一种自动抑制离群点的子空间学习方法

庞彦伟[①②]　刘政凯[①]
[①](中国科学技术大学电子工程与信息科学系　合肥　230027)
[②](天津大学电子信息工程学院　天津　300072)

摘　要：子空间学习如主成分分析是有效的数据降维方法。但这类方法计算的基向量受离群(outlier)数据的影响很大，导致降维后的数据不能准确地刻画数据的真实分布。为了减少离群数据的影响，该文提出了一种改进的子空间学习方法。该方法不需要直接探测离群数据的位置，而且子空间的求解可归结为特征值分解问题，具有全局最优解。仿真数据上的试验表明该方法是有效的。

关键词：子空间；降维；离群数据
中图分类号：TP391.41　　**文献标识码**：A　　**文章编号**：1009-5896(2008)01-0176-04

Automatically Outlier-Resisting Subspace Learning

Pang Yan-wei[①②]　Liu Zheng-kai[①]
[①](*Dept. of EEIS, University of Science and Technology of China, Hefei* 230027, *China*)
[②](*School of Electronic Information Engineering, Tianjin University, Tianjin* 300072, *China*)

Abstract: Subspace learning is an effective dimensionality reduction method. However, the resulting basis vectors are significantly biased due to the presence of outlier points. Consequently, the transformed data in the subspace cannot faithfully describe the intrinsic distribution of the original data. To tackle this problem, a modified subspace learning algorithm is proposed. In the algorithm it is not necessary to detect outliers. Moreover, the algorithm is reduced to an eignenvalue problem which has a globally optimal solution. Experiments on synthetic data demonstrate the effectiveness of the proposed algorithm.

Key words: Subspace learning; Dimension reduction; Outlier data

1　引言

在计算机视觉、模式识别和数据挖掘中常常需要计算高维数据 $\boldsymbol{x}\in R^D$ 的低维表示 $\boldsymbol{y}\in R^d$, 其中 $d<D$。例如把大小为 100×100 像素的人脸图像用向量表示，它的维数高达 10000。这不仅使人脸识别的计算量很大，而且也导致严重的“维数灾难”。因此，一般都要对高维数据进行降维。子空间方法是常用的降维手段。它基于这样的事实：虽然原始数据的维数很高，但一类数据一般通常具有某种属性，使得它们并不能充满整个高维空间，而是仅位于嵌套在高维空间中一个很小的子空间中。子空间方法的目的就是根据某种准则，计算出这个子空间的基向量 $\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_d]$，从而得到数据的低维表示 $\boldsymbol{y}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}$。

常用的子空间方法[1]有主分量(成分)分析(PCA)[2]，线性判别分析(LDA)[3]，独立主分量分析(ICA)[4]等。它们的区别在于各自不同的目标函数。PCA 是在均方重建误差最小意义下最优的子空间方法。ICA 是 PCA 的推广，但与 PCA 不同，ICA 不需要假设数据服从高斯分布，而且能分离出数据的高阶统计量。PCA 和 ICA 都属于无监督的子空间学习方法，而 LDA 则是一种有监督的方法。LDA 的优化目标是使得类间距离与类内距离之比最大。

虽然这些方法在不同的场合得到了成功应用，但它们受离群数据点的影响很大。在计算机视觉的应用中，图像噪声、配准误差和遮挡等因素都可能导致离群数据，从而“污染”训练集合，使得到的子空间的基向量偏离理想的基向量。这是由于在目标函数中离群点的贡献过大。图 1 以 PCA 为例，假设理想的 4 个数据点分布在一条水平直线上，显然主分量(基向量)的方向就是这条直线的方向。然而，实际的数据常常由于某种原因而受到离群点的“污染”。例如图 1 中有一个离群点，由这个 5 个点组成的数据所得到的基向量会严重偏离理想的基向量，二者之间的夹角达 90° 。因此，减弱离群点对子空间学习的影响是着重要意义。

为了消除离群点的影响，最直接的方法是通过某种检测方法，检测出离群点，然后把它从训练集合中去除。常用的检测算法可分为基于统计的方法、基于距离的方法、和基于密度的方法等[5]。但这种先检测后去除的策略不适合人脸识别等训练样本十分稀疏(小样本)的情况。首先检测算法的精度有限，错误地去除有效样本对子空间学习的损失很大。很多情况下，离群数据是由于图像的局部变化造成的，仅仅因

