ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.7, July 2008, pp.1666−1673
DOI: 10.3724/SP.J.1001.2008.01666

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

邻域参数动态变化的局部线性嵌入*

文贵华[1+], 江丽君[2], 文 军[3]

[1](华南理工大学 计算机科学与工程学院,广东 广州 510641)

[2](华南理工大学 电子材料科学与工程系,广东 广州 510641)

[3](湖北民族学院 理学院,湖北 恩施 445000)

Dynamically Determining Neighborhood Parameter for Locally Linear Embedding

WEN Gui-Hua[1+], JIANG Li-Jun[2], WEN Jun[3]

[1](School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China)

[2](Department of Electronic Material Science and Engineering, South China University of Technology, Guangzhou 510641, China)

[3](School of Mathematical Science, Hubei Insitute for Nationalities, Enshi 445000, China)

+ Corresponding author: E-mail: crghwen@scut.edu.cn

Wen GH, Jiang LJ, Wen J. Dynamically determining neighborhood parameter for locally linear embedding. *Journal of Software*, 2008,19(7):1666−1673. http://www.jos.org.cn/1000-9825/19/1666.htm

Abstract: Locally linear embedding is a kind of very competitive nonlinear dimensionality reduction with good representational capacity for a broader range of manifolds and high computational efficiency. However, they are based on the assumption that the whole data manifolds are evenly distributed so that they determine the neighborhood for all points with the same neighborhood size. Accordingly, they fail to nicely deal with most real problems that are unevenly distributed. This paper presents a new approach that takes the general conceptual framework of Hessian locally linear embedding so as to guarantee its correctness in the setting of local isometry to an open connected subset but dynamically determines the local neighborhood size for each point. This approach estimates the approximate geodesic distance between any two points by the shortest path in the local neighborhood graph, and then determines the neighborhood size for each point by using the relationship between its local estimated geodesic distance matrix and local Euclidean distance matrix. This approach has clear geometry intuition as well as the better performance and stability to deal with the sparsely sampled or noise contaminated data sets that are often unevenly distributed. The conducted experiments on benchmark data sets validate the proposed approach.

Key words: manifold learning; Hessian locally linear embedding; neighborhood size; dimensionality reduction

摘 要: 局部线性嵌入是最有竞争力的非线性降维方法,有较强的表达能力和计算优势.但它们都采用全局一致的邻域大小,只适用于均匀分布的流形,无法处理现实中大量存在的非均匀分布流形.为此,提出一种邻域大小动态确定的新局部线性嵌入方法.它采用 Hessian 局部线性嵌入的概念框架,但用每个点的局部邻域估计此邻域内任意

* Supported by the 2008 Project of Scientific Research Foundation for the Returned Overseas Chinese Scholars (2008年教育部留学回国人员科研启动基金); the Science-Technology Project of Guangdong Province of China under Grant No.2007B030803006 (广东省科技攻关项目); the Science-Technology Project of Hubei Province of China under Grant No.2005AA101C17 (湖北省科技攻关项目)

Received 2006-11-03; Accepted 2007-01-24

点之间的近似测地距离,然后根据近似测地距离与欧氏距离之间的关系动态确定该点的邻域大小,并以此邻域大小构造新的局部邻域.算法几何意义清晰,在观察数据稀疏和数据带噪音等情况下,都比现有算法有更强的鲁棒性.标准数据集上的实验结果验证了所提方法的有效性.

