

抗噪的有监督局部保留投影降维算法*

张建军^a, 王士同^b

(江南大学 a. 物联网工程学院; b. 数字媒体学院, 江苏 无锡 214122)

摘要: 针对数据降维中的噪声干扰问题, 提出基于 L1-norm 有监督局部保留投影算法 SLPP-L1。SLPP-L1 利用 L1-norm 替代了 L2-norm; 因为欧式距离比绝对值距离对噪声更加敏感, 使得 SLPP-L1 抗噪性方面非常有效。实验结果表明, 该方法可以有效地剔除噪声的影响并且提高分类的识别率。

关键词: 数据降维; 绝对值距离; 鲁棒性; 欧式距离; 分类

中图分类号: TP391

文献标志码: A

文章编号: 1001-3695(2012)05-1641-03

doi:10.3969/j.issn.1001-3695.2012.05.010

Anti-noise supervised locality preserving projection algorithm for dimensionality reduction based on L1-norm

ZHANG Jian-jun^a, WANG Shi-tong^b

(a. School of Internet of Things Engineering, b. School of Digital Media, Jiangnan University, Wuxi Jiangsu 214122, China)

Abstract: To solve the noise disturbance of dimensionality reduction, this paper proposed a L1-norm based supervised locality preserving projection algorithm (termed SLPP-L1). It proposed SLPP-L1 by replacing L2-norm with L1-norm. Because Euclidean distance was more sensitive to outlier than absolute distance, SLPP-L1 became robust to outlier effectively. The experiments show that the proposed method can eliminate noise effectively and improve rate of recognition in classification.

Key words: dimensionality reduction; absolute distance; robustness; Euclidean distance; classification

0 引言

降维是分析高维数据的首要步骤之一, 因为高维数据里面包含很多冗余信息, 且数据量庞大不利于研究。降维的本质在于将高维数据嵌入到相对低维的空间, 提取数据集中重要的属性信息, 从而有效地解决了所谓的维数灾难问题。

众多降维算法可以分成线性和非线性两大类。其中主成分分析 (principal component analysis, PCA^[1]) 和线性判别分析 (linear discriminant analysis, LDA^[2]) 是两种常见的线性降维算法。PCA 力求降维后的数据集的方差达到最大, 没有考虑数据的类别属性, 因而属于无监督的降维算法; LDA 则依靠数据的类别信息、最大化类间离散度、最小化类内离散度, 因而属于有监督的降维算法。有一些高维数据, 在降维的同时要求保持局部形状, 这是近些年兴起的流形学习算法所解决的范畴, 如 ISOMAP^[3]、LLE^[4]、Laplacian Eigenmaps、LPP^[5] 等。其中 LPP 就是典型的非线性降维算法, 在降维的同时能够保持数据点流形上的邻域关系。在这些降维算法中, 用来度量数据点之间的相似度 (距离) 时, 通常使用的是欧式距离, 因为数据点距离的平方计算简单方便, 在一定程度上也较好地反映了数据点的相似度。但是由于距离的平方放大了噪声点对数据集降维的影响, 导致如 PCA、LPP 等算法的抗噪性很差。

为了提高 PCA、LPP 等算法的抗噪性能, 可以通过一定的改进, 用 Manhattan 距离^[6]代替 Euclidean 距离^[7], 也就是上文

提到的 L1-norm 替代 L2-norm^[8] 的做法。Manhattan 距离不会放大噪声的影响, 有效地提高了算法的抗噪性能, 如 PCA-L1^[9,10]、LPP-L1^[11]。由于 LPP 是无监督的, 因而文献^[12]提出了 SLPP 用于改善 LPP 在有类别信息的数据集降维时的性能。同样 SLPP 也是基于 L2-norm 的。本文基于 L1-norm 改进 SLPP, 提出了 SLPP-L1, 并且研究 SLPP-L1 在抗噪性能方面的表现。