2006-06-12 收到，2007-03-14 改回
国家自然科学基金(60605005)资助课题

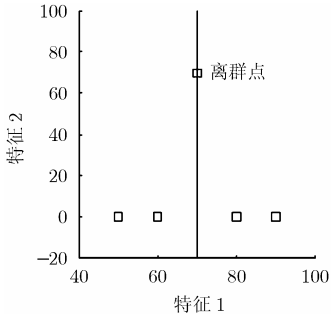

图 1　离群点使得 PCA 得到错误的主分量

为局部原因而放弃使用整个样本，这种做法忽略了其它大部分区域携带的有用信息。基于 M-estimation 的鲁棒主成分分析(RPCA)[6,7]虽然可以很大程度上克服上述缺点，但 RPCA 迭代优化过程复杂，计算量大。例如计算 256 幅 120×160 像素的前 20 个基向量需要 3 个小时[6]。而且这种方法也没有利用类标签信息，不能应用到有监督的子空间学习中。本文提出了一种简单有效的对离群点不敏感的子空间方法。该方法不需要直接探测离群数据的位置，而且子空间的求解可归结为特征值分解问题，具有全局最优解。特别地，它可以很容易地推广到有监督的子空间学习中。

2　增加校正项的 PCA 子空间方法

本节在 PCA 的基础上提出对离群点有抑制特性的子空间学习方法。

给定高维训练数据 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$,其中 $\boldsymbol{x}_i\in R^D$，N 表示样本的个数。设子空间的 d 个基向量组成的矩阵为 $\boldsymbol{U}=[\boldsymbol{u}_1\ \boldsymbol{u}_2\ \boldsymbol{u}_d]$,其中 $\boldsymbol{u}_i\in R^D$，$d<D$。通过变换 $\boldsymbol{y}_i=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_i$ 把高维数据 $\boldsymbol{X}$ 映射为低维数据 $\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N]$，其中 $\boldsymbol{y}\in R^d$。$\boldsymbol{U}$ 对 $\boldsymbol{x}_i$ 的重建为 $\tilde{\boldsymbol{x}}_i=\boldsymbol{U}\boldsymbol{y}_i=\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_i$。PCA 的优化目标是使样本重建的均方误差最小：

$$J_{\mathrm{pca}}(\boldsymbol{U})=\sum_{i=1}^{N}||\boldsymbol{x}_i-\tilde{\boldsymbol{x}}_i||^2=\sum_{i=1}^{N}||\boldsymbol{x}_i-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_i||^2 \tag{1}$$

不失一般性，假设 $\boldsymbol{X}$ 中只有一个离群点 $\boldsymbol{x}_N$。则上式可分解为

$$J_{\mathrm{pca}}(\boldsymbol{U})=\sum_{i=1}^{N-1}||\boldsymbol{x}_i-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_i||^2+||\boldsymbol{x}_N-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_N||^2 \tag{2}$$

离群点距离孤立于其它按照某种概率分布聚合在一起的数据，最小化过程中为了使均方误差最小一定更加关注离群点，也就是说离群点对均方误差之和的贡献比较大。这就导致最后得到的主分量 $\tilde{\boldsymbol{U}}$ (特征值最大的基向量)偏离由无离群点数据$[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{N-1}]$计算出的主分量 $\boldsymbol{U}^*$。

为了校正离群点对均方误差的影响，本文提出在式(1)后增加一个校正项$\varphi(\boldsymbol{U},\boldsymbol{X})$，得到新的代价函数 J：

$$J(\boldsymbol{U})=\sum_{i=1}^{N}||\boldsymbol{x}_i-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_i||^2+\varphi(\boldsymbol{X},\boldsymbol{U}) \tag{3}$$

要求校正项能把 $\tilde{\boldsymbol{U}}$ 尽可能地“牵引”到 $\boldsymbol{U}^*$，即使 $\boldsymbol{u}$ 和 $\boldsymbol{u}^*$ 的夹角 θ 变小。定义校正项为

$$\varphi(\boldsymbol{X},\boldsymbol{U})=\sum_{i=1}^{M}||\boldsymbol{z}_i-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{z}_i||^2 \tag{4}$$