关键词: 流形学习;Hessian局部线性嵌入;邻域大小;降维

中图法分类号: TP181 **文献标识码**: A

1 问题的提出

很多高维数据,如遥感、气候等常常分布于较低维的流形上,自从《Science》在2000年发表最有代表性的方法ISOMAP(isometric feature mapping)和LLE(locally linear embedding)以来[1,2],寻找描述这样低维流形的参数空间就成为最近的研究热点.ISOMAP在降维过程中通过计算点对之间的测地距离,并采用MDS(multi-dimensional scaling)方法来获取全局最优的几何结构,获得了较好的效果,目前已发展了很多改进算法,如基于核方法的ISOMAP、监督ISOMAP[3]、增量式ISOMAP[4]等.LLE在降维嵌入过程中保持局部的几何结构不变,并能避免局部极小,最终获得一个全局的低维嵌入系统,效果也很好.目前的改进算法包括利用Hessian变换改进的算法HLLE(Hessian LLE)[5]、利用数据分类信息改进的监督LLE、增量式LLE[6]、利用Fisher改进的LLE[7]等.目前,国内也展开了较深入的理论研究和应用实践[8].例如,ISOMAP中连续流形与其低维参数空间等距映射的存在性证明[9]、根据放大因子和延伸方向研究高维观测数据与其低维参数空间数据的联系[10]等. ISOMAP的基本假设是全局等距映射和凸的参数空间,这在很多情况下难以满足;而HLLE只要求局部等距映射和开的连通参数空间,有更宽的应用范围.但是,与ISOMAP一样,都极大程度地依赖于局部邻域是否正确地反映了流形的内在结构.现有的k-近邻邻域确定方法对稀疏和噪音数据容易产生扭曲的邻域结构,从而导致短路现象[11].所谓短路是指流形上的折叠面靠得很近,使得某些点的邻域来自不同的折叠面,因而并不是流形上的最近邻,这常常导致显著的性能偏差,因此需要邻域优化.邻域优化方法包括从完全连接图中重复抽取最小生成树来构造连通邻域图的方法[12],以保证降维之后不丢失数据之间的相对位置.利用数据的分类信息重定义距离,进而利用新定义的距离来确定邻域的方法[3],缺点是对无分类信息的数据不适用.目前,也有利用残差和线性重构造系数来自动选择最佳邻域大小的研究[13−15],但一旦确定,每个数据点的邻域大小仍然是相同的.另一种方法是为每个点选择初步邻域,利用PCA(principal component analysis)构造此邻域的主线性子空间,然后从邻域中删除偏离主线性子空间的邻域点[16],当邻域本质上是非线性时,此方法可能不适用,同时,太多的参数使得应用起来较为困难.

我们以前的工作重点是利用聚类技术对数据自动聚类,然后采用有监督的方法改善邻域[17],利用图代数优化邻域[18]等,但邻域大小仍然是全局统一的.考虑到HLLE需要保持局部区域的线性化,当数据流形是非均匀分布时,采用全局统一的邻域大小难以满足,因为若邻域参数取得太大,则容易消除流形的小尺度结构,并不可避免地面临短路问题,相反,则容易导致流形分裂[19].因此,我们曾提出了对整个不均匀分布流形递归分解为近似均匀分布的子流形,并自动计算每个子流形邻域大小的方法,进而改进LLE[20],但是它需要计算所有点之间的测地距离,时间复杂度太高,接近$O(|X|^3)$,而且LLE性能不及HLLE.为此,本文只计算每个点与其附近点之间的近似测地距离,并用它来确定该点邻域的大小,进而提出邻域大小动态改变的Hessian局部线性嵌入算法VK-HLLE(variable k Hessian locally linear embedding),不仅性能显著提高,而且也未增加时间复杂度.

2 Hessian局部线性嵌入

假定有一个参数空间$\Theta\subset R^d$和一个光滑映射$\varphi:\Theta\to R^n$,其中,嵌入空间R^n满足$n>d$,则称$M=\varphi(\Theta)$为流形,流形学习的目的是根据观察数据确定参数空间Θ.ISOMAP采用等距特征映射(isometric feature mapping)来实现流形学习,其基本假设是:① 等距:测地距离在等距嵌入映射下是不变的,流形上任意点之间的测地距离在等距嵌入变换下获得的欧氏空间中仍然保持:$G(x,y)=|\alpha-\beta|,x,y\in M$且$\alpha,\beta\in\Theta$;② 凸性:参数空间$\Theta$是凸的.对任意$\alpha,\beta\in\Theta$,线段$\{(1-t)\alpha+t\beta:t\in(0,1)\}$仍然属于$\Theta$.ISOMAP算法利用等距嵌入的这种不变性,在没有任何关于观察数据测度知识

的基础上构造测地距离,方法是,假定当两点非常近时,测地距离等于欧氏距离,而对较远的点对之间的测地距离则根据近邻点之间测地距离的累加来实现.当观察数据集足够密集且内在的参数空间是凸的时候,则ISOMAP能够成功地获得参数空间.ISOMAP面临的问题是,在很多情况下,参数空间并不是凸的,此时它得不到正确的结果.

