基于样本加权的格拉斯曼平均算法

钟 倩,杨 丽,梁志贞
ZHONG Qian, YANG Li, LIANG Zhizhen

中国矿业大学 计算机科学与技术学院,江苏 徐州 221116
School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China

ZHONG Qian, YANG Li, LIANG Zhizhen. Grassmann average algorithm based on sample weight. Computer Engineering and Applications, 2017, 53(22):126-129.

Abstract: Grassmann average subspace corresponds to the leading principal component for Gaussian data and a scalable principal component analysis is provided. But it assumes that the contribution of each sample is determined by its length, which is sensitive to outliers. For this reason, this paper proposes a novel algorithm based on sample weight and Grassmann average by using local characteristics of data in unsupervised learning or class information in supervised learning. The experimental results on UCI data sets and ORL data set show that the algorithm is more robust and it has improved recognition rate by 1%~2%.

Key words: sample weight; Grassmann average; Principal Component Analysis(PCA); robust

摘 要:格拉斯曼平均子空间对应着高斯数据的主成分,解决了PCA的扩展性问题,但算法假定样本的贡献取决于样本的长度,这可能导致离群点对算法的干扰较强。为此,利用无监督学习数据的局部特性或监督学习中样本的类别信息建立样本的权重,从而提出一种基于样本加权的格拉斯曼平均的算法,在UCI数据集和ORL人脸数据库上的实验结果表明,新算法有好的鲁棒性并且其识别率比已有方法提高1%~2%。

关键词:样本加权;格拉斯曼平均;主成分分析法;鲁棒性
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1605-0416

1 引言

主成分分析(PCA)是统计分析法中的一种重要方法,在系统评价、故障诊断、质量管理和发展对策等许多方面都得到广泛的应用[1],常用来提取数据的主要特征,降低数据特征间的相关性。传统的PCA通过特征值分解来进行特征提取,当样本集不服从高斯分布时,分解出的主成分不相互独立[1],且离群样本极易影响PCA的运算精度。2008年Nojun Kwak提出了一种基于L1范式的最优化技术[2],不同于传统的基于L2范式的PCA算法,上述算法通过迭代算法得出投影向量,得到局部的最优解对,有效地抑制了离群点对运算结果的影响,增强了算法的鲁棒性。2015年Søren Hauberg等人发现在零均值数据集中,每个观测点跨越了单维子空间,与格拉斯曼流形中的一个点相对应,而平均子空间与高斯数据的主成分相对应,于是提出了基于格拉斯平均(Grassmann Average,GA)的PCA算法[3],增强了算法的扩展性,也在一定程度上减少了数据中离群点的影响,然而它的分析过程中没有引入合理的衡量样本重要性的权值度量,其对噪声或离群点的抗干扰能力还有待进一步的提高。

为了进一步增强基于GA的PCA算法的鲁棒性,提出一种基于样本加权的格拉斯曼平均算法。该算法通过格拉斯曼平均值的计算解决了降维处理和模型可扩展性的问题,同时通过计算类内当前样本与其他样本在空间中的相似性,在子空间的学习过程中挑选出一些重要的样本数据,并给予它们较高的权重,降低离群样本

基金项目:国家自然科学基金(No.61403394);中央高校基础研究基金(No.2015XKMS084);全国大学生创新训练计划(No.201510290004)。

作者简介:钟倩(1995—),女,研究领域为机器学习;杨丽(1995—),女,研究领域为机器学习,E-mail:1172179495@qq.com;梁志贞(1976—),男,博士,副教授,研究领域为图像处理、机器学习。

收稿日期:2016-05-29 **修回日期:**2016-12-12 **文章编号:**1002-8331(2017)22-0126-04
CNKI**网络优先出版:**2016-12-14, http://www.cnki.net/kcms/detail/11.2127.TP.20161214.1122.006.html

计算机工程与应用 www.ceaj.org

对算法性能的影响，为建立训练样本的判别子空间提供了有效的依据，提高了算法的识别率和对离群点的鲁棒性。

2　相关工作

将处理的样本矩阵定义为：

$$X=[x_1,x_2,\cdots,x_n]\in R^{d\times n} \tag{1}$$

其中，$x_i(i\in[1,n])$ 为一个列向量，表示每个数据点，各样本的维数为 d。Søren Hauberg等人发现平均子空间即对应为数据集的第一主成分。在接下来定义第一主成分对应的平均子空间。