1 LPP-L1

1) LPP 训练数据集 $X = [x_1, \dots, x_N] \in \mathbb{R}^{n \times N}$, 其中 $x_i \in \mathbb{R}^{n \times 1}$ 代表某个数据点; $U = [u_1, \dots, u_r] \in \mathbb{R}^{n \times r}$ 表示转换矩阵。

$$u = \arg \min_u f(u) = \arg \min_u \sum_{i,j} (u^T x_i - u^T x_j)^2 s_{ij} \quad (1)$$

$$\text{s. t. } u^T X D u X^T = 1 \quad (2)$$

其中: s_{ij} 是数据点之间的相似度量, 通常使用热核式 $s_{ij} = \exp(-\frac{\|x_i - x_j\|_2^2}{t})$, t 为调节参数, $D_{ii} = \sum_j s_{ij}$ 。

2) LPP-L1 引入变量 $d_{ij} = 1 - s_{ij} > 0$, 使得求最小值的问题转换成求最大值的问题。

$$u = \arg \max_u f(u) = \arg \max_u \sum_{i,j} (u^T x_i - u^T x_j)^2 d_{ij} \quad (3)$$

此时可以引入 L1-norm, 问题转换为

$$u = \arg \max_u f(u) = \arg \max_u \sum_{i,j} |u^T x_i - u^T x_j| d_{ij} \quad (4)$$

收稿日期: 2011-11-04; 修回日期: 2011-12-12 基金项目: 国家自然科学基金资助项目 (60903100, 60975027); 江苏省自然科学基金资助项目 (BK2009067)

作者简介: 张建军 (1987-), 男, 安徽巢湖人, 硕士, 主要研究方向为模糊系统建模 (qijieai123456@163.com); 王士同 (1964-), 男, 江苏扬州人, 教授, 博导, 主要研究方向为模式识别、人工智能、生物信息学。

条件式(2)可以简化为 s. t. $\|u\|_2 = 1$ 。

s. t. $\|\mu\|_2 = 1$ (13)

2 SLPP 与 SLPP-L1

2.1 SLPP

LPP 是一种非监督的降维方法,对于有类别信息的降维时,可能效果会很差。例如当两类样本靠得很近,由于 LPP 的局部保持特性,它会把属于两类的数据点混在一起。

LPP 的目标函数可以作以下分析:

$$\sum_{ij} (y_i - y_j)^2 S_{ij} = B + W = \sum_{ij} (y_i - y_j)^2 B_{ij} + \sum_{ij} (y_i - y_j)^2 W_{ij} \quad (5)$$

其中:

$$B_{ij} = \begin{cases} S_{ij} & x_i \text{ 和 } x_j \text{ 属于不同类} \\ 0 & \text{其他} \end{cases}$$

$$W_{ij} = \begin{cases} S_{ij} & x_i \text{ 和 } x_j \text{ 属于同一类} \\ 0 & \text{其他} \end{cases}$$

显然要是能把不同类别的数据给分开,应该尽可能让同一类的数据越近越好,不同类间的数据越远越好,即满足

$$\frac{\sum_{ij} (y_i - y_j)^2 B_{ij}}{\sum_{ij} (y_i - y_j)^2 W_{ij}}$$

。这也就是 SLPP 的本质所在。

2.2 SLPP-L1

从 SLPP 的目标函数可以看出,它仍然使用的是 L2-norm,现在通过改进 SLPP,从而引入 L1-norm。

$$\frac{\sum_{ij} (y_i - y_j)^2 B_{ij}}{\sum_{ij} (y_i - y_j)^2 W_{ij}} = \frac{\frac{1}{2} \sum_{ij} (a^T x_i - a^T x_j)^2 B_{ij}}{\frac{1}{2} \sum_{ij} (a^T x_i - a^T x_j)^2 W_{ij}} =$$

$$\frac{\sum_i a^T x_i^B x_i^T a - \sum_{ij} a^T x_i B_{ij} x_j^T a}{\sum_i a^T x_i D_{ii}^W x_i^T a - \sum_{ij} a^T x_i W_{ij} x_j^T a} =$$

$$\frac{a^T x (D^B - B) x^T a}{a^T x (D^W - W) x^T a} = \frac{a^T x L^B x^T a}{a^T x L^W x^T a} \quad (6)$$

