

基于范畴的数据降维方法

周丽丽 李凡长

(苏州大学计算机科学与技术学院 苏州 215006)

摘 要 范畴理论主要是一些特定数学的对象和映射的概括和抽象,在此利用范畴理论阐述图像分析和识别中的数据降维问题,定义高维数据降维范畴的过程,并以主成分分析范畴和等距映射范畴分别验证了范畴理论应用到图像数据降维问题中的正确性。

关键词 范畴理论,数据降维,线性降维范畴,非线性降维范畴

中图法分类号 TP181 **文献标识码** A

Data Dimension Reduction Based on Category Theory

ZHOU Li-li LI Fan-zhang

(Department of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract The research of category theory is mainly about summary and abstraction of some specific mathematical objects and mapping. The dimensionality reduction problem based on category theory was discussed, and it could solve the problems of image analysis and image recognition. Besides, the process of dimensionality reduction based on category theory was listed. There are two examples, Principal Component Analysis Category and Isomap Category, which verified the correctness of the category theory applied to the dimensionality reduction problem.

Keywords Category theory, Data dimension reduction, Linear dimension reduction category, Nonlinear dimension reduction category

1 引言

近年来,图像识别和图像分析等问题在模式识别和人工智能领域备受关注。经过近半个世纪的发展,图像识别和分析研究领域已经取得了许多成果,基本实现了特定环境下的准确识别。其中,文献[1]针对人脸识别中的高维数据降维及识别问题展开了深入的研究。尽管如此,图像处理技术要达到完全使用水平,还面临着诸多挑战,图像维数过高就是问题之一。

从目前取得的成果来看,图像高维数据的降维算法大致可以分为两类,一类是线性的方法,如主成分分析(PCA)、多维尺度变换(MDS)和独立主成分分析(ICA);另一类是非线性的方法,最有代表性的有等距映射(Isomap)、局部线性嵌入(LLE)、拉普拉斯映射(LE)和局部切空间排列(LTSA);文献[2]分别列出了各种降维算法的具体内容。这些算法虽然很多已经得到了广泛应用,但仍然有各种局限性,对于图像高维数据的降维仍然达不到理想的效果。本文利用范畴理论这一抽象数学工具,结合各种降维算法的相似处和不同点,抽象概括出一种基于范畴论的高维数据降维算法,该方法是一种鲁棒性能和灵活性很好的方法,能够为今后的研究起到指导作用。

2 数据降维范畴

在文献[3]中,我们已经证明了机器学习系统可以用范畴来表示,各种学习算法也可以用范畴的理论表示。由此可知图像数据的降维算法也可以用范畴表示,同理,在数据降维范畴中,可以根据各种对象和态射之间的关系得到图像数据对象到低维空间对象的态射关系,从中可以推导出数据集对象的聚类和空间分布等结构特性;另一方面,从数据降维范畴中也可以得到数据预处理的效果,将高维数据对象转化为比较容易处理的低维数据对象,可以降低各种数据对象之间的态射复杂度。

假设数据降维范畴(Data Reduction Category)为 $\mathcal{C}$,存在一族数据对象(Data Object) $ob\mathcal{C}$,其中,已知对象 $X \in \mathcal{C}$ 为原始高维数据集对象,且存在一族态射(Morphism),使任意一对对象 A, B 都对应一个集合 $\mathcal{C}(A, B)$ 。

数据降维范畴的主要目的是寻求数据流形对象的合适的低维坐标描述,即是将原始高维数据集对象由各种态射映射到低维空间,得到低维数据集对象,获得原始高维数据集对象的低维简洁表示。

因此,高维数据降维问题通过数据降维范畴 $\mathcal{C}$ 可表示为:

1) 原始高维数据对象 $X: \Omega^D \subset X$, 其中, Ω^D 为数据样本集

到稿日期:2010-11-20 返修日期:2011-01-27 本文受东吴学者计划(14317360),国家自然科学基金项目(61033013),国家自然科学基金项目(60775045)资助。

周丽丽(1986—),女,硕士,主要研究方向为智能化信息处理;李凡长(1964—),男,教授,博士生导师,主要研究方向为机器学习和动态模糊逻辑, E-mail: lfzh@suda.edu.cn(通信作者)。

