac{\vec{x}_{i,j}^N}{|\vec{x}_{i,j}^N|^2}\right|. \tag{6}$$

差向量的模标识数据间距离的远近. 式(6)中, 将差向量的模取平方是为了确保$\boldsymbol{x}_i$的自然邻居集合中距离其越近的数据对最优密度方向的影响越大. 在图3所示示例中, 红色箭头表示了$\boldsymbol{x}_i$未归一化的最优密度方向, 结合前文分析可知$\boldsymbol{x}_i$的最优密度方向将指向$\boldsymbol{x}_i$的局部流形方向并且具有稳定性, 下文将利用最优密度方向对$\boldsymbol{x}_i$与其最近邻的位置关系进行缩放, 并设计缩放因子的大小与噪声对数据的影响程度相关, 进而降低噪声影响, 还原数据与其近邻间的原始分布情况.

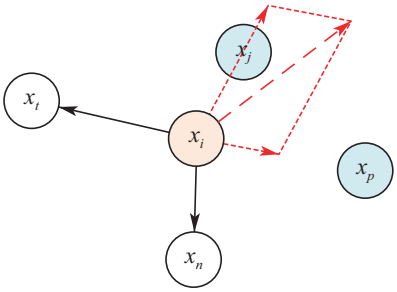



图 3　未归一化的最优密度方向

Fig. 3　Unnormalized optimal density direction

### 3.2　基于最优密度方向的距离缩放矩阵

设$\vec{x}_{i,j}^k$为$k(\vec{x}_i)$中向量$\vec{x}_j$与$\vec{x}_i$的差向量, 则差向量$\vec{x}_{i,j}^k$与最优密度方向向量$\hat{x}_i$夹角的余弦$\cos\gamma_j$可表达为

$$\cos\gamma_j = \frac{\langle \vec{x}_{i,j}^k, \hat{x}_i \rangle}{|\vec{x}_{i,j}^k|}. \tag{7}$$

进而可获得差向量$\vec{x}_{i,j}^{k}$在最优密度方向向量$\hat{x}_i$上的相对投影长度$\tilde{x}_{i,j}$

$$\tilde{x}_{i,j} = |\vec{x}_{i,j}^{k}|(\varepsilon + \mathrm{abs}(\sin\gamma_j)) = |\vec{x}_{i,j}^{k}|(\varepsilon + \sqrt{1-\cos^2\gamma_j}) = \varepsilon|\vec{x}_{i,j}^{k}| + \sqrt{|\vec{x}_{i,j}^{k}|^2 - \langle\vec{x}_{i,j}^{k}, \hat{x}_i\rangle^2}, \tag{8}$$

式(8)中: $\mathrm{abs}(\cdot)$表示取绝对值, $|\cdot|$表示取向量模, $\varepsilon$为系数调节因子.

$k(\vec{x}_i)$中数据在最优密度方向$\hat{x}_i$上的相对投影长度和夹角求取过程如图4所示.

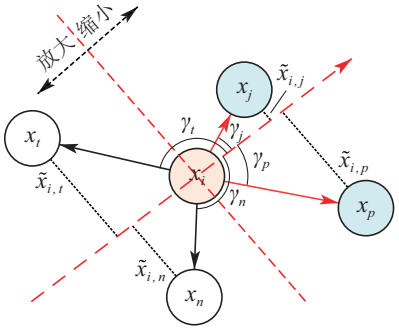



图 4 $K$近邻在最优密度方向上映射

Fig. 4 Projection of $K$-nearest neighbors on optimal density direction

$\cos\gamma_j$的符号代表了$k(\vec{x}_i)$中向量与$\hat{x}_i$的方向关系, 这里进一步定义符号函数$sg$以区分同向与反向数据

$$sg = -\mathrm{sgn}(\cos\gamma_j), \tag{9}$$

其中$\mathrm{sgn}(\cdot)$表示符号函数运算.

则$\boldsymbol{x}_i$相对于$\boldsymbol{x}_j$的密度缩放因子$\hat{x}_{i,j}$可按照下式计算:

$$\hat{x}_{i,j} = \begin{cases} \exp(sg(\tilde{x}_{i,j}\rho^{sg})^{sg}) + \varepsilon, & \cos\gamma_j \neq 0, \\ \exp(\tilde{x}_{i,j}\rho), & \cos\gamma_j = 0, \end{cases} \tag{10}$$

其中: $\exp(\cdot)$表示指数运算, 引入密度缩放因子$\rho$是为了方便控制沿最优密度方向的数据间距离的缩放程度, $\rho$越大缩放程度越剧烈.

遍历$\boldsymbol{X}$中所有数据后, 将得到所有数据相对于其$k$个最近邻的密度缩放因子矩阵$\boldsymbol{DM}_{m\times n}$. $\boldsymbol{X}$中某一数据$\boldsymbol{x}_i$相对于其$k$最近邻$\boldsymbol{x}_j$的密度缩放因子与$\boldsymbol{x}_j$相对于$\boldsymbol{x}_i$的密度缩放因子并不相同, 故$\boldsymbol{DM}$为非对称矩阵. $\boldsymbol{DM}$需要按下式进行对称处理:

$$\boldsymbol{DM}' = \boldsymbol{DM}\cdot\boldsymbol{DM}^{\mathrm{T}}. \tag{11}$$

使用$\boldsymbol{DM}'$对高维数据$\boldsymbol{X}$数据间的无向邻域距离矩阵$\boldsymbol{D}$进行缩放, 可得有向邻域距离矩阵$\boldsymbol{OM}$

$$\boldsymbol{OM} = \boldsymbol{D}\cdot\boldsymbol{DM}'. \tag{12}$$

使用密度缩放因子对高维数据的局部邻域距离进行缩放, 会使得与最优密度方向同向的近邻点之间的距离缩小, 且投影向量与最优密度方向夹角余弦绝对值越小或与当前高维数据点距离越近收缩程度越大; 相反的, 与最优密度方向反向的近邻点之间的距离会被放大, 且投影向量与最优密度方向夹角余弦绝对值越大或与当前高维数据点距离越远则放大程度越大. 图5示例中展示了数据缩放后的距离与图1所示原始距离的对比.

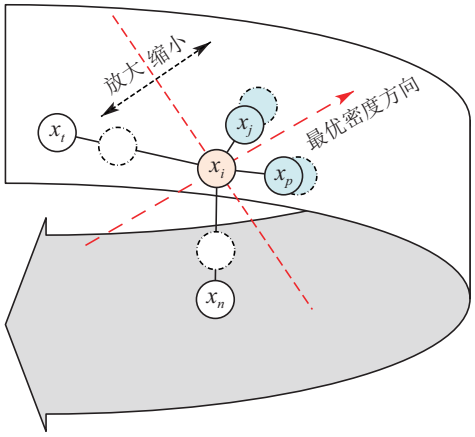



图 5 缩放后距离

Fig. 5 Scaled distance

### 3.3 ODD–ISOMAP算法步骤

**步骤 1** 输入待处理原始高维数据矩阵$\boldsymbol{X}$, 设定最近邻集合的个数$k$, 设定密度缩放因子$\rho$及系数调节因子$\varepsilon$ (一般取$10^{-4} < \varepsilon \leqslant 10^{-1}$), 设定密度缩放因子矩阵$\boldsymbol{DM}_{m\times n}$为全1矩阵. 从$\boldsymbol{X}$中任选一高维数据$\boldsymbol{x}_i$作为待处理数据.