HLLE采用局部线性方法实现流形学习.它只要求局部等距映射的子集是开的且连通,而不必是凸的.其理论依据来源于流形切空间上的Hessian变换.假定流形$M\subset R^n$是光滑的,其任意点$m\in M$都有切空间$\Gamma_m(M)$,在此空间引入欧氏空间的内积就可以建立局部坐标系统,且有原点$O\in\Gamma_m(M)$.设N_m是$m\in M$的邻域,对任意$m'\in N_m$都有唯一的最近点$v'\in\Gamma_m(M)$使得映射$m'\to v'$是光滑的,因此,N_m具有局部坐标系统.设$f:M\to R$在m附近C^2光滑的,$g:U\to R,U\subset R^d$是零点O的一个邻域.令$g(x)=f(m)$,因$m'\to x$是光滑的,g是C^2光滑的,则f的Hessian变换为

$$(H_f^{\tan}(m))_{i,j}=\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}g(x)|_{x=o}.$$

二次型$H(f)=\int_M\|H_f^{\tan}(m)\|_F^2 d_m$定义了$f:M\to R$在$M$上的平均弯曲率,其中,$\|\ \|_F^2$是Frobenius范数.可以证明,若参数空间$\Theta$是开的连通子集,则$H(f)$有一个$(d+1)$维的零空间(null space),参数空间能够通过计算此零空间的一个合适的基坐标来发现,这就是HLLE的核心.由此可以发现:① 框架上HLLE与LLE一致;② HLLE需要对每个数据点计算$n\times n$次偏导数,当观察数据的维非常高时,计算量不小;③ 每个点的邻域要求是线性的,当邻域高度弯曲时,极易面临短路威胁,这是本文要解决的问题.

3 邻域参数的动态确定

现有局部线性嵌入算法采用全局统一的邻域参数,无法处理现实中大量存在的非均匀流形.根据局部线性嵌入的核心思想,只要邻域内的点都在一个线性空间内,则邻域点越多越好,即要取大的邻域参数,如图1(a)中的y点;相反,当流形的不同折叠面靠得很近时,需要采用较小的邻域参数,如图1(b)中的点x,否则将产生短路问题,如图1(a)中的点x,因为u并不是x在流形上的最近邻,v比u更近.显然,当数据流形是非均匀分布时,以上两种情况是矛盾的,唯一的解决方法是根据流形的结构动态确定.如图1(b)所示,对位于极度弯曲流形上的数据点如点x,取较小的邻域参数,否则取较大的邻域参数,如点y.因此,关键是如何判断数据流形的弯曲性及其与邻域参数的计算关系.

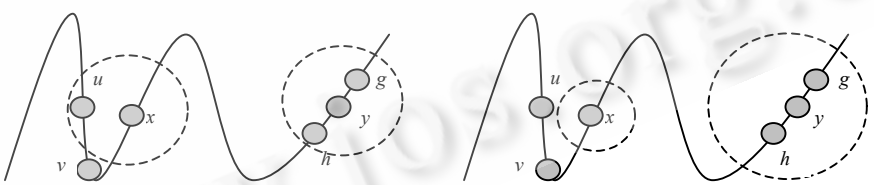

Fig.1 Relationship between neighborhood size and manifold structure

图1 邻域大小与流形结构的关系

我们的方法是采用测地距离与欧氏距离之间的关系来动态确定每个点的邻域大小,其几何意义如图2所示,图中A和B之间的测地距离是AEB曲线长度l_{AB},A和B之间的欧氏距离是AB直线长度d_{AB}.不难看出,$d_{AB}/l_{AB}<d_{CD}/l_{CD}$,且曲线$AEB$所在流形的弯曲度比曲线$CFD$所在流形的弯曲度要大,因此,两点之间的欧氏距离与它们之间的测地距离的比例越小,则位于这两点之间的局部流形越弯曲,应取的邻域参数就越小;反之亦然.这种确定邻域参数的方法需要计算测地距离.直接计算所有输入数据之间测地距离的时间复杂度太高,接近$O(|X|^3)$,为此,这里改变策略,只计算任意点与其附近点之间的近似测地距离,并用它来确定邻域的大小参数,进而提出邻域参数动态改变的Hessian局部线性嵌入算法VK-HLLE,不仅性能显著提高,而且也未增加时间复杂度.

