

基于点密集度的非线性流形学习算法

黄淑萍

(南开大学 经济学院, 天津 300071)

摘 要: 提出一种样本点密集度的非线性流形学习算法. 该算法提出了一个有效的数据点密集参数, 能够很好地对非均匀数据的低维嵌入进行约束, 其嵌入结果明显优于 LLE 算法. 在人工和人脸数据集上的实验结果表明, 新算法产生了较好的嵌入及分类结果.

关键词: 流形学习; 组合投资; 局部线性嵌入; 密集度

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)06-0010-04

Non-linear Manifold Learning Algorithm Based on Intensity of Points

HUANG Shu-ping

(School of Economics, Nankai University, Tianjin 300071, China)

Abstract: This paper presents a non-linear manifold learning algorithm based on intensity of sample points. It proposes an effective intensity parameter of sample points, which constraints the low-dimensional embedding of uneven data well. There is a better embedding result than LLE. The experimental results on the artificial and face datasets show that the new algorithm yields a better embedding and classification result.

Key words: manifold learning; portfolio investment; Local Linear Embedding (LLE); intensity

1 引言

流形学习是近年来新兴的非线性降维技术, 其目的是要研究和模拟人类的感知能力, 实质是挖掘嵌入在高维数据中的低维线性的非线性流形^[1-2], 这种嵌入保留了原始数据的几何特性. 为了寻找这种内在的低维流形几何结构, 很多著名的降维技术被提出, 例如, 局部线性嵌入^[3] (Locally Linear Embedding, LLE)、Isomap^[4]、局部切空间排列^[5] (Local Tangent Space Alignment, LTSA)、拉普拉斯特征映射^[6] (Laplacian Eigenmaps) 以及基于张量的降维方法^[7] 等等.

LLE 是文献^[3] 中提出的一种非线性流形学习方法, 它能够实现高维输入数据点到低维坐标系的全局映射, 同时能够保留邻接点之间的几何拓补结构关系. 然而, LLE 仅仅适合处理样本点均匀分布的数据, 而不能处理分布不均匀的数据. 因此本文提

出一种样本点密集度的非线性流形学习算法, 该算法提出了一个有效的数据点密集参数, 能够很好地对非均匀数据的低维嵌入进行约束, 其嵌入结果明显优于 LLE 算法. 在人工和人脸数据集上的实验结果表明, 新算法产生了较好的嵌入及分类结果.

2 LLE 算法

局部线性嵌入 (LLE) 的基本思想是假设每个样本点与它的近邻点处于流形的一个线性或近似线性区域中, 将全局非线性转化为局部线性, 而相互重叠的局部邻域能够提供全局结构的信息. 设给定的高维观测数据点集为 $X = [x_1, \dots, x_N]$, $x_i \in R^D$, 采样自 d 维流形, 求地低维坐标 $Y = [y_1, \dots, y_N]$, $y_i \in R^d$, 且 $d \ll D$. LLE 在高维空间中寻找每个样本点 x_i 的 k ($k < N$) 个近邻点, 通过解决约束的最小二乘问题来获得样本点的局部重建权值矩阵 W :

$$\begin{cases} \min_{\epsilon}(W) = \sum_{i=1}^N \left\| x_i - \sum_{j=1}^k w_{ij} x_j \right\|^2 \\ \text{s. t.} \quad \sum_{j=1}^k w_{ij} = 1 \end{cases} \quad (1)$$

式中, $x_j (j=1, 2, \dots, k)$ 为 x_i 的 k 个近邻点; $W = \{w_1, w_2, \dots, w_N\}$, $w_i = [w_{i1}, w_{i2}, \dots, w_{ik}]^T (1 \leq i \leq N)$ 为第 i 个样本点的局部重建权值, w_{ij} 是 x_i 与 x_j 之间的权值。

LLE 通过解决优化问题(2)将高维空间中的 X 映射到低维空间:

$$\begin{cases} \min \Phi(Y) = \sum_{i=1}^N \|YI_i - Yw_i\|^2 \\ = \sum_{i=1}^N \|Y(I_i - w_i)\|^2 = \min \text{tr} YMY^T \\ \text{s. t.} \quad YY^T = I \end{cases} \quad (2)$$