**步骤 2** 按照式 (4) 计算 $\boldsymbol{x}_i$ 的自然邻居集合$\psi(\boldsymbol{x}_i)$. 若$\psi(\boldsymbol{x}_i)$为空集, 则$\boldsymbol{x}_i$为离群点, 若实际情况允许可将$\boldsymbol{x}_i$剔除, 否则可将$k(\boldsymbol{x}_i)$视为$\psi(\boldsymbol{x}_i)$.

**步骤 3** 基于$\psi(\boldsymbol{x}_i)$按照式(5)–(6)获取$\boldsymbol{x}_i$最优密度方向向量$\hat{x}_i$.

**步骤 4** 从$k(\vec{x}_i)$中任选一向量$\vec{x}_j$, 计算$\vec{x}_j$与$\vec{x}_i$的差向量$\vec{x}_{i,j}^{k}$, 按照式(7)–(8)分别计算差向量$\vec{x}_{i,j}^{k}$与最优密度方向向量$\hat{x}_i$夹角的余弦$\cos\gamma_j$及在$\hat{x}_i$上的相对投影$\tilde{x}_{i,j}$. 重复此步骤, 直到获取$k(\vec{x}_i)$中所有向量与$\hat{x}_i$夹角的余弦及所有向量在$\hat{x}_i$上的相对投影集合.

**步骤 5** 按照式(9)–(10)计算$\boldsymbol{x}_i$相对于$k(\boldsymbol{x}_i)$中每一个近邻数据的密度缩放因子.

**步骤 6** 将$\boldsymbol{x}_i$相对于$k(\boldsymbol{x}_i)$的密度缩放因子存入矩阵$\boldsymbol{DM}$对应位置. 从$\boldsymbol{X}$中任选一没有处理过的高维数据, 重复步骤2–5, 直到$\boldsymbol{X}$中所有数据被处理完毕.

**步骤 7** 按照式(11)对$\boldsymbol{DM}$对称处理, 在获取有向邻域距离矩阵$\boldsymbol{D}$基础上按照式(12)计算距离缩放后的有向邻域距离矩阵$\boldsymbol{OD}$.

**步骤 8**　基于$\boldsymbol{OD}$使用Dijkstra或Floyd最短路径算法获取测地距离矩阵, 最后使用多维尺度变换算法MDS得到原始高维数据矩阵$\boldsymbol{X}$降维后数据矩阵$\boldsymbol{Y}$.

需要说明的是, 文中最近邻及自然邻居的求取仍然使用欧式距离. 而随着数据集维度的增加该距离计算方式将呈现出一定的弊端.

### 3.4　ODD–ISOMAP算法复杂度分析

ODD–ISOMAP算法步骤2中需要计算原空间数据的自然邻居, 应用了KD树优化后的自然邻居选择时间复杂度为$\mathrm{O}(m\log m)$[16]. 步骤3计算所有数据的最优密度方向的复杂度为$\mathrm{O}(m)$. 步骤4与步骤5中计算空间所有数据与其$k$个最近邻的差向量、相对最优密度方向夹角余弦和投影以及计算所有数据相对其$k$个近邻的密度缩放因子的复杂度均为$\mathrm{O}(km)$, 故此步骤总体时间复杂度为$\mathrm{O}(4km)$. 步骤6中获取$\boldsymbol{DM}$矩阵的复杂度为$\mathrm{O}(1)$. 步骤7与步骤8是经典的ISOMAP算法步骤, 该步准确复杂度为$\mathrm{O}(m^3+dm^2\log m)$[17].

故ODD–ISOMAP算法准确时间复杂度为$\mathrm{O}(m\cdot\log m+m+4km+1+1+dm^2\log m+m^3)$, 考虑到$k\ll m$及$d\ll m$, 则ODD–ISOMAP总体时间复杂度与经典ISOMAP算法的时间复杂度大致相同, 为$\mathrm{O}(m^3)$.

另外, 经典ISOMAP算法中获取无向邻域矩阵$\boldsymbol{D}$(也称为构建邻域图)需要筛选数据的$k$个最近邻, ODD–ISOMAP算法中自然邻居的选取只需要在此基础上进一步筛选出自然邻居即可, 额外付出的计算单位为$\mathrm{O}(m)$. 故相对经典 ISOMAP 算法而言, ODD–ISOMAP 算法中需要额外付出的总体计算单元为$\mathrm{O}(2m+4km+1+1)$. 相比$\mathrm{O}(m^3)$的总体时间复杂度, ODD–ISOMAP算法付出了较少的额外算力消耗取得了较好的噪声抗干扰能力.

## 4　实验及效果分析

为检验文章所提算法的算法降维效果, 本节在Swiss roll, Gauss, Iris, Seeds, Wine, Vertebral column-2c, QCM sensor alcohol共7种数据集上对ODD–ISOMAP算法进行了测试, 并将降维效果与其他6种常见降维算法HLLE, LTSA, LEIGS, LLE, PCA, ISOMAP进行了对比, 实验环境为MATLAB 2019a.

### 4.1　测试数据集与评价指标

7种数据集中前2种为人工合成数据集, 后5种来自UCI machine learning实测数据库, 去除异常值后各数据集基本参数如表1所示.

为直观展示各算法降维效果, 将所有数据降维为2维, 并使用统一参数设置的经典K-mediods算法对第2至7种多类别数据集降维后数据进行聚类分析. 由于数据集的类别数量已知, 因此若降维算法能够保持高维数据的拓扑结构, 经各算法降维后的2维数据聚类结果应该与真实数据类标相同, 否则说明算法降维效果较差. 基于以上规律, 引入聚类正确率指标(clustering accuracy, CA)定量描述聚类结果与数据真实类别相似程度, CA表示为

$$\mathrm{CA}=\frac{\{\sum\limits_{p=1}^{t}\sum\limits_{i=1}^{m}[\boldsymbol{x}_i,\boldsymbol{y}_i]|\boldsymbol{x}_i\in C_p,\boldsymbol{y}_i\in C_p\}}{m},\quad(13)$$

上式中: $\boldsymbol{y}_i$表示高维数据$\boldsymbol{x}_i$映射后对应的低维数据, $[\boldsymbol{x}_i,\boldsymbol{y}_i]$表示高低维数据对的数据量, $C_p$表示类标, $t$为数据簇类别数量, $m$为数据簇数据的总量. CA表示所有类别中映射前后属于同一个类标的高低维数据对数据量总和与数据簇总数据量的比值. CA的值介于0到1之间, 取值越大表示聚类结果与高维数据真实类别相似度越高, 即算法降维效果越好.