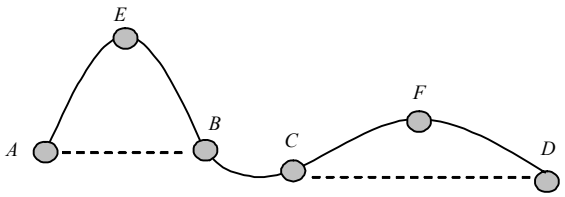

Fig.2 Relationship between the curvature of the manifold and the ration of Euclidean distance and geodesic distance

图 2 欧氏距离与测地距离比例同流形弯曲的关系

算法 1. 确定每个点的邻域参数 *ComputeNbrSizes*(*X*,*k*).

/*输入中,$\boldsymbol{X}$是高维观察数据,k为初始邻域大小;输出:每个点$x_i \in X$的邻域参数k_i*/

1) 用欧氏距离计算X中任意点$\boldsymbol{x}_i$的k-邻域,并由此k-邻域构成局部数据集X_i.

2) 采用ISOMAP的方法计算局部数据集X_i中任意两点之间的局部测地距离,主要包括两步:

- 根据X_i和k确定每个点的k-邻域,然后构造权重图$G=(V,E)$.V对应于X_i中的数据,E为连接V中两点的边集合,$(\boldsymbol{x}_i,\boldsymbol{x}_j)\in E$,若$x_i$是$x_j$的$k$-最近邻,$x_i$与$x_j$之间的距离为欧氏距离$d_e(x_i,x_j)$.
- 通过求G上任意两点之间的最短距离来估计X_i所形成的局部流形上的所有点对之间的测地距离$d_g(x_i,x_j)$.首先对所有$(x_i,x_j)\in E$,令$d_g(x_i,x_j)=d_e(x_i,x_j)$;否则,令$d_g(x_i,x_j)=\infty$.然后利用所有$t$,迭代计算所有的$d_g(x_i,x_j)=\min\{d_g(x_i,x_j),d_g(x_i,x_t)+d_g(x_t,x_j)\}$.

3) 计算X_i内所有点之间的欧氏距离之和与局部测地距离之和的比例,以其作为$x_i \in X_i$所在的局部数据流形弯曲的测量: $\lambda_i = \sum_{x_t,x_j\in X_i} d_e(x_t,x_j) \Big/ \sum_{x_t,x_j\in X_i} d_g(x_t,x_j)$.

4) 计算x_i的邻域参数,基本思想是具有所有λ_i平均值的数据点应取k为邻域参数,其他数据点应以k为中心进行调节: $k_i = k\times\lambda_i \Big/ \left(\left(\sum_{t=1}^{N}\lambda_t\right)\Big/N\right)$.

算法的第 1 步采用欧氏距离d_e确定初始邻域,与所有局部线性嵌入算法相同.第 2 步只计算k邻域内点之间的测地距离,而k为常数,时间复杂度为$O(N)$.第 3 步和第 4 步的时间复杂度都是$O(N)$.因此,算法增加的时间复杂度是$O(N)$.而且邻域参数的动态改变优化了数据的邻域结构,从而加快了后继的嵌入过程.因此,整体上并未增加时间复杂度,这可以从后面的实验结果得到证明.另外,算法中的初始邻域大小k的取值会影响局部测地距离的估计,若取得太小,容易产生不连通的邻域图,导致局部测地距离的估计偏差,进而使得局部邻域参数的计算不准确.但我们可以采用文献[12]的方法来构造连通的邻域图,从而保证算法在任意情况下都可成功运行.

根据 *ComputeNbrSizes*(*X*,*k*)算法,我们可以得出 HLLE 的改进算法 VK-HLLE.为保持算法的清晰性,我们给出完整的 VK-HLLE 算法如下:

算法 2. VK-HLEE($\boldsymbol{X}$,k,d).

/*输入中,$\boldsymbol{X}$ 是高维观察数据,k 为初始邻域大小,d 是低维参数空间的维数;输出是低维参数空间 $\boldsymbol{W}$*/

1) 采用*ComputeNbrSizes*(*X*,*k*)计算每个点$\boldsymbol{x}_i$的邻域参数k_i,并根据k_i和欧氏距离修正$\boldsymbol{x}_i$的k-邻域为k_i-邻域,然后将k_i-邻域表达为中心化的行向量$k_i\times n$矩阵M^i.

2) 采用奇异值分解每个邻域矩阵M^i,将其正交向量V的前d个分量作为其切空间.

