ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.8, August 2009, pp.2153−2159
doi: 10.3724/SP.J.1001.2009.03391

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种解决大规模数据集问题的核主成分分析算法*

史卫亚+, 郭跃飞, 薛向阳

(复旦大学 计算机科学与技术系,上海 200433)

Efficient Kernel Principal Component Analysis Algorithm for Large-Scale Data Set

SHI Wei-Ya+, GUO Yue-Fei, XUE Xiang-Yang

(Department of Computer Science and Technology, Fudan University, Shanghai 200433, China)

+ Corresponding author: E-mail: wyshi@fudan.edu.cn, http://www.fudan.edu.cn

Shi WY, Guo YF, Xue XY. Efficient kernel principal component analysis algorithm for large-scale data set. *Journal of Software*, 2009,20(8):2153−2159. http://www.jos.org.cn/1000-9825/3391.htm

Abstract: A covariance-free method of computing kernel principal components is proposed. First, a matrix, called Gram-power matrix, is constructed with the original Gram matrix. It is proven by the theorem of linear algebra that the eigenvectors of newly constructed matrix are the same as those of the Gram matrix. Therefore, each column of the Gram matrix can be treated as the input sample for the iterative algorithm. Thus, the kernel principle components can be iteratively computed without the eigen-decomposition. The space complexity of the proposed method is only $O(m)$, and the time complexity is reduced to $O(pkm)$. The effectiveness of the proposed method is validated by experimental results. More importantly, it still can be used even if traditional eigen-decomposition technique cannot be applied when faced with the extremely large-scale data set.

Key words: KPCA (kernel principal component analysis); Gram matrix; large-scale data set; covariance-free; eigen-decomposition

摘 要: 提出一种大规模数据集求解核主成分的计算方法.首先使用Gram矩阵生成一个Gram-power矩阵,根据线性代数的理论可知,新形成的矩阵和原先的Gram矩阵具有相同的特征向量.因此,可以把Gram矩阵的每一列看成核空间迭代算法的输入样本,这样,无须使用特征分解即可迭代地计算出核主成分.该算法的空间复杂度只有 $O(m)$;在大规模数据集的情况下,时间复杂度也降低为 $O(pkm)$.实验结果表明了所提出算法的有效性.更为重要的是,在大规模数据集的情况下,当传统的特征分解技术无法使用时,该方法仍然可以提取非线性特征.

关键词: 核主成分分析;Gram 矩阵;大规模数据集;协方差无关;特征分解

中图法分类号: TP181 **文献标识码**: A

主成分分析是一种用于特征提取和维度减少的经典方法[1].该方法一般使用有较大方差的主成分而忽略较少的重要成分.尽管主成分分析成功用于维度减少,但是在非线性数据分布情况下该方法不够理想,通常使用

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z176 (国家高技术研究发展计划(863)); the Key Project of the Ministry of Education of China under Grant No.104075 (国家教育部科学技术研究重点项目); the National Key Technology R&D Program of China under Grant No.2007BAH09B03 (国家科技支撑计划)

Received 2008-04-10; Accepted 2008-06-03

核方法[2]将其推广到核空间使用.其主要思想是,把数据映射到高维的特征空间,在映射的特征空间中,可以使用传统的线性算法实现非线性特征的特征提取.其中,不必知道映射函数而采用核技巧就可以计算数据之间的内积.所提取的非线性特征被用于许多复杂的应用中,例如人脸识别、图像压缩等.

在标准的核主成分计算过程中,需要储存所有数据形成的 Gram 矩阵[3],其空间复杂度为 $O(m^2)$.其中,m 表示样本数.另外,对 Gram 矩阵进行特征分解的时间复杂度为 $O(m^3)$.因为有限的内存容量,在大规模数据集情况下,该方法是不可行的.Zheng 等人[4]提出把数据集分成若干小的数据集,然后分别处理.贪婪的核主成分分析方法[5]通过采样训练集而近似求解特征向量.但是,这些方法仍然存在较大的储存问题.在文献[6]中,通过把传统的广义 Hebbian 算法扩展到核空间迭代的求解核主成分,可以避免储存问题,但是这种方法的收敛性不能保证.