表 1　测试数据

Table 1　Test datasets

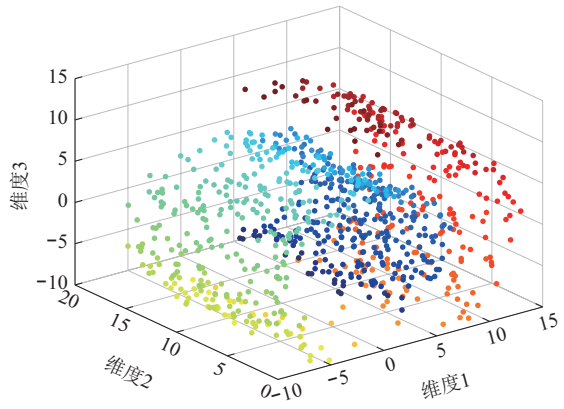
| 序号 | 数据集 | 数据量 | 变量维度 | 类别数量 |
|---|---|---|---|---|
| 1 | Swiss roll | 1000 | 3 | 1 |
| 2 | Gauss | 965 | 3 | 5 |
| 3 | Iris | 150 | 4 | 3 |
| 4 | Seeds | 201 | 7 | 3 |
| 5 | Wine | 178 | 13 | 3 |
| 6 | Vertebral column 2c | 309 | 6 | 2 |
| 7 | QCM3 Sensor Alcohol | 25 | 10 | 5 |

### 4.2　合成数据实验

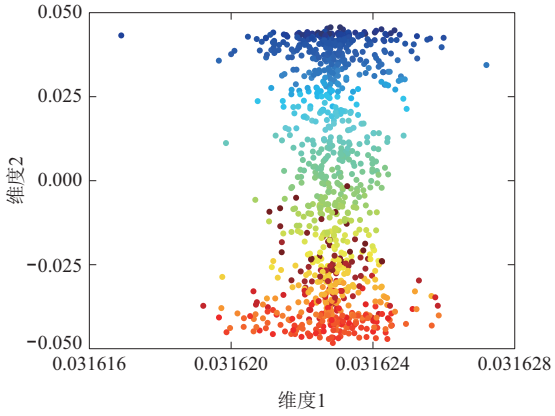
Swiss roll数据为3维人工合成数据, 常用来进行流形降维测试, 本节为了模拟较大曲率及噪声对数据的影响, 对1000组Swiss roll数据的维度3进行了适度压缩, 并添加了60%的随机噪声, 最终生成的Swiss roll数据如图6(a)所示. 使用本文提出的ODD–ISOMAP算法及其他6种降维算法将该数据降为2维, 其中各算法近邻数$k$取最优, 各算法降维效果如图6(b)–6(h)所示.

分析图 6 可知, 在模拟强噪声和大曲率情况下, HLLE算法降维效果出现了重叠, LTSA, LLE, PCA算法降维效果未能保持 3 维数据流形, LEIGS 与 ISOMAP算法的降维效果受到了一定影响, 只有ODD–ISOMAP保持了较好的降维效果.

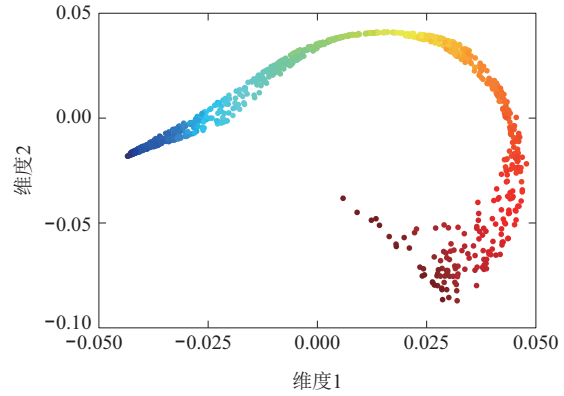
Gauss数据为3维5簇人工合成数据, 各数据簇边缘贴近但不重叠, 如图7(a)所示. 图7(b)–7(h)展示了各算法在该数据上的降维表现, 从效果来看 ODD–ISOMAP算法能够有效促使各数据簇向簇中心靠拢, 进而减少降维后簇间数据重叠度, 而其他算法降维后都不可避免的出现了数据重叠.



