ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.4, April 2009, pp.815−824
doi: 10.3724/SP.J.1001.2009.03425

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

基于测地线距离的广义高斯型 Laplacian 特征映射*

曾宪华[1,2+], 罗四维[1], 王 娇[1], 赵嘉莉[1]

[1](北京交通大学 计算机与信息技术学院,北京 100044)

[2](西华师范大学 计算学院,四川 南充 637002)

Geodesic Distance-Based Generalized Gaussian Laplacian Eigenmap

ZENG Xian-Hua[1,2+], LUO Si-Wei[1], WANG Jiao[1], ZHAO Jia-Li[1]

[1](School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

[2](School of Computer, China West Normal University, Nanchong 637002, China)

+ Corresponding author: E-mail: xianhuazeng@gmail.com

Zeng XH, Luo SW, Wang J, Zhao JL. Geodesic distance-based generalized Gaussian Laplacian eigenmap. *Journal of Software*, 2009,20(4):815−824. http://www.jos.org.cn/1000-9825/3425.htm

Abstract: The conventional Laplacian Eigenmap preserves neighborhood relationships based on Euclidean distance, that is, the neighboring high-dimensional data points are mapped into neighboring points in the low-dimensional space. However, the selections of neighborhood may influence the global low-dimensional coordinates. In this paper, both the geodesic distance and generalized Gaussian function are incorporated into Laplacian eigenmap algorithm. At first, a generalized Gaussian Laplacian eigenmap algorithm based on geodesic distance (GGLE) is proposed. The global low-dimensional coordinates obtained by GGLE have different clustering properties when different generalized Gaussian functions are used to measure the similarity between the high-dimensional data points. Then, this paper utilizes these properties to further propose the ensemble-based discriminant algorithm of the above-motioned GGLE. The main advantages of the ensemble-based algorithm are: The neighborhood parameter K is fixed and to construct the neighborhood graph and geodesic distance matrix needs one time only. Finally, the recognition experimental results on wood texture dataset show that it is an efficient ensemble discriminant algorithm based on manifold.

Key words: manifold learning; Laplacian eigenmap; generalized Gaussian function; geodesic distance; ensemble

摘 要: 传统的 Laplacian 特征映射是基于欧氏距离的近邻数据点的保持,近邻的高维数据点映射到内在低维空间后仍为近邻点,高维数据点的近邻选取最终将影响全局低维坐标.将测地线距离和广义高斯函数融合到传统的 Laplacian 特征映射算法中,首先提出了一种基于测地线距离的广义高斯型 Laplacian 特征映射算法(geodesic

* Supported by the National Natural Science Foundation of China under Grant Nos.60773016, 60373029 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z168 (国家高技术研究发展计划(863)); the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20050004001 (国家教育部博士点基金); the Scientific Research Foundation of Sichuan Provincial Education Department of China under Grant No.07ZA121 (四川教育厅重点项目)

Received 2008-03-21; Accepted 2008-07-24

distance-based generalized Gaussian LE,简称 GGLE),该算法在用不同的广义高斯函数度量高维数据点间的相似度时,获得的全局低维坐标呈现出不同的聚类特性;然后,利用这种特性进一步提出了它的集成判别算法,该集成判别算法的主要优点是:近邻参数 K 固定,邻接图和测地线距离矩阵都只构造一次.在木纹数据集上的识别实验结果表明,这是一种有效的基于流形的集成判别算法.

关键词: 流形学习;Laplacian 特征映射;广义高斯函数;测地线距离;集成

中图法分类号: TP181 **文献标识码**: A

流形学习是一个具有基础性和前瞻性的研究方向,由于有着广阔的应用前景,近年来已成为机器学习、模式识别、数据挖掘等领域的研究热点之一,涌现出一批参数少、运算快、易求全局最优解的非线性流形学习算法,如等距映射(isometrical mapping,简称 ISOMAP)算法[1]、Laplacian 特征映射(Laplacian eigenmap,简称 LE)算法[2]、局部切空间对齐(local tangent space alignment,简称 LTSA)算法[3]、局部线性嵌入(locally linear embedding,简称 LLE)算法[4]等.这些算法都要通过构造邻接图来表示高维数据的局部几何结构,然后在不同假设条件下确定数据点之间的某种关系(如 ISOMAP 是估计数据点之间的测地线距离,LE 是近邻数据点之间的相似度,LTSA 是将每个数据点的邻域数据点投影到局部切空间上并寻求局部切坐标整合到全局坐标的仿射变换关系,LLE 是寻找每个近邻点和它的近邻数据点之间的线性组合关系),利用这些不同关系构造全局低维坐标.它们所面临的一个共同问题是选择近邻构造邻接图,近邻选取最终将影响全局低维坐标.詹德川和周志华提出集成 ISOMAP 的流形学习算法[5],该算法通过选择多个近邻参数 K,对每一个 K 值都运行 ISOMAP 获取低维坐标,然后将这些低维坐标加权平均得到最终低维坐标.张军平等人提出基于集成的判别流形学习算法[6],每一个近邻参数运行 ULLELDA(united LLE and linear discriminant analysis)[7]产生独立的子空间集合,独立学习分类器,获得了较好的集成分类结果.这两种算法都要多次选择近邻参数,多次构造邻接矩阵,集成 ISOMAP 要多次计算测地距离矩阵,集成 ULLELDA 却要多次构造重建矩阵,也就是说,这两种集成流形学习在取得好的效果的同时,时间消耗也很大.

