第 37 卷 第 21 期
Vol.37 No.21

计 算 机 工 程
Computer Engineering

2011 年 11 月
November 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)21—0138—03 文献标识码：A 中图分类号：TP311.13

# 基于密度刻画的降维算法

李燕燕，闫德勤，刘胜蓝

(辽宁师范大学计算机与信息技术学院，辽宁 大连 116081)

**摘 要：**针对 LLE 算法在数据密度变化较大时很难降维的问题，提出一种基于密度刻画的降维算法。采用 cam 分布寻找数据点的近邻，并在低维局部重建时对数据点加入密度信息。对手写体数字图像进行字符特征的降维，再对降维后的特征进行分类识别。实验结果表明，该方法能区分字符，具有较好的识别率，能够发现高维空间的低维嵌入流形。

**关键词：**流形学习；降维；密度信息；手写体识别；cam 分布

# Dimensionality Reduction Algorithm Based on Density Portrayal

LI Yan-yan,YAN De-qin, LIU Sheng-lan

(School of Computer and Information Technology, Liaoning Normal University, Dalian 116081, China)

**【Abstract】**In order to improve the correctness of dimensionality reduction algorithms based on Locally Linear Embedding(LLE) caused by data density change, a novel approach based on density is proposed in this paper. It adapts cam distribute to find the data's nearest neighbor, meanwhile, adds the data's density information during the low dimensional local reconstruction. The proposed algorithm is used to reduce the dimensionality of input feature, and the reduced feature is classified by simple classifier. Experimental result indicates that the method can effectively improve the recognition rate of handwritten digits and can dig the manifold embedded in the high dimensional space.

**【Key words】**manifold learning; dimensionality reduction; density information; handwriting recognition; cam distribution

**DOI:** 10.3969/j.issn.1000-3428.2011.21.047

## 1 概述

随着信息时代的到来，对高维数据的处理成为迫切需要解决的问题。如何对高维数据进行降维[1]、有效地减轻维数灾难[2]、促进高维数据的分类、可视化及压缩，已成为人工智能与机器学习等领域的研究热点。目前已经提出了许多降维方法，可以分为线性和非线性两大类，非线性降维方法比线性降维方法能更好地发掘隐藏在高维数据中的流形分布，但各种算法本身还存在着局限性，例如流形本征维数的估计、流形本身带有空洞、维数不固定，以及数据点密度信息很难把握等问题[3]。因此，需要在现有的非线性降维方法上做进一步改进。

本文秉承了流形学习[4]的思想，从数据密度入手，提出一种非线性降维的改进算法——基于密度的局部线性嵌入(Density-based Weighted Locally Linear Embedding, DWLLE)算法。

## 2 DWLLE 算法的提出

### 2.1 相关算法介绍

传统的LLE算法虽然能够有效地学习非线性流形的全局结构，但对于分布不均匀的数据，利用欧式距离来选择 $k$ 近邻时容易造成信息选取方向的缺失[5]，影响映射效果。针对 LLE 的这一不足，文献[6]在其基础上提出了一种基于权重的非线性降维算法 WLLE(Weighted Locally Linear Embedding)，在 WLLE 中引入 cam 分布，增加了样本点间的权重信息，这样在选择近邻点时，避免了信息选取方向的缺失。但 WLLE 算法没有很好地考虑到数据间的密度信息，在处理数据密度变化较大的流形时很难正确降维。图 1 展示了 2 种算法的近邻选择情况。其中，求星形点的 $k$ 近邻，实线内的点是利用 LLE(Locally Linear Embedding)求得的的 $k$ 近邻，虚线内的点是利用 WLLE 求得的 $k$ 近邻。

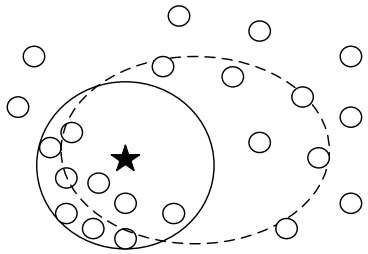


**图 1 LLE 和 WLLE 算法的 $k$ 近邻选择**

### 2.2 基于密度的局部线性嵌入

本文将密度信息的思想引入 WLLE 算法中，提出了一种新的基于密度信息的局部线性嵌入算法(DWLLE)。DWLLE 的主要思想是结合样本本身的流形结构信息，利用 cam 分布来调整样本点之间的距离，在低维线性重建的过程中利用密度信息来调整权值距离矩阵进行降维。

给定下述高维观测数据集为 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_N\}$，$x_i\in R^D$，求低维坐标 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_N\}$，$y_i\in R^d$，$(d\ D)$。DWLLE 算法的具体步骤如下：