下面的式(12)等同于

$$\max a^T x L^B x^T a \quad (7)$$

$$\text{s. t. } a^T x L^W x^T a = 1 \quad (8)$$

令 $U = x L^B x^T$, $V = x L^W x^T$ 。

对 U 和 V 分别进行 Hess 分解可以得到:

$$[P, Q] = \text{hess}(U)$$

$$PQP^T = U$$

$$U = PQP^T = P(Q^{\frac{1}{2}})^T Q^{\frac{1}{2}} P^T = (Q^{\frac{1}{2}} P^T)^T Q^{\frac{1}{2}} P^T$$

$$a^T x L^B x^T a = a^T (Q^{\frac{1}{2}} P^T)^T Q^{\frac{1}{2}} P^T a = (Q^{\frac{1}{2}} P^T a)^T Q^{\frac{1}{2}} P^T a$$

同理,

$$[R, S] = \text{hess}(V)$$

$$a^T x L^W x^T a = (S^{\frac{1}{2}} R^T a)^T S^{\frac{1}{2}} R^T a$$

式(7)和(8)可以写成:

$$\max \|Q^{\frac{1}{2}} P^T a\|_2 \quad (9)$$

$$\text{s. t. } \|S^{\frac{1}{2}} R^T a\|_2 = 1 \quad (10)$$

经过进一步变形:

$$\max \|Q^{\frac{1}{2}} P^T (S^{\frac{1}{2}} R^T)^{-1} \mu\|_2 \quad (11)$$

$$\text{s. t. } \|\mu\|_2 = 1$$

此时引入 L1-norm 规则可以得到:

$$\max |Q^{\frac{1}{2}} P^T (S^{\frac{1}{2}} R^T)^{-1} \mu| \quad (12)$$

经过改进后的 SPP,完全符合 L1-norm 的基本思想,也就是把欧式距离变成绝对值距离。下面用实验来研究 SPP-L1 的抗噪性能。

3 实验结果与分析

3.1 实验 1

从最简单的开始做起,准备 200 个二维点分成两类,用 LPP 和 SLPP 把二维点投影到一维空间。第一组实验两类点尽可能分开,实验结果如图 1 所示;第二组实验两类点有部分点重叠,实验结果如图 2 所示。

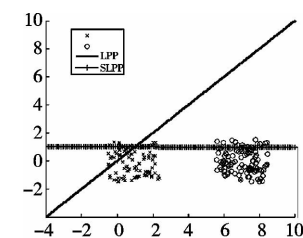


图1 无混合两类数据集分类

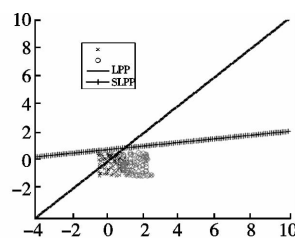


图2 有混合两类数据集分类

由以上实验可以看出,当不同类数据点靠近时,SLPP 能够较准确地进行分类,而 LPP 却不行。

在上述实验的基础上,对 SLPP 和 SLPP-L1 进行对比实验。第一组实验加入少量噪声,噪声的类别在两类中随机挑选,实验结果如图 3 所示,当随机噪声过多时,分类变得没有意义。第二组实验不加入噪声,而是给原来数据集加入一定量的扰动,数据集缓慢地变化,研究变化前后 SLPP 和 SLPP-L1 的投影向量变化率大小。第二组实验进行多次,求平均投影向量变化率。