传统的 Laplacian 特征映射是基于欧氏距离 K 近邻或 ε 邻域的近邻数据点的保持,近邻的高维数据点映射到内在低维空间后仍为近邻点[2,8,9].那么,多大程度的近邻点以及多少数量的近邻点应该在低维空间需要保持呢?当不同数量的近邻点需要保持时,Laplacian 特征映射需要重新构造邻接图和邻接权矩阵.同时,对于位于嵌入在高维空间中的低维流形上的数据点来说,如若增加近邻点的数量,仍然用欧氏距离来度量近邻关系是不合适的,测地距离应该更为合理.本文将测地线距离和广义高斯函数融合到传统的 Laplacian 特征映射算法中,首先提出了一种基于测地线距离的广义高斯型 Laplacian 特征映射算法(GGLE),该算法在用超高斯、高斯和次高斯函数度量数据点之间的相似度时所强调的局部近邻保持的程度是不同的,获得的全局低维坐标也呈现出不同的聚类特性.然后,利用这种特性进一步提出了 GGLE 的集成判别算法.该集成判别算法的显著不同是近邻参数 K 固定,邻接矩阵和测地线距离矩阵都只构造一次,只需要多次选择广义高斯型函数构造多个 Laplacian 矩阵,获取多个独立的全局低维坐标集合,独立学习分类器,集成分类识别,这是一种更为高效的集成流形学习算法.最后,在人造数据集和真实的木纹数据集上的实验结果也说明了这两种算法的有效性.

1 Laplacian 特征映射

2001 年,Belkin 和 Niyogi 提出了 Laplacian 特征映射算法[2,8],该算法寻求一个能在平均意义上保持流形局部特性的映射,直观看来,近邻的高维数据点映射到内在低维空间后仍为近邻点.若 M 是一个光滑的 m 维黎曼流形,在 D 维欧式空间中的嵌入映射为 $f:M\to R^D$,对于流形上的两个近邻点 $x,y\in M$,光滑映射 f 在 x 处按一阶泰勒展开并利用范数的性质,得到下列不等式:

$$\| f(y)-f(x)\| \le dist_M(x,y)\,\|\nabla f(x)\| + o(dist_M(x,y)) \tag{1}$$

上式中,$\|\nabla f\|$ 表明映射 f 把流形上的近邻点映射到低维欧氏空间后欧氏距离度量的偏离程度.如若约束映射 $\|f\|_{l^2(M)}=1$,则保持局部近邻特性的映射可以通过下列目标函数来实现:

$$\arg\min_{\|f\|_{L^2(M)}=1}\int_M \|\nabla f(x)\|^2 \tag{2}$$

其中,积分是 Riemannian 流形上的标准测度.由于定义 Laplacian-Beltrami 算子 $L(f)\overset{def}{=}-div\nabla(f)$,其中 div 表示向量场的散度,因此,根据 Stokes 公式有 $\int_M\|\nabla f(x)\|^2=\int_M L(f)f$.于是,公式(2)的目标函数等价于最小化下列准则函数:

$$\begin{cases}\min\limits_f \frac{1}{2}\sum_{i,j}(f_i-f_j)^2W_{ij}=\min\limits_Y tr(Y^TLY)\\ \text{subject to: } Y^TY=1 \text{ or } Y^TDY=1\end{cases} \tag{3}$$

其中,$Y=(f_1,f_2,\ldots,f_n)$, $D_{ii}=\sum_j W_{ji}$, $L=D-W$ 是对称的半正定 Laplacian 矩阵.由拉格朗日乘数法,上述优化问题就转化为求 L 的特征函数问题.令特征值按升序排列为 $0=\lambda_0\le\lambda_1\le\lambda_2\le\ldots$,$f_i$ 为对应于 λ_i 的特征函数.显然,f_0 是一个常数函数,把流形上所有的数据点映射为一个点.为了避免这种情况的出现,选择和 f_0 正交的嵌入映射,所以,f_1 就是最优嵌入映射.

根据谱图理论,如果数据均匀采样自高维空间中的低维流形,流形 Laplacian-Beltrami 算子可以由图上的 Laplacian 算子逼近,而流形上 Laplacian-Beltrami 算子对应特征函数的离散逼近,就是图的 Laplacian 矩阵最小的几个特征值对应的特征向量.算法具体步骤如下:

Step 1. 使用 K 近邻或ε邻域的方法构建邻接图 G;若两点 i,j 近邻,则 $G_{ij}=1$.

Step 2. 定义邻接权矩阵 W.权值使用热核(heat kernel)方法,即如果 $G_{ij}=1$,则 $W_{ij}=\exp\{-\|x_i-x_j\|^2/t\}$,否则为 0.或者使用简单方法,如果 $G_{ij}=1$,令权值 $W_{ij}=1$,否则为 0.

Step 3. 特征映射.假设图 G 为连接图(否则对每一个连接部分),计算方程 $L\xi=\lambda D\xi$ 的 $d+1$ 个最小特征值及对应特征向量.令这 $d+1$ 个特征向量为$\{\xi(0),\xi(1),\ldots,\xi(d)\}$,分别对应于从小到大排列的特征值.除去特征值 0 对应的特征向量以外,x_i 在低维空间 R^d 的像可以由$(\xi_i(1),\ldots,\xi_i(d))$给出.