本文提出了一种大规模数据集情况下有效的求解核主成分的方法.主要思想是,利用线性代数中对称矩阵的性质,首先使用初始的 Gram 矩阵创建一个新的 Gram-power 矩阵.因为新构成的矩阵和原先的矩阵具有相同的特征向量,所以我们可以把 Gram 矩阵的每一列看成核空间迭代算法[7]的输入样本.通过若干迭代后,可以很容易地求出核主成分.该方法不需要事先存储 Gram 矩阵,空间复杂度从 $O(m^2)$减少到 $O(m)$.另外,算法的快速收敛性已经得到证明[8].更为重要的是,在处理非常大规模的数据集情况下,传统的特征分解技术无法使用,而我们所提出的方法仍然可以较好地提取非线性特征.通过在人工合成的数据集以及真实的数据上进行实验,充分证明了所提出算法的有效性.

本文第 1 节对核主成分的实现过程进行回顾,同时把常用的一些迭代算法进行比较.第 2 节详细描述本文所提出的方法.第 3 节给出实验结果,证明所提出算法的有效性.第 4 节进行总结.

1 核主成分分析的回顾和迭代算法分析

1.1 核主成分分析的回顾

假定 $X=(x_1,x_2,\ldots,x_m)$是输入空间的数据集,其中,$x_i,i=1,2,\ldots,m$ 是 d 维向量,m 是数据集中的总样本数.存在一个函数ϕ把数据映射到高维(甚至无限维)的希尔伯特空间.

$$\begin{cases}\phi: & \Re^d \to \mathbb{F} \\ & x_i \mapsto \phi(x_i)\end{cases} \tag{1}$$

使用这个映射函数ϕ,我们可以得到特征空间的数据集$\Phi(X)=(\phi(x_1),\phi(x_2),\ldots,\phi(x_m))$.一个核函数被用于计算这些映射数据样本之间的内积,这个核函数定义为$\kappa(.,.)=\kappa(x_i,x_j)=\phi(x_i)^T\phi(x_j)$.这样,在映射的特征空间协方差定义为

$$C=\frac{1}{m}\sum_{i=1}^{m}\phi(x_i)\phi(x_i)^T \tag{2}$$

它符合特征方程:

$$Cv=\lambda v \tag{3}$$

其中,v和λ分别对应协方差矩阵的特征向量和特征值.根据定义[2],特征向量v可以用全部数据集的$\Phi(X)=(\phi(x_1),\phi(x_2),\ldots,\phi(x_m))$张程表示为

$$v=\sum_{i=1}^{m}\alpha_i\phi(x_i) \tag{4}$$

将公式(3)和公式(4)代到公式(2)中,可以推导出下面的公式(5):

$$K\alpha=m\lambda\alpha \tag{5}$$

其中,α是张程系数;K 是 Gram 矩阵,定义为 $K=\Phi(X)^T\Phi(X)$.该矩阵的元素为 $k_{ij}=k(x_i,x_j)$.理论证明[9],Gram 矩阵是半正定的.为了计算核主成分,一般都是对 Gram 矩阵采用特征分解的方法.这样得到特征向量α后,可以使用公式(4)计算核主成分 v.对于任意一个测试样本 x,其非线性特征为

$$(v,\phi(x))=\sum_{i=1}^{m}\alpha_i(\phi(x_i)\cdot\phi(x))=\sum_{i=1}^{m}\alpha_i k(x_i,x) \tag{6}$$

上面的推导中是假定所有数据都具有零均值的,如果不是,则可以得到 $\tilde{K}=K-1_m K-K1_m+1_m K1_m$. 其中,$1_m=\left(\frac{1}{m}\right)_{m\times m}$ [3].