格拉斯曼流型 $Gr(k,d)$ 是空间 R^d 的所有 k 维线性子空间的集合，样本 x_i 是 $Gr(1,d)$ 上的一个点。$Gr(1,d)$ 可以改写成单位球面的形式，即 $\frac{S^{d-1}}{\{\pm1\}}$[3]。

将 $Gr(1,d)$ 上的点可记作如下形式：

$$[x_i]=\{x_i,-x_i\} \tag{2}$$

其中 $[x_i]=[-x_i]$。对于给出的观测点，可将它的跨越子空间记作 $[u_i]=\pm x_i/\|x_i\|$，而它的符号是未知的。

对于零均值数据集 $x_{1:n}\subset R^d$，它的加权平均通常被定义成使加权平方和距离最小的量[3]：

$$[q]=\arg\min_{[v]\in Gr(1,d)}\sum_{i=1}^{n}w_n dist^2_{Gr(1,d)}([u_n],[v]) \tag{3}$$

其中 $dist_{Gr(1,d)}$ 表示 $Gr(1,d)$ 上的距离。将公式（3）中的 $dist_{Gr(1,d)}$ 改写成 $dist_{S^{d-1}}$ 的形式可以得到以的公式[7]：

$$dist^2_{Gr(1,d)}([u_1,u_2])=\min\{dist^2_{S^{D-1}}(v_1,v_2)\} \tag{4}$$

其中 $v_1\in[u_1],v_2\in[u_2]$，对于任意的满足式（2）的 $[q]\in Gr(1,d)$，其中任意 $q\in[q]\subset S^{d-1}$ 都必存在 $u_{1:n}\subset[u_{1:n}]\subset S^{d-1}$，因此 q 可以改写成用 $dist_{S^{d-1}}$ 表示的形式：

$$q=\arg\min_{v\in S^{d-1}}\sum_{i=1}^{n}\|x_i\|dist^2_{S^{d-1}}(u_i,v) \tag{5}$$

为了提高算法的处理速度，Søren Hauberg等人将球面距离 $dist_{S^{d-1}}(u_1,u_2)$ 定义为：

$$dist^2_{S^{d-1}}(\boldsymbol{u}_1,\boldsymbol{u}_2)=\|\boldsymbol{u}_1,\boldsymbol{u}_2\|^2/2=1-\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{u}_2 \tag{6}$$

对于任意 $u_{1:n}\subset S^{d-1}$ 的加权平均可表示成[8]：

$$q=\frac{\mu(w_{1:n},u_{1:n})}{\|\mu(w_{1:n},u_{1:n})\|} \tag{7}$$

其中的 μ 为欧几里德加权平均：

$$\mu(w_{1:n},u_{1:n})=(\sum_{i=1}^{n}w_i)^{-1}\sum_{i=1}^{n}w_iu_i \tag{8}$$

在将各观测量当作子空间前，需要对 $x_i(i\in[1,n])$ 进行归一化，因此选取权重：

$$w_i=\|x_i\| \tag{9}$$

由公式（6）和公式（7）可以推出：

$$dist^2_{Gr(1,d)}([\boldsymbol{u}_1],[\boldsymbol{u}_2])=1-|\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{u}_2| \tag{10}$$

$$[q]=\arg\max_{v\in S^{d-1}}\sum_{i=1}^{n}|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}| \tag{11}$$

GA算法每次迭代的时间复杂度为 $O(n,d)$。Søren Hauberg等人经实验证明，GA算法在主成分提取上比普通PCA算法的鲁棒性更高。上述算法只计算数据集的第一主成分，在这里规定数据集的各成分之间相互正交，通过去除样本集中已估计出的主成分，可进一步估计出剩余数据的下一主成分。

