



文章编号:1001-9081(2017)03-0680-04

DOI:10.11772/j.issn.1001-9081.2017.03.680

基于 KL 散度和近邻点间距离的球面嵌入算法

张变兰, 路永钢*, 张海涛

(兰州大学 信息科学与工程学院, 兰州 730000)

(*通信作者电子邮箱 ylu@lzu.edu.cn)

摘要:针对现有球面嵌入算法在非近邻点间的距离度量不准确或缺失的情况下,不能有效地进行低维嵌入的问题,提出了一种新的球面嵌入算法,它能够只利用近邻点间的距离,将任何尺度的高维数据嵌入到单位球面上,同时求出适合原始数据分布的球面半径。该算法从一个随机产生的球面分布开始,利用 KL 散度衡量每对近邻点间的归一化距离在原始空间和球面空间中的差异,并基于此差异构建出目标函数,然后再用带有动量的随机梯度下降法,不断优化球面上点的分布,直到结果稳定。为了测试算法,模拟产生了两类球面分布数据:分别是球面均匀分布和球面正态分布的数据。实验结果表明,对于球面均匀分布的数据,即使在近邻点个数很少的情况下,仍然能够将数据准确地嵌入球面空间,嵌入后的数据分布与原始数据分布的均方根误差(RMSE)低于 0.000 01,且球面半径的估算误差低于 0.000 001;而对于球面正态分布的数据,在近邻点个数较多的情况下,该算法也可以将数据较准确地嵌入球面空间。因此,在非近邻点间距离缺失的情况下,所提方法仍然可以较准确地对数据进行低维嵌入,这非常有利于数据的可视化研究。

关键词:球面嵌入;KL 散度;随机梯度下降法;最近邻

中图分类号: TP181 **文献标志码:**A

Spherical embedding algorithm based on Kullback-Leibler divergence and distances between nearest neighbor points

ZHANG Bianlan, LU Yonggang*, ZHANG Haitao

(School of Information Science and Engineering, Lanzhou University, Lanzhou Gansu 730000, China)

Abstract: Aiming at the problem that the existing spherical embedding algorithm cannot effectively embed the data into the low-dimensional space in the case that the distances between points far apart are inaccurate or absent, a new spherical embedding method was proposed, which can take the distances between the nearest neighbor points as input, and embeds high dimensional data of any scale onto the unit sphere, and then estimates the radius of the sphere which fit the distribution of the original data. Starting from a randomly generated spherical distribution, the Kullback-Leibler (KL) divergence was used to measure the difference of the normalized distance between each pair of neighboring points in the original space and the spherical space. Based on the difference, the objective function was constructed. Then, the stochastic gradient descent method with momentum was used to optimize the distribution of the points on the sphere until the result is stable. To test the algorithm, two types of spherical distribution data sets were simulated: which are spherical uniform distribution and Kent distribution on the unit sphere. The experimental results show that, for the uniformly distributed data, the data can be accurately embedded in the spherical space even if the number of neighbors is very small, the Root Mean Square Error (RMSE) of the embedded data distribution and the original data distribution is less than 0.000 01, and the spherical radius of the estimated error is less than 0.000 001; for spherical normal distribution data, the data can be embedded into the spherical space accurately when the number of neighbors is large. Therefore, in the case that the distance between points far apart are absent, the proposed method can still be quite accurate for low-dimensional data embedding, which is very helpful for the visualization of data.

Key words: spherical embedding; Kullback-Leibler (KL) divergence; stochastic gradient descent method; nearest neighbor

0 引言

近年来,数据可视化分析已经成为处理大数据的重要方法之一。研究表明,人们从外界接收的各种信息中 80% 以上

是通过视觉获得的^[1]。通过对大数据可视化,人们可以对数据产生直观的理解,以便对其进行分析和研究,因此,数据可视化在大数据分析中正起着越来越重要的作用。为了避免维数灾难带来的影响,以及更好地对大数据进行可视化分

收稿日期:2016-09-19;修回日期:2016-11-11。

基金项目:国家自然科学基金面上项目(61272213);中央高校基本科研业务费专项资金资助项目(lzujbky-2016-k07,lzujbky-2016-142)。