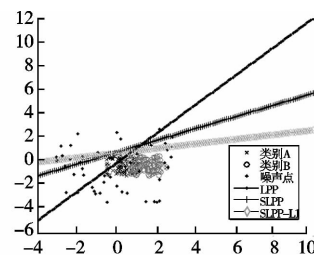


图3 有混合两类数据集噪声污染后的分类

为了方便度量投影向量的变化率,引入

$$\Delta_i = \|w_i - w_0\| / \|w_0\| + |b_i - b_0| / |b_0| \quad (14)$$

对于 SLPP,利用原始数据 w_0 和 b_0 ,然后连续做 20 次实验作为一组,求得 $\Delta_{slpp}(w_0, b_0) = \frac{1}{20} \sum_{i=1}^{20} \Delta_i$,同时求得 $\Delta_{slpp-l1}(w_0, b_0) = \frac{1}{20} \sum_{i=1}^{20} \Delta_i$ 。表 1 为具体实验结果。

表 1 SLPP 与 SLPP-L1 投影向量变化率

实验序号	实验组数	扰动系数	Δ_{slpp}	$\Delta_{slpp-l1}$
1	20	.005	0.152	0.124
2	20	.008	0.196	0.165
3	20	.01	0.254	0.187
4	20	.02	0.342	0.248

3.2 实验 2

用常见的 segment、letter recognition、landsat satellite 数据集,观察 SLPP 和 SLPP-L1 进行降维后的可视化结果,比较扰动前后的低维图像变化情况,结果如图 4~6 所示。

由图4~6可以看出,LPP 甚至连分类都分不开,更不用谈及抗噪性了;数据集进行一定的扰动后发现 SLPP 前后的类别分布变化较大,而 SLPP-L1 的前后类别分布变化不大,说明 SLPP-L1 较之于 LPP 的抗噪性能好。

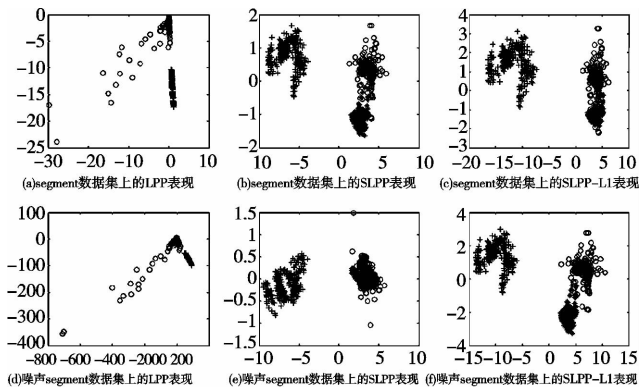


图4 Segment数据集下LPP、SLPP、SLPP-L1分类对比

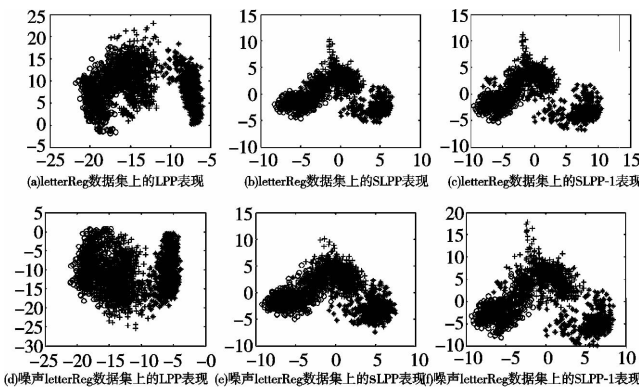


图5 LetterReg数据集下LPP、SLPP、SLPP-L1分类对比

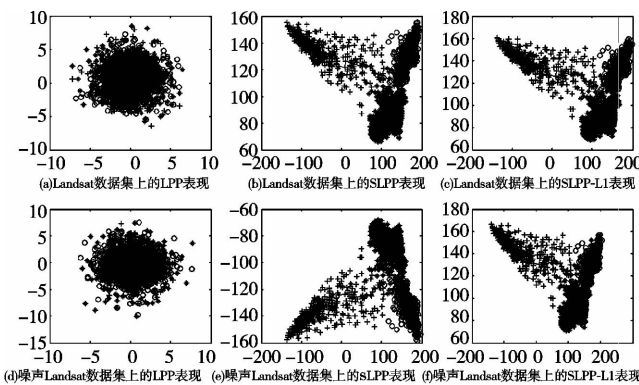


图6 Landsat数据集下LPP、SLPP、SLPP-L1分类对比

3.3 实验3

从 Yalefaces 库中任意选五个人的11张照片,图像大小为 64×64 ,如图7所示。



图7 Yalefaces人脸样本

任意选取每个人的8张照片作为训练样本,剩下的3张照片作为测试样本,训练样本中随机选取20%的照片按照图8那样加入随机噪声,最后进行多次实验记录 SLPP 和 SLPP-L1 在识别率方面的表现,结果如表2所示。