(1)局部近邻选取。在高维空间中利用权重距离寻找每个数据点 $x_i$ 的 $k$ ( $k<N$ )个近邻点。

(2)计算数据点的局部重建权重矩阵 $\boldsymbol{W}$，定义误差函数：

$$\min\varepsilon(\boldsymbol{W})=\sum_{i=1}^{N}\left\|x_i-\sum_{j=1}^{k}W_{ij}x_j\right\|^2 \quad (1)$$

**基金项目：**辽宁省教育厅高等学校科学研究基金资助项目(2008344)；中国科学院自动化研究所复杂系统与智能科学重点实验室开放课题基金资助项目(20070101)

**作者简介：**李燕燕(1986－)，女，硕士研究生，主研方向：数据降维，模式识别；闫德勤，教授、博士；刘胜蓝，硕士研究生

**收稿日期：**2011-05-26 **E-mail：**liyanyan1016@126.com

其中，$x_j$ ( $j=1,2,\cdots,k$ )为 $x_i$ 的 $k$ 个近邻点；$W_{ij}$ 是 $x_i$ 与 $x_j$ 之间的权值，$\boldsymbol{W}_i=[W_{i1},W_{i2},\cdots,W_{ik}]^{\mathrm{T}}$ 为第 $i$ 个数据点的局部重建权值。

(3)计算每个样本点的密度信息：

$$\alpha_i^*=\max\left\|x_i-x_j\right\|^2 \tag{2}$$

$$\alpha_i=1-\alpha^*\Big/\sum_{j=1}^{k}\left\|x_i-x_j\right\|^2 \tag{3}$$

并以 $\alpha_i$ 校正上步中求得的重建权值 $\boldsymbol{W}_i$：

$$\boldsymbol{W}^*=[\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_N]\cdot[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}} \tag{4}$$

并且满足：

$$\sum_{j=1}^{k}W_{ij}\alpha_i=1 \tag{5}$$

(4)利用权值矩阵 $\boldsymbol{W}^*$ 对原数据点重构，寻找数据集 $\boldsymbol{X}$ 的低维嵌入 $\boldsymbol{Y}$。通过最小化重构误差函数：

$$\min\varphi(\boldsymbol{Y})=\sum_{i=1}^{N}\left\|y_i-\sum_{j=1}^{N}W_{ij}y_j\right\|^2 \tag{6}$$

来实现。

引进密度信息，结合式(3)，将式(6)变为：

$$\min\varphi(\boldsymbol{Y})=\sum_{i=1}^{N}\left\|y_i-\alpha_i\sum_{j=1\ i}^{N}W_{ij}y_j\right\|^2=\min tr\left(\boldsymbol{YMY}^{\mathrm{T}}\right) \tag{7}$$

相应的优化问题转化为下列约束条件问题：

$$\begin{cases}\min\varphi(\boldsymbol{Y})=\min tr(\boldsymbol{YMY}^{\mathrm{T}})\\ \text{s.t.}\quad \boldsymbol{YY}^{\mathrm{T}}=\boldsymbol{I}\end{cases} \tag{8}$$

其中，$\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W}^*)^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W}^*)$，相应的输出结果 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_N\}$，$y_i\in R^d$。

## 3　时间复杂度分析

对 DWLLE 算法进行时间复杂度分析，并计算经典算法 PCA、LTSA、LLE、WLLE 的时间复杂度，通过比较来进一步说明新算法的良好计算性能。

降维算法的复杂度主要由数据点的个数 $N$、原始维数 $D$ 以及近邻点个数 $k$ 来确定。

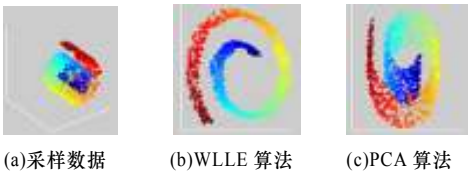
PCA、LTSA、LLE 算法的时间复杂度分别为 $O\left(D^3\right)$、$O\left(pN^2\right)$、$O\left(pN^2\right)$(其中，$p$ 是稀疏矩阵中非零元和零元的比率)。WLLE 同 LLE 在算法执行中唯一的不同在于寻找近邻点时要计算数据点间的权重距离，所需的时间为 $O\left(ND+2DN^2\right)$，可知 WLLE 算法总的时间复杂度亦为 $O\left(pN^2\right)$。DWLLE 比 WLLE 算法多考虑了数据点间的密度信息，计算数据间密度信息时所需的时间 $O(kN)$，对密度信息的计算并不影响 DWLLE 总的时间复杂度。因此，DWLLE 的时间复杂度也为 $O\left(pN^2\right)$，可见 DWLLE 的时间复杂度与 LLE、LTSA、WLLE 是同阶的，只是比线性降维算法 PCA 的时间复杂度略高。