2 基于测地线距离的广义高斯型 Laplacian 特征映射算法

2.1 问题分析与算法描述

传统的 Laplacian 特征映射是基于欧氏距离的 K 近邻或ε邻域的近邻数据点的保持,近邻的高维数据点映射到内在低维空间后仍为近邻点.那么,多大程度的近邻点以及多少数量的近邻点应该在低维空间需要保持呢?当不同数量的近邻点需要保持时,Laplacian 特征映射需要重新计算邻接图构造邻接权矩阵.同时,对于位于嵌入在高维空间中的低维流形上的数据点来说,如若增加近邻点的数量,仍然用欧氏距离来度量近邻关系是不合适的,测地距离应该更为合理.

实验验证如图 1 所示,Swiss_roll 曲面上的 500 个 3 维数据点(如图 1(a)所示),当用欧氏距离度量每一数据点的 40 个近邻点时,通过传统 LE 算法学习到的内在低维坐标如图 1(c)所示,流形两端的数据点几乎混在一起,然而实际上它们应该离得更远;然而,当采用测地线距离(实验时在 5 最近邻图上获得近似测地线距离)度量每一数据点的 40 个近邻点时,通过 LE 算法学习到的内在低维坐标如图 1(d)所示,流形上近邻的点尽可能地近,流形上测地距离远的点仍然较远.目前,ISOMAP 算法是保持任意两点间测地线距离的经典流形学习算法,但又需要避免 ISOMAP 算法中稠密测地线距离矩阵的分解,正如 Belkin 在文献[2,8]中指出,ISOMAP 等度规地保持全局测地距离只对平坦流形(即曲率张量为 0)理论上可行;比如,3 维空间中的 Gauss 曲面映射到 2 维平面上,测地距离就不能得以保持,而会发生不同程度的撕裂与挤压.因此,我们结合在流形上用测地线距离度量数据点之间关系的优点和广义高斯函数的特性(见第 2.2 节的分析)并融合到传统的 Laplacian 特征映射算法中,提出了一种基于测地线距离的广义高斯型 Laplacian 特征映射算法,该算法可以调整结点间的相似度,根据问题需要选择超高斯、高斯或者次高斯函数来体现多大程度的近邻局部特性需要保持,不需要重新计算邻接矩阵.而且,当需要保持近邻关系的数据点邻域增大时,采用测地线距离可以避免欧氏距离度量不合理的缺陷.综上所述,一种更合理

的流形学习算法——基于测地线距离的广义高斯型 Laplacian 特征映射算法描述如下:

Step 1. 使用 K 近邻或ε邻域的方法构建欧氏邻接图 $\boldsymbol{G}$;若两点 i,j 近邻,则该邻接边的距离定义为 Euclidean 距离 $d(i,j)$.

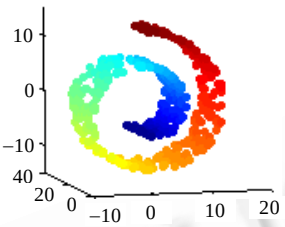
Step 2. 应用 Dijkstra 算法计算图 G 上任意两点间的最短距离构建近似测地线距离矩阵 S.

Step 3. 利用广义高斯型函数计算成对结点之间的相似度,构成相似度矩阵 W.相似度为

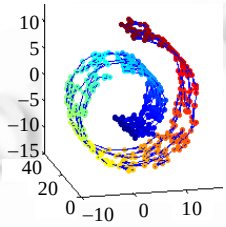
$$W_{ij}=\frac{1}{\Gamma}\exp(-(S_{ij}/\sigma)^{\beta}),$$

其中,Γ是归一化因子;σ表示尺度因子;S_{ij} 表示结点间的测地线距离;β衡量广义高斯曲线的走势,当 $0<\beta<2$ 时,表示用超高斯函数度量相似度,$\beta=2$ 表示用高斯函数度量相似度,$\beta>2$ 表示用次高斯函数度量相似度.

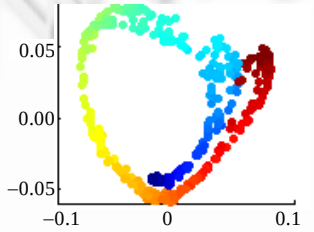
Step 4. 计算内在低维坐标,即计算特征方程 $L\xi=\lambda D\xi$得到 $d+1$ 个最小特征值及对应特征向量.令这 $d+1$ 个特征向量为$\{\xi(0),\xi(1),\ldots,\xi(d)\}$,分别对应于从小到大排列的特征值.除去特征值 0 对应的特征向量以外,x_i在低维空间 R^d 的像可以由$(\xi_i(1),\ldots,\xi_i(d))$给出.