对象, $\Omega^D = \{x_n\}_{n=1}^N \subset \mathbb{R}^D$, D 为数据的维数。

2) 降维后的低维数据对象 Y , 即数据的低维表示空间: $\Omega^d \subset Y$, 其中, Ω^d 为高维数据样本集对象 Ω^D 通过一系列态射、映射得到的低维数据对象; $d \leq D$ 为数据降维后的维数。

3) 数据降维态射 $f: X \rightarrow Y, f \in \mathcal{C}(X, Y), X, Y \in \text{ob}\mathcal{C}$ 。以样本数据对象为例可记为: $f(\Omega^D): \Omega^D \rightarrow \Omega^d$, 称 Ω^d 为 Ω^D 的降维表示。

同理, 我们也可以由数据降维范畴 $\mathcal{C}$ 得到相应的数据恢复过程, 即由降维后的低维数据对象 Y , 经过相应态射的变换, 得到原有高维数据对象 X , 表示为如下态射:

数据恢复态射, $f^{-1} \in \mathcal{C}(Y, X), X, Y \in \text{ob}\mathcal{C}$ 。以样本数据对象降维后的对象来表示可记为: $f^{-1}(\Omega^d): \Omega^d \rightarrow \Omega^D$ 。

需要说明的是数据恢复态射 f^{-1} 需要满足一定的条件, 当条件不满足时, 数据恢复态射就没有意义。一般在线性的数据降维范畴 $\mathcal{C}$ 中, 数据恢复态射 f^{-1} 是有意义的, 而在非线性的数据降维范畴中, 数据恢复态射 f^{-1} 无意义。

在数据降维范畴 $\mathcal{C}$ 中, 寻找数据降维态射 f 是整个范畴形成的关键, 然而此数据降维态射 f 可以有多种形式, 目前从数学思想的角度来讲主要有线性表示的态射和非线性表示的态射之分。因此, 在描述数据降维范畴 $\mathcal{C}$ 时, 可以从这两个角度出发。

当数据降维范畴 $\mathcal{C}$ 中的数据降维态射选用线性表示时, 可以将基于数据降维范畴的降维算法描述如下:

输入: 原始高维数据对象 X

输出: 低维坐标数据对象 Y

1) 构造关联对象 M 。态射 g 为构造关联对象的过程, $g: X \rightarrow M$ 。其中关联对象 M 和态射 g 需要根据高维数据对象 X 和已知关系对象 Δ 确定, 不同的关系对象 Δ 对应不同的关联对象 M , 得到最终输出的低维坐标数据对象 Y 的表示意义不同。

2) 计算关联对象 M 的特征值对象 Z 和特征向量对象 W 。 $h: M \rightarrow Z, \varphi: Z \rightarrow W$, 根据数据降维范畴的复合运算律 (composition law) 可得 $\varphi h: M \rightarrow W$ 。

3) 分析数据降维范畴 $\mathcal{C}$ 中已知对象的特征和关系, 求得低维坐标数据对象 Y 。其中需要自定义算子 $\cdot$ 来确定中间过渡对象 T , 使得 $\varphi: T \rightarrow Y$ 。

将数据降维范畴 $\mathcal{C}$ 中的降维算法用图表交换的形式表示, 如图 1 所示。从图 1 中可以直观地看到数据降维范畴中各对象及其态射之间的关系, 以及由已知的对象和一系列构造的态射得到所求对象的过程, 完成了算法从输入到输出的一个流程。

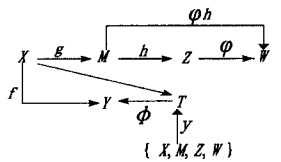


图 1 数据降维范畴表示的线性降维算法

当数据降维范畴 $\mathcal{C}$ 中的数据降维态射选用非线性表示时, 可以将基于数据降维范畴的降维算法描述如下:

输入: 原始高维数据对象 X

输出: 低维嵌入数据对象 Y

1) 确定原始高位数据对象 X 中各元素的邻域点, 构造邻

域对象 G 。态射 l 为计算各个元素邻域点, 构造邻域对象的过程, $l: X \rightarrow G$ 。

2) 构造矩阵对象 H 。邻域对象 G 中包含原始高位数据对象 X 中各元素及其属性和关系 (例如距离或权值等), 利用邻域对象 G 中已知元素, 寻找合适的态射 p 映射得到矩阵对象 $H, p: G \rightarrow H$ 。其中矩阵对象 H 的意义与所选邻域对象 G 中元素的属性有关。