去除掉当前数据集中的第一主成分，将数据集更新成如下形式：

$$\hat{X}=X-\boldsymbol{q}^{\mathrm{T}}(\boldsymbol{q}X) \tag{12}$$

将 $\hat{X}$ 作为新的样本集代入上述GA算法就可以迭代求出其余的主成分。

3　基于样本加权的格拉斯曼平均算法

传统PCA算法没有对样本进行加权处理，离群样本与其他样本对算法产生的影响是相同的。要通过为离群点赋予很小的权值[4]，提高PCA的鲁棒性，降低异常值对算法运算精度的影响。采用高斯函数求出样本的权重矩阵 $\boldsymbol{W}_{n\times n}$，计算公式如下：

$$W_{ij}=\exp(-\|x_i-x_j\|^2/\tau) \tag{13}$$

其中 τ 是一个超参数。$D(i)$ 被定义为在 d 维空间中数据点 x_i 与其他数据点相似性的总和，而矩阵 $\boldsymbol{W}$ 为一对称矩阵，从而得到 $D(i)$ 的计算公式如下：

$$D(i)=\sum_{j=1}^{n}W_{ij}=\sum_{j=1}^{n}ji \tag{14}$$

其中 $\boldsymbol{D}$ 是一个 n 维向量，该向量的每一个成分都表示了数据点 x_i 的全局密度，而数据点的权重 $p(x_i)$ 则与对应点的密度成正比[5]：

$$p(x_i)=D(i)/\sum_{i=1}^{n}D(i) \tag{15}$$

虽然等式（14）和（15）提供了一种权重策略，但是它考虑了数据的所有可能数据点。还需要考虑数据的内在特性，即数据的分布是不平均的。即使是同类数据，它们之间的分布也可能比较分散，甚至也有不同类数据分布比较接近的可能，即位于边界上的数据点[6-7]。对于个 x_i 数据点，搜索它的 k 个最近邻点，表示为 $y_1,y_2,\cdots,y_k$。当类别信息提供时，搜索同类内的 k 个最近邻点。利用这些信息计算每个样本的局部权重，表示为：

$$\hat{P}_i(x_i)=\{\sum_{k=1}^{k}\exp(-\|x_i-y_j\|^2/\tau)\}/k \tag{16}$$

当类别信息没有提供时，并没有采用等式（15）去归化权重，直接利用了等式（16）。很显然 $\hat{P}(x_i)$ 位于[0，1]区间。但是对于分类问题，由于类别信息是已知的，所以考虑在等式（16）的基础上继续归化权重，使得每类样本的权重之和等于1。因为在原算法中，式（8）中的权重如式（9）所示，即选取了数据的模，仅完成了数据的归一化，从而得到等式（11）的优化模型，很显然等式（11）没

计算机工程与应用 www.cejo.org

有对各数据的离群点进行判别和抑制，致使这种子空间的投影没有去除掉异常样本的影响。为了找到上述的格拉斯曼平均子空间，在权重的基础上，提出了下面的优化模型，表示为：

$$[q]=\arg\max_{v\in S^{d-1}}\sum_{i=1}^{n}\hat{p}_i(\boldsymbol{x}_i)|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{v}| \tag{17}$$

对于模型(17)提出了下面算法。

算法1　加权的格拉斯曼平均

(1)对于每个数据点，计算每个数据点的 k 最近邻点，利用等式(16)计算每个样本的权重。

(2) $q(0)$ 初始化为随机的 d 维列向量，令 $q(0)=q(0)/\|q(0)\|$，并将迭代次数 j 初始化为1。

(3)任意样本 x_i 带上权重 $p(x_i)$ 后取符号位 $\partial_i(j)=\mathrm{sign}(p(x_i)\boldsymbol{x}_i^{\mathrm{T}}\cdot q(j-1))$。

(4) $q(j)=\dfrac{(\sum_{i=1}^{n}\partial_i(j-1)\cdot p(x_i)\cdot w_i\frac{x_i}{\|x_i\|})}{\left\|(\sum_{i=1}^{n}\partial_i(j-1)\cdot p(x_i)\cdot w_i\frac{x_i}{\|x_i\|})\right\|}$，$j=j+1$，其中 $w_i=\|x_i\|$。

(5)若 $\|q(j)-q(j-1)\|\geqslant\varepsilon$ 且小于最大迭代次数 M，执行第(3)步，否则，提取成分 $q(j)$ 作为当前数据的第一主成分。

