基于聚类的核主成分分析在特征提取中的应用*)

王和勇[1] 姚正安[2] 李 磊[1]

(中山大学软件研究所 广州510275)[1] (中山大学数学与计算科学学院 广州510275)[2]

摘 要 本文分析了一般主成分分析在处理非线性问题上的不足,阐述了核主成分分析方法及其计算速度的缺陷,提出了基于聚类的核主成分分析方法。试验结果显示:基于聚类的核主成分分析方法具有好的特征提取性能,相比核主成分分析大大提高了特征提取的速度。

关键词 图像检索,KPCA,特征提取

The Application of Feature Extraction on Using Kernel Principal Component Analysis Based on Clustering

WANG He-Yong[1] YAO Zheng-An[2] LI Lei[1]

(Institute of Software, Zhongshan University, Guangzhou 510275)[1]

(College of Mathematics and Computer Science, Zhongshan University, Guangzhou 510275)[2]

Abstract This paper points out the drawbacks of the general principal component analysis(PCA) when it is used to solve nonlinear problem firstly. The kernel principal component analysis(KPCA) and its drawbacks on computing are explained secondly. KPCA based on clustering is introduced in the end. The research result shows that the KPCA based on clustering has excellent performance of feature extraction.

Keywords Image retrieval, KPCA, KPCA based on clustering, Feature extraction

1 引言

主成分分析(PCA)是最为常用的特征提取方法[1],被广泛应用到各领域,如图像处理、综合评价、语音识别、故障诊断等。通过对原始数据的加工处理,简化问题处理的难度并提高数据信息的信噪比,以改善抗干扰能力。然而,从本质上讲PCA是一种线性映射算法,在处理非线性问题时,往往不能取得好的效果。为此,大量文献提到非线性的核子空间(KNS)[2]方法。在KNS中,当输入空间的向量为非线性时,使用核主成分分析(KPCA)[3]方法得到主成分,此主成分的获得不是在原来的空间,而是通过变换后的高维空间获得。KPCA计算使用核矩阵K,核矩阵的维数等于样本点的数量,明显得到的矩阵是大矩阵。

在这篇文章里,我们采用一种计算优化的方法来解决问题,而不是直接采用原来的数据点定义核矩阵计算主成分,我们提出了基于聚类的代表点的方法来计算核矩阵。因为聚类具有满足全局分布结构的特性,使用聚类的方法,对无效分类数据点使用排除的方法来缩减核矩阵的阶数,而没有明显降低最后显示的效果。试验显示本文介绍的方法在几乎没有影响检索效果的前提下可以大大缩减KPCA的计算时间。

2 基于聚类的核主成分分析

下面介绍PCA、KPCA、聚类的方法和基于聚类的KPCA方法。

2.1 PCA方法

设$x_i\in R^p(i=1,2,\cdots,n)$为样本点,$x_i=\{x_{i1},x_{i2},\cdots,x_{ip}\}$,$x_i^*$是$x_i$的标准化后的$p$维向量。计算矩阵$R=XX^T$,$X$是$x_i^*(i=1,2,\cdots,n)$组成的$p\times n$矩阵,$R$是$p\times p$矩阵,根据

$$RU_i=\lambda U_i \tag{1}$$

求矩阵的特征值和特征值对应的特征向量,按一定的标准(前几个特征值占总特征值的比例≥85%),取前$m(m<p)$个特征值和对应的标准化后的特征向量$\alpha_1,\alpha_2,\cdots,\alpha_m(\alpha_i\in R^p,i=1,2,\cdots,m)$,此时计算每个样本$x_i$分别在$\alpha_k(k=1,2,\cdots,m)$上的投影:$g_k(x)=(\alpha_k\cdot x_i),(k=1,2,\cdots,m)$。将所有的投影值形成一个矢量$g(x)=(g_1(x),g_2(x),\cdots,g_m(x))$作为样本$x_i$的特征值,就可以把$x_i$原来的$p$维降为$m$维。

