基于密度缩放因子的ISOMAP算法

李香元 蔡 骋 何进荣
(西北农林科技大学信息工程学院 陕西 杨凌 712100)

摘 要 等度量映射(ISOMAP)算法是一种被广泛应用的非线性无监督降维算法,通过保持各个观测样本间的测地距离进行等距嵌入,从而实现高维空间向低维空间的坐标转换。但在实际应用中,观测数据无可避免地会存在噪声,由于测地距离的计算对噪声比较敏感,并且也没有考虑数据集的密度分布,导致ISOMAP算法降维后低维坐标表示存在几何变形。针对这一缺点,根据局部密度的思想,提出一种基于密度缩放因子的ISOMAP(Density Scaling Factor Based ISOMAP,D-ISOMAP)算法。在传统的ISOMAP算法框架下,首先,针对每个观测样本计算一个局部密度缩放因子;然后,在测地距离的计算过程中,将直接相邻的两个样本之间的测地距离除以这两个样本密度缩放因子的乘积;最后,通过最短路径算法求得改进后的距离矩阵,并对其进行降维处理。改进的测地距离在密度较大的区域被缩小,而在密度较小的区域被放大,这样可以减小噪声对降维效果的影响,提升可视化和聚类效果。人工数据集和UCI数据集上的实验结果表明,在数据集的可视化和聚类效果方面,D-ISOMAP算法较经典的无监督降维算法具有一定的优势。

关键词 ISOMAP,流形学习,密度缩放因子,噪声数据,降维
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.07.036

Density Scaling Factor Based ISOMAP Algorithm

LI Xiang-yuan CAI Cheng HE Jin-rong
(College of Information Engineering, Northwest A & F University, Yangling, Shaanxi 712100, China)

Abstract ISOMAP is a widely used nonlinear unsupervised dimensionality reduction algorithm, and preserves the coordinate transformation from high-dimensional space to low-dimensional space by keeping the geodesic distance between the observation samples. However, practical data are inevitably corrupted by noise. Since the calculation of geodesic distance is sensitive to noise and does not consider the density distribution of dataset, the geometric deformation of the ISOMAP algorithm is happening after dimensionality reduction. For this shortcoming, according to the idea of local density, a density scaling factor based ISOMAP algorithm called D-ISOMAP was proposed. In the framework of traditional ISOMAP algorithm, local density scaling factor is first calculated for each observation sample. And then the original geodesic distance of two samples is divided by the product of their density scaling factors. Finally, the improved distance matrix is obtained by the shortest path algorithm. Since the improved geodesic distance is reduced in the larger density region and enlarged in the smaller density region, it can achieve a good effect in noisy situations. The experimental results on artificial dataset and UCI dataset show that the D-ISOMAP algorithm is better than the classical unsupervised dimensionality reduction algorithms in terms of visualization and clustering of data sets.

Keywords ISOMAP, Manifold learning, Density scaling factor, Noise data, Dimensionality reduction

1 引言

随着传感器技术的快速发展,数据的来源呈现出多样化特点,如文本数据、图像数据、音频数据和视频数据,数据的维度也越来越高。实际应用领域中获取的数据都是高维的,如信息检索[1]、文本挖掘[2]、生物信息[3]和人脸识别等领域。维数的增高给数据处理带来了很多困难,在高维情形下出现的数据样本稀疏、距离计算复杂度大等问题是所有机器学习方法面临的共同难题,通常将这一困难称作维数灾难[4]。如何快速、有效地处理高维数据是目前机器学习领域的一个重要研究方向,而对数据进行降维处理是克服维数灾难问题的一种有效方法。在大多数实际问题中,尽管观测数据集中的样本维数较高,但与学习任务密切相关的属性特征只是某个低维分布,即高维空间的一个低维嵌入[5]。因此高维数据总是

到稿日期:2017-04-19 返修日期:2017-07-04 本文受国家自然科学基金项目(61202188),西北农林科技大学博士科研启动基金项目(2452015302),杨凌示范区科技计划项目(2016NY-31)资助。

李香元(1993—),男,硕士生,主要研究领域为机器学习;**蔡 骋**(1979—),男,博士,教授,硕士生导师,CCF会员,主要研究领域为计算机视觉与机器学习,E-mail:chengcai@nwsuaf.edu.cn(通信作者);**何进荣**(1984—),男,博士,讲师,CCF会员,主要研究领域为计算机视觉与数据降维。

包含许多冗余信息，各个特征属性之间通常具有较强的相关性，这种现象在几何学上表现为数据分布在低维流形上，或者分布在低维流形附近。因此，通过某种数学变换将原始高维属性空间转变为一个低维子空间，在这个低维子空间中样本的密度被大幅度提高，距离计算也更加容易，数据处理比较方便。通过降维对高维数据进行预处理，是数据进行可视化分析、聚类或分类任务的关键一步。对于可视化任务，降维的目标是将数据集映射到一个保持数据内部判别结构的 2 维或 3 维空间。对于聚类或分类任务，降维的目标是将输入数据映射到一个各种类别的数据流形都可以清晰分开的特征空间。