3) 求切空间的Hessian矩阵.当d=2 时,根据切空间中的点形成如下矩阵:X^i=[1 $V_{\cdot,1}$ $V_{\cdot,2}$ $(V_{\cdot,1})^2$ $(V_{\cdot,2})^2$ $(V_{\cdot,1}\times V_{\cdot,2})$],其中,$V_{\cdot,1}$表示切空间中所有点的第 1 个维的值.对$d$>2 采用相同的方法创建 1+$d$+$d$($d$+1)/2 列的矩阵,然后用Gram-Schmidt正交化X^i产生新的正交矩阵,并将其转置后取最后的d(d+1)/2 列构成Hessian矩阵H^i.

4) 构造二次型 $H_{ij}=\sum_l\sum_r((H^l)_{r,i}(H^l)_{r,j})$,对$H$=($H_{ij}$)进行特征分析,获取其($d$+1)个最小特征值对应的($d$+1)维子空间,第 1 个特征值 0 对应于常函数,接下来的 d 个特征向量就构成 d 维空间,对其选择一

个正交基,变换就可获得要恢复的参数空间 W.

算法VK-HLLE中的第 1 步调用*ComputeNbrSizes*(X,k)计算任意点x_i的新邻域参数k_i和新邻域,余下的步骤与HLLE相同,因此,复杂的数学推导和更详细的算法描述见HLLE原文[5].VK-HLLE与HLLE一样都只需要N个$k_i \times k_i$稀疏的特征问题计算,而ISOMAP需要一个$N \times N$密集的特征问题求解.如果N很大,则VK-HLLE比ISOMAP在时间上有更显著的优越性.而从后面的实验结果来看,VK-HLLE的降维性能比HLLE,ISOMAP和LLE都要好.

4 实验分析

实验比较 VK-HLLE 与 HLLE,LLE 和 ISOMAP 方法的嵌入性能和时间复杂度.4 种方法均采用 matlab 实现,其中,HLLE,LLE 和 ISOMAP 采用原作者提供的 matlab 代码,VK-HLLE 由作者实现.HLLE,LLE 和 ISOMAP 的实验参数选择各自提供的原始参数,它们应该是原作者经过实验选择的较好参数,即 LLE 和 HLLE 设置邻域 k=12, ISOMAP 设置 k=7.VK-HLLE 中的邻域参数 k 与 LLE 和 HLLE 相同,以保持严格的可比性.实验数据是 Swiss Roll Surface,它是 HLLE,LLE,ISOMAP 等都采用的标准测试数据.下面的若干实验全部采用 HLLE 的方法和代码从 Swiss Roll Surface 上采样数据规模为若干个点的长方形但从其中心移去一个小长方形的非凸数据.

4.1 性能分析

实验 1. 比较 4 种方法在数据的参数空间非凸且密集无噪音情况下的性能.我们从 Swiss Roll Surface 采样数据规模为 800 个点的多个数据集,然后运行 4 种方法.分析发现,HLLE 在部分情况下能够较完美地将数据嵌入在二维空间,但不够稳定,ISOMAP 是稳定的,但总是将去除的区域强烈膨胀,并扭曲其余的数据点.LLE 在性能上是最差的,绝大多数情况下都得不到正确结果.而 VK-HLLE 是稳定的,在绝大多数情况下能够较完美地将数据嵌入在二维空间,其中心移去的一个小长方形也能在嵌入的二维空间中正确体现.图 3 是其中的一个结果,不难看出,VK-HLLE 表现最好.