2.2 KPCA方法

PCA为特征向量由高维降为低维提供了一个很好的方法,但现实的特征向量之间往往是线性不可分的,为此人们提出了KPCA方法。KPCA的基本思想可以概括为:首先通过非线性变换将输入空间变换到高维空间,然后在高维空间就线性可分。而这种非线性变换是通过定义适当的内积函数实现。

设$x_i\in R^p(i=1,2,\cdots,n)$为样本点,把输入空间$R^p$通过非线性变换$\phi$映射到特征空间$F$,即:

$$\phi:R^p\to F \tag{2}$$

F中的样本点记为$\phi(x_i),(i=1,2,\cdots,n)$。$F$空间中样本的协方差矩阵为:

$$C=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi(x_i)\phi(x_j)^T \tag{3}$$

根据

$$Cv=\lambda v \tag{4}$$

求C的特征值λ和λ所对应的特征向量$V\in F\backslash\{0\}$,C的特征值均为非负。由(3)式看到,计算C需要知道$\phi(x_i)$和ϕ

*)基金项目:国家自然科学基金项目(10171113)。王和勇 博士生,主要研究领域为模式识别、数据分析、计算机通信;姚正安 教授,博士生导师,主要研究领域为计算机通信、数据分析、模式识别;李 磊 教授,博士生导师,主要研究领域为数据库与知识库、计算机通信、数据分析、模式识别。

(x_i)，而 ϕ 是未知的，所以无法计算 C。不失一般性，设 C 的特征值为 $0\leqslant\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$，对应的特征向量分别记为 $\upsilon_1,\upsilon_2,\cdots,\upsilon_n$。另外，$\upsilon_1,\upsilon_2,\cdots,\upsilon_n$ 可由 F 空间中的样本 $\phi(x_i)$ 张成。

记

$$\upsilon_r=\sum_{i=1}^{n}\alpha_r^i\phi(x_i) \tag{5}$$

考虑等式

$$\phi(x_i)\cdot C\upsilon=\lambda(\phi(x_i)\cdot\upsilon) \tag{6}$$

将(3)(5)代入(6)并令

$$K=(k_{ij})_{n\times n}=(\phi(x_i)\cdot\phi(x_j))(i,j=1,2,\cdots,n)$$

得

$$K\alpha=n\lambda\alpha \tag{7}$$

其中 K 称为核矩阵，是 $n\times n$ 矩阵。$n\lambda_r$ 是 K 的特征值，$\alpha_1,\alpha_2,\cdots,\alpha_n$ 是对应的特征向量。按一定的标准(前几个特征值占总特征值的比例≥85%)，取前 $m(m<n)$ 个特征值和对应的标准化后特征向量 $\alpha_1,\alpha_2,\cdots,\alpha_m$，其中 $\alpha_r=\{\alpha_r^1,\alpha_r^2,\cdots,\alpha_r^n\}(r=1,2,\cdots,m)$，此时对 F 空间中样本 $\phi(x_j)(j=1,2,\cdots,n)$ 在 υ_r 上投影：

$$g_r(x_j)=(\varphi(x_j)\cdot\upsilon_r)=\sum_{i=1}^{n}\alpha_r^i(\phi(x_j)\cdot\phi(x_i)),(r=1,2,\cdots,m) \tag{8}$$

称 $g_r(x_j)$ 为对应 ϕ 的第 r 个非线性主元分量。将所有的投影值形成一个矢量 $g(x_j)=(g_1(x_j),g_2(x_j),\cdots,g_m(x_j))$ 作为样本的特征值。根据 Mercer[4]定理，用核函数 $K(x_i,x_j)=(\phi(x_i)\cdot\varphi(x_j))$ 代替 F 空间中的内积运算，(8)式可写为：

$$g_r(x_j)=(\phi(x_j)\cdot\upsilon_r)=\sum_{i=1}^{n}\alpha_r^iK(x_i,x_j) \tag{9}$$