作者简介:张变兰(1991—),女,山西吕梁人,硕士研究生,主要研究方向:模式识别;路永钢(1974—),男,甘肃陇南人,教授,博士,CCF 会员,主要研究方向:模式识别、人工智能、生物信息;张海涛(1986—),男,甘肃兰州人,博士,主要研究方向:模式识别、软件工程。



析^[2],数据降维方法常被用来产生数据的一个低维可视化表示。

在计算机视觉和模式识别中,许多问题都是基于样本点间的距离的,例如手势识别和形状识别等。在这些问题中,只知道样本点间的相似性或者距离度量,而不知道样本在原始空间的坐标或者其对应的特征向量。这时,可以使用嵌入算法来得到样本点在对应空间中的坐标分布。在嵌入低维的情况下,也可以通过降维得到样本的可视化表示。处理该类问题的嵌入算法有多维尺度分析(MultiDimensional Scaling, MDS)^[3]、最大方差展开(Maximum Variance Unfolding, MVU)^[4]、等距映射(Isometric Mapping, IsoMap)^[5]和t分布随机邻域嵌入(t-Distributed Stochastic Neighbor Embedding, t-SNE)^[6]等。它们都是利用所有样本间的相似性或者距离信息来构建样本的低维表示,并使得样本在低维空间中的结构与高维空间的分布尽量保持一致^[3,7]。

然而,这类算法大部分都是将数据嵌入线性空间。在计算机视觉中,对于很多类型的数据,其样本都是分布在高维非线性空间中的,因此,将这些数据嵌入至低维线性空间是不可行的^[8-9]。针对上述问题,出现了很多基于曲面的嵌入算法,例如将数据嵌入环形表面或者球面,以便对此类数据进行可视化研究。其中,关于球面嵌入的研究更为广泛,而且球面嵌入算法有很多实际应用,例如在地球模型表面的数据表示,或类球状物表面的纹理贴图等^[8,10]。文献[10-11]中的算法都能够有效地将数据嵌入至球面空间,它们都是基于MDS算法的改进,最终成功将数据嵌入到非线性空间^[8]。这类算法的关键步骤是优化过程,它们首先定义一个衡量嵌入质量的目标函数^[8,10-11],然后通过优化算法不断调整低维空间中样本点的位置来优化目标函数。文献[11]中提出了一种球面MDS的嵌入算法,这是最早提出球面嵌入算法的论文之一。它用球面极坐标来表示样本点,用克鲁斯克系数(Kruskal Stress)^[11]构造目标函数。算法采用了最速下降法进行优化,通过调整点在球面的位置来使目标函数最小。文献[8]中,提出了一种曲面流形嵌入算法,将已知的所有点对间距离的数据集嵌入恒定曲率的曲面空间,如球面或双曲面,并且可求出该曲面空间的曲率半径;该算法还可以将数据嵌入超球面空间。该算法的最大优点是,在无需任何优化的情况下,根据已知的对称距离矩阵可以快速有效地将数据嵌入曲面空间,并估计出曲面空间的曲率半径;但是,现有的球面嵌入算法的共同缺点是,必须利用所有点间的相似性或距离信息来进行嵌入。

而对于许多高维数据来说,只有近邻点间的相似性度量是比较可靠的,所以大多数非线性降维算法只采用近邻点间的距离进行低维嵌入,例如MVU^[4]、IsoMap^[5]和局部线性嵌入(Locally Linear Embedding, LLE)^[12]等。这些算法的本质是,先找到每个样本点的前K个近邻点,通过优化目标函数,使得近邻点间的距离尽量保持不变,从而将非线性数据嵌入至线性空间。

本文提出了一种新的球面嵌入算法,能够在只知道近邻点间距离的情况下将数据集嵌入到单位球面上,并尽量保持近邻点间的结构。这样就实现了只利用近邻点间的相似性信息,将非线性数据嵌入至球面空间。据考证,目前还没有类似的算法,而本文是首次提出了基于近邻点间距离的球面半径未知情况下的球面嵌入算法。该方法用KL散度^[13-14]来计算嵌入球面前后每对近邻点间的相对分布差异,并基于此差

异构建出目标函数。然后利用带有动量的随机梯度下降法^[15-16]进行优化,使得所有近邻点间相对分布的差异之和最小。这样就可以将任意尺度的高维数据嵌入到单位球面上。最后,利用嵌入前后所有近邻点间的距离之和的比值,就可估计出适合原始数据分布的球面半径。