3) 根据所得矩阵对象 H 分析计算低维嵌入数据对象 Y 。利用态射 ψ 完成从 H 到 Y 的映射, 即: $\psi: H \rightarrow Y$ 。其中态射 ψ 的形式及其方法取决于矩阵对象 H 的意义, 不同的 ψ 的过程需要构造不同的对象和态射。

同上, 我们也可以将数据降维范畴 $\mathcal{C}$ 中的非线性降维算法用图表交换的形式表示, 如图 2 所示。图 2 中从对象 H 到对象 Y 的态射中有虚线箭头连接, 表示态射 $\psi \in \mathcal{C}(H, Y)$ 是由复合运算律和结合律等态射的性质和运算得到的。

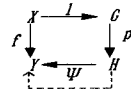


图 2 数据降维范畴表示的非线性降维算法

3 算法验证

为了验证上一节介绍的数据降维范畴 $\mathcal{C}$ 及其降维过程的正确性和合理性, 利用已有算法进行验证。选择线性降维的经典算法主成分分析算法 (Principal Component Analysis) 和非线性降维算法等距映射法 (Isomap) 进行验证。

例 1 假设主成分分析算法所在的范畴定义为主成分分析范畴 $\mathcal{C}_1$, 已知原始高维数据对象记为 X , 选择对象 X 样本的向量表示 $\{x_1, x_2, \dots, x_m\} \subset X$ 。为了使原始数据中心化, 优化对象 X 中的原始数据, 将对象 X 映射为优化后的数据对象 X' , 由态射 g_1 表示, $g_1(x) = x_i - \frac{1}{m} \sum_{i=1}^m x_i, i=1, \dots, m$, 即:

$$g_1: X \rightarrow X' \quad (1)$$

构造关联矩阵对象 M , 此处的关联矩阵对象为协方差矩阵对象, 由协方差态射 g_2 得到, $g_2(x) = \frac{1}{m} \sum x_i' (x_i')^T$, 其中, $x_i' \in X'$, 即:

$$g_2: X' \rightarrow M \quad (2)$$

求关联矩阵对象 M 的特征值对象 Z , 态射 h 可由 $\det(M - \lambda I) = 0$ 变换得到, 其中, $\lambda_i \in Z$ 。可得:

$$h: M \rightarrow Z \quad (3)$$

求关联矩阵对象 M 的特征向量对象 W , 态射 φ 可由 $Mx_i = \lambda_i w_i$ 变换得到, 其中, $\lambda_i \in Z, w_i \in W, Mx_i$ 为关联矩阵对象 M 的元素。可得:

$$\varphi: Z \rightarrow W \quad (4)$$

定义算子 $\cdot, \gamma: Z \cdot W \rightarrow T$, 其中, $\lambda_i \cdot w_i = \sqrt{\lambda_i} w_i$ 。由特征向量对象 W 的正交特征, 选择前 d 个特征值对象中的元素及其对应的特征向量对象中的元素, 利用态射 ϕ 得到低维数据表示对象 $Y, \phi = [\sqrt{\lambda_1} w_1, \sqrt{\lambda_2} w_2, \dots, \sqrt{\lambda_d} w_d]^T, \sqrt{\lambda_i} w_i \in T$ 。

$$\phi: T \rightarrow Y \quad (5)$$

图 3 展现了主成分分析范畴 $\mathcal{C}_1$ 中各对象和态射的关系, 以及原始数据降维的过程, 并与数据降维范畴表示的线性降维算法做比较。

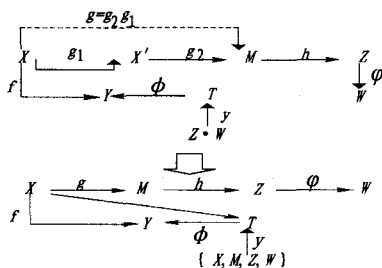


图3 基于主成分分析范畴的数据降维过程

例2 假设等距映射算法所在的范畴定义为等距映射范畴 $\mathcal{C}_2$,已知原始高维数据对象,记为 X ,选择对象 X 样本的向量表示 $\{x_1, x_2, \dots, x_m\} \subset X$ 。首先确定原始高位数据对象 X 中各元素的邻域点,构造邻域对象 G ,利用求距离态射 $l_1 = d_E(i, j)$ 确定各元素间的距离, $l_1: X \rightarrow D$,对象 D 为距离对象。取最近的 k 个元素作为邻域,通过态射 l_2 构造无向图对象 G_0 ,即: $l_2: X \rightarrow G_0$ 。结合距离对象和无向图对象,可得近邻图对象 G ,其中对象 G_0 中的各连接线的权值为对象 D 中对应的元素,可以表示为:

$$l_3: D \cdot G_0 \rightarrow G \quad (6)$$

构造矩阵对象 H ,通过对象 G 求得 X 中每个元素之间的测地线距离,组成矩阵对象 H ,转换态射 p 可以利用dijkstra算法得到,使得 $p: G \rightarrow H$,其中, $p = d_G(i, j) = \min_{1 \leq k \leq m} \{d_G(i, j), d_G(i, k) + d_G(k, j)\}$ 。

对构造矩阵对象求特征值对象 Z 和特征向量对象 W ,最后根据对象 Z 和对象 W 得到低维嵌入数据对象 Y ,如例1所述,可得:

$$\phi_1: H \rightarrow Z \quad (7)$$

$$\phi_2: Z \rightarrow W \quad (8)$$

$$\phi_3: Z \cdot W \rightarrow Y \quad (9)$$

图4表示等距映射范畴 $\mathcal{C}_2$ 中各对象和态射的关系,以及原始数据降维的过程,并与数据降维范畴表示的非线性降维算法做比较。

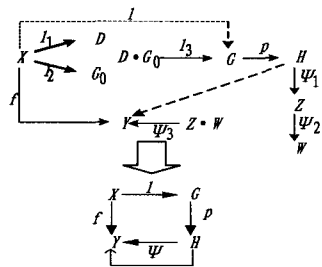


图4 基于等距映射范畴的数据降维过程

4 实例及结果

以数据降维范畴中的非线性降维算法为模型,经分析得到新的数据降维算法,并以其为例,验证了数据降维范畴的可行性和正确性。

首先确定态射 l ,求得邻域对象 G ,具体态射过程利用基于中心点的邻域选择算法完成,其中邻域对象 G 包含 m 个子集集合 $\bigcup_{i=1}^m S_i$;然后根据邻域对象 G 构造矩阵对象 H ,主要利用样本协方差;最后由矩阵对象 H 中各元素的特征值对象和特征向量对象,经态射 ϕ 得到低维嵌入数据对象 Y 。

现以Frey_rawface数据库作为原始高维数据对象 X ,经过数据降维范畴的处理,得到二维降维数据对象 Y ,如图5所

示。图6表示等距映射范畴的降维效果图,可以看到人脸按照一定的规律变化,充分说明了数据降维范畴的可行性和正确性。

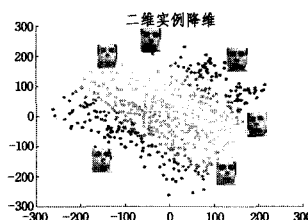


图5 Frey_rawfac数据实例降维效果图

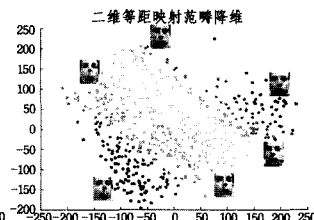


图6 Frey_rawfac数据等距映射范畴降维效果图

以ORL标准人脸数据库为原始高位数据对象 X 做进一步的验证和对比,选取40个人中的5个人的人脸数据作为输入数据,每一个人选取前5张图片,共25张图片作为样本数据,其余25张作为测试数据。分别通过数据降维范畴实例和等距映射范畴降维得到的效果如图7和图8所示。不同的图形不同的颜色表示不同的人脸。