目前，降维算法可分为线性降维和非线性降维两种类型。线性降维算法主要有主成分分析(PCA)[6]算法、独立成分分析(ICA)[7]算法、线性判别分析(LDA)[8]算法和多维尺度变换算法(MDS)[9]等。PCA 的目标是寻找新的变量，使它们反映事物的主要特征，压缩原有数据矩阵的规模。每个新变量是原有变量的线性组合，体现了原有变量的综合效果，具有一定的实际含义。ICA 算法与 PCA 算法相似，其成分被认为是相互独立的。LDA 是一种有监督降维算法，其基本思想是将高维的数据样本投影到最佳鉴别矢量空间，保证投影后的数据样本在新的子空间中具有最大的类间距离和最小的类内距离。MDS 算法以数据样本间的欧氏距离矩阵作为算法的输入，可以将数据降低到一个低维的空间，并且能够尽可能地保持原始数据的本质信息。但实际上，大多数数据是非线性结构的，并且还会存在噪声等情况。为了克服这些限制，一些非线性降维算法被提出，如等距映射(ISOMAP)[10]算法、局部线性嵌入(LLE)算法[5]和核组成成分分析(KPCA)[11]算法等。ISOMAP 算法是一种具有代表性的非线性降维算法，主要是通过分析现有高维流形得到高维流形对应的低维嵌入，从而使高维流形上数据点间的近邻结构在低维嵌入中得到较完整的重现。将 MDS 算法作为分析工具，但在计算高维流形上数据点间的距离时，以近似的测地距离而非欧氏距离作为 MDS 的输入，使得在嵌入空间中原始样本间的近似测地距离得以保持，从而在从高维到低维的流形学习过程中最大化地重构所有样本点之间的流形距离。另外，测地距离能很好地反映数据的拓扑结构。

但是，在实际情况下，数据集中会存在一定程度的噪声。当输入数据中存在噪声时，ISOMAP 算法的拓扑稳定性就会遭受破坏。基于此，对传统的 ISOMAP 算法进行改进，通过局部密度的思想[12-13]提出一种基于密度缩放因子的 ISOMAP 算法，使用基于密度的距离度量进行降维处理能提高 ISOMAP 算法对噪声的鲁棒性，因为 ISOMAP 算法的一个关键步骤就是计算数据之间的距离度量。传统的 ISOMAP 算法使用的测地距离虽然能反映出数据的拓扑结构，但是若数据中存在噪声，则测地距离的计算会受到较大的影响。而 D-ISOMAP 算法首先通过构建的 k 邻近图为每个数据样本计算出一个局部密度缩放因子，在计算数据间的距离时，通过密度缩放因子对原始的测地距离进行放缩，即用直接相邻的两个样本点之间的欧氏距离除以这两个样本点密度缩放因子的乘积，最后通过最短路径算法求出改进后的距离度量矩阵。改进后的距离使大密度区域数据间的距离被缩短，数据变得更加紧密；使小密度区域数据间的距离被增大，数据变得更加稀疏。通过计算样本点的局部密度，结合样本点之间的距离，可以有效地识别出样本点的聚类结构和噪声点[13]，基于此进行降维，能很好地反映数据的密度分布，减小数据中噪声的影响，对数据集中噪声的存在具有一定的鲁棒性。ISOMAP 算法还有很多其他改进方法，如有监督的 ISOMAP(S-ISOMAP)算法[14]、正交边际约束的 ISOMAP(M-ISOAMP)算法[15]、多流形判别 ISOMAP(MMD-ISOMAP)算法[16]等，同时也有其他降维算法，如基于判别投影的数据集降维算法[17]等。这些算法都能提取出具有较好判别性能的低维特征，但是它们是有监督的降维算法，对于无标签的数据集并不适用。可视化和聚类实验表明，与经典的无监督降维算法相比，文中所提 D-ISOMAP 算法具有一定优势。

本文第 2 节简要概述 ISOMAP 算法，介绍测地距离的计算方法；第 3 节提出 D-ISOMAP 算法，介绍密度缩放因子的计算方法，同时讨论它与原始 ISOMAP 算法的区别与联系；第 4 节给出实验结果与分析；最后总结全文并提出未来值得研究的几个问题。

2 传统的 ISOMAP 算法

传统的等度量映射(ISOMAP)算法的基本出发点是认为低维流形嵌入到高维空间之后，直接在高维空间中计算直线距离具有误导性，因为高维空间的直线距离在低维嵌入流形上是不可达的。ISOMAP 算法采用由数据点形成的无向加权邻近图中的点与点之间的最短路径的长度来刻画流形数据上点与点之间的距离[18]。在邻接图上计算两点之间的最短路径时，可采用著名的 Dijkstra 算法或 Floyd 算法，亦可得到任意两点间的距离，即测地距离。得到测地距离之后，可以通过 MDS 算法来获得样本点在低维空间中的坐标。测地距离的计算主要分为两个步骤：1)为原始数据构造一个 k 邻近图，在该图中，每个观测样本都被看作一个结点并且与它最近的 k 个结点直接相连，用欧氏距离表示连接之间的权重；2)通过构造好的 k 邻近图来计算点与点之间的最短路径，最后得到的最短路径矩阵即为各个观测样本点间的测地距离。