1 球面嵌入算法

1.1 球面上的距离计算

在球面坐标系中,球面上的点的坐标为 $x_i = (\theta_i, \varphi_i)$,极角 θ_i 表示向量 x_i 与z轴的夹角,方位角 φ_i 表示向量 x_i 与x轴的夹角。在球面上,两点间的距离为两向量间夹角对应的球面上的弧长。若在半径为r的球面上,两点间的夹角记为 Θ_{ij} ,则它们在此球面上的距离可表示为:

$$d_{ij} = r\Theta_{ij} \quad (1)$$

$$\Theta_{ij} = \cos^{-1}(\cos \theta_i \cos \theta_j + \sin \theta_i \sin \theta_j \cos(\varphi_i - \varphi_j)) \quad (2)$$

1.2 球面嵌入算法

首先,该算法将输入的所有近邻间的距离整体归一化。对于样本点 x_i 和点 x_j ,其归一化距离为 p_{ij} :

$$p_{ij} = d_{ij}/\left(\sum_{m \neq n} w_{mn} d_{mn}\right) \quad (3)$$

$$d_{ij} = \|x_i - x_j\| \quad (4)$$

嵌入单位球面空间后,用同样的归一化方法,可得到点 y_i 与点 y_j 的归一化距离 q_{ij} :

$$q_{ij} = \Theta_{ij}/\left(\sum_{m \neq n} w_{mn} \Theta_{mn}\right) \quad (5)$$

$$\Theta_{ij} = \|y_i - y_j\| \quad (6)$$

其中: Θ_{ij} 表示嵌入到单位球面上的两点间的距离,也就是两点对应的向量间的夹角。另外,该算法中定义了一个系数因子w,当点 x_i 和点 x_j 为近邻时, $w_{ij} = 1$,否则 $w_{ij} = 0$ 。算法将只利用 $w_{ij} = 1$ 的这部分归一化距离进行数据嵌入。

对于任意 $w_{ij} = 1$ 对应的两个近邻点,如果嵌入单位球面后的归一化距离 q_{ij} 和原始样本点间的归一化距离 p_{ij} 相等,就意味着嵌入前后这两点的相对分布一致,因此,该算法的目标就是在嵌入的球面空间中调整近邻点的位置分布,使得每对近邻点之间 p_{ij} 和 q_{ij} 的差异最小,进而使得所有近邻点间的归一化距离在嵌入前后的差异之和达到最小。本文利用KL散度作为衡量 p_{ij} 和 q_{ij} 间差异的指标,因此,所有近邻点之间的KL散度之和构成目标函数:

$$C = \sum_i \sum_j w_{ij} (p_{ij} \ln \frac{p_{ij}}{q_{ij}}) \quad (7)$$

此目标函数的梯度为:

$$\nabla C = \left(\frac{\partial C}{\partial \theta_i}, \frac{\partial C}{\partial \varphi_i} \right) \quad (8)$$

$$\frac{\partial C}{\partial \theta_i} = \frac{\partial C}{\partial \Theta_{ij}} \frac{\partial \Theta_{ij}}{\partial \theta_i} \quad (9)$$

$$\frac{\partial C}{\partial \varphi_i} = \frac{\partial C}{\partial \Theta_{ij}} \frac{\partial \Theta_{ij}}{\partial \varphi_i} \quad (10)$$

$$\frac{\partial C}{\partial \Theta_{ij}} = \frac{q_{ij} - p_{ij}}{\Theta_{ij}} \quad (11)$$

$$\frac{\partial \Theta_{ij}}{\partial \theta_i} = \frac{\sin \theta_i \cos \theta_j - \cos \theta_i \sin \theta_j \cos(\varphi_i - \varphi_j)}{\sin \Theta_{ij}} \quad (12)$$

$$\frac{\partial \Theta_{ij}}{\partial \varphi_i} = \frac{\sin \theta_i \sin \theta_j \sin(\varphi_i - \varphi_j)}{\sin \Theta_{ij}} \quad (13)$$