其中 $\boldsymbol{z}_i\in R^D$ 由 $\boldsymbol{X}$ 中的 3 个样本($\boldsymbol{x}_j$，$\boldsymbol{x}_k$，$\boldsymbol{x}_l$)产生，$\boldsymbol{z}_i$ 是样本 $\boldsymbol{x}_j$ 在直线 $\overline{\boldsymbol{x}_k\boldsymbol{x}_l}$ 上的投影点，满足 $\boldsymbol{z}_i=\boldsymbol{x}_k+t(\boldsymbol{x}_k-\boldsymbol{x}_l)$，其中 $t=[(\boldsymbol{x}_j-\boldsymbol{x}_l)^{\mathrm{T}}(\boldsymbol{x}_{\mathrm{k}}-\boldsymbol{x}_l)]/[(\boldsymbol{x}_k-\boldsymbol{x}_l)^{\mathrm{T}}(\boldsymbol{x}_k-\boldsymbol{x}_l)]$。固定 $\boldsymbol{x}_j$，从剩下的 N-1 个点中可以组合 $C_{N-1}^2=(N-1)\times(N-2)/2$个无序对($\boldsymbol{x}_k$，$\boldsymbol{x}_l$), $k\neq l=j$。遍历 $\boldsymbol{x}_j$ 则整个数据集合 $\boldsymbol{X}$ 可产生 $M=NC_{N-1}^2=N(N-1)\times(N-2)/2$个投影点。设 $\boldsymbol{Z}=[\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_M]$。因为 $\boldsymbol{z}_i$ 不是由实际的物理过程产生的，因此称它为虚拟样本。下面计算式(2)的最优解，校正项φ 作用的直观解释见第 4 节。

把式(4)代入式(3)中，并令矩阵 $\boldsymbol{A}_{D\times(M+N)}=(\boldsymbol{X},\boldsymbol{Z})$得到：

$$\begin{aligned}J(\boldsymbol{U})&=\sum_{i=1}^{N}||\boldsymbol{x}_i-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}_i||^2+\sum_{i=1}^{M}||\boldsymbol{z}_i-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{z}_i||^2\\&=\sum_{i=1}^{N+M}||\boldsymbol{a}_i-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{a}_i||^2\\&=\sum_{i=1}^{N+M}(\boldsymbol{a}_i-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{a}_i)^{\mathrm{T}}(\boldsymbol{a}_i-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{a}_i)\\&=\mathrm{trace}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\boldsymbol{A}\end{aligned} \tag{5}$$

为了求得唯一解增加正交约束项 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}$，并利用拉格朗日乘子，于是有：

$$L(\boldsymbol{U})=\mathrm{trace}\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\boldsymbol{A}+\sum_{j=1}^{d}\lambda_j e_j^{\mathrm{T}}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}-\boldsymbol{I}) \tag{6}$$

其中 $\boldsymbol{e}_i\in R^d$,表示第 i 个元素为 1 但其它元素都为 0 的向量。求 $L(\boldsymbol{U})$对 $\boldsymbol{U}$ 的导数并令其为 0，得到如下的特征值分解形式：

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{U}\boldsymbol{\Lambda} \tag{7}$$

其中 $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_d)$。

式(7)的形式和传统的 PCA[2]一样，区别在于式(7)是对扩充的样本 $\boldsymbol{A}$ 的协方差矩阵 $\boldsymbol{A}\boldsymbol{A}^T$进行分解，而传统的 PCA 是对原始样本 $\boldsymbol{X}$ 的协方差求解。由于在本方法中，并没有直接寻找离群点的位置。因此是一种自动的方法。而且子空间的求解可归结为标准特征值分解问题，具有全局最优解。