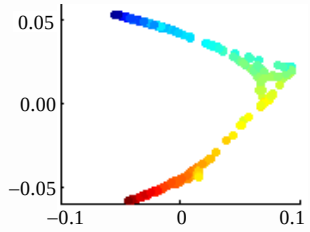
(a) 500 Swiss_roll data points
(a) 500 个 Swiss_roll 数据点



(b) Neighborhood graph (K=5)
(b) 邻接图,K=5



(c) 2D coordinates by using conventional LE (K=40, heat-kernel parameter t=30)
(c) 传统的 LE 获得的二维坐标(K=40,热核参数 t=30)



(d) 2D coordinates by using geodesic distance and LE (K=40, heat-kernel parameter t=30)
(d) 采用测地线距离和 LE 获得的二维坐标(K=40,热核参数 t=30)

Fig.1 Reasonableness of using geodesic distance in LE
图 1 在 LE 算法中采用测地距离的合理性

2.2 GGLE算法分析

在上述学习算法中,Step 3采用自变量为测地线距离的广义高斯型函数计算出的相似度矩阵 W是一个稠密矩阵,但是由于广义高斯型函数在[0,+∞]上单调下降,且随着测地距离的增加以指数形式下降到一个很小的范围内,即测地距离越大,相似度越小.显然,任意两点间相似度都不为 0.我们认为,位于同一流形上的点之间是有联系的,也是有一定相似度的,因此,广义高斯型函数被用作计算成对结点之间的相似度是合理的.而且,超高斯、高斯和次高斯函数度量的相似度随着测地距离的变化曲线发生变化也有显著的不同,随着β的增加,近邻点的相似度增高,测地线距离小于尺度因子σ(类似于方差)的近邻点的相似度逐渐升高,如图 2 所示.特别地,

$$\lim_{\beta\to+\infty}W_{ij}=\lim_{\beta\to+\infty}\frac{1}{\Gamma}\exp(-(S_{ij}/\sigma)^{\beta})=\begin{cases}1, & \text{当}S_{ij}<\sigma,\ \Gamma=1\\ 1/e, & \text{当}S_{ij}=\sigma,\ \Gamma=1\\ 0, & \text{当}S_{ij}>\sigma,\ \Gamma=1\end{cases}\tag{4}$$

尺度因子σ确定算法能够保持的近邻点的最大容量,在最大容量范围内,不同的β实现不同程度的近邻保持;当$\beta\to+\infty$时,类似于近邻数为最大容量,每个点与其所有近邻点的相似度都趋近于最大值 1.因此,可以根据实际需要强调局部近邻保持程度的不同选择β,即需要保持较大区域的近邻点时选择次高斯(β>2),需要保持较小区域的近邻点时选择超高斯(β<2),两者之间的需求选择高斯(β=2).另外,在实际应用中,为了提高计算效率,对一些很小的相似度赋值为 0,即在算法的 Step 3 中,当相似矩阵 W 的元素小于某阈值δ时赋值为 0(按统计学规律,阈值δ一般取 2σ对应的相似度),这样,W 就是一个稀疏矩阵,Step 4 就是一个稀疏半正定矩阵的特征分解,时间复杂度就从满稠密矩阵特征分解的 $O(n^3)$下降到求稀疏矩阵的 d 个最小特征值对应的特征向量的 $O(dpn^2)$,其中,n 为矩阵的阶数,d 为低维维数,p 为非零元素的比例.

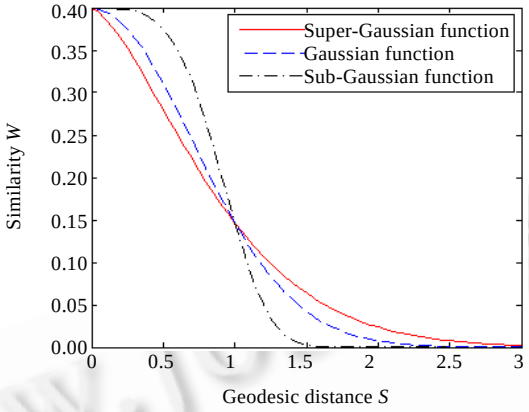


Fig.2 Generalized Gaussian curves versus geodesic distance

图 2 定义在测地线距离上广义高斯函数曲线

3 基于测地线距离的广义高斯型 Laplacian 特征映射的集成判别算法(EGGLE)

3.1 问题分析和算法描述

集成学习[10,11]是使用多个学习器来学习同一问题的学习系统,可以取得比单一学习器更好的性能.集成学习技术广泛应用于神经网络及许多模式识别领域中.詹德川和周志华提出了集成 Isomap 流形学习算法[5],该算法通过选择多个近邻参数 K,对每一个 K 值都运行 Isomap 获取低维坐标,然后将这些低维坐标加权平均得到最终低维坐标,在可视化实验中取得了较好的效果.张军平等人提出基于集成的判别流形学习算法[6],该算法核心是针对 LLE 的,并且论述了 LLE 获取的最小 d 个特征值和特征向量很小并且很接近,同一个近邻参数多次运行 LLE 的主要特征可能发生交替,因此 LLE 是不稳定的.张军平等人从而提出了一种集成方式:每一个近邻参数运行 ULLELDA[7]产生独立的子空间集合,独立学习分类器,获得了较好的集成分类结果.这两种算法都要多次选择近邻参数,多次构造邻接矩阵,集成 Isomap 要多次计算测地线距离矩阵,集成 ULLELDA 却要多次构造重建矩阵.也就是说,这两种集成流形学习在取得好的效果的同时,时间消耗也很大.本文上一节提出的基于测地线距离的广义高斯型 Laplacian 特征映射算法,通过在 Step 3 选择广义高斯型函数,以测地线距离为自变量计算成对结点之间相似度.正如上一节分析,超高斯、高斯和次高斯函数在度量相似度时强调局部近邻保持的程度是不同的,低维坐标呈现出不同的聚类特性,从而它们的不同 Laplacian 矩阵的特征分解所得到的低维坐标,对于分类效果也应该是不同的.因此,本节选择多个广义高斯型函数获取多个相似度矩阵,每个相似度矩阵对应的 Laplacian 矩阵获得独立的低维子空间,独立学习分类器,再集成确定分类结果,这就是本节将提出的基于测地线距离的广义高斯型 Laplacian 特征映射的集成判别算法.如上所述,该集成学习算法与文献[5,6]中的集成流形学习算法的显著不同是:固定的近邻参数 K,邻接矩阵和测地线距离矩阵都只构造一次,只需要多次选择广义高斯型函数构造多个 Laplacian 矩阵,获取多个独立的低维空间坐标集合,独立学习分类器,集成分类识别,这是一种