在该球面嵌入算法中,首先将单位球面上随机产生的样



本点分布作为嵌入空间中的初始分布,然后采用带有动量的随机梯度下降法进行优化,具体的迭代过程为:

$$\mathbf{y}_i^{(k)} = (\theta_i^{(k)}, \varphi_i^{(k)}) \quad (14)$$

$$\mathbf{y}_i^{(k+1)} = \mathbf{y}_i^{(k)} + \Delta \mathbf{y}_i^{(k+1)} \quad (15)$$

$$\Delta \mathbf{y}_i^{(k+1)} = -\rho^{(k)} \frac{\nabla C(\mathbf{y}_i^{(k)})}{\|\nabla C(\mathbf{y}_i^{(k)})\|} + \alpha \Delta \mathbf{y}_i^{(k)} \quad (16)$$

$$\alpha = \begin{cases} 0.5, & k < 250 \\ 0.8, & k \geq 250 \end{cases} \quad (17)$$

$$\rho^{(k)} = \frac{\sum_{i=1}^N \|\nabla C(\mathbf{y}_i^{(k)})\|}{\sum_{i=1}^N \frac{\nabla C(\mathbf{y}_i^{(k)})^T \mathbf{D}(\mathbf{y}_i^{(k)}) \nabla C(\mathbf{y}_i^{(k)})}{\|\nabla C(\mathbf{y}_i^{(k)})\|^2}} \quad (18)$$

式(16)中, α 表示动量; k 表示迭代次数; $\Delta \mathbf{y}_i$ 表示在每次迭代后样本点 i 的位置的变化量, 带动量的随机梯度下降法每次都记录这个位置变化, 并利用梯度和前一次的位置变化量的组合得出新的位置变化量; $\rho^{(k)}$ 表示第 k 次迭代的最佳步长, 确定最佳步长的计算过程见式(18)。在式(18)中, $\mathbf{D}(\mathbf{y}_i)$ 为 $C(\mathbf{y}_i)$ 的二阶偏导数矩阵, 详细计算过程为:

$$\mathbf{D} = \begin{pmatrix} \frac{\partial^2 C}{\partial \theta_i^2} & \frac{\partial^2 C}{\partial \theta_i \partial \varphi_i} \\ \frac{\partial^2 C}{\partial \theta_i \partial \varphi_i} & \frac{\partial^2 C}{\partial \varphi_i^2} \end{pmatrix} \quad (19)$$

其中:

$$\frac{\partial^2 C}{\partial \theta_i^2} = \frac{\partial^2 C}{\partial \Theta_{ij}^2} \left(\frac{\partial C}{\partial \theta_i} \right)^2 + \frac{\partial C}{\partial \Theta_{ij}} \frac{\partial^2 \Theta_{ij}}{\partial \theta_i^2} \quad (20)$$

$$\frac{\partial^2 C}{\partial \theta_i \partial \varphi_i} = \frac{\partial^2 C}{\partial \Theta_{ij}^2} \frac{\partial \Theta_{ij}}{\partial \theta_i} \frac{\partial \Theta_{ij}}{\partial \varphi_i} + \frac{\partial C}{\partial \Theta_{ij}} \frac{\partial^2 \Theta_{ij}}{\partial \theta_i \partial \varphi_i} \quad (21)$$

$$\frac{\partial^2 C}{\partial \varphi_i^2} = \frac{\partial^2 C}{\partial \Theta_{ij}^2} \left(\frac{\partial \Theta_{ij}}{\partial \varphi_i} \right)^2 + \frac{\partial C}{\partial \Theta_{ij}} \frac{\partial^2 \Theta_{ij}}{\partial \varphi_i^2} \quad (22)$$

$$\frac{\partial^2 C}{\partial \Theta_{ij}^2} = \frac{p_{ij} - q_{ij}^2}{\Theta_{ij}^2} \quad (23)$$

$$\frac{\partial^2 \Theta_{ij}}{\partial \theta_i^2} = \frac{\cos \theta_i \cos \theta_j + \sin \theta_i \sin \theta_j \cos(\varphi_i - \varphi_j)}{\sin \Theta_{ij}} - \left(\frac{\partial \Theta_{ij}}{\partial \theta_i} \right)^2 \frac{\cos \Theta_{ij}}{\sin \Theta_{ij}} \quad (24)$$

$$\frac{\partial^2 \Theta_{ij}}{\partial \theta_i \partial \varphi_i} = \frac{\cos \theta_i \sin \theta_j \sin(\varphi_i - \varphi_j)}{\sin \Theta_{ij}} - \frac{\partial \Theta_{ij}}{\partial \theta_i} \frac{\partial \Theta_{ij}}{\partial \varphi_i} \frac{\cos \Theta_{ij}}{\sin \Theta_{ij}} \quad (25)$$

$$\frac{\partial^2 \Theta_{ij}}{\partial \varphi_i^2} = \frac{\sin \theta_i \sin \theta_j \cos(\varphi_i - \varphi_j)}{\sin \Theta_{ij}} - \left(\frac{\partial \Theta_{ij}}{\partial \varphi_i} \right)^2 \frac{\cos \Theta_{ij}}{\sin \Theta_{ij}} \quad (26)$$