式中, I 表示 N 维单位矩阵. $M = (I - W)^T(I - W)$. 使用 Lagrange 乘子法, 解得 $MY^T = \lambda Y^T$, 则 M 的特征向量即所求的嵌入坐标。

3 基于点密集度的非线性流形学习算法

本文提出的基于数据点分布密集度的非线性流形学习方法, 改进和扩展了原始的 LLE 算法。

由(1)知, 重建权值 $\{w_{ij}\}$ 的确定等价于调整合适的权值 w_{ij} 以确定 x_i 与其 k 个近邻点形成凸集合的最小距离, 但 x_i 的 k 个近邻点分布的密集信息没有被考虑. 这样, 利用 LLE 算法中优化问题(2)来获取低维空间的嵌入表示时, 只能认为在高维空间中的数据是均匀分布的, 当数据分布变化较大时, LLE 算法就很难正确降维。

为了解决上述分析中的缺陷, 必须考虑样本点分布的均匀密集程度, 提出一种基于数据点密集度的非线性流形学习算法——ILLE 算法, 该算法提出了采样数据点分布的密集参数:

$$\theta_i = 1 - \left(\frac{\theta_i^*}{\sum_{j=1}^k \theta_j^*} \right) \quad (3)$$

式中, $i=1, 2, \dots, N$, $\theta_i^* = \max \|x_i - x_j\|$. 根据式(1)求得 w_{ij} , 并以 θ_i 乘以第 i 个样本点的局部重建权值 w_i , 则权值矩阵校正为

$$W^* = [w_1 \ w_2 \ \dots \ w_N] \cdot [\theta_1, \theta_2, \dots, \theta_N]^T \quad (4)$$

ILLE 算法执行的具体步骤如下:

第一步: 寻找每个样本点 x_i 的 $k (k < N)$ 个近邻

点, 距离公式为 $d_{ij} = \sqrt{\sum_{p=1}^D |x_{ip} - x_{jp}|^2}$, 其中, x_{ik} 为

样本点 x_i 第 p 维的元素;

第二步: 通过(1), 计算 $\{w_{ij}\}$, 结合 $\sum_{j=1}^k w_{ij} = 1$, 式(1)可改写为

$$\begin{aligned} \min_{\epsilon}(W) &= \sum_{i=1}^N \left\| \sum_{j=1}^k w_{ij} (x_i - x_j) \right\|^2 \\ &= \sum_{i=1}^N \| (x_i - x_j) w_i \|^2 \\ &= \sum_{i=1}^N (w_i)^T Z_i w_i \end{aligned} \quad (5)$$

式中, $Z_i = (x_i - x_j)^T (x_i - x_j)$. 通过拉格朗日乘子法有

$$\begin{aligned} L(W) &= \sum_{i=1}^N (w_i)^T Z_i w_i + \lambda \left(\sum_{j=1}^k w_{ij} - 1 \right) \Rightarrow \\ \frac{(\partial L)}{\partial w_{ij}} &= 2Z_i w_i + \lambda \times I \Rightarrow Z_i w_i \end{aligned} \quad (6)$$

第三步: 计算每个样本点的密集参数 θ_i , 用 θ_i 校正已求得的重建权值 w_i : $W^* = [w_1 \ w_2 \ \dots \ w_N] \cdot [\theta_1, \theta_2, \dots, \theta_N]^T$, 并满足 $\sum_{j=1}^k w_{ij} \theta_i = 1$;

第四步: 以修正的权值矩阵 W^* 构建低维嵌入 Y , 这等价于最小化下面约束问题:

$$\begin{cases} \min \Phi(Y) = \sum_{i=1}^N \|YI_i - \theta_i Yw_i\|^2 \\ = \sum_{i=1}^N \|Y(I_i - \theta_i w_i)\|^2 \\ = \min \text{tr}(YMY^T) \\ \text{s. t.} \quad YY^T = I \end{cases} \quad (7)$$

式中, $M = (I - W^*)(I - W^*)^T$. 使用拉格朗日乘子法, 解得 $MY^T = \lambda Y^T$, 则 M 的特征向量即所求的嵌入坐标. 取 M 最小的 d 个非零特征值所对应的特征向量作为低维坐标 Y .