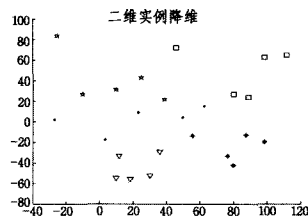


图7 ORL数据实例降维效果图

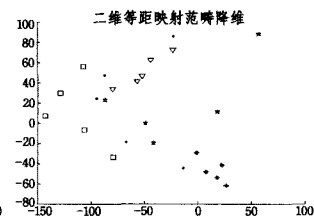


图8 ORL数据等距映射范畴降维效果图

分别利用数据降维范畴的实例算法和等距映射范畴对Frey_rawfac数据集和ORL数据集进行降维,对降维后的结果用半监督学习算法进行识别,得到正确识别率的结果,如图9所示。

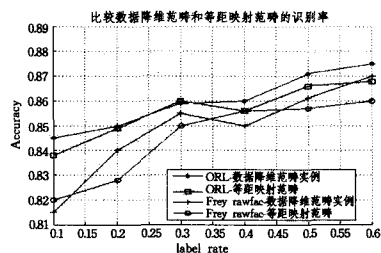


图9 比较数据降维范畴实例和等距映射范畴利用半监督学习方法得到的识别率

结束语 综上所述,本文利用范畴理论详细阐述了图像分析和识别中的数据降维问题,给出了高维数据降维范畴的具体过程,并且分别以线性降维范畴和非线性降维范畴来说明了这个过程。为了验证高位数据降维范畴的正确性和可行性,本文以主成分分析范畴和等距映射范畴为例加以说明,充分验证了范畴理论应用到图像数据降维问题中的正确性。在以后的研究中,我们将利用范畴中的函子及其性质对学习问题做进一步研究。

参考文献

- [1] 刘翠响. 人脸识别中高位数据特征分析[D]. 河北工业大学, 2008

(下转第247页)

离。

根据模式序列的动态时间弯曲距离可以求任意两个长度的模式序列间的距离。模式序列间的动态时间弯曲距离满足正定性、对称性和三角不等式性,因而,可用它定义两个模式序列的相似性。

定义 4 压缩时间序列 $X=(x_1, x_2, \dots, x_n)$ 和 $Y=(y_1, y_2, \dots, y_m)$ 之后得到的模式序列分别为:

$$S_X=\{M_1, M_2, \dots, M_t\}$$

$$S_Y=\{N_1, N_2, \dots, N_p\}$$

称 $D_{dwt}(S_X, S_Y)$ 为模式序列 S_X 和 S_Y 的相似性函数,即有:

$$\text{Sim}(S_X, S_Y)=D_{dwt}(S_X, S_Y)$$

在判断两个模式序列是否相似时,只需要计算出它们之间的动态时间弯曲距离,再给定正常数 ϵ ,如果相似性函数值小于 ϵ 时,那么它们就相似,否则它们不相似。

定义 5 把时间序列 $X=(x_1, x_2, \dots, x_n)$ 转换成模式序列,结果分别为:

$$S_X=\{M_1, M_2, \dots, M_t\}$$

任意序列模式 $S_{X1}, S_{X2} \in S_X, S_{X1}=\{M_j, M_{j+1}, \dots, M_{j+l_1}\}, S_{X2}=\{M_i, M_{i+1}, \dots, M_{i+l_2}\}$, 称 $D_{dwt}(S_{X1}, S_{X2})$ 为序列模式 S_{X1} 和 S_{X2} 的相似性函数,即有:

$$\text{Sim}(S_X, S_Y)=D_{dwt}(S_X, S_Y)$$

式中, $1 \leq i, i+l_1 \leq t, 1 \leq j, j+l_2 \leq t$ 。

4 仿真实验

用 2006 年沪深 300 指数的全部数据,以及 2007 年沪深 300 指数的全部数据作为实验对象,如图 2、图 3 所示。分别把它们转换成模式序列,分别含有 80 和 88 个元模式,应用模式序列动态时间弯曲距离方法计算它们的距离为 28.4。因而这两个时间序列的模式序列并不具有较好的相似性,这是因为 2006 年沪深 300 指数温和增长,而到了 2007 年几乎全民进入股市,促使 2007 年沪深 300 指数急剧增长,使它们具