更为高效、合理的集成流形学习算法.

在识别任务中,已标记的训练样本集记为 $\{x_j,t_j\}_{j=1}^{l}$,其中,t_j 为类别标记,未标记样本集记为 $\{x_j\}_{j=l+1}^{l+u}$,假定这些数据都来自于嵌入在高维空间的低维流形上.在本文第 2 节提出的基于测地线距离的广义高斯型 Laplacian 特征映射中,Step 3 度量相似度的广义高斯型函数 $\frac{1}{\Gamma}\exp\{-(S_{ij}/\sigma)^{\beta}\}$ 的参数 β 取值体现了不同的函数类型:$0<\beta<2$,对应超高斯函数;$\beta=2$,对应高斯函数;$\beta>2$,对应次高斯函数.不同的 β 值,GGLE 获得的低维坐标所强调的局部近邻保持的程度也是不同的.因此,本文提出的基于测地线距离的广义高斯型 Laplacian 特征映射的集成判别算法(ensemble-based discriminant algorithm of GGLE,简称 EGGLE)描述如下:

Step 1. 对所有的($l+u$)个样本使用 K 近邻或 ε 邻域的方法构建欧氏邻接图 G;若两点 i,j 近邻,则该邻接边的距离定义为 Euclidean 距离 $d(i,j)$.

Step 2. 应用 Dijkstra 算法计算图 G 上任意两点间的最短距离,构建近似测地线距离矩阵 S.

Step 3. 训练第 n 个学习器(初时值为 1,终止为 N),如果 $n>N$,转 Step 5,否则执行以下几步:

a) 选择一个 β,计算相似度矩阵 W,其中,$W_{ij}=\frac{1}{\Gamma}\exp\{-(S_{ij}/\sigma)^{\beta}\}$.

b) 计算内在低维坐标 $\{\xi_j\}_{j=1}^{l+u}$,其中,低维坐标由计算特征方程 $L\xi=\lambda D\xi$ 的 $d+1$ 个最小特征值对应特征向量的第 2 到第 $d+1$ 个特征向量构成,ξ_j 表示第 j 个数据点的低维坐标.

c) 对标记训练样本对应的低维表示 $\{\xi_j,t_j\}_{j=1}^{l}$ 训练分类器 T_n.

d) 分类器 T_n 识别 $\{\xi_j\}_{j=l+1}^{l+u}$ 的类别 $L_n=\{t_j\}_{j=l+1}^{l+u}$.

Step 4. $n=n+1$,转 Step 3.

Step 5. 多数投票法确定最终分类结果,即是对 N 个分类器 $T_1,\ldots,T_N$ 对应的所有识别结果 $L_n=\{t_j\}_{j=l+1}^{l+u}$,$n=1,\ldots,N$,采用多数投票法得到未标记样本的类别标记.

3.2 EGGLE算法复杂度分析

集成流形学习技术的时间复杂度由观测数据点个数 n、观测数据的维数 D、内在维数(目标维数)d、最近邻参数 K 以及集成的学习器数目 N 等几个因素决定.本节就当前的几种集成流形学习算法 Ensemble-Isomap[5],En-ULLELDA[6]和本文所提出的 EGGLE 算法在时间复杂度上进行比较.

由于 3 种算法解决问题的目标及采用的辅助技术不完全相同——Ensemble-Isomap 主要是为了高维数据的可视化,En-ULLELDA 却融入了 LDA 技术,因此,为了表明清晰的可比性,只将 3 种集成技术各学习 N 个学习器发现 N 种低维表示的时间复杂度进行比较.

Isomap,LLE 及 LE 这 3 种算法都由 3 步构成,且第 1 步相同,计算复杂度上都是第 3 步对 $n\times n$ 阶的对称矩阵进行特征分解,时间复杂最高[2,12−14],Isomap 是对非稀疏矩阵特征分解,其时间复杂度是 $O(n^3)$;LLE 和 LE 都是对稀疏矩阵的特征分解,如若非零元素的比例是 $p(<1)$,则 LLE 和 LE 的时间复杂度都可表示为 $O(dpn^2)$.