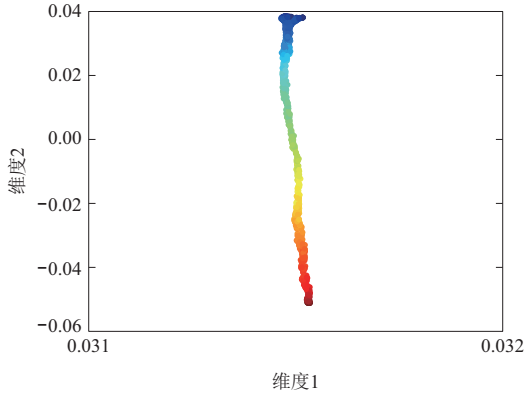

(a) Swiss roll data



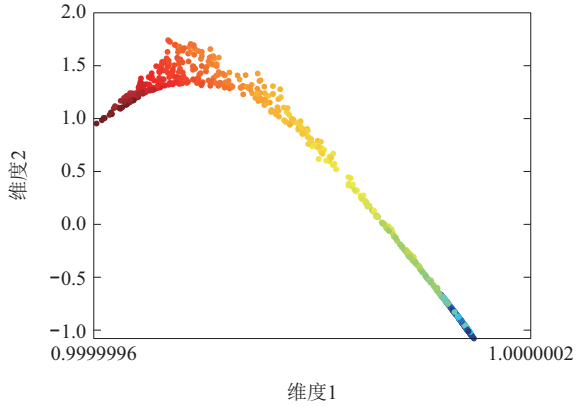

(b) HLLE



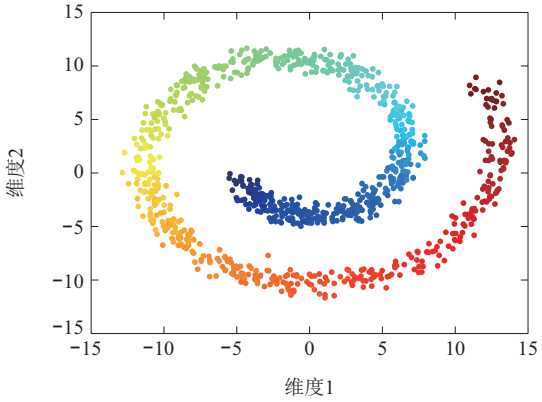

(c) LTSA



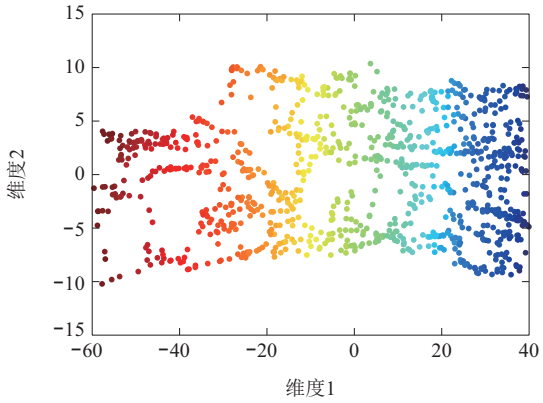

(d) LEIGS



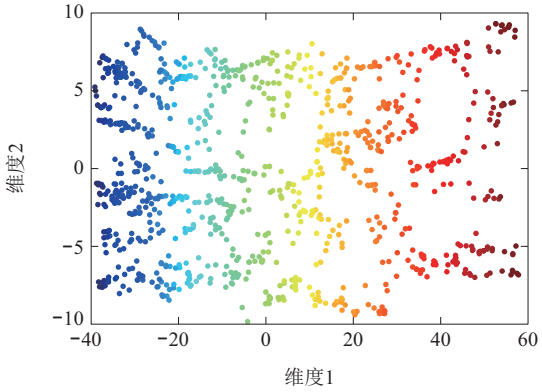

(e) LLE




(f) PCA



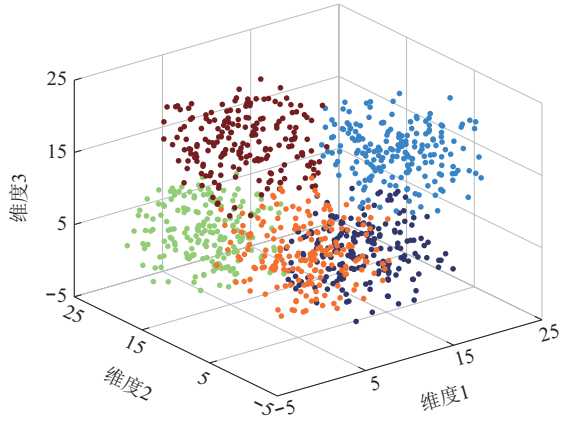

(g) ISOMAP



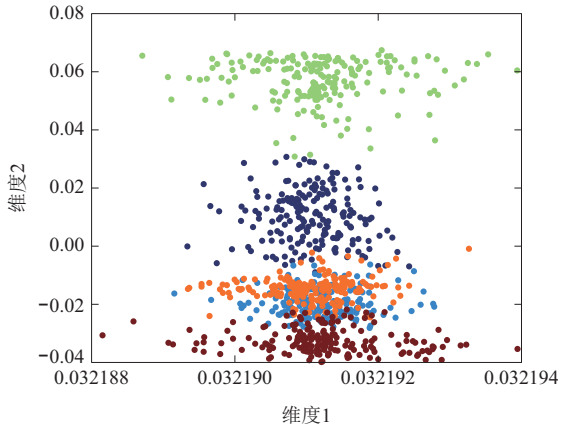

(h) ODD–ISOMAP

图 6 各算法在Swiss roll数据上降维表现

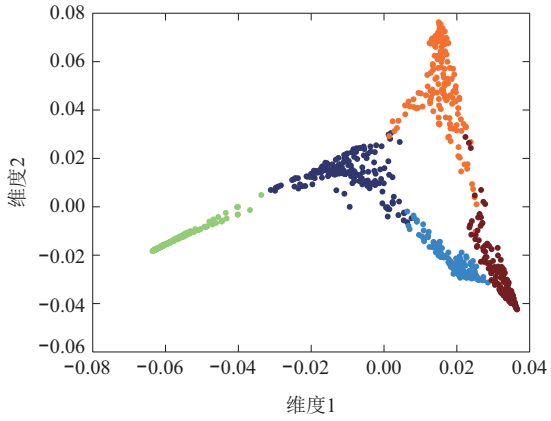
Fig. 6 Dimensionality reduction performance of each algorithm on Swiss roll dataset