1.2 迭代算法分析

因为传统的特征分解方法需要非常大的内存.为了克服这个不便,一些计算主成分的迭代方法被提出来,例如 GHA[10],APEX[11],但是这些方法一般收敛速度较慢.Weng 等人[7]利用统计学上的效能估计概念提出了一种增量的协方差无关的方法.该方法不是像传统方法一样特征分解协方差矩阵$(x_1,\ldots,x_m)(x_1,\ldots,x_m)^T$,而是循环地输入样本向量 x_i,与其他迭代方法相比,收敛速度快而且计算复杂度很低.因此,我们选用这种方法作为所提出算法的迭代工具.

2 我们提出的核主成分分析算法

一般情况下,映射函数是不知道的.因此,特征空间的样本向量没法明确表示.这样,在核空间就没有办法使用上述迭代方法计算核主成分.为了解决这个问题并给出我们的方法,首先给出一个线性代数的定理[12,13].

定理. 矩阵 H 和 H^2 具有相同的特征向量和不同的特征值.

证明:假定ω和λ分别是矩阵的特征向量和特征值,

$$H\omega=\lambda\omega \tag{7}$$

$$H^2\omega=HH\omega=\lambda H\omega=\lambda^2\omega \tag{8}$$

因此,矩阵 H^2 的特征向量和特征值即为ω和λ^2. □

2.1 我们提出的方法

因为 Gram 矩阵 K 是半正定的,可以得到 $K=K^T$.因此,可以构造一个矩阵 Gram-power,定义为 $G=KK^T=K^2$.根据前面的定理,新构造的矩阵 G 的特征向量$\{U_G\}$与 Gram 矩阵 K 的特征向量$\{U_K\}$相同,但有不同的特征值λ_G和λ_K ($\lambda_G=(\lambda_K)^2/m$).

$$G=KK^T=\begin{pmatrix}\kappa_{11}&\cdots&\kappa_{1m}\\ \vdots&\ddots&\vdots\\ \kappa_{m1}&\cdots&\kappa_{mm}\end{pmatrix}\begin{pmatrix}\kappa_{11}&\cdots&\kappa_{1m}\\ \vdots&\ddots&\vdots\\ \kappa_{m1}&\cdots&\kappa_{mm}\end{pmatrix}^T=(K(x_1),\ldots,K(x_m))(K(x_1),\ldots,K(x_m))^T=\sum_{i=1}^{m}K(x_i)K(x_i)^T \tag{9}$$

其中,$K(x_i)=(\kappa_{i1},\kappa_{i2},\ldots,\kappa_{mi})^T$.因此,我们可以把矩阵 K 的每一列 $K(x_i)$作为迭代算法[7]的输入样本.这样,经过一些迭代后就可以迅速得到矩阵 G 的特征向量,而不用像传统方法那样对 Gram 矩阵 K 特征分解.因为在实际计算中,我们每次只需计算 $K(x_i)$,从而有效地解决了大样本需要存储 Gram 矩阵的问题.

下面我们具体给出求解矩阵 G 的特征向量的算法过程.根据定义,新构造的 Gram-power 矩阵 G 的特征向量$\{U_G\}$和特征值λ_G符合下面的公式(10):

$$\omega(n)=\lambda_G U_G=GU_G \tag{10}$$

其中,$\omega(n)$是特征向量在第 n 时刻估计.在得到特征向量后,可以很容易地计算特征向量 $U_G=\omega/\|\omega\|$和特征值$\lambda_G=\|\omega\|$.因为现在的输入样本为$\{K(x_1),\ldots,K(x_m)\}$,因此可以依次将样本 $K(x_i)$输入到迭代算法中.这样,在 n 时刻,第 i 阶核主成分的估计$\omega_i(n)$可以表示为