显然算法1仅仅取得了一个投影向量，除去计算权重时的复杂性，算法1的计算复杂性是 $O(ndM)$，其中 M 是最大的迭代次数。同时，算法的尺度性较好。在实际情况下，往往利用 $\bar{x}_i=x_i-\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}x_i$ 处理每个数据点来得到新的数据点，然后在新的数据点的基础上得到另外一个投影向量。当一个投影向量得到后，$\|\bar{x}_i\|^2$ 可以表示第 i 个数据点的重建误差，根据 $\|\bar{x}_i\|^2$ 的大小，就可以找到样本集中包含的离群点[8]。这样可以根据重建误差对每个数据点排序，去除重建误差比较大的数据点，进而尽可能地去除掉离群点[9]。当然可以利用数据点的重建误差重新计算权重，从而提高计算权重的鲁棒性。

4　实验结果

4.1　人脸重建

人脸重建实验用到的数据库是ORL人脸库，有40人，总共400张照片。为了提高速度，把分辨率为92×112像素的图像减小到56×46像素，并在其中的10幅图像中引入随机噪声。对于每张照片，每个像素都可被视为一个输入变量，从而组成了维度为2 576的空间[10]。

对于得到的2 576×400的数据集，取特征向量个数20，并分别采用PCA、GA、SWGA三种方法进行实验。实验结果如图1所示。图1第一列为带有随机噪声的图像，第二列为PCA的重建结果，第三列为GA的重建结果，第四列为本文提出算法的重建结果。从图1可以看出，本文的方法对处理随机噪声有很好的结果。

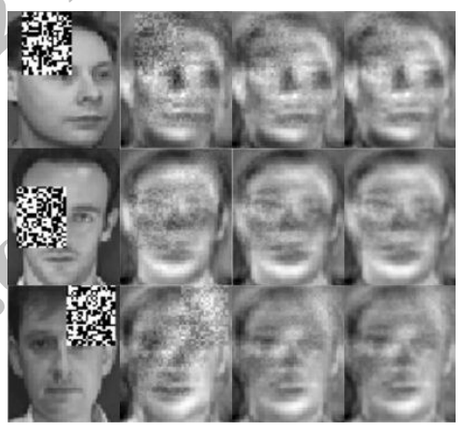

图1　原图像和各种算法的重建图像

通过上述实验，已经实现了通过一部分的特征完成人脸的重建，然后再利用原始的未被遮挡的图像计算平均重建误差 $\bar{e}(m)$[11]：

$$\bar{e}(m)=\frac{1}{n}\sum_{i=1}^{n}\left\|x_i^{\mathrm{org}}-\sum_{k=1}^{m}\boldsymbol{w}_k\boldsymbol{w}_k^{\mathrm{T}}x_i\right\|_2 \tag{18}$$

其中 m 为提取的特征向量的个数，n 是样本总数，在本实验中 m 取20，n 取400，x_i^{org} 和 x_i 分别是第 i 个原始的未被遮挡的图像和相对应的第 i 个被遮挡的图像。图2表示了GA和本文方法的重建曲线。由图可见，改进的基于样本加权的GA算法的平均重建误差小于GA算法的平均重建误差。

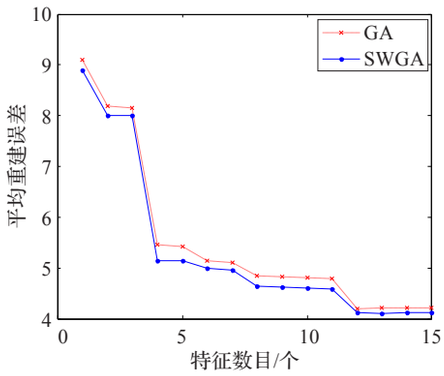

图2　ORL数据集的平均重建误差

检查算法的收敛性，其收敛性结果如图3所示。即随着迭代次数的增大，目标值是几乎稳定的。由图可知，它的收敛速度较快。

4.2　UCI数据分类

实验使用UCI数据学习库中的6个数据集iris、wine、Isonosphere、heart、sona、glass对PCA、GA、SWGA进行性能比较。实验中，SWGA的初始化投影样本矩阵的最大2范数投影，分别考虑了每一类内样本的权重。将样本权重中的 τ 限制在一定的区间，并选择最好的参数。对每一个数据集，运用5倍交叉验证进行实验，并