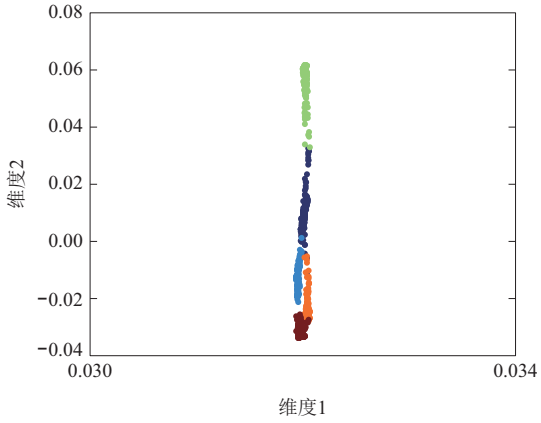

(a) Gauss data



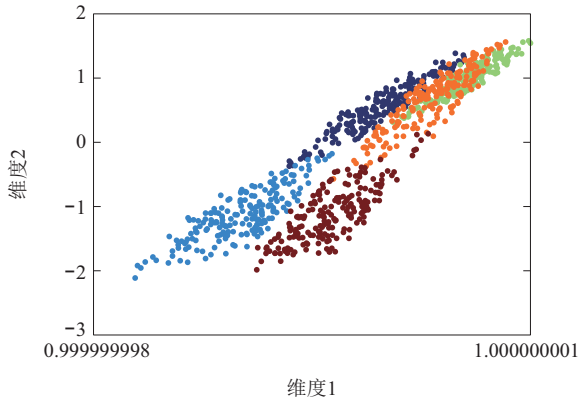

(b) HLLE



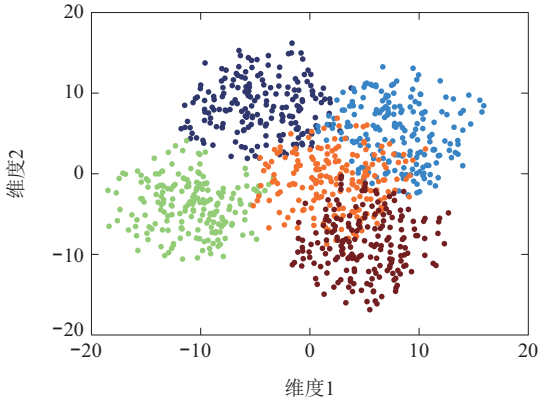

(c) LTSA



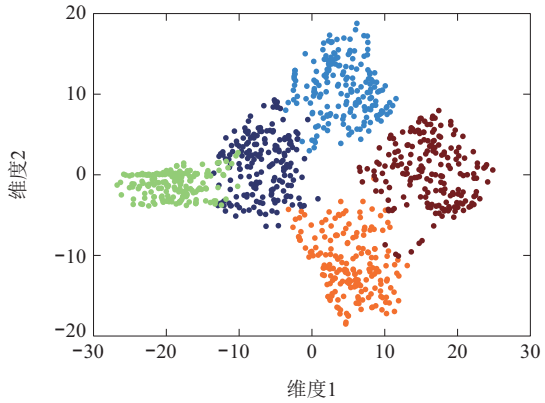

(d) LEIGS



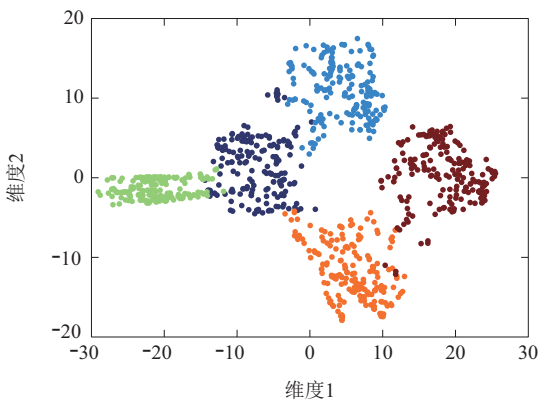

(e) LLE




(f) PCA




(g) LLE




(h) PCA

图 7　各算法在Gauss数据上降维表现

Fig. 7　Dimensionality reduction performance of each algorithm on Gauss dataset

为进一步研究各算法对噪声的敏感度差别, 对Gauss数据集逐步添加3%至15%的高斯噪声, 统计各算法在各噪声幅度下的CA值, 如表2所示. 图8以曲线形式展示了Gauss数据各算法CA值随噪声幅度变化规律, 分析表2与图8可知, 随着高斯噪声幅度的增加各算法对应的CA值呈下降趋势, 说明噪声确实对降维效果有负面影响, 但不同算法受影响程度不同. 在不同幅度噪声影响下ODD–ISOMAP算法的CA值均高于经典ISOMAP算法. 表明在最优密度方向引导下, 噪声干扰被密度缩放因子有效修正, 在一定程度上保持了数据间的真实位置关系. 另外, 在噪声幅度不大于15%情况下, 只有LEIGS与ODD–ISOMAP算法的CA值保持在0.8以上, 并且总体来看后者优于前者; 其他算法均受到了较大影响, 其中LLE算法效果最差. 从噪声角度来看, 只有在添加3%噪声时LEIGS算法降维效果略优于ODD–ISOMAP算法, 其他情况下后者效果均优于其他6种算法, 说明尽管噪声影响了数据的原始分布, 但是多个数据间近邻关系存在一定的稳定性, 而利用了此规律的ODD–ISOMAP算法对噪声有较好的抵抗能力.

表 2 Gauss数据值随CA噪声变化

Table 2 CA values of algorithms vary with noises on Gauss dataset

| 算法 | CA/$k$/$\rho$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 3% | 6% | 9% | 12% | 15% |
| **ODD–ISOMAP** | **0.965/9 /0.17** | 0.905/9 /0.17 | **0.897/9 /0.17** | **0.888/9 /0.17** | **0.844/9 /0.17** | **0.820/9 /0.17** |
| ISOMAP | 0.961/9 | 0.851/9 | 0.804/9 | 0.841/9 | 0.830/9 | 0.784/9 |
| LLE | 0.620/10 | 0.618/10 | 0.608/10 | 0.594/10 | 0.532/10 | 0.520/10 |
| HLLE | 0.875/10 | 0.855/10 | 0.847/10 | 0.812/10 | 0.792/10 | 0.782/10 |
| LEIGS | 0.938/6 | **0.911/6** | 0.890/6 | 0.868/6 | 0.837/6 | 0.808/6 |
| LTSA | 0.924/10 | 0.906/10 | 0.861/10 | 0.841/10 | 0.811/10 | 0.796/10 |
| PCA | 0.856 | 0.834 | 0.814 | 0.803 | 0.753 | 0.723 |

注: PCA算法主元贡献率取0.95, $\varepsilon$统一取0.001.

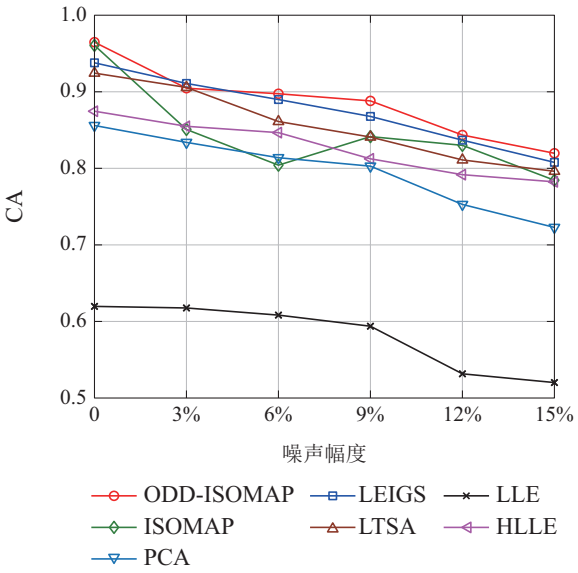



图 8 Gauss数据集各算法CA值随噪声变化曲线

Fig. 8 CA curves of algorithms vary with noises on Gauss dataset

图9展示了ODD–ISOMAP算法与ISOMAP算法在添加3%高斯噪声的Gauss数据上降维后聚类效果. 图中不同形状表示原始高维数据真实类标, 不同颜色表示降维后数据聚类类标, 红色圈中数据为分类错误数据. 从图 9 中可看出, 相比 ISOMAP 算法, ODD–ISOMAP算法能够有效放大各数据簇间距离, 并使各簇数据向簇心收缩, 从而减少分类错误, 较好保持了高维数据原始样式.

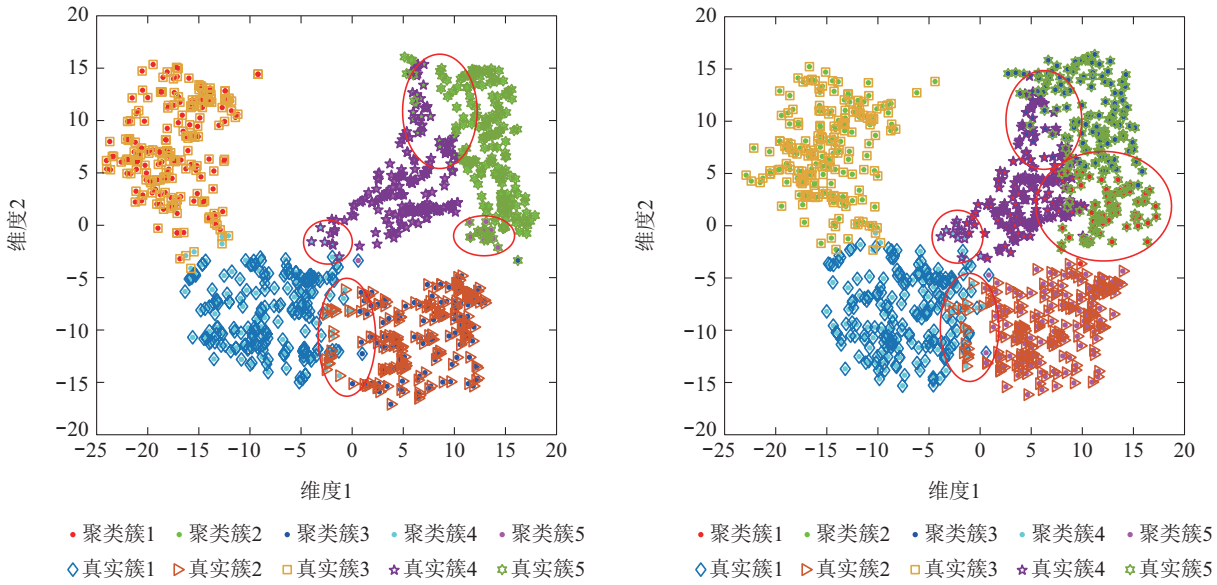
### 4.3 实测数据实验

本节在优化近邻数值$k$、系数调节因子$\varepsilon$和密度缩放因子$\rho$前提下, 对比了 ODD–ISOMAP 算法与 ISOMAP, HLLE, LTSA, LEIGS, LLE, PCA算法在5类实测数据上的降维表现. 表3统计了各算法在实测数据集上的CA值, 图10 至图14分别展示了各数据集上CA值最高的前2种算法降维效果. 各图中不同形状表示原始高维数据的真实类标, 不同颜色表示降维后数据聚类类标, 红色圈中数据为错误分类数据.