首先分析一下 Ensemble-Isomap,En-ULLELDA 和本文的 EGGLE 集成的 N 个学习器的前两步,设集成算法中集成了 N 个学习器,Ensemble-Isomap 要计算 N 次邻接矩阵需要 $O(Nn\log n)$,计算 N 个 $n\times n$ 阶的测地线距离矩阵需要 $O(Nn^2\log n)$(注:利用 Fibonacci 堆的迪杰斯特(Dijkstra)算法快速求解 $n\times n$ 阶的测地线距离矩阵的时间复杂度只需 $O(n^2\log n)$);En-ULLELDA 计算 N 次邻接矩阵需要 $O(Nn\log n)$和 N 次重建权需要 $O(NDnK^3)$;本文的 EGGLE 算法集成的每一个学习器都对相似矩阵进行了稀疏化(即对相似度小于某个阈值时赋值为 0,设非零元素的比例为 p),该集成算法中只计算一个邻接矩阵和一个测地矩阵的时间复杂度分别为 $O(n\log n)$和 $O(n^2\log n)$,计算 N 次相似矩阵只需 $O(Ndpn)$.

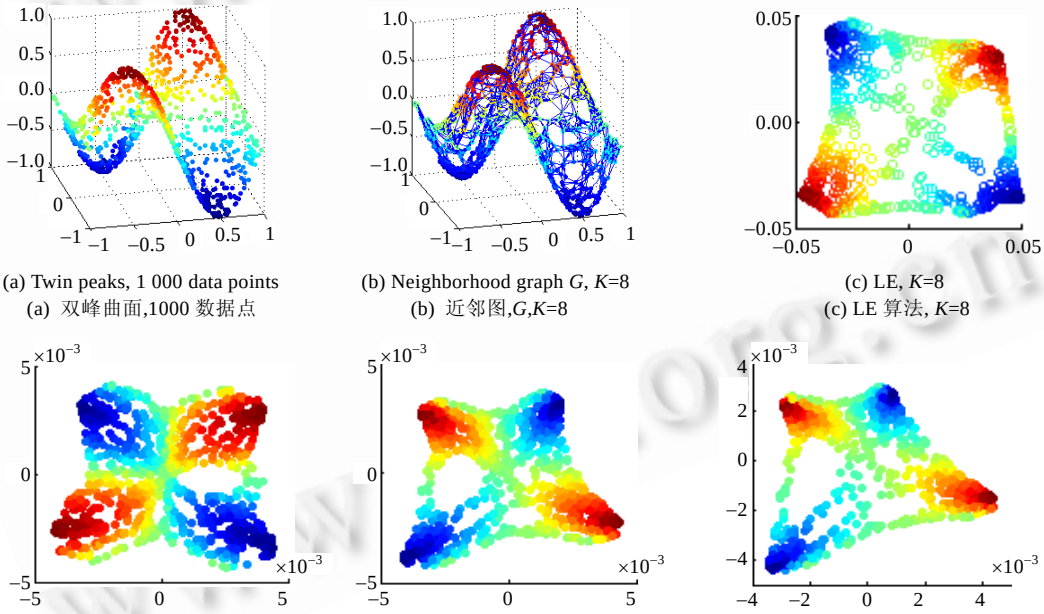
最后,由于以 Isomap,LLE 及 LE 为基础的集成技术对矩阵进行特征分解求低维坐标的时间复杂最高,因此,Ensemble-Isomap 的时间复杂度是 $O(Nn^3)$;En-ULLELDA 不计 LDA 运算量的时间复杂度是 $O(Ndpn^2)$,当本文的 EGGLE 不计提前计算的一个和集成学习器数目 N 无关的 $n\times n$ 阶测地线距离矩阵的计算量时,EGGLE 的时间复

杂度也可表示为 $O(Ndpn^2)$.故 Ensemble-Isomap 的时间复杂度最高,En-ULLELDA 和本文的 EGGLE 相当.通常,En-ULLELDA 计算 N 次重建权的复杂度比 EGGLE 计算 N 次相似矩阵的复杂度要高得多.

4 实 验

4.1 人造数据集上的可视化实验

本实验分别应用 Laplacian 特征映射和我们提出的基于测地线距离的广义高斯型 Laplacian 特征映射算法对 3 维双峰曲面(twin peaks)上的数据集(如图 3(a)所示)进行实验,分析不同程度的近邻数据需要保持时新算法的可视化效果.在 3 维双峰曲面 $z=\sin(\pi x)\tanh(3y)$的定义域$[-1,1]\times[-1,1]$对应的曲面上随机采样 1 000 个样本,实验中固定近邻数 $K=8$,图 3(b)是这些数据点的邻接图,数据点之间的边表示其中一个数据点是另一个数据点的 K 近邻.显然,该曲面的内在维数是 2,所以算法展示了样本集对应的内在二维坐标的可视化效果.图 3(c)是应用 Laplacian 特征映射的二维可视化效果,其中热核参数 $t=1$,图 3(d)~图 3(f)是我们的算法在广义高斯函数$\beta=0.5$, 2,8 这 3 种情况时学习到的二维可视化效果,其中,此实验中方差常数$\sigma=1$.为了降低时间复杂度,将测地线距离大于 2σ的点对之间相似度约束为 0.实验结果表明,新算法无须修改近邻数 K,也不需要重复建邻接矩阵,就能保持不同程度的近邻点,而且采用测地线距离的广义高斯计算结点间的相似度,展示的全局几何结构更为合理,避免了近邻数较多时常规 Laplacian 特征映射用欧氏距离或 1 度量近邻距离的不合理性.同时,不同的β对应超高斯、高斯和次高斯,实现了不同近邻数的保持,低维坐标呈现出不同邻接程度的几何结构,同时也具有不同的聚类效果.另外,我们还进行了大量的实验,发现,当 $0<\beta<2$ 时,保持每个点的近邻程度低,局部聚类明显(如图 3(d)所示),但保持全局几何结构不如次高斯函数.而当$\beta>2$ 时,随着β的增加,局部聚类降低,全局几何结构的保持程度改善;当$\beta>16$ 时,低维几何结构整体上没有较大的区别,正如第 2.2 节分析的那样,算法几乎达到能够保持近邻点的最大容量(尺度因子σ确定).



