保持局部邻域关系的增量 Hessian LLE 算法

高翠珍[1]，胡建龙[1,2]，李德玉[1,2]
（1. 山西大学计算机与信息技术学院，太原 030006;
2. 计算智能与中文信息处理教育部重点实验室，太原 030006）

摘要：Hessian LLE 算法是一种经典的流形学习算法，但该方法是以批处理的方式进行的，当新的数据点加入时，必须重新运行整个算法计算所有数据点低维嵌入，原来的运算结果被全部丢弃。鉴于此，提出了一种保持局部邻域关系的增量 Hessian LLE（LIHLLE）算法，该方法通过保证流形新增样本点在原空间和嵌入空间局部邻域的线性关系不变，用其已有邻域点的低维坐标线性表示新增样本点，得到新增点的低维嵌入，实现增量学习。在 Swiss roll with hole 和 frey_rawface 数据集上的实验表明该方法简便、有效可行。

关键词：流行学习；Hessian LLE；增量学习
中图分类号：TP391.4

Incremental Hessian LLE by preserving local adjacent information between data points

GAO Cuizhen[1], HU Jianlong[1,2], LI Deyu[1,2]
(1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006;
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006)

Abstract: Hessian LLE algorithm is a classical manifold learning algorithm. However, Hessian LLE is a batch mode. If only new samples are observed, the whole algorithm must run repeatedly and all the former computational results are discarded. So, incremental Hessian LLE (LIHLLE) algorithm was proposed, which to preserve local neighborhood relationship between the original space and the embedding space. New sample points were linearly reconstructed with existing embedding results of local neighborhood samples. The proposed method can learn manifold in an incremental way. Simulation results in Swiss roll with hole and frey_rawface database testify the efficiency and accuracy of the proposed algorithms.

Key words: Manifold learning; Hessian LLE; Incremental learning

0 引言

流形学习是一种非线性降维技术，通过分析数据集的外在结构来认识其本质，已经成为机器学习、模式识别、数据挖掘等领域的研究热点之一。近年来，流形学习已经得到了快速的发展，产生了大量的研究成果。Tenenbaum等人[1]提出的Isomap算法首先使用最近邻图中的最短路径得到近似的测地线距离，用其代替不能表示内在流形结构的欧式距离，然后输入到多维尺度分析(MDS)中处理，进而发现嵌入在高维空间的低维坐标。Rowels和Saul[2]提出的LLE算法能够将高维输入数据点映射到一个全局低维坐标系，同时保持了邻接点之间的关

基金项目：国家自然科学基金资助项目(60875040，60970014，61175067)；教育部高等学校博士点基金(200801080006)；山西省自然科学基金资助项目(2010011021-1)；山西省科技攻关项目(20110321027-02)；太原市科技局明星专项(09121001)
作者简介：高翠珍(1986-)，女，硕士研究生，主要研究方向为数据挖掘、流形学习
通信联系人：李德玉(1965-)，男，教授，主要研究方向为粗糙集理论、模式识别、机器学习等. E-mail: lidy@sxu.edu.cn

系，这样原有的几何结构就能够得到保留。基于LLE的发展，人们提出了一些改进的算法，包括利用拉普拉斯(Laplacian)算子变化改进的算法LE[3]、利用赫森(Hessian)变换改进的算法HLLE[4]、利用数据分类信息改进的监督LLE、增量式LLE等。然而，在性能上,HLLE是对LLE的较大改进,甚至在某些情况下超越了ISOMAP的能力。ISOMAP是在假设全局等距和凸参数空间下进行的,这在多数情况下难以满足，而HLLE只要求局部等距映射和开的连通参数空间,因而适用范围更广。

以上算法已经被广泛的应用，然而这些方法都是批处理的模式，当新样本不断地加入时，批处理方法必须重新计算所有的样本，计算是复杂的。为了克服此问题，一些科研工作者已经致力于研究增量学习算法，2005年Martin and Anil提出了ISOMA的增量算法[5]该方法首先更新测地距离，然后将问题转化为子空间的特征分解来实现全局的增量学习；Kouropteva等人在假定原有特征值不变的基础上提出了LLE的增量算法[6]，在低维空间实现最优化；2006年Liu等人提出了LTSA的增量算法；2009年Peng Jia等人用邻域点的低维坐标表示新增点的坐标，提出了LE的增量算法[7]；2010年李厚森和成礼智提出了增量的HLLE算法[8]等等。