从图10来看, 对Iris数据降维效果最好的为ODD–ISOMAP算法和ISOMAP算法. 对比图10(a)与图10(b)的聚类簇1, 可以发现ODD–ISOMAP算法能够使真实簇更加聚集. 对比两图聚类簇2–3可知, 两种算法降维后聚类效果均出现了重叠, 但ODD–ISOMAP算法能够使两个不同簇沿着各自簇内数据密集方向收拢, 从而减少了错分数据, 提高了分类正确率. 图11至图13展示了Seeds, Wine和Vertebral column2c数据集的结果, 获得了与图10同样的结论. 这是由于ODD–ISOMAP算法能够根据高维数据的原始分布情况沿最优密度方向合理缩放了数据间距离, 使得降维后各数据

集同一簇内数据更加紧密, 不同簇间距离更加分散, 簇边缘数据向簇内靠拢, 较好保持了高维数据真实簇分类特征, 从而有效改善了聚类效果. 图 14 中虽然 ODD–ISOMAP, ISOMAP 和 PCA 算法都能在 QCM Sensor Alcohol数据上取得极高正确率(CA = 1), 但是前者通过数据缩放功能使得分类效果更加显著. 从表3数据角度来看, 在Wine数据集上ODD–ISOMAP算法和HLLE算法均能得到最高的CA值, 其他数据集上ODD–ISOMAP算法的CA值为最高, 说明本文所提算法降维效果最好.

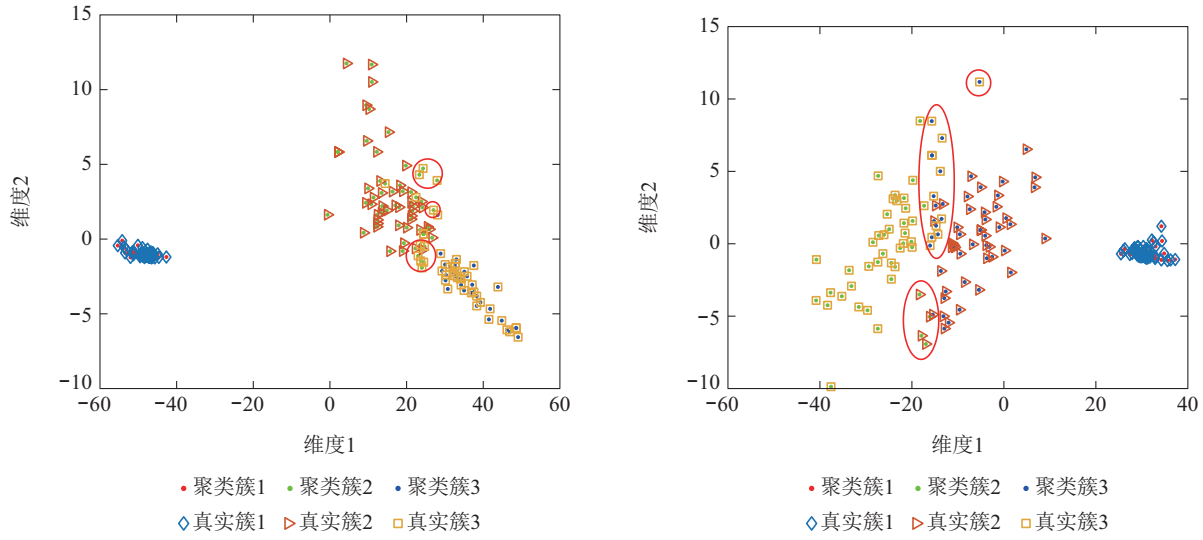
为进一步研究各算法对噪声的敏感度差别, 在Seeds数据集上逐步添加5%至25%的高斯噪声, 统计各算法在各噪声幅度下的CA值, 如表4所示. 图15以曲线形式展示了Seeds 数据集上各算法CA值随噪声幅度变化规律. 从表4据图15可以看出, 除了在25%强噪声干扰下 ODD–ISOMAP 算法降维效果略低于LTSA算法外; 在其他情况下ODD–ISOMAP算法都能保持更好的降维效果.




(a) ODD–ISOMAP　　(b) ISOMAP

图 9　Gauss数据3%噪声下降维后聚类效果

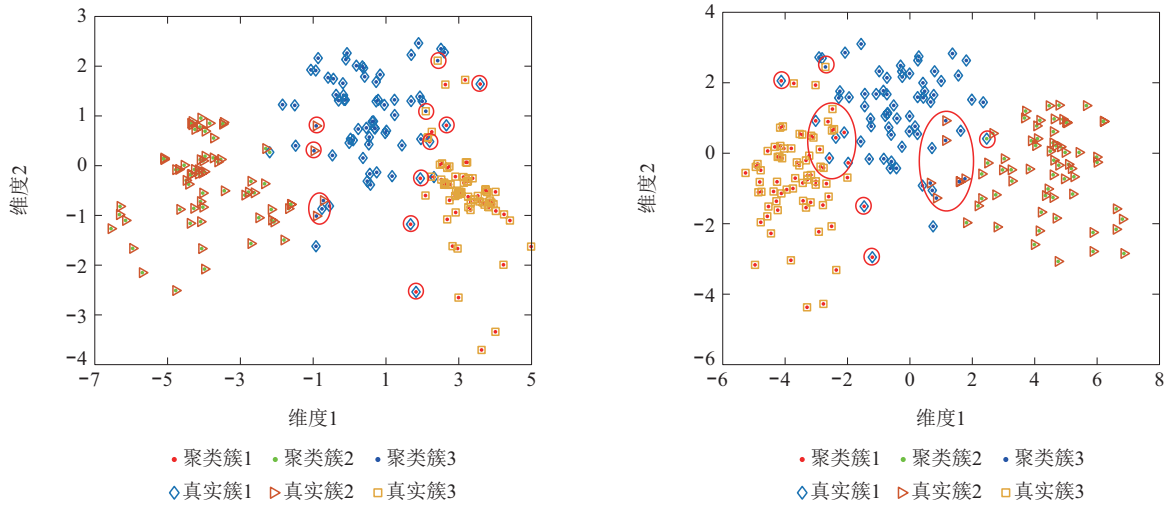
Fig. 9　Clustering effect after dimensionality reduction of Gauss dataset with 3% noise