从上面计算可以看到，PCA 的协方差矩阵与样本点的维数相关，而核矩阵与样本点的个数相关，PCA 不能解决非线性问题，KPCA 虽能解决，但由于样本点的个数比较大，带来核矩阵的维数比较大，造成计算复杂度增加。为了解决核矩阵的计算复杂性，已有的方法有疏散的贪婪矩阵近似(SGA)[5]和 TOM 方法[6]，根据概率空间提出了疏散的核 PCA 方法[7]等，在这篇文章里，我们使用聚类的方法缩减样本点个数来达到降低核矩阵的阶数。

2.3 聚类

文[5～7]都是直接对核矩阵计算进行优化。本文采用减少样本点的个数来降低核矩阵的阶数进行优化计算。为了尽可能多地保持原有的样本点分类信息，使变化后的信息尽量含有原样本点所拥有的信息，可行的方法是采用聚类算法。因为聚类算法是一种多元统计分类方法，这种方法不必事先知道分类对象的分类结构，而是基于整个数据集内部存在若干“分组”或“聚类”为出发点而产生的一种数据描述方法，每个子集中的点具有高度的内在相似性。另外，类均值向量含有大量的分类信息，所以就可以用每类的均值向量即中心点来代表该类，这样，核矩阵的阶数就随着样本点分类而减少，大大降低了核矩阵的计算复杂度。

聚类分析的具体算法很多，有系统聚类法、动态聚类法、神经网络聚类法、模糊聚类法、遗传聚类法等[8]。本文选择基于动态聚类的 K 均值聚类算法[8]进行试验。

3 算法实现

用纹理图像的特征提取验证算法的正确性，具体的算法为：

首先对图像的每一个像素用 $K\times K$ 掩码组成的窗口覆盖(如图1)，每一个窗口包含 $n\times n$ 个像素。测量是在窗口内进行的，构成了 R^{K^2} 维的特征向量。定义特征向量 $Z=(m_1,m_2,\cdots,m_{k^2})$，其中 m_j 是第 j 个窗口的度量。

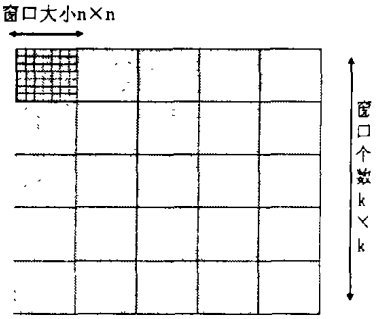

图1

度量使用的是每个窗口的灰度值的标准偏差：

$$m_j=\sqrt{\frac{\left(\sum_{r=1}^{n^2}i_r^2\right)}{n^2}-\left(\frac{\sum_{r=1}^{n^2}i_r}{n^2}\right)^2}$$

$$(j=1,2,\cdots,k^2),(1\leqslant r\leqslant n^2) \tag{9}$$

其中 i_r 代表像素的灰度值，并且 $\sum_{r=1}^{n^2}i_r^2$ 表示第 j 个窗口所有像素的平方和，$\sum_{r=1}^{n^2}i_r$ 表示第 j 个窗口所有像素的和。本文试验图像的宽度为128个像素，高度为128个像素。取多窗口的个数 $(k=5)\times(k=5)$，窗口的大小 $(n=7)\times(n=7)$，因此对每个像素都有 $Z=(m_1,m_2,\cdots,m_{25})$ 维的向量。对图像的每个像素分别用上述所讲的多窗口来覆盖，所以图像共有128×128个 $(m_1,m_2,\cdots,m_{25})$ 向量。本文把128×128个像素作为样本点，每个样本点 $(m_1,m_2,\cdots,m_{25})$ 是25维向量，对于纹理图像，每个像素的特征变化很大，呈现线性不可分状况。