$$\begin{aligned}\omega_i(n)&=GU_G\\ &=\frac{1}{n}\sum_{t=1}^{n}K_i(x_t)K_i^T(x_t)\frac{\omega_i(t-1)}{\|\omega_i(t-1)\|}\\ &=\frac{1}{n}\sum_{t=1}^{n-1}K_i(x_t)K_i^T(x_t)\frac{\omega_i(t-1)}{\|\omega_i(t-1)\|}+\frac{1}{n}K_i(x_n)K_i^T(x_n)\frac{\omega_i(n-1)}{\|\omega_i(n-1)\|}\\ &=\frac{n-1}{n}\omega_i(n-1)+\frac{1}{n}K_i(x_n)K_i^T(x_n)\frac{\omega_i(n-1)}{\|\omega_i(n-1)\|}\end{aligned} \tag{11}$$

其中,$K_i(x_t)$是计算第 i 阶核主成分时在 t 时刻的输入样本.其他高阶特征向量可以用残留的数据向量计算.而残留的数据向量是用最初数据减去其在低阶特征向量的投影($K_1(x_n)=K(x_n)$).

$$K_{i+1}(x_n)=K_i(x_n)-K_i(x_n)^T\frac{\omega_i(n)}{\|\omega_i(n)\|}\frac{\omega_i(n)}{\|\omega_i(n)\|} \tag{12}$$

这样,通过迭代计算就可以得到需要的特征向量和特征值.

算法概括如下:

1) 使用前 k 个样本初始化前 k 阶特征向量.
2) *Iteration*=1:p 进行下面的计算.
3) t=1:m 进行下面的计算.
4) i=1:k 进行下面的计算.
5) 对每个输入样本 x_t,计算相应的 $K_i(x_t)$,作为输入向量.
6) 使用公式(11)和公式(12)计算前 k 阶主成分.
7) 转到步骤 4).
8) 转到步骤 3).
9) 转到步骤 2).
10) 输出特征向量$\{U_G\}$和特征值λ_G.

2.2 算法的复杂度分析

整个计算过程中,不需要特征分解 Gram 矩阵,因为这样空间复杂度是 $O(m^2)$,时间复杂度是 $O(m^3)$.相反,在每一步只处理样本 $K(x_i)$,其复杂度只有 $O(m)$,这样经过一些迭代后就可以得到核主成分.因为算法需要经过一些迭代,总的时间复杂度为 $O(pkm)$,其中,p,k,m 分别为迭代次数、特征向量数目和总的样本数.在大样本数据集情况下,迭代次数和提取向量的数目远小于样本数目,因此所提出算法的时间复杂度也大为降低.

3 实验结果和讨论

为了验证所提出算法的有效性,我们使用标准的 KPCA(kernel principal component analysis)算法和我们提出的方法在一个具有 3 聚类、2 维的简单问题上进行实验.除此之外,还使用 USPS(The United States Postal Service)数据集在图像降噪和分类上验证提出算法的可行性.实验中如果没有明确说明,核函数使用高斯核 $k(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$($\sigma$是核函数的宽度参数,一般需要通过交叉验证而确定).

图 1 给出了实验结果,它表征了两种方法提取的前 3 个核主成分的轮廓图,灰度值代表测试样本投影的非线性特征值.所得到的对应特征值在表 1 中列出.其中,λ_G 和 λ_K 分别是用我们提出的方法和标准的方法得到的特征值.从结果可以清晰地看到,我们的方法可以得到与标准方法类似的结果.另外,我们还比较两种方法产生的前 3 个特征向量的平均内积随迭代次数的变化,结果如图 2 所示.从图中可以观察到,经过一些迭代后,所提出的方法得到的特征向量非常好地收敛到标准方法产生的特征向量.

模拟问题:2 维的数据集含有 3 个聚类,每个聚类有 30 个样本,都具有高斯分布,其均值分别为[−0.5,−0.2],[0,0.6],[0.5,0],标准方差为 0.1,核参数σ为 0.1.

USPS 数据集:接下来,我们在一些标准数据集上测试提出的方法.USPS 数据集是 256 维的字符向量,含有 7 291 个训练样本和 2 007 个测试样本.在这个实验中,核参数σ设置为$\sigma=d\cdot c$.其中,d 是数据向量的维度,c 设置为数据平均方差的 2 倍.