2.1 构造 k 邻近图

测地距离计算的第一步就是构造一个表示原始数据的邻近图。邻近图拥有许多不同类型，如 ε 邻近图和 k 邻近图[19]，本文选择 k 邻近图，因为它在实际中应用广范并且对邻近图的构建和局部信息的描述是相对简单且有效[20]的。观测样本 x_i 的 k 邻近可以表示为：

$$k(x_i)=\{x_j \mid \| x_i-x_j \|_2 \leqslant d_i^k\} \tag{1}$$

其中，$\| x_i-x_j \|_2$ 表示两个观测样本点之间的欧氏距离；d_i^k 表示距离观测样本点 x_i 欧氏距离最近的 k 个样本点的欧氏距离。

通过 k 邻近的表达式，观测样本 x_i 的对称邻近集 $\Omega(x_i)$ 可表示为：

$$\Omega(x_i)=\{x_j \mid x_j \in k(x_i) \text{ or } x_i \in k(x_j)\} \tag{2}$$

如果观测样本 x_i 包含于 $k(x_i)$ 或包含于 $k(x_j)$，那么 x_j 为 x_i 的一个对称邻近点。k 邻近图的表示如图 1 所示，对称邻近集如表 1 所列。

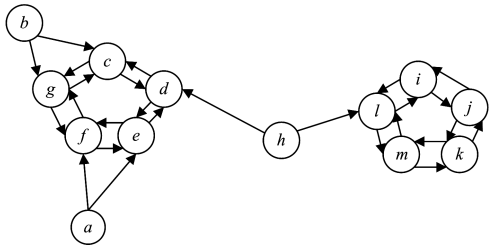


图 1　空间点的几何描述($k=2$)

Fig. 1　Geometric description of data points($k=2$)

图 1 中存在 13 个高维空间点，其中 c,d,e,f,g 和 i,j,k,l,m 为两个不同的数据类，a,b,h 为噪声点。

表 1　对称邻近集($k=2$)

Table 1　Symmetry neighborhood set($k=2$)

序号	观测样本	对称邻接数据集
1	a	e,f
2	b	c,g
3	c	b,d,g
4	d	c,e,h
5	e	a,d,f
6	f	a,e,g
7	g	b,c,f
8	h	d,l
9	i	j,l
10	j	i,k
11	k	j,m
12	l	h,i,m
13	m	l,k

2.2　测地距离的计算

在 k 邻近图中，两观测样本间的边 e_{ij} 可表示为：

$$e_{ij}=\begin{cases}1, & \text{if } x_j\in k(x_i) \text{ or } x_i\in k(x_j)\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

将直接相连的两点在邻接矩阵中的值设置为 1，不相连的两点设置为 0。根据 e_{ij} 的值计算点与点之间的权重 w_{ij}。w_{ij} 为两点之间的欧氏距离，计算公式可表示为：

$$w_{ij}=\begin{cases}\| x_i-x_j \|_2, & \text{if } e_{ij}=1\\ \infty, & \text{otherwise}\end{cases} \tag{4}$$

得到 w_{ij} 之后，通过 Djkstra 最短路径算法能计算出各个样本点之间的最短路径 g_{ij}，计算公式可表示为：

$$g_{ij}=\begin{cases}w_{ij}, & \text{if } e_{ij}=1\\ \min\{w_{ij},w_{il}+w_{lj}\}, & \text{otherwise}\end{cases} \tag{5}$$

当观测样本 i 与 j 之间直接相连时，测地距离则为它们之间的权重；否则通过中间样本点 $l(l=1,2,\cdots,n)$ 求取两样本点之间的测地距离。最后得到一个关于所有观测样本点的测地距离矩阵。测地距离虽然适用于非线性的数据集，能有效地反映出数据的拓扑结构，但是并没有考虑数据的密度分布，并且在大多数实际情况中数据存在噪声，测地距离对数据周围的噪声比较敏感[12]。因此，对于存在噪声的非线性数据集，ISOMAP 算法的性能有待提高。

3　D-ISOMAP 算法

3.1　D-ISOMAP 算法的基本思想

传统的 ISOMAP 算法中测地距离的计算并没有考虑数据的密度，且还会受到噪声的影响，为了降低噪声的影响，我们采用密度的思想来计算测地距离，得到一个新的距离度量，进而提出了 D-ISOMAP 算法。在传统的 ISOMAP 算法框架下，将每个观测样本看成一个结点，通过原始测地距离的计算方法，在计算对称邻近集时转向求共同邻近集；通过求得的共同邻近集，为每个观测点都计算出一个局部密度缩放因子，该值越大，表明该值所在区域的密度越大，反之，该值所在区域的密度越小。通过该密度缩放因子来减小或增加数据之间的权重，使得在样本密度高的区域样本变得更加紧密，而在样本密度低的区域样本变得更加稀疏。如此求得的距离矩阵反映了观测样本的密度，对数据中的噪声具有鲁棒性。最后，使用该距离矩阵在 MDS 算法下获得低维坐标，完成降维处理。所提 D-ISOAMP 算法在可视化和聚类效果上均有一定提升。

3.2　D-ISOMAP 算法的步骤

D-ISOMAP 基于传统的 ISOMAP 算法框架，改变了距离度量的计算，其余步骤均与原方法一致。基于密度缩放因子的 ISOMAP 的距离计算主要分为 3 个步骤：1)为原始数据构造一个 k 邻近图，并根据邻近图求得每个样本的对称邻近集和共同邻近集；2)根据共同邻近集计算每个观测样本的密度缩放因子；3)通过构造好的 k 邻近图和密度缩放因子来计算点与点之间的最短路径。