最后, 在求得嵌入单位球面的样本之后, 即可利用嵌入前后近邻点间的距离之和的比值, 求出原始样本分布的球面半径 R , 公式如下:

$$R = \left(\sum_{i \neq j} w_{ij} d_{ij} \right) / \left(\sum_{i \neq j} w_{ij} \Theta_{ij} \right) \quad (27)$$

2 实验和结果分析

为了验证本文提出的球面嵌入算法的正确性, 文中设计了两类模拟数据进行测试, 一类是球面均匀分布的数据集, 另一类是球面正态分布的数据集。下面将在 2.1 节中详细介绍产生这两类模拟数据的过程, 在 2.2 节中详细介绍实验过程和评价结果。

2.1 模拟数据的产生

下面介绍两类模拟数据集: 球面均匀分布的数据集和球面正态分布的数据集的产生过程。

2.1.1 球面均匀分布的模拟数据集

每个样本点可表示为 $x_i = (\theta_i, \varphi_i), i = 1, 2, \dots, N$, 其中 $\theta_i \in [0, \pi], \varphi_i \in [0, 2\pi]$, N 为样本总数。首先模拟产生了随机均匀分布于单位球面的 $N = 2000$ 个样本, 然后利用式(2)计算出这些样本两两间的夹角 Θ_{ij} , 设半径 r 为 0.5, 利用式(1), 即可得到均匀分布于半径为 0.5 的球面上的数据对应的距离矩阵。

2.1.2 球面正态分布的模拟数据集

本实验用 Kent 分布^[17] 模拟产生了位于单位球面上的正态分布数据。这个数据集($N = 913$) 主要由三部分组成, 一部分是呈圆形的正态分布, 另两部分都是呈椭圆形的正态分布, 而且这两个椭圆形分布的数据, 其分布大小和密度都不同。之后, 得到一个分布于半径为 2 的球面上的包含 3 个不同 Kent 分布的数据集对应的距离矩阵。

2.2 实验结果

在实验中, 先取每个样本点和其前 nn ($nn \in [0, N]$) 个近邻点的距离构成稀疏距离矩阵, 将此作为球面嵌入算法的输入。算法的输出为所有样本点在单位球面上的坐标。通过此坐标可以计算出嵌入单位球面空间后样本点间的夹角 Θ_{ij} , 然后利用式(27) 计算出适合原始数据分布的球面半径 R 。最后以均方根误差(Root Mean Square Error, RMSE) 为指标, 衡量所有的原始数据两两间的夹角 d_{ij}/r 与嵌入球面后对应的数据两两间的夹角 Θ_{ij} 间的误差, 见式(28)。用近邻均方根误差(Root Mean Square Error between Nearest Neighbors, NN_RMSE) 来表示原始数据的近邻点间的夹角与嵌入球面后对应的夹角之间的误差, 见式(29)。另外, 半径的估算误差(Radius estimation Error, R_Error) 计算见式(30)。

$$RMSE = \sqrt{\frac{1}{N^2} \sum_{i,j=1}^N (d_{ij}/r - \Theta_{ij})^2} \quad (28)$$

$$NN_RMSE = \sqrt{\frac{1}{\sum_{i,j=1}^N w_{ij}} \sum_{i,j=1}^N w_{ij} (d_{ij}/r - \Theta_{ij})^2} \quad (29)$$

$$R_Error = |R - r| \quad (30)$$

若这三个值越小, 则说明将样本嵌入球面空间的效果越好。

对于球面均匀分布的数据集, 设置近邻点个数 $nn = \{N, 0.75N, 0.5N, 0.25N, 0.05N\}$ 进行实验, 由于该算法的初始化是随机的, 因此在每个参数设置下同一个实验都重复运行 3 次。最后, 对于半径为 $r = 0.5$ 的数据的实验结果汇总于表 1。