通过以上分析，不难发现大多数已有的增量学习方法都是将原空间的特征分解问题转化为低维空间的特征分解来实现的，这些方法虽然降低了算法的复杂性，但每次增加一个样本点均需重新计算所有样本点的低维嵌入。事实上，由 LLE、HLLE 等局部流形学习算法的特性可知，当有单个新样本点加入时，大多数样本点的结果几乎不发生改变，因此这样的计算也造成了浪费。鉴于此，本文基于人的认知，利用流形结构的局部线性特性，提出了保持局部邻域关系的增量 HLLE 算法（LIHLLE）。当有新的样本加入时，保证流形新增样本点在原空间和嵌入空间局部邻域的权值不变，用其邻域点的低维坐标线性表示新增样本点，得到新增点的嵌入结果。

1 Hessian 局部线性嵌入（HLLE）

HLLE 算法是 LLE 算法的一种改进算法，将 LLE 的局部带权线性表示方法用局部等距代替，实现了数据降维。具体的算法步骤如下：

⑴ 邻域选取。获取每个样本点 x_i 的邻域点。记 $X_i=[x_{i1},...x_{ik}]$ 为样本点 x_i 的 k 个最近的邻域点。

⑵ 计算切空间坐标。对每个样本点的邻域，计算中心化矩阵 $X_i-\overline{x_i}1_k^T$ 的前 d 个最大的特征值对应的特征向量，并将这 d 个特征向量组成矩阵 V_i。

⑶ 估计Hessian矩阵。有 $M_i=[1,V_i,(V_i(:,s)*V_i(:,I))_{1\le s\le I\le d})]$，

其中矩阵共有 $1+d+d(d+1)/2$ 列：前 $d+1$ 列由分量为 1 的列向量和 V_i 组成，$V_i(:,s)*V_i(:,I)$ 表示矩阵 V_i 的第 s 列和第 I 列的点积。对矩阵 M_i 进行 Gram-Schmidt 正交化，得到列正交阵 $\overline{M_i}$，则 Hessian 矩阵 $H^i=\overline{M_i}(:,d+1:1+d+d(d+1)/2)^T$。

⑷ 构造二次项。利用每个邻域的Hessian矩阵 H^i，$i=1,...,N$ 来构造对称矩阵 H，其元素为

$$H_{ij}=\sum_{s=1}^{N}\sum_{I=1}^{d(d+1)/2}(H^s)_{I,i}(H^s)_{I,j}$$

⑸ 计算 H 的零空间。计算 H 的 $d+1$ 个最小特征值对应的特征向量 $u_1,...,u_d$，则 $U=[u_2,...,u_{d+1}]$ 就是所求的零空间。

⑹ 计算低维嵌入.记矩阵 $R_{ij}=\sum_{I\in J_1}U_{I,i}U_{I,j}(i,j=1,...,d)$ 其中 J_1 表示某个样本点的邻域，则 $T=R^{-1/2}U^T$ 为低维嵌入。

由以上算法可以发现：1)HLLE 框架上与 LLE 一致,不同的是 HLLE 用 Hessian 变换取代

了 LLE 的局部带权线性表示；2)每个点的切空间要求其邻域是线性的,即选择合适的邻域，保证邻域的局部线性特性，这是 HLLE 需要克服的一个问题；3)HLLE 需要对每个数据点计算 $d \times d$ 次偏导数,当观察数据的维数非常高时,计算量较大，这是 HLLE 需要克服的另一个问题。因此，针对增量的 HLLE 算法，本文着重考虑这两个问题，保证随着数据点的增多，邻域的大小也自适应地发生变化，同时尽可能降低算法的复杂度。

2 增量 Hessian LLE 算法（LIHLLE）

在假设原样本点计算结果准确的基础上，本文提出的保持局部邻域关系的增量 HLLE 算法，分三步来计算新样本的低维嵌入：①根据流形结构的局部线性特性，从而自适应的确定新增点的邻域，邻域的线性程度用 PCA 来度量；②通过使得新增点在原空间中的邻域线性表示和其本身的误差最小，计算新增点在原空间中的邻域权值；③利用原样本的嵌入结果，保持原空间和嵌入空间邻域权值不变，计算新增点的投影值。