观测样本 x_i 的共同邻近集 $\Psi(x_i)$ 可表示为：

$$\Psi(x_i)=\{x_j \mid x_j\in k(x_i) \text{ and } x_i\in k(x_j)\} \tag{6}$$

如果观测样本 x_j 既包含于 $k(x_i)$ 又包含于 $k(x_i)$，那么 x_j 就为 x_i 的一个共同邻近点。共同邻近集是指两个样本点都互为对方的邻近点，它比对称邻近数据集的要求更加严格[21]。图 1 中的共同邻近集如表 2 所列，其中“—”表示无共同邻接数据集。

表 2　共同邻近集($k=2$)

Table 2　Mutual neighborhood set ($k=2$)

序号	观测样本	共同邻接数据集
1	a	—
2	b	—
3	c	d,g
4	d	c,e
5	e	d,f
6	f	e,g
7	g	c,f
8	h	—
9	i	j,l
10	j	i,k
11	k	j,m
12	l	i,m
13	m	l,k

观测样本点的密度系数 γ_i 可表示为：

$$\gamma_i=\sum_{j\in\Psi(x_i)}e_{ij}+\sum_{\substack{j,k\in\Psi(x_i)\\ j<k}}e_{jk} \tag{7}$$

式(7)由两部分组成：第一部分表示结点 i 到其共同邻近点的边的数目；第二部分表示结点的共同邻近点之间相互连接的边数。例如，结点 3 的密度缩放因子通过下式进行计算：

$$\gamma_3=e_{34}+e_{37}+e_{47}=2$$

这种密度缩放因子的计算方法不仅考虑了一个观测点到其共同邻近点的连接，也考虑了其共同邻近点之间的连接。因此，对于一个位于密集区域的观测点，它的密度缩放因子也相对较大。图 1 中观测样本的密度缩放因子如表 3 所列。

表 3 密度缩放因子

Table 3 Density scaling factor

序号	观测样本	密度缩放因子
1	*a*	0
2	*b*	0
3	*c*	2
4	*d*	2
5	*e*	2
6	*f*	2
7	*g*	2
8	*h*	0
9	*i*	2
10	*j*	2
11	*k*	2
12	*l*	2
13	*m*	2

对其进行标准化，可得密度缩放因子：

$$\delta_i = \exp\left(\frac{\gamma_i - \bar{\gamma}}{s_\gamma}\right)$$

其中，$\bar{\gamma}$ 为样本均值，且 $\bar{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\gamma_i$；$s_\gamma$ 为样本方差，且 $s_\gamma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\gamma_i - \bar{\gamma})^2}$；$n$ 为样本的数量。标准化后的密度缩放因子如表 4 所列。

表 4 标准化后的密度缩放因子

Table 4 Standardized density scaling factor

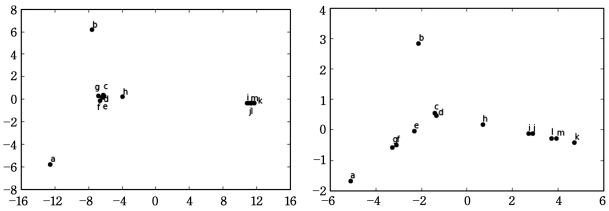
序号	观测样本	密度缩放因子
1	*a*	0.16
2	*b*	0.16
3	*c*	1.73
4	*d*	1.73
5	*e*	1.73
6	*f*	1.73
7	*g*	1.73
8	*h*	0.16
9	*i*	1.73
10	*j*	1.73
11	*k*	1.73
12	*l*	1.73
13	*m*	1.73

根据得到的密度缩放因子，测地距离 d_{ij} 可表示为：

$$d_{ij} = \begin{cases} \dfrac{w_{ij}}{\delta_i \delta_j}, & \text{if } e_{ij} = 1 \\ \min\left\{\dfrac{w_{ij}}{\delta_i \delta_j}, \dfrac{w_{il}}{\delta_i \delta_l} + \dfrac{w_{lj}}{\delta_l \delta_j}\right\}, & \text{otherwise} \end{cases} \tag{9}$$

由式(9)可知，原始的测地距离已经通过密度缩放因子实现了缩放。对于密度缩放因子乘积大于 1 的两个点，它们之间的测地距离会变小，乘积小于 1 的两点之间的测地距离会变大。相比于原测地距离，密度大的区域样本更加集中，密度小的区域样本更加疏远，这样就保持了原始数据在空间中的几何形状和密度分布，减小了噪声的干扰。

为了验证添加密度缩放因子后测地距离的缩放效果和对噪声的鲁棒性，用 ISOMAP 和 D-ISOMAP 将图 1 中的空间点降到 2 维进行可视化展示，分别如图 2(a)和图 2(b)所示，图中横纵轴表示样本点在 2 维平面上的坐标值。由图 2 可以看出，ISOMAP 算法的降维效果明显受到了噪声点的影响，其拓扑结构被破坏。而采用 D-ISOMAP 算法降维后仍然保持了原来的拓扑结构，各聚簇类区分清晰，对噪声点具有鲁棒性，能提高算法的可视化和聚类效果。