其次，按照 KPCA 的方法降维时，根据上面介绍核矩阵的阶数与样本点的个数有关，即为16384×16384阶，在求其特征值和特征向量时，计算比较复杂。采用基于聚类的方法，首先对样本点进行聚类，把类的中心点作为新的样本点，这样核矩阵的阶数只与原样本点的分类的个数有关，大大降低求特征值和特征向量的复杂度。本试验对128×128个样本点按照动态聚类的 K 均值聚类算法求出类的中心点的特征向量，然后再按照 KPCA 的方法进行降维。

最后按马氏距离的方法进行检索。

4 试验结果

从上面的分析可以看到，选择类的个数越多，检索的效果就越好，但是也带来计算的复杂度，图2是分类和检索精度的分析，图3是类的个数和特征提取时间关系。通过实验可以看到在计算精度比较接近，而又考虑计算复杂度问题时，合适的类的个数为50类。

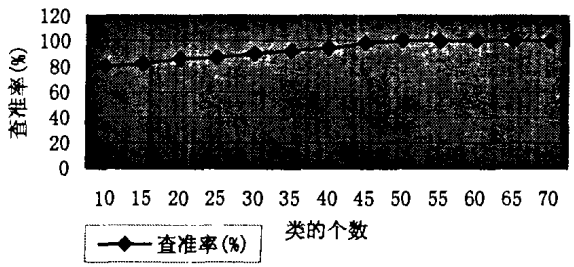

图2

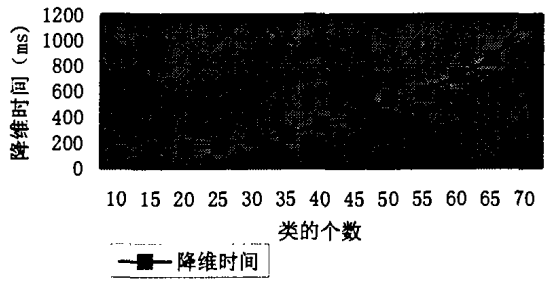

图3

另外，核矩阵的形式比较多[3]，本文使用高斯核 $K(x,y)=\exp(-\|x-y\|^2/\sigma^2)$，通过实验分析 $\sigma=1000$比较合适。

下图是 PCA 方法、KPCA 方法、基于聚类的 KPCA 方法的对比。降维时间如图4；效果趋势如图5。

方法	时间(ms)
PCA	47
KPCA	2200
基于聚类的 KPCA	510

图4

从图5可以看到，基于聚类的 KPCA 方法，大大提高了 KPCA 计算核矩阵的速度，而且随着图像个数的增多，几乎不影响原来的检索精度。

结论 本文首先分析了 PCA 和 KPCA 的不足，着重讲述了 KPCA 针对样本点计算带来的复杂度分析，介绍了基于聚类的 KPCA 方法，该方法不是对整个样本点做主成分分析，而是对代表样本点每类的均值向量做主成分分析，大大缩减了核矩阵的阶数，通过试验验证了算法的优越性。未来继续研究的工作是其他核分析方法。

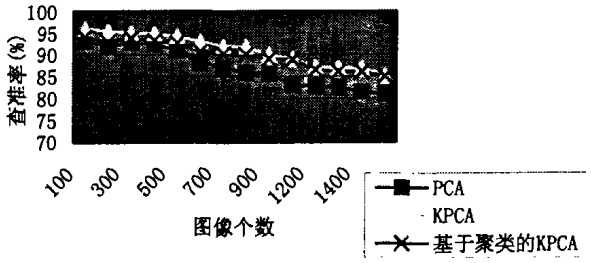

图5

参 考 文 献

1 Carreira-Perpian M A. A Review of Dimension Reduction Techniques: [Technical Report CS-96-09]. Department of Computer Science, University of Sheffield, 1997