3　对传统 PCA 的正则化

注意到训练矩阵 $\boldsymbol{A}_{D\times(M+N)}=(\boldsymbol{X},\boldsymbol{Z})$由原始的训练样本 $\boldsymbol{X}$ 和虚拟样本 $\boldsymbol{Z}$ 组合而成,共有 $M+N=N(N-1)\times(N-2)/2+N$ 个样本，因此矩阵 $\boldsymbol{A}$ 占用很大的存储空间，而且使得协方差矩阵 $\boldsymbol{C}'=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 的计算复杂度比传统的 PCA 协方差矩阵 $\boldsymbol{C}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 高很多。本节将表明上一节所提出的方法等价于对传统 PCA 的一个正则化处理，虚拟点的信息可以包含在一个稀疏矩阵中，从而可以大大节约存和计算开销。如果把 $\boldsymbol{C}$ 表示成 $\boldsymbol{C}=\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}=\boldsymbol{X}\boldsymbol{I}\boldsymbol{X}^{\mathrm{T}}$，其中 $\boldsymbol{I}_{N\times N}$是单位阵，则有：

$$\boldsymbol{C}'=\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{X}(\boldsymbol{I}+\boldsymbol{R})\boldsymbol{X}^{\mathrm{T}} \tag{8}$$

其中 $\boldsymbol{R}_{N\times N}$在本文称为正则矩阵。显然计算 $\boldsymbol{X}(\boldsymbol{I}+\boldsymbol{R})\boldsymbol{X}^{\mathrm{T}}$ 比计算 $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 所需的存储量和计算量要少很多。更重要的是 $\boldsymbol{X}(\boldsymbol{I}+\boldsymbol{R})\boldsymbol{X}^{\mathrm{T}}$ 表明本文提出的对离群点鲁棒的子空间方法实质上是对传统的 PCA 方法的一个正则化，即对单位矩阵的正则化。下面阐述矩阵 $\boldsymbol{R}$ 的计算过程即式(8)的推导。

不妨设投影点 $\boldsymbol{z}_1$ 是由原始样本在 $\overline{\boldsymbol{x}_1\boldsymbol{x}_2}$ 上的投影得到的，

即 $\boldsymbol{z}_1=\boldsymbol{x}_1+t(\boldsymbol{x}_2-\boldsymbol{x}_1)$，其 $t=[(\boldsymbol{x}_3-\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2-\boldsymbol{x}_1)]/[(\boldsymbol{x}_2-\boldsymbol{x}_1)^{\mathrm{T}}(\boldsymbol{x}_2-\boldsymbol{x}_1)]$。下式表明可以用矩阵 $\boldsymbol{X}$ 来表示 $\boldsymbol{z}_1$:

$$\begin{aligned}\boldsymbol{z}_1 &= \boldsymbol{x}_1 + t(\boldsymbol{x}_2 - \boldsymbol{x}_1)\\ &= (1-t)\boldsymbol{x}_1 + t\boldsymbol{x}_2 + 0\boldsymbol{x}_3 + \cdots + 0\boldsymbol{x}_N = \boldsymbol{X}\boldsymbol{s}_1\end{aligned} \tag{9}$$

其中向量$(\boldsymbol{s}_1)_{N\times 1}=[(1-t)\ t\ 0\ \cdots\ 0]^{\mathrm{T}}$最多有两个非零元素，称它为选择向量。同样地，每个 $\boldsymbol{z}_i$ 都一个 $\boldsymbol{s}_i$ 与之对应，使得 $\boldsymbol{s}_i=\boldsymbol{X}\boldsymbol{s}_i, i=1,2,\cdots,M$。因此有 $\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{S}$ 成立。称 $\boldsymbol{S}_{N\times M}=[\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_M]$为选择矩阵。$\boldsymbol{S}$ 是一个稀疏矩阵，它的每一列最多有两个非零向量。利用选择矩阵，可以简化协方差矩阵 $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$:

$$\begin{aligned}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} &= (\boldsymbol{X},\boldsymbol{Z})(\boldsymbol{X},\boldsymbol{Z})^{\mathrm{T}}\\ &= \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + (\boldsymbol{X}\boldsymbol{S})(\boldsymbol{X}\boldsymbol{S})^{\mathrm{T}}\\ &= \boldsymbol{X}(\boldsymbol{I} + \boldsymbol{S}\boldsymbol{S}^{\mathrm{T}})\boldsymbol{X}^{\mathrm{T}}\end{aligned} \tag{10}$$