4 实验结果与性能评价

对 2 组数据进行实验对比分析来验证本文提出的 ILLE 算法的有效性. 第一组为人工构造的非均匀数据集——Swiss Roll 数据集, 随机生成 3 000 个数据点, 分别应用 LLE 和本文提出的 ILLE 算法进行降维至 2 维 (近邻个数均选为 15). Swiss 数据集呈现三维空间中的二维非线性流形, 并采样了数据点分布不均匀程度不同的三组数据, 即样本点的密集度分别为 $\xi=0.3, 0.6, 0.9$, ξ 越大, 表示数据点分布越不均匀. 降维结果如图 1 所示 ILLE 嵌入结果明显优于 LLE 算法。

下面对中间呈现一个小空洞的 Swiss Roll 数据

集进行实验评价,将本文提出的 ILLE 算法和 Iso-map、LLE 算法进行比较(近邻个数均选为 10),图 2 表明: Isomap 和 LLE 算法的嵌入明显地扭曲了原

始数据的内部几何拓扑结构,而 ILLE 算法得到了很好的嵌入结果。

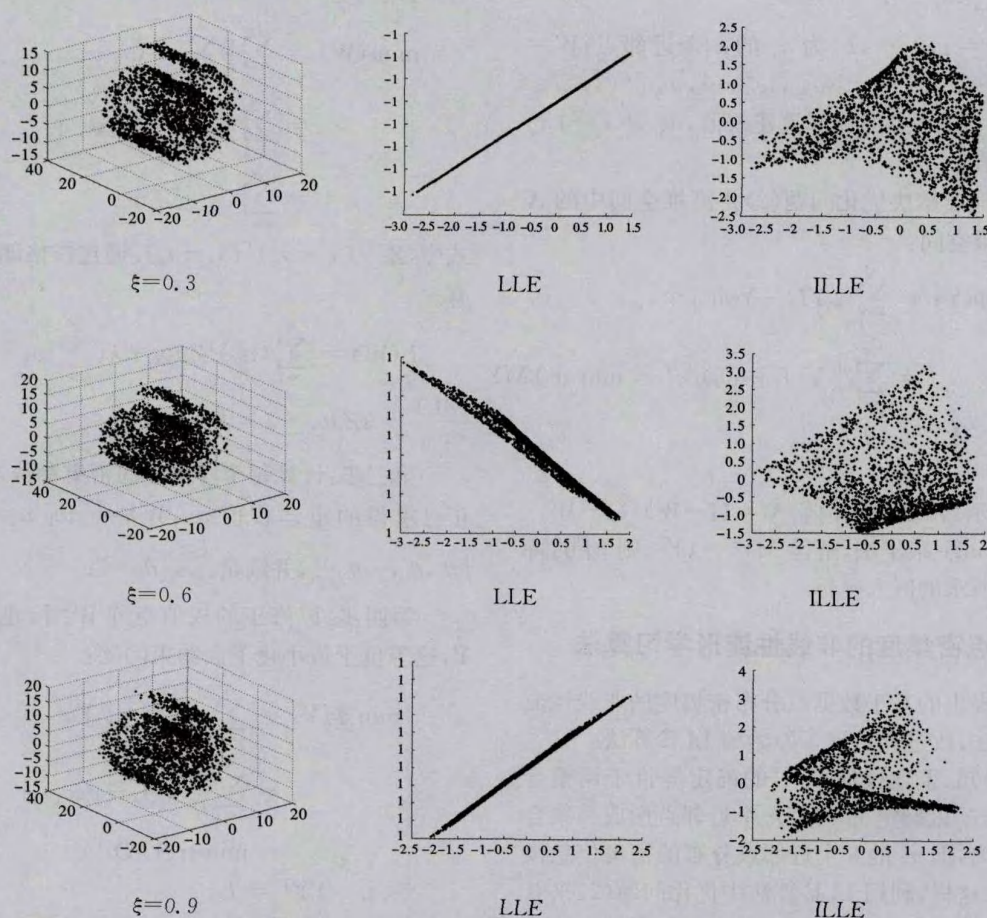


图 1 LLE 及 ILLE 对密度分布不均匀数据集的二维嵌入结果

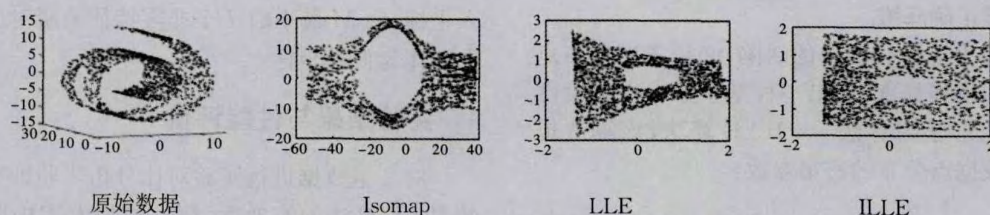


图 2 三种算法的嵌入结果

第二组实验选择 Isomap 人脸数据集(<http://www.cs.toronto.edu/roweis/data.html>), 包含 698 幅人脸图像,从中随机选出 630 幅人脸,视为不均匀数据集,每幅图像为 $64 \times 64 = 4\,096$ 维. 应用 ILLE 方法将其降至 2 维,散点图如图 3 所示:从四个路径描绘了姿势从上到下、从左到右以及不同程度照明的人脸图像,其捕捉到的姿势和照明以一个

连续的方式变化。

5 结束语

本文提出一种改进的基于数据点密集度的非线性降维算法(ILLE),该算法克服了传统 LLE 算法的缺陷. 在人工构造的非均匀数据集和随机抽取的人脸数据集上的实验证实了 ILLE 的有效性。

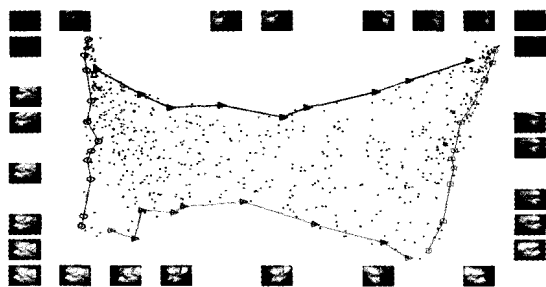


图 3 映射到 2 维空间的 Isomap 人脸图像

参考文献:

- [1] Roweis S T, Saul L K. An introduction to locally linear embedding[EB/OL]. [2006-07-06]. <http://isomap.stanford.edu/data/sets.html>.
- [2] 高小方. 流形学习方法中的若干问题分析[J]. 计算机科学, 2009, 36(4): 25-29.
- [3] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.
- [4] Tenenbaum J B, Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduc-

tion[J]. Science, 2000, 290(5500): 2319-2323.

- [5] Zhang Z Y, Zha H Y. Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment[J]. SIAM Journal of Scientific Computing, 2004, 26(1): 313-338.
- [6] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering[J]. Neural Information Processing Systems, 2002(14): 585-591.
- [7] Mordohai P, Medioni G. Dimensionality estimation, manifold learning and function approximation using tensor voting[J]. Journal of Machine Learning Research, 2010, 10(1): 411-450.

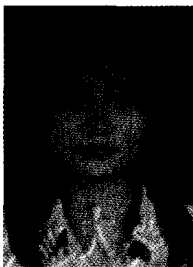
作者简介:



黄淑萍 女,(1977—),博士研究生.研究方向为数据降维、数据挖掘以及国际投资与金融等.

(上接第 9 页)

作者简介:



张园园 女,(1983—),博士研究生.研究方向为网络体系结构、路由协议、选播.

余镇危 男,(1942—),博士生导师.研究方向为网络体系结构、覆盖网、下一代互联网.