(a) ODD–ISOMAP　　(b) ISOMAP

图 10　Iris数据二维映射

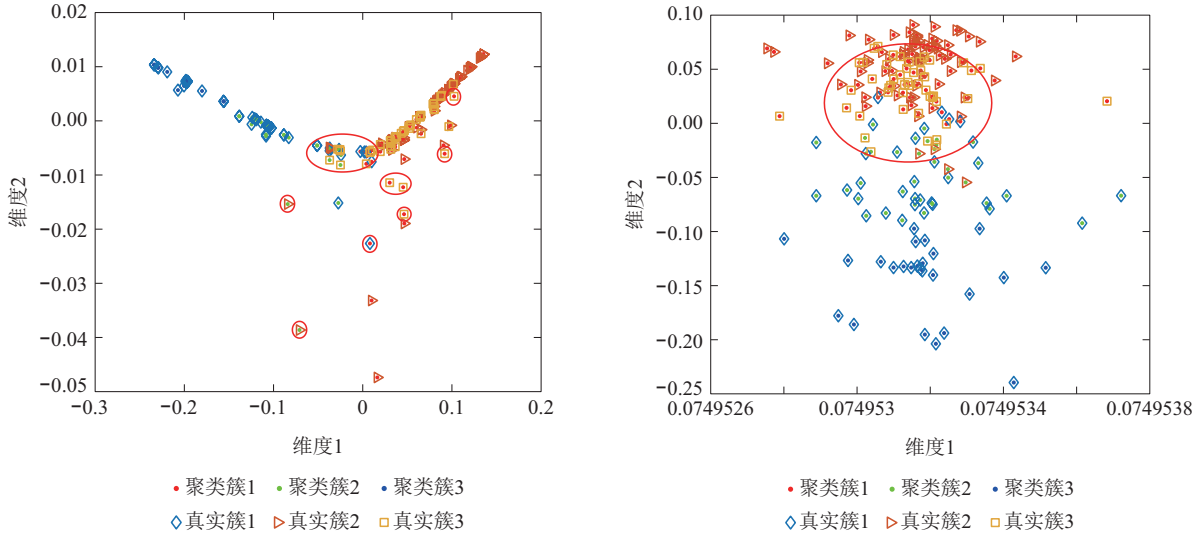
Fig. 10　Two-dimensional maps of Iris dataset




(a) ODD–ISOMAP (b) ISOMAP

图 11 Seeds数据二维映射

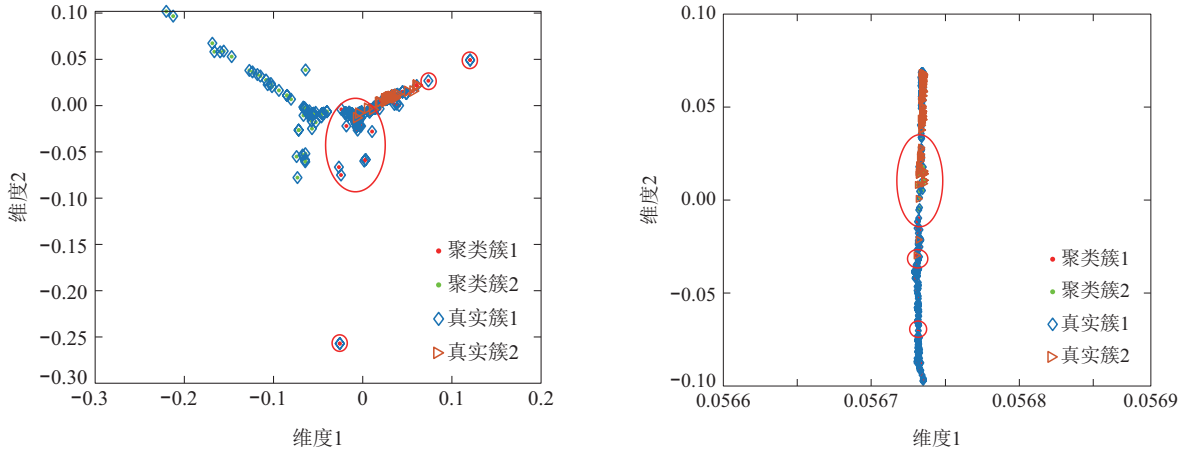
Fig. 11 Two-dimensional maps of Seeds dataset




(a) ODD–ISOMAP (b) HLLE

图 12 Wine数据二维映射

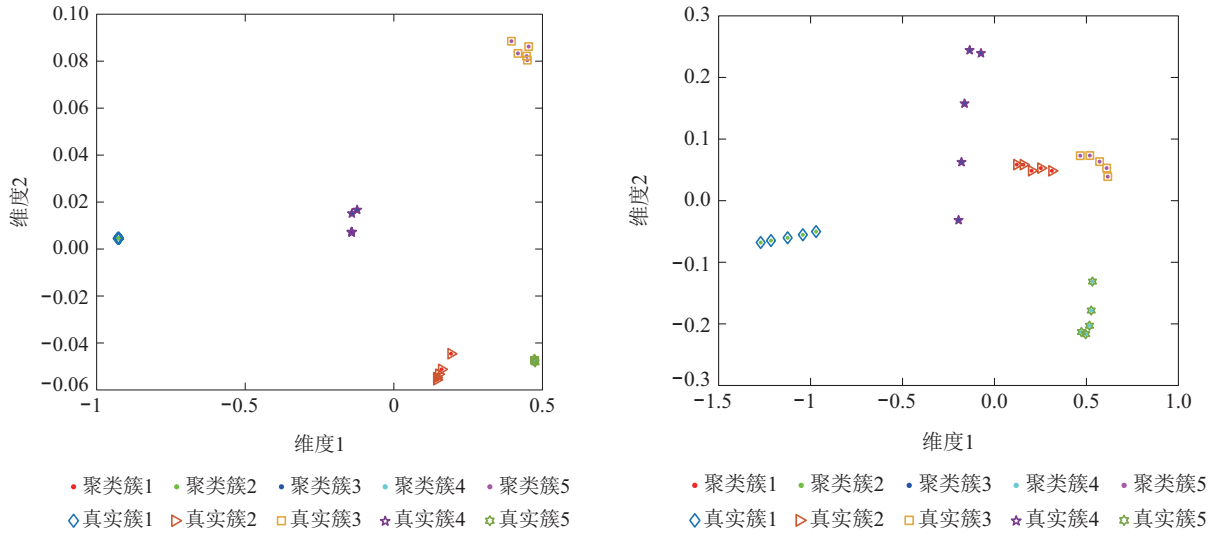
Fig. 12 Two-dimensional maps of Wine dataset




(a) ODD–ISOMAP (b) LEIGS

图 13 Vertebral column 2c数据二维映射

Fig. 13 Two-dimensional maps of Vertebral column 2c dataset




(a) ODD–ISOMAP　　(b) ISOMAP

图 14　QCM Sensor Alcohol数据二维映射

Fig. 14　Two-dimensional maps of QCM Sensor Alcohol dataset

表 3　各算法在多个实测数据上的CA值

Table 3　CA values of algorithms on difference datasets

| 算法 | CA/$k$/$\rho$ | | | | |
|---|---|---|---|---|---|
| | Iris | Seeds | Wine | Vertebral column 2c | QCM Sensor Alcohol |
| **ODD–ISOMAP** | **0.906/25 /0.1** | **0.930/25 /0.1** | **0.787/10 /0.004** | **0.809/7 /0.08** | **1.000/10 /0.025** |
| ISOMAP | 0.893/30 | 0.915/25 | 0.708/10 | 0.728/7 | **1.000/10** |
| LLE | 0.900/30 | 0.880/30 | 0.708/15 | 0.738/15 | 0.920/10 |
| HLLE | 0.867/26 | 0.876/40 | **0.787/10** | 0.670/45 | 0.920/10 |
| LEIGS | 0.773/23 | 0.910/20 | 0.730/10 | 0.751/3 | 0.880/5 |
| LTSA | 0.853/25 | 0.891/15 | 0.764/28 | 0.702/60 | 0.960/6 |
| PCA | 0.886 | 0.915 | 0.702 | 0.676 | **1.000** |

注: PCA算法主元贡献率取0.95, $\varepsilon$统一取0.001.