从表 1 中可以看出, 针对不同的近邻点个数设置, 本文提出的嵌入算法都能得到较准确的结果, 所有的均方根误差(RMSE) 基本都小于 0.00001, 并且, 当每个样本点拥有的近邻点数目越多, 则算法嵌入的效果越好, 得到的整体数据在单位球面上的分布与原始空间中的分布的一致性也越高。

另外, 对于半径 $r = 3.2$ 的球面均匀分布的数据也做了相同的实验, 并得到了类似的测试结果。可见对均匀分布于球面的数据, 该算法即使在非近邻点间距离信息缺失较多的情况下, 仍然能够较准确地还原出球面空间中数据的分布结构; 而且算法还可以较精确地估算出适合数据分布的球面半径。

接着, 对球面正态分布(Kent 分布) 的数据也进行了类似的测试, 其球面半径的设置为 $r = 2$, 并取近邻点个数 $nn =$

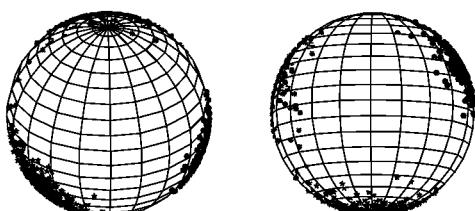


$\{N, 700, 500, 300\}$, 在每个参数设置下都重复运行3次, 实验结果见表1。在 $nn = 300$ 时, 第一次运行的球面嵌入结果如图1(a)所示。作为参照, 图1(b)显示了原始数据的分布。

表1 嵌入算法对两类模拟数据的处理结果

Tab. 1 Processing results of two kinds of simulated data by the proposed embedding algorithm

数据	近邻数	nn	误差	第一次	第二次	第三次
				NN_RMSE	RMSE	R_Error
球面均匀分布	2000	NN_RMSE	1.90E-10	1.93E-10	1.96E-10	
		RMSE	1.90E-10	1.93E-10	1.96E-10	
		R_Error	4.05E-15	4.03E-14	1.90E-14	
	1500	NN_RMSE	2.24E-10	2.19E-10	3.65E-02	
		RMSE	1.95E-10	1.91E-10	4.19E-02	
		R_Error	1.18E-14	1.74E-14	2.24E-04	
球面正态分布	1000	NN_RMSE	2.72E-10	2.72E-10	2.68E-10	
		RMSE	1.94E-10	1.94E-10	1.91E-10	
		R_Error	7.88E-15	2.57E-14	1.45E-13	
	500	NN_RMSE	6.30E-10	3.79E-10	3.74E-08	
		RMSE	5.58E-10	1.91E-10	4.19E-08	
		R_Error	3.50E-13	4.16E-13	1.48E-10	
	100	NN_RMSE	2.48E-05	2.14E-05	6.09E-07	
		RMSE	9.80E-05	9.09E-05	2.64E-06	
		R_Error	1.25E-06	9.55E-07	2.99E-08	
	913	NN_RMSE	6.89E-02	2.81E-10	3.12E-02	
		RMSE	6.89E-02	2.81E-10	3.12E-02	
		R_Error	5.43E-04	2.71E-10	3.62E-05	
	700	NN_RMSE	5.69E-02	1.27E-01	9.90E-02	
		RMSE	5.81E-02	1.28E-01	1.01E-01	
		R_Error	3.07E-04	4.75E-04	6.31E-04	
	500	NN_RMSE	7.08E-02	1.06E-01	8.33E-02	
		RMSE	7.92E-02	1.26E-01	9.40E-02	
		R_Error	9.90E-03	1.30E-02	1.14E-02	
	300	NN_RMSE	1.13E-01	6.28E-02	1.10E-01	
		RMSE	4.02E-01	3.31E-01	4.02E-01	
		R_Error	2.63E-01	1.76E-01	2.56E-01	