有很大的差异性。

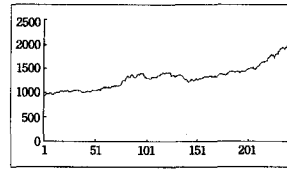


图 2 2006 年沪深 300 指数

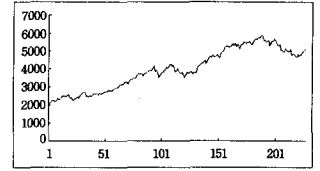


图 3 2007 年沪深 300 指数

参考文献

- [1] 翁颖钧,朱仲英.基于分段线性动态时间弯曲的时间序列聚类算法研究.研究与设计[J].微型电脑应用,2003,19(9)
- [2] Berndt D J, Clifford J. Finding patterns in time series; DB, 2000. A dynamic programming approach[Z]. Advances in Knowledge Discovery and Data Mining, 1996
- [3] Berndt J, Clifford D. Using dynamic time warping to find patterns in time series[C]// AAAI-94 Workshop on Knowledge Discovery in Database. 1994; 229-248
- [4] 曲文龙,张德政,杨炳儒.基于小波和动态时间弯曲的时间序列相似匹配[J].北京科技大学学报,2006,28(4)
- [5] Berndt D, Clifford J. Using dynamic time warping tp find patterns in time series[C]// AAAI Workshop on Knowledge Discovery in Databases. 1994; 229-248
- [6] Aach J, Church G. Alidning gene expression time series with time warping algorithms[J]. Bioinformatics, 2001, 17
- [7] Kim S W, Park S, Chu W W. Efficient processing of similarity search under time warping in sequence databases; an index-based approach[J]. Inf. Syst., 2004, 29(5); 405-420
- [8] Das G, Lin K I, Marmila H, et al. Rule Discovery from Time Series [A]// Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining[C]. [S. l.]: AAA I Press, 1998; 16-22
- [9] 张保稳. 时间序列数据挖掘研究[D]. 西安: 西北工业大学, 2002
- [10] 王亮,姜丽红. 快速挖掘最大频繁模式算法[J]. 计算机工程与应用, 2006(17)
- [11] 周勇. 时间序列时序关联规则挖掘研究[D]. 成都: 西南财经大学, 2008(6); 40-60

(上接第 229 页)

- [10] Sim K M, Guo Y Y, Shi B Y. BLGAN: Bayesian learning and genetic algorithm for supporting negotiation with incomplete information [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(1); 198-211
- [11] 李剑,牛少彰. 一种基于混合遗传算法的双边多议题协商[J]. 北京邮电大学学报, 2009, 32(2); 1-4
- [12] Ng S C, Sulaiman M N, Selamat M H. Machine learning approach in optimizing negotiation agents for E-Commerce [J]. In-

formation Technology Journal, 2009, 8(6); 801-810

- [13] 程昱,高济,古华茂,等. 基于对手态度学习的协商决策模型[J]. 浙江大学学报:工学版, 2008, 42(10); 1676-1680
- [14] 程昱,高济,古华茂,等. 基于机器学习的自动协商决策模型[J]. 软件学报, 2009, 20(8); 2160-2169
- [15] Real C, Kersten G E, Vahidov R. Predicting opponent's moves in electronic negotiations using neural networks[J]. Expert Systems with Applications, 2008, 34(2); 1266-1273

(上接第 244 页)

- [2] 卢小甫. 切丛流行学习算法及其应用研究[D]. 苏州: 苏州大学, 2010
- [3] Zhou Li-li, LI Fan-zhang. Research on Mapping Mechanism of Learning Expression[C]// Jian Yu, et al. Proceeding of Rough Set and Knowledge Technology. Beijing, China, October 2010; 298-303
- [4] 贺伟. 范畴论[M]. 北京: 科学出版社, 2006
- [5] 李凡长,何书萍,钱旭培. 李群机器学习研究综述[J]. 计算机学报, 2010, 33(7); 115-1126
- [6] 卢小甫. 切丛流行学习算法及其应用研究[D]. 苏州: 苏州大学, 2010

- [7] Jolliffe I T. Principal Component Analysis [M]. New York: Springer, 1989
- [8] Tenenbaum J B, de Silva V, et al. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290 (5500); 2319-2323
- [9] 赵连伟,罗四维,等. 高位数据流形的低维嵌入及嵌入维数研究[J]. 软件学报, 2005, 16(8); 1423-1430
- [10] Roweis ST, Saul LK. Nonlinear dimensionality analysis by locally linear embedding[J]. Science, 2000, 290(12); 2323-2326
- [11] Tenenbaum J B, de Silva V, et al. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290 (5500); 2319-2323