2 Maeda E, Murase H. Multi-category classification by kernel based nonlinear subspace method. In: Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 99), IEEE Press, 1999

3 Scholkopf B, Smola A J, Muller Klaus-Robert. Kernel Principal Component Analysis [M]. Advances in Kerrel Methods, MIT Press, 1998. 327～352

4 Vapnik V N. Statistical learning theory. AT&T Research, London University, 1998

5 Smola A J, Scholkopf B. Sparse greedy matrix approximation for machine learning. In: Proc. of ICML'00, Bochum, Germany, Morgan Kaufmann, 2000. 911～918

6 Williams C, Seeger M. Using the Nystrom method to speed up kernel machines. Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2001, 13

7 Tipping M. Sparse kernel principal component analysis. In Advances in Neural Information Processing Systems 13, MIT Press, Cambridge, MA, 2001. 633～639

8 边肇祺，张学工，等. 模式识别. 清华大学出版社

（上接第63页）

方程解的解析表达是极其困难的，因此，只能采用一种近似的数值方法，这种方法在假定方程组(2.2.8)的解存在并且满足正则性，同时还必须假定精确解连续依赖于初始值，这样，根据输入初始值的不同，在计算过程中沿着搜索方向就会陷入不同的局部最小，从而得到不同周期的振荡曲线。

我们再分析圆形主曲线的半径 $r=(R_1+R_2)/2$。由于环形概率密度的分布是对称的，理论上必然得到圆形主曲线，并且在直觉上，所得到的圆形主曲线的半径 $r=(R_1+R_2)/2$。但是，我们所得到的圆形主曲线的半径并不等于环形分布的内外半径的平均值，而是 $r>(R_1+R_2)/2$。由于我们所建立的微分方程是基于主曲线的自相合性，而根据自相合性，生成曲线的外侧区域内的数据点多于内侧区域内的数据点，这样经过计算得到的实际圆形主曲线必然向外偏离半径为 $r=(R_1+R_2)/2$的圆。

由图 2还可以看到，环形均匀分布的主曲线不止一条，并且它们是相交的。这与线性主成分十分相似，数据分布一般也有很多主曲线，并且这些主曲线的相交性类似于线性主成分的正交。

我们选取不同的内外半径的比值和圆形主曲线的半径初始值 r_0，通过大量计算，概括出以下结论：环形均匀分布的主曲线不唯一且相交，并且主曲线的数目随内外半径比值的增加而增加，而周期 T 随比值的增加而减小。

结束语 我们所做的工作是对主曲线的某些性质进行初步的分析和研究，研究的数据分布仅针对特殊的均匀分布，对于平凡分布甚至是未知分布的主曲线性质的研究还有待进一步的研究，并且研究的空间仅限于二维空间，有待于向高维空间发展。

参 考 文 献

1 Hastie T. Principal curves and surfaces. Laboratory for computational Statistics: [Technical Report 11]. Stanford University, Dept. of Statistics, 1984

2 Kégl B. Principal curves: learning, design, and applications: [Ph. D. Thesis]. Concordia University, Canada, 1999

3 Duchamp T, Stuetzle W. Extremal properties of principle curves in the plane. Annals of Statistics, 1996, 24(4): 1511～1520

4 Hermann T, Meinicke P, Ritter H. Principal curve sonification. In: Proc. of Intl. Conf. on Auditory Display, 2000. 81～86

5 Seung H S, Lee D D. The manifold ways of perception? Science, 2000, 12: 2268～2269

6 Banfield J D, Raftery A E. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. Journal of the American Statistical Association, 1992, 87(417): 7～16

7 张军平，王珏. 主曲线研究综述. [J] 计算机学报，2003，26(2)：129～146

8 张红云，苗夺谦. 基于主曲线的相似字符模糊分类方法. 模式识别与人工智能，2004，4

9 唐庆适，苗夺谦. 基于主曲线的指纹细节特征提取方法. 计算机科学，2005，32(1)