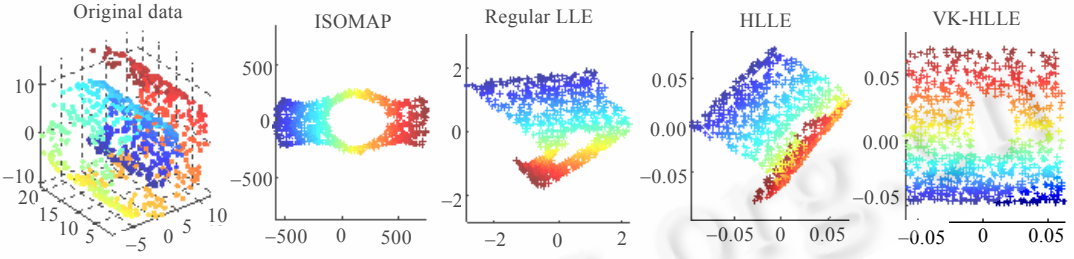

Fig.3 Embedding results on non-convex well sampled data sets

图 3 非凸密集数据集上的嵌入结果

实验 2. 比较 4 种方法在数据稀疏情况下的性能.现实中的很多数据都难以获得足够多的采样,因而是稀疏的,现有的很多算法都难以处理.我们从 Swiss Roll Surface 上随机采样数据规模为 400 点的稀疏数据集,结果如图 4 所示.很明显,HLLE 和 LLE 是混乱的,ISOMAP 将去除的区域强烈膨胀,并扭曲其余的数据点,而 VK-HLLE 则能够较完美地将数据嵌入在二维空间.

实验 3. 比较 4 种方法在数据密集但有噪音情况下的性能.我们从 Swiss Roll Surface 上随机采样 800 个点,然后叠加均值为 0 和方差为 0.4 的高斯噪音,4 种方法的结果如图 5 所示.很明显,HLLE 和 LLE 的效果不理想,ISOMAP 将去除的区域膨胀,并扭曲其余的数据点,而 VK-HLLE 则能够较好地将数据嵌入在二维空间.经过多次实验还发现,VK-HLLE 和其他方法一样都受噪音的影响,不够稳定,在部分情况下也不能正确嵌入,原因是噪声影响了局部测地距离的估计,导致最终的嵌入偏差.

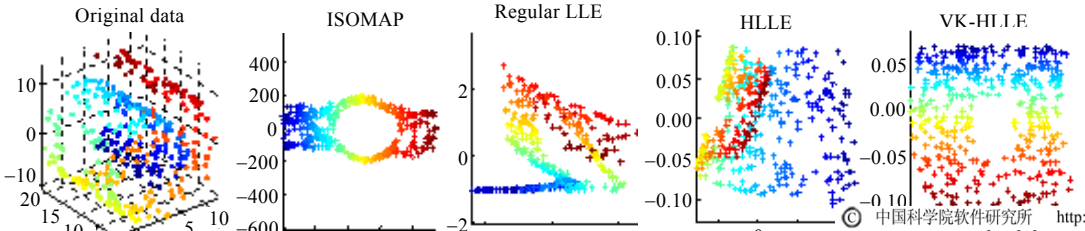

Fig.4 Embedding results on sparse data sets

图 4 在稀疏数据上的嵌入结果

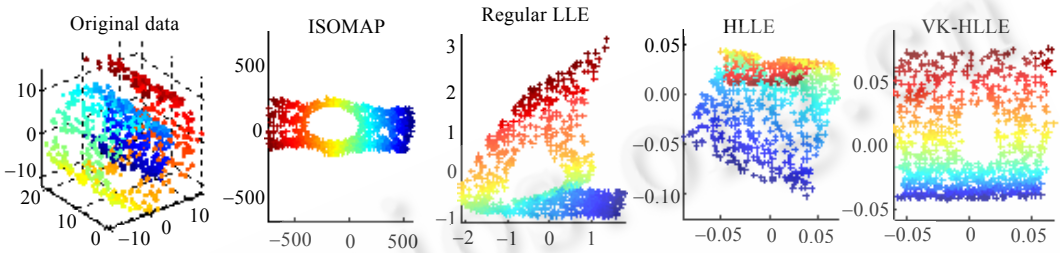

Fig.5 Embedding results on well sampled noisy data sets

图 5 在密集但含噪音数据集上的嵌入结果

实验 4. 分析 HLLE 和 VK-HLLE 对邻域大小的敏感性.我们从 Swiss Roll Surface 上随机采样 800 个点的数据,并叠加均值为 0 和方差为 0.3 的高斯噪音.多次运行发现,VK-HLLE 对邻域大小有更强的鲁棒性,受影响的程度明显比 HLLE 要小,图 6 是其中的一个样例.