(a) 取每个点近邻数为300时,
嵌入单位球面的结果
(b) 原始数据的分布

图1 球面正态分布的数据

Fig. 1 Data of spherical normal distribution

实验结果表明,对于球面正态分布的数据,从图1和表1都可以看出,其嵌入球面后的整体分布与原始分布比较接近,但是,整体嵌入后的误差都明显比表1中球面均匀分布数据的误差大很多。另外,随着近邻点数目的减少,算法将其嵌入单位球面空间后,虽然可以较好地保持其近邻点结构,但是非近邻点间的分布却与原始数据中的分布相差较大。例如表1中,对于球面正态分布的数据 $nn = 300$ 时,NN_RMSE都小于0.113,然而RMSE的值则都大于0.331;同时,对于适合原始数据分布的球面半径的估算误差也随近邻数的减小而增大。

此外,由于初始化是随机的,本文提出的算法有时会陷入局部极小,因此导致实验结果的不稳定。例如,表1中,对于球

面均匀分布的数据 $nn = 1500$ 时,三次运行结果波动很大,第三次实验的运行结果中RMSE和NN_RMSE比前两次对应的误差分别高了8个数量级。另外表1中,对于球面正态分布的数据 $nn = 913$ 时,第二次运行结果的RMSE和NN_RMSE明显比其他两次运行结果的误差低了8个数量级。所以,为保证实验结果的准确性和正确性,每个实验都要经过多次运算。

3 结语

本文首次提出了一种针对球面半径未知且原始数据间的非近邻距离缺失情况下的球面嵌入算法。该算法能够在只已知近邻点间距离的情况下,将任意尺度的数据嵌入至单位球面,还可以估算出适合原始数据分布的球面半径。

本文提出的算法对于球面均匀分布的数据,在非近邻点间距离信息缺失较多的情况下,仍然能得到较准确的球面嵌入结果;但是,对于非均匀分布的数据,嵌入球面空间后,虽然近邻点间的相对位置可以较好地保持,但是无法准确地还原非近邻点间的相对位置,因此对于非均匀分布的数据,球面嵌入算法还有待改进。

参考文献 (References)

- [1] 田守财,孙喜利,路永钢.基于最近邻的随机非线性降维[J].计算机应用,2016,36(2):377-381.(TIAN S C, SUN X L, LU Y G. Stochastic nonlinear dimensionality reduction based on nearest neighbors [J]. Journal of Computer Applications, 2016, 36(2): 377-381.)
- [2] 郝晓军,闫京海,樊友谊.大数据分析过程中的降维方法[J].航天电子对抗,2014(4):58-60.(HAO X J, YAN J H, FAN Y Y. Dimensionality reduction of large volumes of data analysis [J]. Aerospace Electronic Warfare, 2014(4): 58-60).
- [3] COX M A A, COX T F. Multidimensional scaling [J]. Econometric Institute Research Papers, 2014, 46(2): 1050-1057.
- [4] WEINBERGER K Q, SAUL L K. Unsupervised learning of image manifolds by semidefinite programming [C]// Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2004: 988-995.
- [5] TENENBAUM J B, DE SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(5500): 2319-2323.
- [6] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.
- [7] VAN DER MAATEN L J P, POSTMA E O, VAN DEN HERIK H J. Dimensionality reduction: a comparative review [EB/OL]. [2016-03-08]. https://static.aminer.org/pdf/PDF/000/272/419/comparative_investigation_on_dimension_reduction_and_regression_in_three_layer.pdf.
- [8] WILSON R C, HANCOCK E R, PEKALSKA E, et al. Spherical and hyperbolic embeddings of data [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11): 2255-2269.
- [9] WILSON R C, HANCOCK E R. Spherical embedding and classification [C]// Proceedings of the 2010 Joint IAPR International Conference on Structural, Syntactic, and Statistical Pattern Recognition. Berlin: Springer, 2010: 589-599.

(下转第690页)