令 $\boldsymbol{R}\equiv\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}$，就可由式(10)得到式(8)。因为 $\boldsymbol{S}$ 是十分稀疏的矩阵，所以式(9)的存储和计算开销比较小。

4 直观解释和仿真试验

虽然上两节分别表明本文所提方法是在 PCA 基础上增加一个校正项或对 PCA 的正则化得到的。但还没有从理论上显式地证明该方法对离群点的鲁棒性。尽管如此，直观解释和仿真试验有助于理解所提方法的合理性，并对将来的理论证明有启发作用。

注意到本文方法的主要特点是增加了虚拟点 $\boldsymbol{z}_i$。因此，需要观察虚拟点在子空间主分量计算中的作用。若真实的主分量方向为 $\boldsymbol{u}$，而实际求得的方向为 $\boldsymbol{v}$，则可以用 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的夹角 θ 衡量 $\boldsymbol{v}$ 的准确度。θ 越小，越准确。在图 2 中，为了便于说明问题，假设理想的 4 个二维数据点(用“□”表示)位于一条直线上，显然水平方向就是它们的主分量 $\boldsymbol{u}$。由于某种扰动，产生了一个离群点，它位于 4 个理想点之上。这 5 个点产生的虚拟点 $\boldsymbol{z}_i$ 用“*”表示。在图 2(a)和 2(b)中，离群点偏移量较小，影响不大，所以两种方法都能找到正确的主分量。但是随着偏移量的增大(如图 2(c)和 2(d))，离群点的影响也就变大(公式(2))，传统的 PCA 得到的 $\boldsymbol{v}$(虚线表示)严重偏离水平方向 $\boldsymbol{u}$，二者的夹角达 90°。而在本文的方法中，由于虚拟点主要集中在真实主分量 $\boldsymbol{u}$ 附近，使得最终计算出的 $\boldsymbol{v}$(实线表示)与 $\boldsymbol{u}$ 的夹角达 0°，表明本方法的能够自动探测到离群点，并消除它的影响。需要指出，本方法对离群点的抑制作用是有限度的。例如当离群点偏离理想数据过远时(如图 2(e))，本方法和传统的 PCA 方法一样都不能得到最优的主分量。这是由于离群点引起的平方项在式(5)中占的比重太大，以至于引入的虚拟点(校正项)不足以抵消离群点的作用。

除了从虚拟点的分布情况来解释外，还可以从最近邻特征线[8]的角度理解。虽然最近邻特征线分类方法(NFL)和本文方法分属子空间学习(数据降维)和分类器两个不同范畴，但二者都使用了虚拟点，增强了算法的泛化能力。NFL 使用虚拟点增强样本的表达能力，而本文则使用虚拟点来抵消离群点的影响。