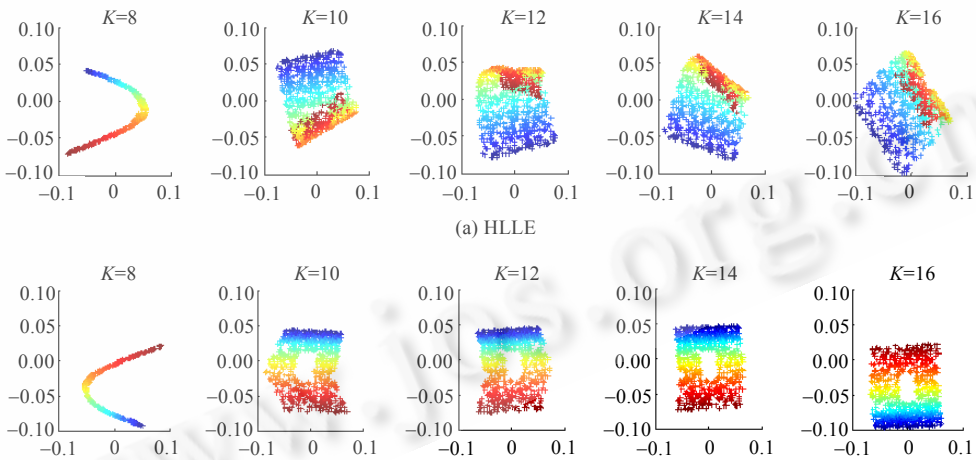

Fig.6 Embedding results of HLLE and VK-HLLE against neighborhood sizes

图 6 HLLE 和 VK-HLLE 随邻域大小变化的嵌入效果

4 类实验一致证实了 VK-HLLE 的有效性,无论在参数空间非凸、数据带噪声还是稀疏的情况下,VK-HLLE 都一致地比 HLLE,LLE 和 ISOMAP 优越,而且 VK-HLLE 比 HLLE 对邻域大小有更强的鲁棒性.

4.2 时间分析

我们从Swiss Roll Surface上依次采样 500,1 000,1 500,2 000,2 500 个点的 5 类规模的数据样本,每类规模的样本随机采样 5 次,记录LLE,HLLE,ISOMAP和VK-HLLE等 4 种方法分别运行这 5 个样本的平均时间作为该规模的时间,则 4 种方法在 5 类规模数据上的平均时间见表 1.不难发现,VK-HLLE与HLLE很接近,并有减少趋势,

规模越大越明显.主要原因是,计算邻域参数增加的时间很少,而优化后的邻域却加速了后继的嵌入过程[17].虽然LLE性能是相对较差的算法,但运算速度最快.而ISOMAP对数据规模十分敏感,时间增长幅度最大,对大规模数据可能不太适用.

Table 1 Average running time of the four approaches on the five data sets with the different size (s)

表 1 4 种方法在 5 种样本规模上的平均运行时间比较(秒)

Data set size	500	1 000	1 500	2 000	2 500
LLE	0.587 6	1.653 0	3.540 8	6.312 6	10.250 0
HLLE	**2.540 6**	**18.237 6**	**59.681 0**	**139.674 6**	**279.850 0**
VK-HLLE	**2.522 0**	**18.250 0**	**59.565 8**	**139.065 6**	**274.200 0**
ISOMAP	7.174 8	55.537 2	182.172 0	442.328 0	862.050 0

5 结 论

提出一种邻域参数动态确定的新局部线性嵌入方法 VK-HLLE.它采用 Hessian 局部线性嵌入的概念框架,但用每个点的局部邻域估计此邻域内任意点之间的近似测地距离,然后根据近似测地距离与欧氏距离之间的关系动态确定该点的邻域大小,并以此邻域大小构造新的局部邻域.算法几何意义清晰,能够处理现实中大量存在的非均匀分布流形,特别是在观察数据稀疏和观察数据带噪音等情况下都比现有算法有更强的鲁棒性.而且与 HLLE 相比,未增加算法的时间复杂度,对大规模数据还有减少趋势;与 ISOMAP 相比,时间上的优势更加明显.VK-HLLE 的缺陷是,与 HLLE 一样,对观察数据的维数敏感,对文本之类高达上万维的观察数据不合适,需要采取进一步的措施.另外,其抗噪音能力仍有待进一步提高.

References:

[1] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000,290(5500): 2319−2323.

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

[3] Geng X, Zhan DC, Zhou ZH. Supervised nonlinear dimensionality reduction for visualization and classification. IEEE Trans. on Systems, Man and Cybernetics, 2005,35(6):1098−1107.

[4] Law MHC, Jain AK. Incremental nonlinear dimensionality reduction by manifold learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006,28(3):377−391.