通过以上的分析可知：相比其他的非线性降维算法，DWLLE 算法在考虑到数据间信息密度后时间复杂度上并没有增加。

## 4　实验结果及分析

### 4.1　仿真实验测试结果

仿真实验采用的是经典的瑞士卷、双峰图 2 种密度不同的流形数据集，其分布在 3 维空间上，具有 2 维本质结构。每一类流形均采样数据点个数分别为 1 000 和 2 000，近邻点 $k$=12。将 PCA、LTSA、LLE、WLLE 作为对比算法来对新算法的性能进行比较说明。图 2~图 5 呈现了各种算法的 2 维嵌入结果。



(a)采样数据　(b)WLLE 算法　(c)PCA 算法

(d)LLE 算法　(e)LTSA 算法　(f)DWLLE 算法

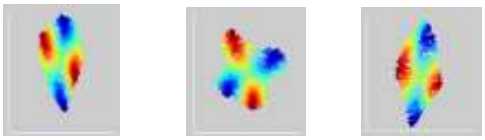
**图 2　瑞士卷采样 1 000 点效果图**

(a)采样数据　(b)WLLE 算法　(c)PCA 算法

(d)LLE 算法　(e)LTSA 算法　(f)DWLLE 算法

**图 3　瑞士卷采样 2 000 点效果图**

(a)采样数据　(b)WLLE 算法　(c)PCA 算法



(d)LLE 算法　(e)LTSA 算法　(f)DWLLE 算法

**图 4　双峰图采样 1 000 点效果图**

(a)采样数据　(b)WLLE 算法　(c)PCA 算法

(d)LLE 算法　(e)LTSA 算法　(f)DWLLE 算法

**图 5　双峰图采样 2 000 点效果图**

从图 2~图 5 中得出以下事实：对于采样密度均匀的瑞士卷流形结构，PCA 显然失效，无法正确揭示瑞士卷的低维结构，WLLE、DWLLE 也无法揭示原始数据集潜在的低维结构，具有严重的扭曲变形问题，相对来说，LTSA、LLE 具有明显优势，均良好地恢复出流形的本质结构。可见，本文算法 DWLLE 对于分布密度均匀的流形数据集进行降维时，效果不是很理想。但是当对双峰图这种分布不均匀且数据间密度变化较大的流形数据集进行降维时，LLE、LTSA、PCA 所揭示的低维结构具有明显的奇异性，而 DWLLE 和 WLLE 的效

果相对较好。尤其是 DWLLE 在考虑到数据间的密度信息后，降维效果比 WLLE 相对更佳。

### 4.2 手写体分类识别应用

为了进一步验证所提算法的性能，在数据库 minst 上将 DWLLE 与 PCA、LTSA、WLLE、LLE 算法进行比较。图 6 为该手写体字符库中部分字符的图像。将 DWLLE 应用到有监督学习中，采用 minst 数据库中的手写体 0~9，每个样本都只经过简单的二值化并归一化为大小为 28×28 像素的图像，在训练集中选取 10 类，每类随机选取 10 个样本，并在测试集中相应的每类随机选取 20 个样本进行测试。实验中未对 minst 字符做任何预处理，仅将字符的二值特征作为原始信息进行降维，为了证明 DWLLE 的降维效果，分类器只选用最简单的最近邻分类器。图 7 为本文所提算法同其他算法约简到不同维数的识别率比较。

**图 6 部分手写体字符图像**

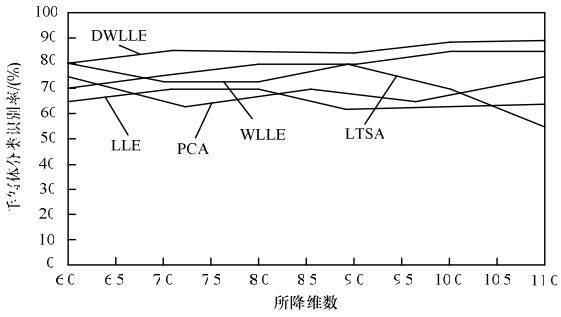



**图 7 不同算法在 minst 库上识别率与维数 $d$ 的关系**

从图 7 的总体识别效果来看，LTSA 随识别维数的增加，特别在高于 9 维后，信息损失比较严重。LLE 和 PCA 的识别率始终保持在 70%左右，比较平稳。WLLE 和 DWLLE 在考虑到数据点间的分布情况以及流形的几何结构的同时，通过权重距离来选择 $k$ 近邻，都能很快地达到较高的识别率，取得较理想的识别效果。特别是 DWLLE，在考虑到数据间的密度信息后，相比 WLLE 识别率有所提升。由此可见，数据点间的密度信息在一定程度会导致降维结果出现偏差，影响整体结果。