(a) Twin peaks, 1 000 data points
(a) 双峰曲面,1000 数据点

(b) Neighborhood graph G, K=8
(b) 近邻图,G,K=8

(c) LE, K=8
(c) LE 算法, K=8

(d) GGLE, K=8, β=0.5
(d) GGLE 算法, K=8,β=0.5

(e) GGLE, K=8, β=2
(e) GGLE 算法,K=8,β=2

(f) GGLE, K=8, β=8
(f) GGLE 算法,K=8,β=8

Fig.3 2-Dimensional visualization on synthetic Twin-Peaks dataset

图 3 人造数据集 Twin-Peaks 上的二维可视化

4.2 GGLE的集成判别算法在木纹识别上的实验

为了测试提出的基于流形的集成判别算法在识别方面的能力,在固定近邻参数 K 的情况下,将本文的基于流形的集成判别算法和文献[9]的半监督流形学习算法框架下的 Laplacian 特征映射做了木纹识别的对比实验.实验中采用的数据集是 USC-SIPI 图像数据库[15]中木纹图像分割获得的木纹数据集,该数据集包含 4 种木纹朝向的 1 024 张 32×32 灰度图像(如图 4 所示),分别是 0°,45°,90°,135°这 4 个朝向的木纹图像,每种木纹有 256 张,它们的灰度值被归一化到[0,1]区间,并且每幅图像被看成是一个 1 024 维的观测数据点.

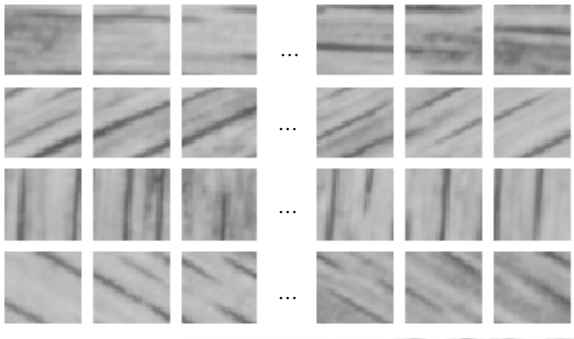


Fig.4 Wood texture dataset

图 4 木纹数据集

实验中固定近邻参数 K=6,由于只使用了近邻信息,分类器都采用最近邻分类器(nearest neighbor classifier);又由于实验数据集是 4 个朝向的木纹数据集,因此,所有实验中假定内在低维数为 4 维.常规 Laplacian 特征映射采用热核函数构造相似度矩阵(其中热核参数 t=1).EGGLE 算法中,固定常数σ=4,是通过在所有测地线距离的 2 倍标准差附近多次重复实验的最好经验值;学习器分别取β={0.5,1,2,4,8,16,32,64},让β的取值尽可能地展示了流形的不同侧面.构造了 8 个学习器,每个学习器都通过基于测地线距离的广义高斯型 Laplacian 特征映射算法来发现每一幅图像的内在低维坐标,然后应用最近邻分类器作为识别,最后通过投票法确定集成分类结果.图 5 表示的是整个木纹数据集通过本文的 GGLE 算法在参数σ=4 和β=0.5 时获得的 2 维可视化流形,呈现出清晰的流形结构.鉴别实验中,对 4 类朝向的木纹数据随机选择训练样本,每类训练样本数如图 6 中的横坐标所示,每次实验将剩下的木纹数据作为测试样本,每次固定训练样本数目的实验进行了 10 次随机选择训练样本.

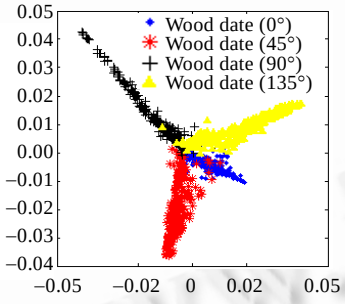


Fig.5 2D visualization manifold of wood texture dataset by using GGLE algorithm

图 5 木纹数据由 GGLE 算法获得的 2 维可视化流形

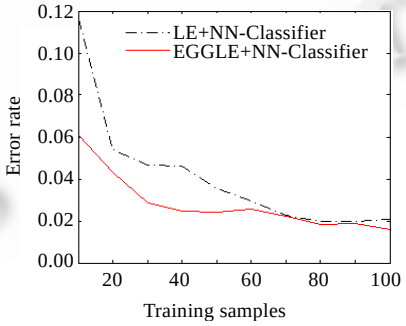


Fig.6 Recognition error rates versus the number of training samples per class by using LE and EGGLE respectively

图 6 LE 和 EGGLE 随每类训练样本数变化对应的识别错误率

图 6 中的识别错误率是 10 次实验的平均识别错误率,EGGLE+NN-Classifier 在固定训练样本数的 10 次实

验的识别错误率变化标准差小于 0.002 6,LE+NN-Classifier 在固定训练样本数的 10 次实验的识别错误率变化标准差小于 0.003 2.图 6 表示随着每类训练样本数的增加,LE 和 EGGLE 对应的平均识别错误率,表明 EGGLE 的识别错误率都低于 LE 与最近邻分类器的组合,即固定的近邻几何结构.EGGLE 采用基于测地线距离的不同广义高斯函数度量数据点间相似度实现对流形上近邻数据点的不同程度保持,在它们所对应的低维空间中,集成识别取得了更好的效果.