2.1 自适应的邻域选择

设原样本集为 $X=\{x_1,...,x_n\}$，用经典的HLLE算法得到的 X 的嵌入结果为 $Y=\{y_1,...,y_n\}$,新增样本点记为 x_{n+1}。用 $KNN(x_{n+1})$ 表示 x_{n+1} 的 K 最近邻，用 $RN(x_{n+1})$ 表示 x_{n+1} 的合理邻域集，初始值为空。新增样本 x_{n+1} 加入后，其合理邻域的计算过程如下：

⑴ 计算距离其最近的 $K(K>d)$ 个点，按边长度递增的顺序排列，记为 $KNN(x_{n+1})=\{x_{n+1,1},...,x_{n+1,K}\}$，同时令 $RN(x_{n+1})=\{x_{n+1,1},...,x_{n+1,d}\}$；

⑵ 用PCA方法度量其局部线性特性，满足邻域线性条件的点为合理邻域点，即依次计算 $x_{n+1,i} \cup RN(x_{n+1})(i=d+1:K)$ 样本集的协方差矩阵的特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$，如果特征值满足 $\sum_{i=1}^{i=d}\lambda_i / \sum_{i=1}^{i=D}\lambda_i > \eta$（$\eta$ 为经验值），$x_{n+1,i}$ 就记为 x_{n+1} 的合理邻域，$RN(x_{n+1})=x_{n+1,i} \cup RN(x_{n+1})$。

2.2 计算邻域权值

通过 $RN(x_{n+1})$ 来计算新增样本点 x_{n+1} 的重构权值 $w_{n+1,i}$，代价函数为

$$\varepsilon(W)=\left|x_{n+1}-\sum_{i=1}^{size(RN(x_{n+1}))} w_{n+1,i}x_{n+1,i}\right|^2，\text{其中 } x_{n+1,i}\in RN(x_{n+1}) \qquad (1)$$

权值 $w_{n+1,i}$ 满足条件 $\sum_i w_{n+1,i}=1$ (2)

这样，求最优权值就是对于公式(1)在约束条件(2)下求解最小二乘问题。

2.3 计算低维嵌入

邻域权值 $w_{n+1,j}$ 确定后，新样本点 x_{n+1} 的低维表示可以通过以下公式得到：

$$x_{n+1}\rightarrow y_{n+1}=\sum_{i=1}^{size(RN(x_{n+1}))} w_{n+1,i}y_{n+1,i}，\text{其中 } y_{n+1,i}\in Y \qquad (3)$$

以上算法可以计算出单个新增点的低维嵌入，当有多个样本同时加入时，循环计算每个数据点，将每次计算出的结果加入已有的嵌入结果中作为原始点，依次计算出所有新增点的低维嵌入。（本实验中 K 取 30，η 取 0.93）

3 实验结果及分析

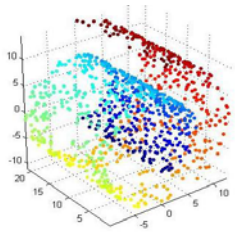
3.1 实验

实验选取 Swiss roll with hole 和 frey_rawface 两个数据集，分别用批处理 HLLE 和增量 HLLE 两种方法计算样本点的投影值，通过比较两种方法的实验结果验证本算法的有效性。Swiss roll with hole 数据集为 3 维空间的面包圈，将其嵌入到 2 维空间；frey_rawface 数据集由 1965 幅，像素为 20*28，不同姿态和表情的人脸图像组成，即原始空间为 20*28=560 维，嵌入到 2 维空间分别表示人脸姿态和表情。图 1 是 frey_rawface 数据集的一些样本。为了比较批处理 HLLE 算法和增量 HLLE 两种算法嵌入值的差别，定义误差来度量。误差的表示形式为：

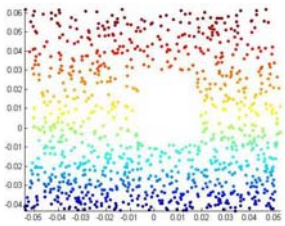
$$error = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\frac{||y_i - \hat{y}_i||^2}{||\hat{y}_i||^2}} \qquad (4)$$