表 4　各算法在不同噪声幅度Seeds数据上的CA值

Table 4　CA values of algorithms vary with noises on seeds dataset

| 算法 | CA/$k$/$\rho$ | | | | |
|---|---|---|---|---|---|
| | 5% | 10% | 15% | 20% | 25% |
| **ODD–ISOMAP** | **0.881/25 /0.15** | **0.816/30 /0.3** | **0.726/14 /0.3** | **0.687/64 /0.3** | 0.647/80 /0.3 |
| ISOMAP | 0.851/25 | 0.706/25 | 0.572/25 | 0.572/25 | 0.507/25 |
| LLE | 0.801/30 | 0.697/30 | 0.597/30 | 0.537/30 | 0.477/30 |
| HLLE | 0.771/40 | 0.667/40 | 0.597/40 | 0.602/40 | 0.567/40 |
| LEIGS | 0.776/20 | 0.627/20 | 0.587/20 | 0.572/20 | 0.537/20 |
| LTSA | 0.821/15 | 0.786/15 | 0.701/15 | **0.687/15** | **0.667/15** |
| PCA | 0.841 | 0.711 | 0.522 | 0.562 | 0.483 |

注: PCA算法主元贡献率取0.95, $\varepsilon$统一取0.001.

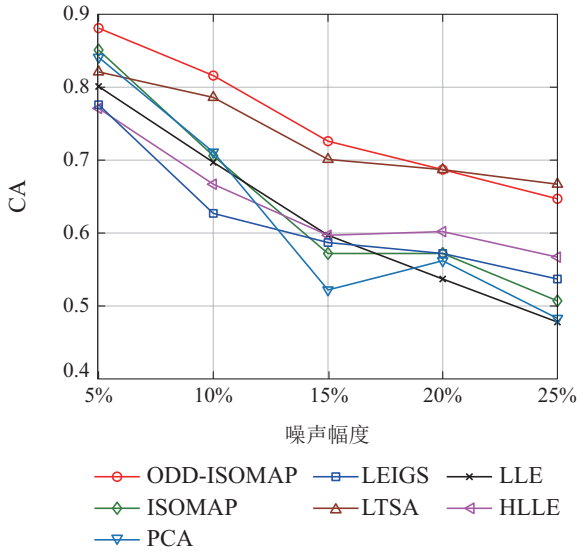



图 15 Seeds数据集各算法CA值随噪声变化曲线

Fig. 15 CA curves of algorithms vary with noises on Seeds dataset

## 5 结论

由于数据的多个近邻间的关系能够反映数据的本质分布情况, 在一定噪声的影响下, 数据与其自然邻居间的关系存在一定的稳定性, 并且由于自然邻居的获取过程取决于近邻数据间的相互位置关系, 因此数据的自然邻居一般沿着高维流形方向分布. 本文利用了此种规律引入了最优密度方向概念, 提出了具有较好噪声抗干扰能力的ODD–ISOMAP算法. 该算法从高维数据点的$k$个最近邻中筛选出自然邻居集合, 进一步得到了各个数据的最优密度方向, 此最优密度方向指向流形方向且不易受噪声干扰. 之后通过计算高维数据的$k$个最近邻在最优密度方向上投影的角度、方向和长度, 获取各近邻数据相对高维数据距离的密度缩放因子. 此密度缩放因子的大小与噪声对数据的影响程度密切相关, 因此使用该密度缩放因子对数据间的局部邻域距离进行缩放, 能够在一定程度上还原数据与其近邻间的原始分布情况, 从而降低噪声对降维过程的影响. 在实验环节对比了 ODD–ISOMAP 算法与ISOMAP, HLLE, LTSA, LEIGS, LLE和PCA算法在2类人工合成数据集和5类实测数据集上的降维表现, 并测试了各算法在不同噪声压力下的降维性能, 验证了本文所提算法的实用性和有效性.

## 参考文献:

[1] WANG Guoqiang, LI Longxing, GUO Xiaobo. Sparsity preserving discriminant embedding for face recognition. *Chinese Journal of Scientific Instrument*, 2014, 35(2): 305 – 312.
(王国强, 李龙星, 郭晓波. 基于稀疏保持判别嵌入的人脸识别. 仪器仪表学报, 2014, 35(2): 305 – 312.)

[2] LEW M S, SEBE N, DJERABA C, et al. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)*, 2006, 2(1): 1 – 19.

[3] BAR-JOSEPH Z, GITTER A, SIMON I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 2012, 13(8): 552 – 564.

[4] DONOHO D L. High-dimensional data analysis: The curses and blessings of dimensionality. *ACM Math Challenges Lecture*. Los Angeles, 2000: 1 – 32.

[5] HARDLE W K, HLAVKA Z. Principal component analysis. *IEEE Trans on Automatic Control*, 2015, 29(1): 163 – 183.

[6] MARTINEZ A M, KAK A C. PCA versus LDA. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2001, 23(2): 228 – 233.

[7] COX M AA, COX T F. Multidimensional scaling. *Journal of the Royal Statistical Society*, 2008, 46(2): 1050 – 1057.

[8] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2019, 15(6): 1373 – 1396.

[9] ROWEIS S Tx, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, 290(5500): 2323 – 2326.

[10] TENENBAUM J B, SILVA D V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, 290(5500): 2319 – 2323.

[11] YU J, KIM S B. Density-based geodesic distance for identifying the noisy and nonlinear clusters. *Information Science*, 2016, 360: 231 – 243.

[12] LI Deyu, GAO Cuizhen, ZHAI Yanhui. An orderly adaptive neighborhood selection algorithm based on manifold curvature. *Journal of Shanxi University (Natural Science Edition)*, 2012, 35(2): 219 – 223.
(李德玉, 高翠珍, 翟岩慧. 基于流形弯曲度的有序自适应邻域选择算法. 山西大学学报(自然科学版), 2012, 35(2): 219 – 223.)

[13] GENG X, ZHAN D C, ZHOU Z H. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics*, 2005, 35(6): 1098 – 1107.

[14] YANG B, XIANG M, ZHANG Y. Multi-manifold discriminant ISOMAP for visualization and classification. *Pattern Recognition*, 2016, 55: 215 – 230.

[15] LI Xiangyuan, CAI Cheng, HE Jinrong. Density scaling factor based ISOMAP algorithm. *Computer Science*, 2018, 45(7): 207 – 213.
(李香元, 蔡骋, 何进荣. 基于密度缩放因子的ISOMAP算法. 计算机科学, 2018, 45(7): 207 – 213.)

[16] HUANG J L, ZHU Q S, YANG L J, et al. A non-parameter outlier detection algorithm based on natural neighbor. *Knowledge-Based Systems*, 2016, 92: 71 – 77.

[17] SILVA V D, TENENBAUM J B. Global versus local methods in nonlinear dimensionality reduction. *Neural Information Processing Systems*, 2003, 15: 705 – 712.

作者简介:

**梁少军** 硕士, 讲师, 目前研究方向为流形学习、故障诊断、模式识别, E-mail: sjliang@wu.edu.cn;

**张世荣** 博士, 副教授, 目前研究方向为复杂工业系统能效优化, E-mail: srzhang@whu.edu.cn;

**孙澜琼** 学士, 助教, 目前研究方向为故障诊断、流形学习, E-mail:sunlanq@foxmail.com.