[5] Donoho DL, Grimes C. Hessian eigenmaps: Locally linear embedding, techniques for high-dimensional data. PNAS, 2003,100(10): 5591−5596.

[6] Kouropteva O, Okun O, Pietikainen M. Incremental locally linear embedding. Pattern Recognition, 2005,38(10):1764−1767.

[7] de Ridder D, Loog M, Reinders MJT. Local fisher embedding. In: Proc. of the 17th Int'l Conf. on Pattern Recognition, Vol.2. 2004. 295−298. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1334176

[8] Xu R, Jiang F, Yao HX. Overview of manifold learning. CAAI Trans. on Intelligent Systems, 2006,1(1):44−51 (in Chinese with English abstract).

[9] Zhao LW, Luo SW, Zhao YC, Liu YH. Study on the low-dimensional embedding and the embedding dimensionality of manifold of high-dimensional data. Journal of Software, 2005,16(8):1423−1430 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/1423.htm

[10] He L, Zhang JP, Zhou ZH. Investigating manifold learning algorithms based on magnification factors and principal spread directions. Chinese Journal of Computers, 2005,28(12):2000−2009 (in Chinese with English abstract).

[11] Balasubramanian M, Schwartz EL. The ISOMAP algorithm and topological stability. Science, 2002,295(5552):7.

[12] Yang L. Building *k* edge-disjoint spanning trees of minimum total length for isometric data embedding. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(10):1680−1683.

[13] Saxena A, Gupta A, Mukerjee A. Non-Linear dimensionality reduction by locally linear ISOMAPs. LNCS 3316, Springer-Verlag, 2004. 1038−1043. http://www.springerlink.com/content/14754g3k7gp4lt2y/

[14] Kouropteva O, Okun O, Pietikainen M. Selection of the optimal parameter value for the locally linear embedding algorithm. In: Proc. of the 1st Int'l Conf. on Fuzzy Systems and Knowledge Discovery. Singapore, 2002. 359−363. http://citeseer.ist.psu.edu/kouropteva02selection.html

[15] Samko O, Marshall AD, Rosin PL. Selection of the optimal parameter value for the ISOMAP algorithm. Pattern Recognition Letters, 2006,27(9):968−979.

[16] Hou YX, Wu JY, He PL. Locally adaptive non linear dimensionality reduction. Computer Applications, 2006,26(4):895−897 (in Chinese with English abstract).

[17] Wen GH, Jiang LJ, Wen J, Shadbolt NR. Clustering-Based nonlinear dimensionality reduction on manifold. LNAI 4099, Springer-Verlag, 2006. 444−453. http://www.springerlink.com/content/k036115h1w031077/

[18] Wen GH, Jiang LJ, Shadbolt NR. Using graph algebra to optimize neighborhood for isometric mapping. In: Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence (IJCAI 2007). 2007. 2398−2403. http://www.ijcai.org/papers07/Papers/IJCAI07-386.pdf

[19] Yang L. Distance-Preserving projection of high-dimensional data for nonlinear dimensionality reduction. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004,26(9):1243−1247.

[20] Wen GH, Jiang LJ, Wen J, Shadbolt NR. Performing locally linear embedding with adaptable neighborhood size on manifold. LNAI 4099, Springer-Verlag, 2006. 985−989. http://www.springerlink.com/content/958007m6vn192017/

附中文参考文献:

[8] 徐蓉,姜峰,姚鸿勋.流形学习概述.智能系统学报,2006,1(1):44−51.

[9] 赵连伟,罗四维,赵艳敞等.高维数据的低维嵌入及嵌入维数研究.软件学报,2005,16(8):1423−1430. http://www.jos.org.cn/1000-9825/16/1423.htm

[10] 何力,张军平,周志平.基于放大因子和延伸方向研究流形学习算法.计算机学报,2005,28(12):2000−2009.

[16] 侯越先,吴静怡,何丕廉.基于局域主方向重构的适应性非线性维数约减.计算机应用,2006,26(4):895−897.

文贵华(1968－),男,湖北利川人,博士,副研究员,主要研究领域为创新计算,数据挖掘与知识发现,机器学习,认知几何计算.

文军(1964－),男,副教授,主要研究领域为创新计算,机器学习,智能软件.

江丽君(1971－),女,讲师,主要研究领域为创新教育,智能 CAD.