首先使用提取的特征进行降噪实验.随机取 3 000 个训练样本用标准的 KPCA 和我们提出的方法进行训练,提取前 64 个主成分.测试样本用均值为 0、方差为 0.5 的高斯噪声迭加.利用提取的非线性特征进行重组图像,实验结果如图 3 所示.从图中可以看到,两种方法重组的图像都获得了很好的降噪效果.

为了进一步验证我们所提出的方法,我们使用所有的训练样本来提取非线性特征,实验中使用多项式核

$\kappa(x,y)=(x^Ty)^d$(d 为多项式度).在这种情况下,Gram 矩阵的大小为 7291×7291.如果用标准的 KPCA 方法,则因为样本数目太大,无法进行计算.但是,我们提出的方法仍然可以提取核主成分.测试样本在提取的前 64 个主成分上进行投影,然后用最近邻方法进行分类.表 2 给出了分类的误差率.数据表明,我们的方法在大样本情况下达到了较好的分类效果.

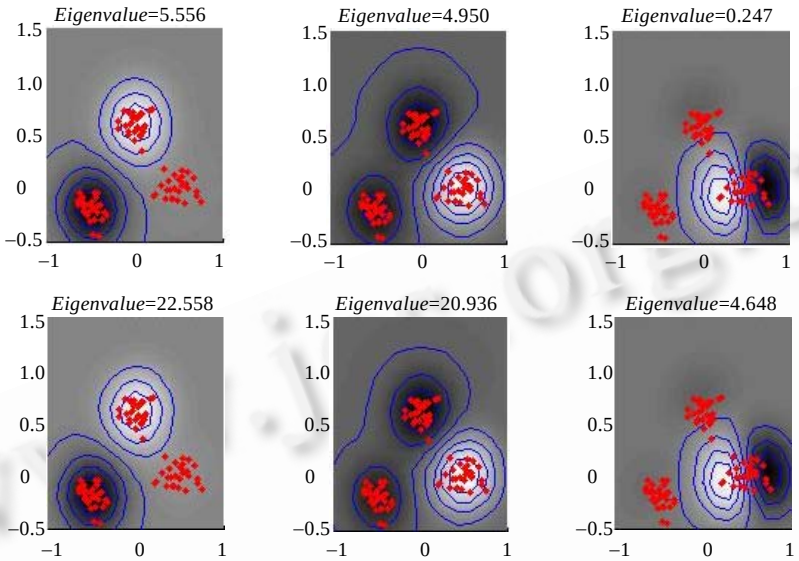

Fig.1 Contour image of first 3 principal components obtained from the proposed method (the top row) and standard KPCA (the bottom row)

图 1 使用我们提出的方法(上行)与标准 KPCA 方法(下行)产生的前 3 个主成分的轮廓图

Table 1 Experimental results of the first 3 eigenvalues using two methods ($\lambda_G=(\lambda_K)^2/m$)

表 1 使用两种方法得到的前 3 个特征值($\lambda_G=(\lambda_K)^2/m$)

	The first 3 eigenvalues		
	1	2	3
λ_G	5.556	4.95	0.247
λ_K	22.558	20.936	4.648

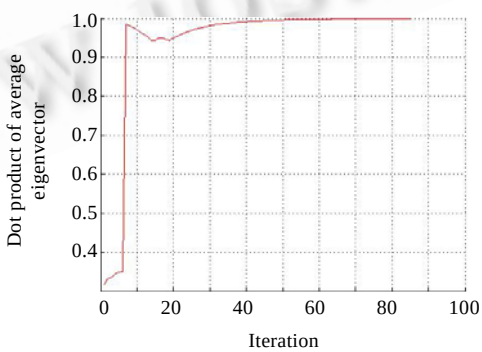

Fig.2 Average dot product of first 3 principal components obtained from standard KPCA and the proposed method