(a)ISOMAP 降维效果 (b)D-ISOMAP 降维效果

图 2 空间点的降维效果图

Fig. 2 Dimension reduction results of space points

D-ISOMAP 算法的具体操作过程如算法 1 所示。

算法 1 D-ISOMAP 算法

输入：样本集 $D=\{x_1, x_2, \cdots, x_m\}$，邻近参数 k，低维空间维数 d′

输出：样本集 D 在低维空间的投影 $Z=\{z_1, z_2, \cdots, z_m\}$

过程：

Step 1 for i= 1,2,…,m do

Step 2 确定 x_i 的 k 邻近；

Step 3 计算 x_i 的密度缩放因子 δ_i；

Step 4 将 x_i 与 k 邻近点之间的距离设置为欧氏距离除以密度缩放因子的乘积，将与其他点的距离设置为无穷大；

Step 5 end for

Step 6 调用最短路径算法计算任意两样本点间的距离 $dist(x_i, x_j)$；

Step 7 将 $dist(x_i, x_j)$ 作为 MDS 算法的输入；

Step 8 return MDS 算法的输出

4 实验结果与分析

为了验证 D-ISOMAP 算法的性能，本文的目标是将原始数据映射到一个维度较低的特征空间中，通过 K-Means 算法在低维特征空间中完成聚类任务，并将其聚类效果与 MDS，LLE，KPCA，PAC 和 ISOMAP 这些经典的无监督降维算法进行比较；同时，与目前最好的无监督聚类方法——Density Peaks Clustering[13] 进行实验对比，以验证所提算法是否能提升聚类效果；另外，通过展示一些可视化效果图来表明聚类效果提高的原因以及实验中不同强弱的噪声对降维后聚类效果的影响。实验编程环境为 Python 2.7。

4.1 数据集说明

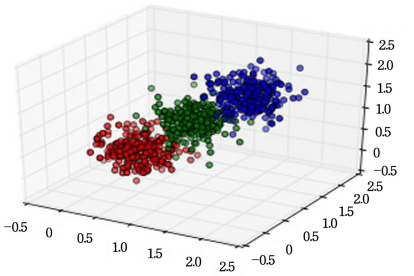
实验采用的数据集包括人工数据集和真实数据集，这些数据集的属性均为数值型。人工数据集如图 3(a)和图 3(b)所示，真实数据集来源于 UCI 机器学习数据库中的 6 个数据集，数据集去除了带有缺失属性值的样本。经过处理后的数据集大小及属性个数如表 5 所列。

表 5 数据集信息

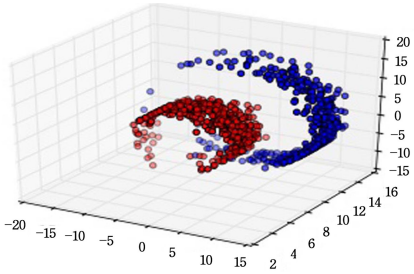
Table 5 Information of datasets

数据集	大小	属性	类别
Synthetic Control Chart	600	60	6
Ionosphere	351	33	2
Banknote	1372	4	2
Semeion handwritten	1593	256	10
Vertebral Column	310	6	2
Breast Cancer	569	30	2
Gaûss	750	3	3
Swiss roll	800	10	2

构造的人工数据集均是带有噪声干扰的数据，图 3(a)为 Gauss 数据集，由 3 类带有高斯噪声的三维点构成，每类有 250 个样本点；图 3(b)为流形学习的典型数据集 Swiss roll，本文构造了 2 类，每类 400 个三维样本点。选择的 UCI 数据集是在真实情况下采集的数据，难免会带有噪声。本文通过这些数据集来验证所提算法的性能。



(a)Gauss



(b)Swiss roll

图 3　人工数据集

Fig. 3　Artificial datasets

4.2　实验设置

为清楚地了解算法在降维后对数据的聚类效果和噪声对降维算法的影响，通过以下 3 类实验来进行对比验证：1)降维前后的聚类效果实验；2)可视化效果实验；3)不同噪声强度对降维算法的影响。使用 Rand Index(RI)指标来对比评估降维后聚类实验的效果，该指标用于衡量所得结果与真实结果的相似程度。Rand Index 的输入是真实的类标签 π 和聚类后的类标签 C，RI 的计算式如下：

$$RI(\pi,C)=\frac{n_{00}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}=\frac{2(n_{00}+n_{11})}{n(n-1)} \tag{10}$$

其中，n_{00} 为类标签和聚类标签都相同的观察样本对数；n_{01} 为类标签不同但聚类标签相同的观察样本对数；n_{10} 为类标签相同但是聚类标签不同的观察样本对数；n_{11} 为类标签和聚类标签都不同的观察样本对数。兰德系数的值介于 0 到 1 之间，0 表示聚类结果的标签与真实类样本标签完全不匹配，1 表示聚类结果的标签与真实类标签完全一致。兰德系数的值越接近 1，表明聚类结果与真实结果越接近。