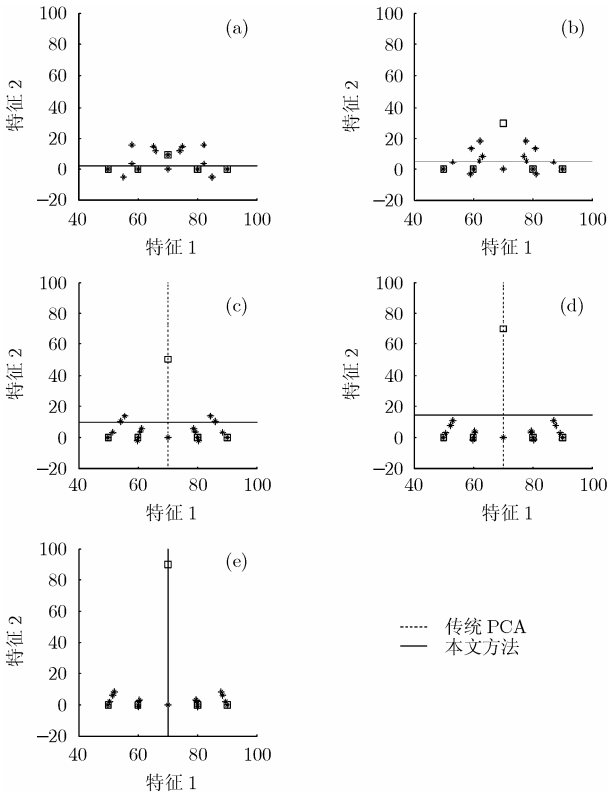

图 2 本文方法与传统 PCA 对离群点的敏感性对比，虚拟点 $\boldsymbol{z}_i$ 用“*”表示

5 在 LDA 上的推广

Fisher 线性判别分析 LDA 也易于受离群点的影响。这一节把增加虚拟点以削弱离群点影响的方法引入到LDA中，使新的 LDA 更加稳定。

5.1 在 LDA 上推广

给定 C 类问题，每类有 N_i 个样本。第 i 类样本的标签为 $l(i)$，均值为 $\boldsymbol{u}_i$。所有样本的均值为 $\boldsymbol{u}$。经典的 LDA 在正交约束下，使类间散布与类内散布之比最大:

$$J_{\mathrm{lda}}(\boldsymbol{u}) = \frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{u}}$$

其中

$$\boldsymbol{S}_b = \sum_{i=1}^{C} N_i(\boldsymbol{u}_i-\boldsymbol{u})(\boldsymbol{u}_i-\boldsymbol{u})^{\mathrm{T}},\ \boldsymbol{S}_w = \sum_{i=1}^{C}\boldsymbol{S}_w^i$$

$$\boldsymbol{S}_w^i = \sum_{\boldsymbol{x}_j\in l(i)} (\boldsymbol{x}_j-\boldsymbol{u}_i)(\boldsymbol{x}_j-\boldsymbol{u}_i)^{\mathrm{T}}$$

上式中 $\boldsymbol{S}_b$ 称为类间散布矩阵，$\boldsymbol{S}_w$ 为类内散布矩阵，$\boldsymbol{S}_w^i$ 为第 i 类样本的类内散布矩阵。为了增强 LDA 对离群点鲁棒性，借鉴第 2 节的方法，首先对每类样本增加虚拟样本扩充训练样本，然后在扩充后的训练样本上实现上述 LDA。当然，实际计算时可以像第 3 节提出的方法那样，不需要显示地计算出虚拟样本，而是把虚拟点的信息涵盖在一个稀疏矩阵中。

图 3 示意出了本方法的优越性。在该图中有两类样本，分别用“Δ”和“□”表示。每类均有 4 个正常的数据。显然，用正常数据求出的鉴别向量 $\boldsymbol{u}$ 为与水平方向平行。现在给第一类增加一个离群点，例如图 3(a)的离群点在左上角。对“污

染后”的数据分别采用传统的 LDA 和本文的 LDA(增加虚拟点)计算鉴别向量 $\boldsymbol{u}$ 和 $\boldsymbol{u}'$。图中 $\boldsymbol{u}$ 和 $\boldsymbol{u}'$分别用实线和虚线表示。可以看到在图 3(a)和 3(b)中，$\boldsymbol{u}'$和水平方向的夹角非常小，表明本文的方法受离群数据的影响很少。对比之下，$\boldsymbol{u}$ 和水平方向的夹角却非常大，说明离群点对传统 LDA 影响十分明显。在图 3(c)中，虽然 $\boldsymbol{u}'$和水平方向之间有相当大的夹角，但它还是显著小于 $\boldsymbol{u}$ 和水平方向的夹角。总之，本文的方法无论对无监督学习的PCA和有监督学习的LDA在对离群点的抑制方面都有显著的改善。

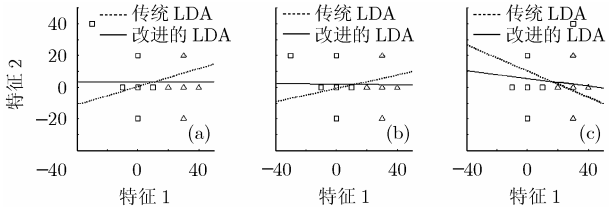

图 3　本文改进的 LDA 与传统 LDA 对离群点的敏感性对比

5.2 在人脸识别中的应用

实验采用了 FERET 人脸数据库[9]的一个子集。该子集由 197 个人的 1394 幅图像组成。每人 7 幅(如图 4 所示，已归一化)，它们的文件名分别标以字符“ba”, “bj”,“bk”, “be”, “bf”, “bd”, 和“bg”。从 7 幅图像中随机选取 3 幅用作训练，其余的 4 幅用作测试。我们做了 15 次上述随机实验，于是有 15 个不同的训练集和测试集。最后取它们的平均识别率。