其中 y_i 和 $\hat{y}_i$ 分别表示第 i 个点用 HLLE 和 LIHLLE 计算的低维嵌入值，误差越小，LIHLLE 的嵌入结果越接近于 HLLE 的嵌入结果。同时通过比较两种方法所耗的时间来度量算法的效率。

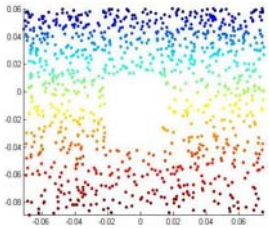
图 1 frey_rawface 数据库中的 10 个样本
Fig. 1 ten samples in the database frey_rawface



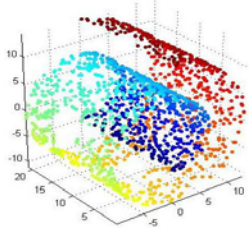
(a) N=1000



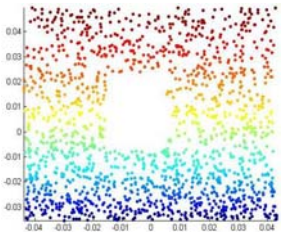
(b) HLLE



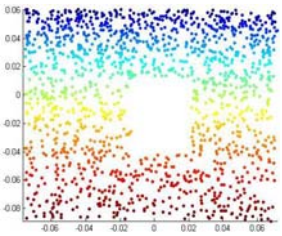
(c) LLIHLLE



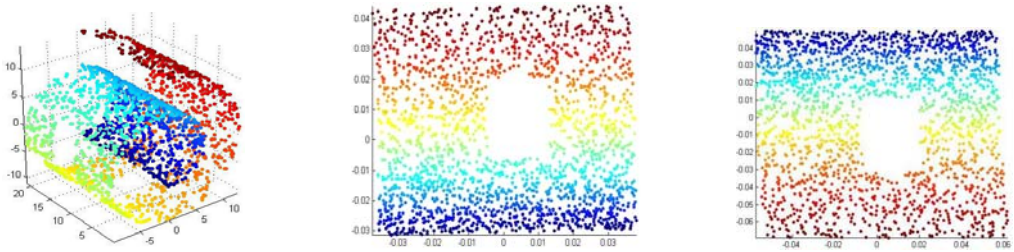
(d) N=1500



(f) LIHLLE



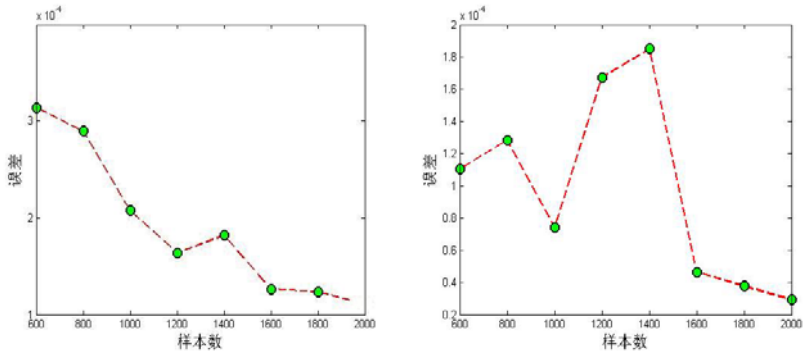
(e) HLL E



(g) N=2000 (h) HLLE (i) LIHLLE

图 2 原始图，HLLE 嵌入图，LIHLLE 嵌入图（样本点分别为 1000、1500、2000）

Fig. 2 raw figure, HLLE embedding figure and LIHLLE embedding figure（sample points, respectively 1000, 1500, 2000）



(a) Swiss roll with hole (b) frey_rawface

图 3 LIHLLE 算法的误差曲线图

Fig. 3 error curve of LIHLLE algorithm

先在每个数据集上各随机抽取500个点作为原始样本点，用HLLE计算它们的低维嵌入，然后再将新样本点一个接一个地加入，用 LIHLLE 算法计算它们的投影值，直到 2000 个数据点。