图8 Yalefaces加随机噪声的人脸样本

其中 d 为提取的特征数,由于计算量很大,提取的特征数不能太多。SLPP-L1 在抗噪声方面优于 SLPP 的表现。

表2 SLPP 与 SLPP-L1 平均识别率对比

实验序数	子空间维数/ d	SLPP	SLPP-L1
1	7	0.84	0.86
2	7	0.82	0.88
3	5	0.72	0.80
4	4	0.70	0.76

4 结束语

本文将有监督的局部保留算法与 L1-norm 准则相结合,使得算法具有较好的抗噪性能。实验结果证明,较之于原先的 SLPP,SLPP-L1 的确能够抗噪声,表现为抗随机噪声和抗数据集扰动两个方面。本文从 L1-norm 的一般形式和成功案例入手^[11],指导 SLPP-L1 的推导。针对 SLPP-L1 的实验只采用了小样本,因为针对 SLPP-L1 的迭代过程计算量很大,还没有使用大样本集进行验证,后续工作将对算法进行改进和优化,以加快运算速度,在大样本集上进行验证研究。

参考文献:

- [1] CHEN Fu-bing, CHEN Xiu-hong, WANG Wen-sheng, *et al.* The generalization of PCA for human face recognition[J]. *Computer Engineering and Applications*, 2010, 27(7): 34-38.
- [2] HAO Hong-wei, ZHANG Lei. Face recognition based on SVD and LDA [J]. *Application Research of Computers*, 2007, 24(12): 377-379.
- [3] MENG De-yu, XU Chen, XU Zong-ben. A new manifold reconstruction method based on isomap[J]. *Chinese Journal of Computers*, 2010, 33(3): 545-554.
- [4] ZHU Ming-han, LUO Da-yong, YI Li-qun, *et al.* Incremental locally linear embedding algorithm based on orthogonal iteration method[J]. *Acta Electronica Sinica*, 2009, 37(1): 132-136.
- [5] ZHANG Du-zhen. A general survey of locality preserving feature transformation [J]. *Computer Engineering and Science*, 2010, 31(1): 80-83.
- [6] LI Jun-hua, PENG Li. New approach to facial expression classification Manhattan distance[J]. *Computer Engineering and Applications*, 2008, 44(2): 74-79.
- [7] SHANG Zhen-hong, FU Xiao-dong, YUAN Mei-yu. Line detection based on Euclidean distance [J]. *Journal of Computer Applications*, 2008, 28(1): 177-180.
- [8] FEI Yan, KRYSTIAN M, JOSEF K. A comparison of L1-norm and L2-norm multiple kernel SVMs in image and video classification [C]//Proc of the 7th International Workshop on Content-Based Multimedia Indexing, 2009: 7-12.
- [9] KWAK N. Principal component analysis based on L1-norm maximization[J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2008, 30(9): 1672-1680.
- [10] DONOHO D L, TSAIG Y. Fast solution of L1-norm minimization problems when the solution may be sparse[J]. *IEEE Trans on Information theory*, 2008, 54(11): 4789-4812.
- [11] PANG Yan-wei, YUAN Yuan. Outlier-resisting graph embedding [J]. *Neurocomputing*, 2010, 73(4-6): 968-974.
- [12] SHEN Zhong-hua, PAN Yong-hui, WANG Shi-tong. A supervised locality preserving projection algorithm for dimensionality reduction [J]. *Pattern Recognition and Artificial Intelligence*, 2008, 21(2): 233-239.