## 5 结束语

本文在对 LLE、WLLE 算法进行分析后，提出了一种基于密度刻画的降维算法——DWLLE，通过考虑数据点间的密度信息，DWLLE 算法能很好地处理分布不均匀、密度变化较大的流形数据，实验证明其有效、可行。并且在对手写体字符识别的实际应用中，在考虑到数据点间的密度信息后，本文算法不仅具有较好的降维效果、较高的稳定性，而且获得了较高的识别率，实际应用价值较高。下一步的研究任务是如何利用 DWLLE 算法减少噪声的干扰、提高算法的鲁棒形。

## 参考文献

[1] 姜 维, 杨炳儒. 基于流形学习的维数约简算法[J]. 计算机工程, 2010, 36(12): 25-27.

[2] Pavlenko T. Curse of Dimensionality and Error Probability in Discriminant Analysis[J]. Journal of Statistical Planning and Inference, 2003, 115(2): 565-584.

[3] 王庆刚. 流形学习算法及若干应用研究[D]. 重庆: 重庆大学, 2009.

[4] 高小方, 梁吉业. 基于采样密度和流形弯曲度的动态邻域算法[J]. 计算机工程, 2010, 36(12): 17-21.

[5] Zhou Changyin, Chen Yanqin. Improving Nearest Neighbor Classification with Cam Weighted Distance[J]. Pattern Recognition, 2006, 39(4): 635-645.

[6] Pan Yaozhang, Ge S S, Mamum A A. Weighted Locally Linear Embedding for Dimension Reduction[J]. Pattern Recognition, 2009, 42(5): 798-811.

编辑 任吉慧

(上接第 137 页)

较快。图 5(b)是采用改进粒子滤波算法[5]的跟踪过程，该过程采用 4 000 个粒子；图 5(c)是本文提出的算法的跟踪过程，该过程采用 50 个粒子，模板大小分别为 70×70 和 30×30 。图 5(b)、图 5(c)中标识的小圆圈为跟踪到的笔尖。从图 5(b)可以看出，改进粒子滤波算法虽然使用了较多的粒子数，但还是易受外界光线、阴影的影响，使跟踪结果不理想，甚至在书写速度较快时出现脱靶现象。从图 5(c)可以看出，本文提出的算法虽然使用的粒子数较少却具有较好的跟踪结果。

## 4 结束语

笔尖跟踪是基于视觉的签名识别系统所需解决的首要问题。本文利用笔尖的形状特点对笔尖进行自动检测与定位，在此基础上提出一种基于粒子滤波的多模板笔尖跟踪算法，该算法在采用少量粒子数的情况下，仍能取得较理想的跟踪效果，且较大程度地克服了因光线、阴影及书写速度过快等因素导致的跟踪不准确等现象。在后续的研究工作中，可结合人类运动学模型进一步指导笔尖的跟踪，使其跟踪更具准确性和鲁棒性。

## 参考文献

[1] Impedovo D, Pirlo G. Automatic Signature Verification: The State of the Art[J]. IEEE Trans. on Systems, Man, and Cybernetics (Part C): Applications and Reviews, 2008, 38(5): 609-635.

[2] 周圣鑫, 周 军, 宋 利, 等. 一种针对小目标的跟踪算法[J]. 计算机工程, 2010, 36(16): 186-188.

[3] 刘袁缘. 基于帧间运动能量差小目标运动检测算子模型研究[D]. 武汉: 华中科技大学, 2007.

[4] Munich M E, Perona P. Visual Input for Pen-based Computers[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002, 24(2): 313-328.

[5] Yasuda K, Muramatsu D, Shirato S, et al. Visual-based Online Signature Verification Using Features Extracted from Video[J]. Journal of Network and Computer Applications, 2010, 33(3): 333-341.

[6] Suzuki S, Abe K. Topological Structural Analysis of Digital Binary Images by Border Following[J]. Computer Vision, Graphics and Image Processing, 1985, 30(1): 32-46.

[7] Arulampalam M S, Maskell S, Gordon N, et al. A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking[J]. IEEE Trans. on Signal Processing, 2002, 50(2): 174-188.

编辑 顾姣健