图 4　ERET 人脸图像举例

表 1 列出了 PCA、LDA 和本文方法在使用最近邻(NN)分类器和最近邻特征线(NFL)分类器的识别率。实验中 PCA 使用的特征数为 430，LDA 使用了 196 个特征，而本文所提方法的特征维数为 108。数据表明本文的降维方法结合 NFL 分类器的识别率比 PCA+NFL 高 24.16%，比 LDA+NFL 也要高 12.6%。显然，这种改进的 LDA 方法优于其它方法。

表 1　人脸识别率比较 (%)

PCA+NN	PCA+NFL	LDA+NN	LDA+NFL	本文方法 +NFL
49.22	57.51	67.30	69.41	81.67

6　结束语

为了减弱离群点对子空间学习的影响，提出了对 PCA 的代价函数增加一个修正项的方法。该方法等效于增加很多虚拟点，由于这些虚拟点聚集在理想的主分量附近，因此把实际计算得到的主分量拉向理想的主分量。为了减少增加虚拟点带来的额外存储和计算开销，提出了用一个稀疏矩阵来涵盖虚拟点的方法。而且仿真试验和人脸识别实验均表明，该思想也可以改善 LDA 的性能。与本文思想比较接近的是最近邻特征线分类方法[8]，虽然二者分属子空间学习(数据降维)范畴和分类器范畴，但二者都使用了虚拟点，增强了算法的泛化能力。今后将进一步探究本方法的理论机理，并且在更多的数据集上测试它的有效性。

参 考 文 献

[1] 刘青山, 卢汉清, 马颂德. 综述人脸识别中的子空间方法[J], 自动化学报, 2003, 29(6): 900-911.
Liu Qing-shan, Lu Han-qing, and Ma Sang-de. A Survey: subspace analysis for face recognition [J]. *Acta Automatica Sinica*, 2003, 29(6): 900-911.

[2] Turk M and Pentalnd A. Eigenfaces for recognition [J]. *Journal of Cognitive Neuroscience*, 1991, 3(1): 71–86.

[3] Belhumeur P, Hespanha J, and Kriengman D. Eigenfaces vs. fisherfaces: recognition using class specific linear projections [J]. *IEEE Trans. on PAMI*, 1997, 19(7): 771–720.

[4] Bell A and Sejnowski T. An information-maximization approach to blind separation and blind deconvolution [J], *Neural Computation*, 1995, 7(6): 1129–1159.

[5] 孙焕良, 鲍玉斌, 于戈, 等..一种基于划分的孤立点检测算法 [J]. 软件学报, 2006, 17(5): 1009–1016.
Jiang Hang-liang, Bao Yu-bin, and Yu Ge, *et al.*. An algorithm based on partition for outlier detection [J]. *Journal of Software*, 2006, 17(5): 1009–1016.

[6] Fernado T and Michael J. Robust principal component analysis for computer vision [C], Proc. IEEE ICCV, Vancouver, Canada, 2001: 1478–1485.

[7] Xu L and Yuille A. Robust principal component analysis by self organizing rules based on statistical physics approach [J]. *IEEE Trans. on Neural Networks*, 1995, 6(1): 131–143.

[8] Li S, and Ju J. Face recognition using the nearest feature line method [J], *IEEE Trans. on Neural Networks*, 1999, 10(2): 439–443.

[9] Phillips P and Moon H, *et al.*. The FERET evaluation methodology for face recognition algorithms [J]. *IEEE Trans. on PAMI*, 2000, 22(10): 1090–1104

庞彦伟：　男，1976 年生， 博士，副教授，主要从事模式识别、机器学习及计算机视觉的研究.

刘政凯：　男，1940 年生，教授，主要从事遥感数字图像处理、视觉信息检索及模式识别的教学和研究工作.