实验中有两个参数需要确定：构建 k 邻近图中的 k 值和降维的维数。对于所有数据集，均降至 2 维进行实验。k 值的大小在 3 到 30 这个区间进行寻参，最终选取最优的 k 值进行聚类分析。

4.3　结果与分析

在 8 个数据集上将 D-ISOMAP 算法与 5 个经典的无监督降维算法进行降维聚类实验，并与目前无监督聚类效果最好的方法——Density Peaks Clustering 进行实验对比，不同算法在每个数据集上的聚类结果如表 6 所列。

表 6　不同算法在 8 个数据集上的聚类效果对比

Table 6　Comparison of clustering performance in terms of Rand index on eight datasets

数据集	Rand index						
	MDS	LLE	KPCA	PCA	ISOMAP	D-ISOMAP	Density Peaks Clustering
Synthetic Control Chart	0.84	0.86	0.86	0.85	0.86	**0.88**	0.77
Ionosphere	0.55	**0.73**	0.59	0.58	0.57	0.70	0.60
Banknote	0.51	0.79	0.52	0.52	0.64	**0.84**	0.65
Semeion handwritten	0.83	0.87	0.86	0.86	0.87	**0.89**	0.87
Vertebral Column	0.57	0.50	0.55	0.57	0.59	**0.66**	0.63
Breast Cancer	0.77	**0.83**	0.77	0.77	0.76	0.79	0.67
Gauss	0.99	0.96	0.88	0.90	0.99	**1.00**	0.99
Swiss roll	0.66	0.81	0.66	0.66	0.74	0.99	**1.00**

通过表 6 可以看出，文中所提 D-ISOMAP 算法对 8 个带噪声的数据进行降维，其中 6 个数据的聚类效果优于其他降维算法，剩下的 2 个数据集的效果与最优 LLE 算法的效果相差也不大。与传统的 ISOMAP 算法相比，D-ISOMAP 算法的聚类效果有明显提升，整体来看，降维后的聚类效果相对于未降维的 Density Peaks Clustering 聚类效果会有一定提升。这是因为降维后数据在子空间中的密度会变大，使得聚类效果得到提升；而且 Density Peaks Clustering 算法的聚类中心点是人工确定的，对于带噪声的非线性数据集而言，该聚类算法的聚类效果会受到影响。实验表明，对于带有噪声的数据集，D-ISOMAP 算法的整体效果优于或等价于一般的降维算法。在实际情况下，数据集是非线性的并且常带有噪声，这说明文中所提出的 D-ISOMAP 算法的适用范围更广。

根据以上聚类实验分析，文中所提出的 D-ISOMAP 算法在 8 个不同数据集中的聚类效果均较好。本文分别选取 Synthetic Control Chart 数据集、Banknote 数据集和两个人工数据集作为代表，使用 D-ISOMAP 算法、ISOMAP 算法、LLE 算法和 KPCA 算法分别将这些数据集映射到二维特征空间中来可视化数据的情况，实验结果如图 4—图 7 所示。由图 4—图 7 可以看出，对于实验的 4 个数据集，KPCA 算法降维后的数据比较分散，不同流体间的重叠部分也比较大。而 ISOMAP 算法和 LLE 算法降维后的数据虽然并没有分散，但是受到噪声的影响，不同流体间的区分度不佳。从图中可以清晰地看出，所提 D-ISOAMP 算法降维后能较好地聚合数据，不同流体间的数据重叠也较少。由于都是无监督降维算法，降维后存在重叠现象是很正常的，相比于其他无监督降维算法，所提算法已经取得了更好的聚类结果。