5 结 论

本文结合在流形上用测地线距离度量数据点之间关系的优点和广义高斯函数的特性,融合到传统的 Laplacian 特征映射算法中.首先提出了一种基于测地线距离的广义高斯型 Laplacian 特征映射算法,该算法可以调整节点间的相似度,通过选择超高斯、高斯或者次高斯函数来实现不同程度的近邻局部特性的保持,不需要重新计算邻接矩阵,而且当需要保持近邻关系的数据点邻域增大时,采用测地线距离可以避免欧氏距离度量不合理的缺陷;该算法在用超高斯、高斯和次高斯函数度量数据点间的相似度时,局部近邻保持的程度是不同的,低维坐标呈现出不同的聚类特性.然后,利用这种特性进一步提出了基于测地线距离的广义高斯型 Laplacian 特征映射的集成判别算法,该集成判别算法的主要优点是:近邻参数 K 固定,邻接矩阵和测地线距离矩阵都只构造一次,只需要多次选择广义高斯型函数构造多个 Laplacian 矩阵,获取多个独立的低维空间坐标集合,独立学习分类器,集成分类识别;本文在木纹数据集上的识别实验结果表明,这是一种有效的基于流形的集成判别算法.今后我们进一步要做的工作是,在本文提出的集成判别算法中,广义高斯函数参数的选取有一定的技巧性,尺度因子 σ(类似于方差)的选取依赖于实际问题测地距离的统计规律,而指数 β 取值应保持差异性,使其尽可能地展示流形的不同侧面,但还未有可行理论指导.另外,流形的不同部分应当选用不同的 β 和 σ 也是一个值得研究的问题,比如不同曲率的地方采用不同的 β 和 σ.另一个更具挑战性的问题是,应用 Laplacian 特征映射丰富的理论基础对提出的算法进行理论分析,并研究 Out-of-Sample 问题的解决办法.

References:

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

[2] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 2003,15(6): 1373−1396.

[3] Zhang ZY, Zha HY. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM Journal of Scientific Computing, 2004,26(1):313−338.

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

[5] Zhan DC, Zhou ZH. Ensemble-Based manifold learning for visualization. Journal of Computer Research and Development, 2005, 42(9):1533−1537 (in Chinese with English abstract).

[6] Zhang JP, He L, Zhou ZH. Ensemble-Based discriminant manifold learning for face recognition. In: Jiao L, Wang L, Gao X, Liu J, Wu F, eds. Proc. of the 2nd Int'l Conf. on Natural Computation (ICNC 2006). LNCS 4221, Berlin, Heidelberg: Springer-Verlag, 2006. 29−38.

[7] Zhang JP, Shen XH, Zhou ZH. Unified locally linear embedding and linear discriminant analysis algorithm (ULLELDA) for face recognition. In: Li SZ, Lai JH, eds. Advances in Biometric Personal Authentication. LNCS 3338, Berlin, Heidelberg: Springer-Verlag, 2004. 209−307.

[8] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Technical Report, TR 2001-01, Chicago: University of Chicago, 2001.

[9] Belkin M, Niyogi P. Semi-Supervised learning on riemannian manifolds. Journal of Machine Learning, 2003,56(1-3):1−34.

[10] Dietrerich TG. Ensemble learning. In: Arbib MA, ed. The Handbook of Brain Theory and Neural Networks. 2nd ed., Cambridge: MIT Press, 2002.

[11] Hanse LK, Salamon P. Neural network ensembles. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1990,12(10): 993−1001.

[12] Silva VD, Tenenbaum JB. Global versus local methods in nonlinear dimensionality reduction. In: Becker S, Thrun S, Obermayer K, eds. Proc. of the Neural Information Processing Systems 15 (NIPS 2002). Cambridge: MIT Press, 2002. 705−712.

[13] Saul LK, Roweis ST. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research, 2002,4(2):119−155.

[14] Belkin M. Problems of learning on manifolds [Ph.D. Thesis]. Chicago: University of Chicago, 2003.

[15] USC-SIPI image database. http://sipi.usc.edu/services/database/Database.html

附中文参考文献:

[5] 詹德川,周志华.基于集成的流形学习可视化.计算机研究与发展,2005,42(9):1533−1537.

曾宪华(1973－),男,四川攀枝花人,博士生,主要研究领域为流形学习,机器学习,人工智能,模式识别.

罗四维(1944－),男,博士,教授,博士生导师,主要研究领域为神经计算,计算机视觉,机器学习,并行处理,网格计算.

王娇(1982－),女,博士生,主要研究领域为半监督学习,机器学习.

赵嘉莉(1971－),女,博士,高级工程师,主要研究领域为图像处理,模式识别,机器学习.