- Pattern Recognition. Washington, DC: IEEE Computer Society, 2012: 733 – 740.
- [18] FRINTROP S, WERNER T, GARCIA G M. Traditional saliency reloaded: a good old model in new shape [C] // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 82 – 90.
- [19] GONG C, TAO D, LIU W, et al. Saliency propagation from simple to difficult [C] // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 2531 – 2539.
- [20] GOPALAKRISHNAN V, HU Y, RAJAN D. Random walks on graphs to model saliency in images [C] // Proceeding of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2009: 1698 – 1705.
- [21] JIANG B, ZHANG L, LU H, et al. Saliency detection via absorbing Markov chain [C] // Proceedings of the 2013 IEEE International Conference on Computer Vision. Washington, DC: IEEE Computer Society, 2013: 1665 – 1672.
- [22] REN Z, HU Y, CHIA L T, et al. Improved saliency detection based on superpixel clustering and saliency propagation [C] // Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM, 2010: 1099 – 1102.
- [23] YANG C, ZHANG L, LU H, et al. Saliency detection via graph-based manifold ranking [C] // Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2013: 3166 – 3173.
- [24] ZHOU D, WESTON J, GRETTON A, et al. Ranking on data manifolds [J]. Advances in Neural Information Processing Systems, 2004, 16: 169 – 176.
- [25] LI C, YUAN Y, CAI W, et al. Robust saliency detection via regularized random walks ranking [C] // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 2710 – 2717.
- [26] QIN Y, LU H, XU Y, et al. Saliency detection via cellular automata [C] // Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 110 – 119.
- [27] ACHANTA R, SHAJI A, SMITH K, et al. Slic superpixels [EB/OL]. [2016- 01- 07]. https://infoscience.epfl.ch/record/149300/files/SLIC_Superpixels_TR_2.pdf?version=2.
- [28] JIANG H, WANG J, YUAN Z, et al. Salient object detection: a discriminative regional feature integration approach [C] // Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2013: 2083 – 2090.
- [29] LI Y, HOU X, KOCH C, et al. The secrets of salient object segmentation [C] // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2014: 280 – 287.
- [30] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection [C] // Proceeding of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2009: 1597 – 1604.
- [31] HAREL J, KOCH C, PERONA P. Graph-based visual saliency [C] // Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2006: 545 – 552.
- [32] CHENG M M, WARRELL J, LIN W Y, et al. Efficient salient region detection with soft image abstraction [C] // Proceedings of the 2013 IEEE International Conference on Computer Vision. Washington, DC: IEEE Computer Society, 2013: 1529 – 1536.

This work is partially supported by the Key Project of National Natural Science Foundation of China (61133009), the National Natural Science Foundation of China (61472245, U1304616, 61502220).

XIE Chang, born in 1991, M. S candidate. His research interests include digital image processing, pattern recognition, machine learning.

ZHU Hengliang, born in 1981, Ph. D. candidate. His research interests include digital image processing, pattern recognition.

LIN Xiao, born in 1978, Ph. D., associate professor. Her research interests include digital image processing, video processing.

MA Lizhuang, born in 1963, Ph. D., professor. His research interests include computer graphics, digital image processing, computer aided design, scientific data visualization, computer animation.

(上接第 683 页)

- [10] ELAD A, KELLER Y, KIMMEL R. Texture mapping via spherical multi-dimensional scaling [C] // Scale Space and PDE Methods in Computer Vision, LNCS 3459. Berlin: Springer, 2005: 443 – 455.
- [11] COX M A A, COX T F. Multidimensional scaling on the sphere [M] // EDWARDS D, RAUN N E. Compstat. Berlin: Springer, 1988: 323 – 328.
- [12] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323 – 2326.
- [13] KULLBACK S, LEIBLER R A. On information and sufficiency [J]. Annals of Mathematical Statistics, 1951, 22(1): 79 – 86.
- [14] KULLBACK S. Information Theory and Statistics [M]. Hoboken, NJ: John Wiley and Sons, 1959.
- [15] SUTSKEVER I. Training recurrent neural networks [EB/OL]. [2016-02-09]. http://www.cs.utoronto.ca/~ilya/pubs/ilya_sutskever_phd_thesis.pdf.
- [16] SUTSKEVER I, MARTENS J, DAHL G, et al. On the importance of

initialization and momentum in deep learning [EB/OL]. [2016-02-09]. <http://www.cs.toronto.edu/~hinton/absps/momentum.pdf>.

- [17] KENT J T. The Fisher-Bingham distribution on the sphere [J]. Journal of the Royal Statistical Society, 1982, 44(1): 71 – 80.

This work is partially supported by the National Natural Science Foundation of China (61272213), the Fundamental Research Funds for the Central Universities (lzujbky-2016-k07, lzujbky-2016-142).

ZHANG Bianlan, born in 1991, M. S. candidate. Her research interests include pattern recognition.

LU Yonggang, born in 1974, Ph. D., professor. His research interests include pattern recognition, artificial intelligence, bioinformatics.

ZHANG Haitao, born in 1986, Ph. D. Her research interests include pattern recognition, software engineering.