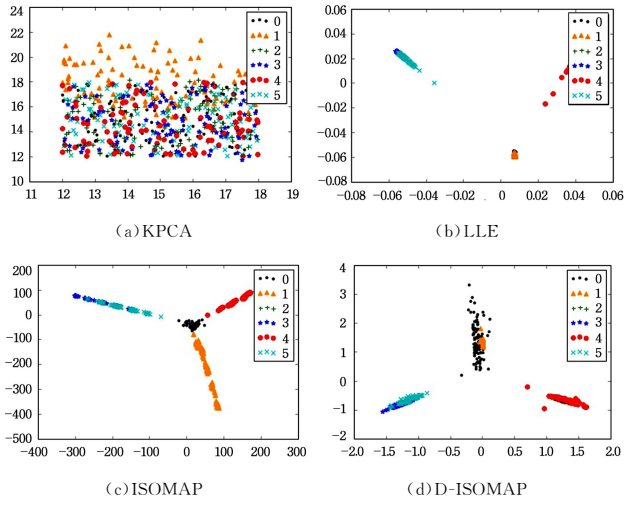


图 4 Synthetic Control Chart 数据集的二维嵌入图

Fig. 4 Two-dimensional embedding results of Synthetic Control Chart dataset

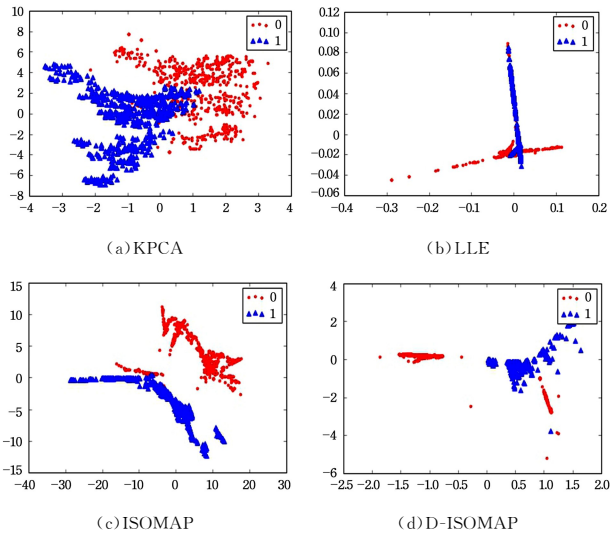


图 5 Banknote 数据集的二维嵌入图

Fig. 5 Two-dimensional embedding results of Banknote dataset

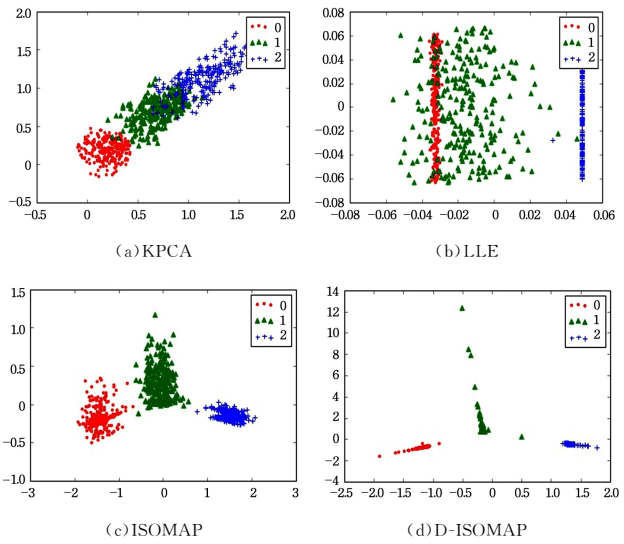


图 6 Gauss 数据集的二维嵌入图

Fig. 6 Two-dimensional embedding results of Gauss dataset

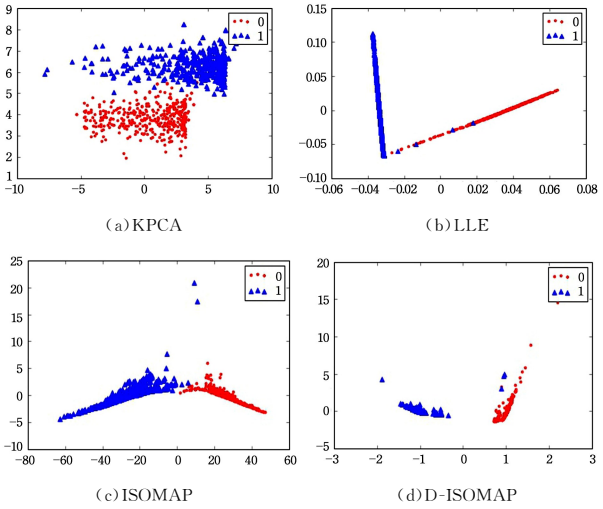


图 7 Swiss roll 数据集的二维嵌入图

Fig. 7 Two-dimensional embedding results of Swiss roll dataset

下面通过不同强度的噪声来验证 D-ISOMAP 算法对噪声的鲁棒性。在图 3(b)所示的数据集上分别加上 1～5 倍的均值为 0、标准差为 0.5 的高斯噪声。将不同算法在这 5 类噪声强度下进行降维聚类。实验中 k 值取 6，数据维数降到 2。聚类结果如表 7 所列。

表 7 不同降维算法在不同噪声强度的 Swiss roll 数据集上的聚类效果

Table 7 Comparison of clustering performance in terms of Rand index on different noise intensity for Swiss roll dataset

噪声倍数	Rand index					
	MDS	LLE	KPCA	PCA	ISOMAP	D-ISOMAP
1	0.66	0.99	0.66	0.66	0.97	**0.99**
2	0.65	0.74	0.66	0.65	0.76	**0.85**
3	0.63	0.63	0.65	0.63	0.64	**0.84**
4	0.61	0.62	0.63	0.63	**0.64**	**0.64**
5	0.59	0.50	0.60	0.58	**0.59**	0.56

由表 7 可以看出，对于噪声，D-ISOMAP 算法具有较好的鲁棒性，ISOMAP 与 LLE 受噪声的影响较大。原因在于，ISOMAP 算法中的测地距离对噪声敏感，其计算很大程度上会受到噪声的影响；而 LLE 算法是基于保持样本间局部线性重构关系指导数据降维的，噪声的存在很容易使得这种线性关系受到影响；D-ISOMAP 算法使用了密度缩放因子，使得高密度区域的数据更加集中，低密度区域的数据更加稀疏，减小了噪声的影响，从而在降维后保证了数据间的距离，使聚类效果更优。

结束语 本文针对传统的 ISOMAP 算法存在的不足，提出了一种基于密度缩放因子的 ISOMAP 算法，即 D-ISOMAP。该算法从数据密度分布的角度出发，在计算距离度量时，通过密度缩放因子对原始的测地距离进行放缩，使降维后的数据在密度大的区域更加紧凑，在密度小的区域数据更加疏散，减小了噪声对降维的影响。与其他经典降维算法相比，D-ISOMAP 算法对噪声的存在相对鲁棒，并且在可视化和聚类任务上都能达到较好的性能。