计算机工程与应用 www.ceaj.com

表1　三种算法的最优分类率及其对应的特征数

数据集	PCA		GA		SWGA	
	最优分类率/%	特征数/个	最优分类率/%	特征数/个	最优分类率/%	特征数/个
iris	96.00	1	96.00	1	96.67	1
wine	97.20	2	97.71	4	98.86	3
Isonospher	88.32	9	88.33	9	89.19	10
heart	79.26	12	79.26	13	79.63	12
sona	88.07	15	87.09	19	88.47	12
glass	67.00	4	67.00	4	68.43	4

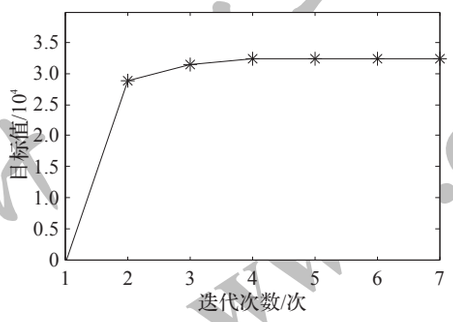

图3　SWGA的收敛性结果

选用1—最近邻分类器进行分类。在进行训练之前，训练集中的每个输入变量都被标准化成均值为0、方差为1的变量[12]。测试集中的变量也运用训练集的均值和方差进行标准化。

表1列出了实验结果，从表中可以发现SWGA性能在这6个数据集中均优于其他两种方法。因此，在很多情况下，SWGA的性能高于PCA和GA，对应的所需提取的特征数也比其他两种方法少，这说明增加样本的类内权重可以较大地提高分类准确度。

5　结束语

PCA[1]是一种进行数据分析和降唯处理的有效方法，然而在数据处理应用方面，它易受离群点的影响，而格拉斯曼平均（GA）为子空间的估算提供了一种简单高效且可扩展的途径。在此基础上，引入了样本加权，为每个数据点赋予不同的权重，成功实现了数据的判别分析和对离群点的抑制，与其他特征提取方法和分类方法相比，本文提出的算法具有更高的识别率[13]，且在人脸重构[14]方面具有更优的重建误差。

参考文献：

[1] 王松，夏绍玮．一种鲁棒主成分分析（PCA）算法[J]．系统工程理论与实践，1998，18（1）：9-13.

[2] Kwak N.Principal component analysis based on ll-norm maximization[J].IEEE Transactions on Pattern Analysis and Machine Intelligence，2008，30（9）：1672-1680.

[3] Hauberg S，Feragen A，Enficiaud R，et al.Scalable robust principal component analysis using grassmann averages[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015.

[4] 刘兵，夏士雄，周勇，等．基于样本加权的可能性模糊聚类算法[J]．电子学报，2012，40（2）：371-375.

[5] Ren Chuanxian，Dai Daoqing，He Xiaofei，et al.Sample weighting：an inherent approach for outlier suppressing discriminant analysis[J].IEEE Trans on Knowl Data Eng，2015，27（11）：3070-3083.

[6] Lee J M.Introduction to smooth manifolds[M].[S.l.]：Springer，2002.

[7] Bridson M R，Haefliger A.Metric spaces of non-positive curvature[M].[S.l.]：Springer Verlag，1999.

[8] Mardia K V，Jupp P E.Directional statistics[M].[S.l.]：John Wiley & Sons，1999.

[9] 李勇，梁志贞，夏士雄．基于核函数的PCA-L1算法[J]．计算机工程，2011，37（22）：174-175.

[10] 刘宁．Lp范数约束的广义主成分分析在人脸识别中的应用[D]．徐州：中国矿业大学，2015.

[11] 杨开睿．基于重建误差的样本加权特征提取算法研究[D]．徐州：中国矿业大学，2014.

[12] 胡永生，冀小平．基于阈值的图像分割方法的研究[J]．科技情报开发与经济，2006，17（2）.

[13] 孙即祥．现代模式识别[M]．长沙：国防科技大学出版社，2002.

[14] 章毓晋．图像处理和分析[M]．北京：清华大学出版社，2003.

计算机工程与应用 www.ceaj.org