图 2 使用标准 KPCA 方法和我们提出的方法所得到的前 3 个特征向量的平均内积

Fig.3 De-Nosing results obtained by different methods (1st row: original image; 2nd row: noisy image; 3rd row: de-noising result by batch KPCA; 4th row: de-noising result by proposed method)

图 3 使用两种方法得到的图像降噪结果(第 1 行:原图像;第 2 行:加噪图像;第 3 行:使用标准 KPCA 方法;第 4 行:使用本方法经过降噪处理后得到的图像)

Table 2 Error rate of 2007 testing sample of USPS data set using the proposed method (%)

表 2 使用本文提出的方法在 USPS 数据集上 2007 个测试样本的误差率(%)

Number of extracting principal components	Error rate				
	d=2	d=3	d=4	d=5	d=6
64	5.88	6.13	6.57	7.06	7.25

4 结束语

本文提出了一种大规模数据集情况下提取核主成分的有效方法.该方法首先利用 Gram 矩阵构造一个 Gram-power 矩阵,因为两个矩阵有相同的特征向量,因此无须像传统方法一样特征分解 Gram 矩阵,而是用 Gram 矩阵的每一列作为每一步迭代算法的输入样本.这样,可以有效解决传统方法在大规模数据集下无法使用的问题.该方法空间复杂度减少为 $O(m)$,时间复杂度也降低为 $O(pkm)$.更重要的是,当样本数目非常大时,本文所提出的方法仍然可以提取非线性特征.

References:

[1] Kirby Y, Sirovich L. Application of the Karhunen-Loeve procedure for the characterization of human faces. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1990,12(1):103−108.

[2] Scholkopf B, Smola A. Learning with Kernel: Support Vector Machines, Regularization, Optimization and Beyond. London: MIT Press, 2002. 25−55.

[3] Scholkopf B, Smola A, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998, 10(5):1299−1319.

[4] Zheng W, Zou C, Zhao L. An improved algorithm for kernel principal components analysis. Neural Processing Letters, 2005,22(1): 49−56.

[5] France V, Hlavac V. Greedy algorithm for a training set reduction in the kernel methods. In: Petkov N, Westenberg MA, eds. Proc. of the Int'l Conf. on Computer Analysis of Images and Patterns. Berlin, Heidelberg: Springer-Verlag, 2003. 426−433.

[6] Kim KI, Franz MO, Scholkopf B. Iterative kernel component analysis for image modeling. IEEE Trans. on Pattern Analysis Machine Intelligent, 2005,27(9):1351−1366.

[7] Weng J, Zhang Y, Huang WS. Candid covariance-free incremental principal component analysis. IEEE Trans. on Pattern Analysis Machine Intelligence, 2003,25(8):1034−1040.

[8] Zhang Y, Weng J. Covergence analysis of complementary candid incremental principal component analysis. Technical Report, MSU-CSE-01-23, East Lansing: Michigan State University, 2001.

[9] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Endland: Cambridge University Press, 2004.

[10] Sander TD. Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Network, 1989,2(6): 459–473.

[11] Kung SY, Diamantaras KI. A neural network learning algorithm for adaptive principal component extraction (APEX). In: Ludeman L, ed. Proc. of the IEEE Conf. on Acoustics, Speech, and Signal, Vol.2. Albuquerque: IEEE Computer Society, 1990. 861–864.

[12] Golub GH, van Loan CF. Matrix Computation. 3rd ed., Baltimore: The Johns Hopkins University Press, 1996. 49–85.

[13] Strang G. Introduction to Linear Algebra. 2nd ed., Wellesley: Wellesley-Cambridge Press, 1998. 245–282.

史卫亚(1973－),男,河南周口人,博士,工程师,主要研究领域为机器学习,神经网络.

郭跃飞(1964－),男,博士,副教授,CCF 会员,主要研究领域为机器学习,神经网络.

薛向阳(1968－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为多媒体信息处理,人工智能.