图 2 显示了在 Swiss roll with hole 数据集上的效果图，(a.b.c),(d.e.f),(g.h.i)三组图分别表示样本点数 N 从 500 增加到 1000、1500、2000 时对应的原始数据图、HLLE 算法的低维嵌入图及 LIHLLE 算法的低维嵌入图。图 3 用误差曲线图定量的表示了两种方法的低维嵌入结果，图 3(a)为 Swiss roll with hole 数据集上，LIHLLE 算法对应的误差曲线图，图 3 (b)为 frey_rawfac 数据集上，LIHLLE 算法对应的误差曲线图，其中横轴均表示样本数，纵轴均表示用两种方法映射产生的误差。同时表 1 记录了原始样本点数 N 从 500 依次增加为 1000、1500、2000 个数据点时，用两种方法映射到低维空间时所消耗的时间。

表 1 HLLE 和 LIHLLE 的运行时间（单位：s）

Tab. 1 running time of HLLE and LIHLLE（unit: s）

	Swiss roll with hole (3 to 2)		frey_rawface (560 to 2)	
	HLLE	LIHLLE	HLLE	LIHLLE
500	1.51		5.96	
1000	837.3	0.14	4013.9	0.46
1500	4951.1	0.37	12573.7	1.32
2000	13530.2	0.65	N/A	N/A

3.2 结果分析

通过图 2 可以直观地发现本文提出的增量 HLLE 算法也可以很好地将 Swiss roll with

hole 数据集展开，和原始 HLLE 算法的展开结果几乎一样。图 3(a)和 3(b)进一步用误差定量的证明了本文中算法的有效性，该方法在 Swiss roll with hole 数据集上的平均误差为 9.7153e-005，在 frey_rawface 数据集上的平均误差为 1.8955e-004，这个误差均很小，可见该方法的处理精度完全满足要求。

表 1 的数据表示了本文的增量 HLLE 算法和原始的 HLLE 算法在两种数据集上的运行时间，由此可知本文中算法的速度明显快于批处理的 HLLE 算法。此外，文献[8]中提出的增量 HLLE 学习方法虽然比原始的 HLLE 算法效率高，但仍需要对每个点计算 Hession 矩阵，计算过程仍然是复杂的，而本文提出的算法巧妙地避免了这些复杂的计算，大大减少了算法运行的时间。因此，本文在保证嵌入结果准确的前提下，大大提高了算法的效率。

4 结束语

本文基于流形的局部线性特性，在假设原来的映射结果准确的基础上提出了增量的 HLLE 算法。该算法避免了重复计算所有样本点的嵌入，利用原有的嵌入结果简单高效的计算出了新增样本点的嵌入结果，实现了 HLLE 的增量学习。在 Swiss roll with hole 和 frey_rawface 数据集上的实验证明，该算法得到的结果与批处理方法得到的结果相近，然而算法的效率得到了很大的提高。

[参考文献] (References)

[1] TENENBAUM J B, DE SILVA V, LANGFORD J C. Aglobal geometric framework for nonlinear dimensionality reduction[J]. Science, 2000,290:2319-2323.
[2] ROWEIS S T, SAUL L K. Nonlinear Dimensionality Reduction by Locally Linear Embedding[J]. Science, 2000,290:2323-2326.
[3] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data rep resentation[J]. Neural Computing, 2003, 15(6):1373-1396.
[4] DONOHO D L, GRIMES C. Hessian Eigenmaps: Locally Linear Embedding Techniques for High Dimensional Data[J]. Proceedings of the National Academy of Sciences of the United States of America,2003,100(10):5591-5596.
[5] MARTIN H C L, ANIL K J. Incremental Nonlinear Dimensionality Reduction by Manifold Learning[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2006,28(3): 377-391.
[6] P. JIA et al. Incremental Laplacian eigenmaps by preserving adjacent information between data points[J]. Pattern Recognition, 2009,30: 1457-1463.
[7] OLGA KOUROPTEVA, OLEG OKUN et al. Incremental locally linear embedding algorithm[J]. Pattern Recognition, 2005,38(10): 1764-1767.
[8] 李厚森,成礼智. 增量 Hessian LLE 算法研究[J] .计算机工程, 2010,37(6):159-161.