但是由于 D-ISOMAP 算法是无监督的降维算法，对于数

据在空间中有重叠的情况，其与其他无监督降维算法一样，也不能将降维后的不同聚簇类分隔开来。另外，在时间复杂度上，所提算法在计算样本间的最短路径和进行特征分解这两步上所花费的时间比其他算法多，时间复杂度较高。对于如何在考虑密度的情况下通过有监督方法改进 ISOMAP 算法，使之有利于数据分类任务，并且降低算法的时间复杂度，是我们后续继续研究的问题。

参 考 文 献

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

[2] ZHANG T，TANG Y Y，FANG B，et al. Document Clustering in Correlation Similarity Measure Space[J]. IEEE Transactions on Knowledge & Data Engineering，2012，24(6)：1002-1013.

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

[4] DONOHO D L. High-Dimensional Data Analysis：The Curses and Blessings of Dimensionality[C]//Lecture — Math Challenges of Century. 2000：178-183.

[5] ROWEIS S T，SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science，2000，290(5500)：2323-2326.

[6] HÄRDLE W K，HLÁVKA Z. Principal Component Analysis [J]. IEEE Trans on Automatic Control，2015，29(1)：163-183.

[7] AMARI S I，CICHOCKI A. A new learning algorithm for blind source separation[C]//Advances in Neural Information Processing Systems. 1996：757-763.

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

[9] COX M AA，COX T F. Multidimensional Scaling[J]. Journal of the Royal Statistical Society，2008，46(2)：1050-1057.

[10] TENENBAUM J B，SILVA V D，LANGFORD J C. A Global Geometric Framework for Nonlinear Dimensionality Reduction [J]. Science，2000，290(5500)：2319-2323.

[11] YANG J，FRANGI A F，YANG J，et al. KPCA plus LDA：a complete kernel Fisher discriminant framework for feature extraction and recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2005，27(2)：230-244.

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

[13] RODRIGUEZ A，LAIO A. Clustering by fast search and find of density peaks[J]. Science，2014，344(6191)：1492-1496.

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

[15] ZHANG Z，CHOW T W S，ZHAO M. M-Isomap：Orthogonal constrained marginal isomap for nonlinear dimensionality reduction[J]. IEEE Transactions on Cybernetics，2013，43(1)：180-191.

[16] YANG B，XIANG M，ZHANG Y. Multi-manifold Discriminant Isomap for visualization and classification[J]. Pattern Recognition，2016，55：215-230.

[17] HE J R，DING L X，LI Z K，et al. Margin Discriminant Projection for Dimensionality Reduction[J]. Journal of Software，2014，25(4)：826-838. (in Chinese)
何进荣，丁立新，李照奎，等. 基于边界判别投影的数据降维[J]. 软件学报，2014，25(4)：826-838.

[18] BERNSTEIN M，SILVA V D，LANGFORD J C，et al. Graph Approximations to Geodesics on Embedded Manifolds[J]. Stanford University，2001，24(9)：153-158.

[19] MAIER M，LUXBURG U V，HEIN M. Influence of graph construction on graph-based clustering[C]//Conference on Neural Information Processing Systems. Vancouver，British Columbia，Canada，DBLP，2008：1025-1032.

[20] FU Y，MA Y. Graph Embedding for Pattern Analysis[M]. New York：Springer，2013.

[21] LEE J S，OLAFSSON S. Data clustering by minimizing disconnectivity[J]. Information Sciences，2011，181(4)：732-746.

(上接第 206 页)

[12] ZHANG L，REN X F，SHI K Q. Intelligent switch-camouflage of information laws and P-law augmented matrices [J]. Journal of Shandong University(Natural Science)，2016，51(8)：90-97. (in Chinese)
张凌，任雪芳，史开泉. 信息规律智能变换-伪装与 P-规律增广矩阵[J]. 山东大学学报(理学版)，2016，51(8)：90-97.

[13] LU Y，LI B Y，SHI K Q. Intelligent mining for P-information with expanded characteristic of disjunctive attribute [J]. Computer Science，2015，42(1)：215-219. (in Chinese)
路英，李备友，史开泉. 具有属性析取扩展特征的内逆 P-信息智能挖掘[J]. 计算机科学，2015，42(1)：215-219.

[14] SHI K Q，YAO B X. Function S-rough sets and law identification [J]. Science in China F：Information Science，2008，51(5)：499-510.

[15] PAL S K，MEHER S K，DUTTA S. Class-dependent rough-fuzzy granular space，dispersion index and classification[J]. Pattern Recognition，2011，45(7)：2690-2707.

[16] ANDRZEJ S，JAROSLAW S，ROMAN S. Modeling rough granular computing based on approximation space [J]. Information Science，2011，184(1)：20-43.

[17] WITOLD P，BARBARA R，GIANCARLO S. Knowledge transfer in system modeling and its realization through an optimal allocation of information granularity [J]. Applied Soft Computing，2012，12(8)：1985-1995.

[18] 张文修，姚一豫，梁怡. 粗糙集与概念格[M]. 西安：西安交通大学出版社，2006.