

# 采用核聚类分析的 KPCA 改进算法

邓 貌<sup>1</sup>, 陈 旭<sup>1</sup>, 陈天翔<sup>2</sup>, 王徽蓉<sup>1</sup>, 鲁华祥<sup>1</sup>

(1. 中国科学院 半导体研究所, 北京 100083; 2. 厦门理工学院 电子与电气工程系, 福建 厦门 361005)

**摘 要:** 为了解决核主分量分析方法处理大训练样本集时计算代价巨大的问题, 在采用子集划分的 KPCA 算法基础上, 提出采用核聚类划分分子集, 并用每个子集的协方差矩阵的特征值累积贡献率作为标准来选取相应的特征向量。分别在人工和实际数据集上测试, 实验结果显示在同一累积贡献率和给定子集个数的条件下, 采用核聚类划分分子集总能得到较小尺寸的核矩阵, 而核矩阵尺寸的减小有助于改善测试样本的特征提取速度以及降低特征分解核矩阵的时间复杂度。

**关键词:** 核主分量分析; 核聚类; 子集划分; 协方差矩阵; 特征向量

**中图分类号:** TP18 **文献标识码:** A **文章编号:** 1673-4785(2010)03-0221-06

## Improved kernel principal component analysis based on a clustering algorithm

DENG Mao<sup>1</sup>, CHEN Xu<sup>1</sup>, CHEN Tian-xiang<sup>2</sup>, WANG Hui-rong<sup>1</sup>, LU Hua-xiang<sup>1</sup>

(1. Institute of Semiconductors, Chinese Academy Sciences, Beijing 100083, China; 2. Department of Electronic and Electrical Engineering, Xiamen University of Technology, Xiamen 361005, China)

**Abstract:** To overcome the computational problems of the standard kernel principal component analysis (KPCA) algorithm, the authors proposed a new method for eigenvector selection by evaluating the cumulative contribution rate of the eigenvalues of the covariance matrix. In addition, a new way to partition the training data set based on kernel clustering was also developed. The influence was then explored of different partitions of training data sets on the size of the final kernel matrix, on the conditions causing a given cumulative contribution rate, and on the number of subsets. Experimental results showed that a smaller kernel matrix can be obtained when kernel clustering method are used to partition the training dataset. The proposed algorithm can be helpful to reduce the time complexity of the eigen decomposition of a kernel matrix and to improve the speed of feature extraction for test samples.

**Keywords:** KPCA; kernel clustering; partition of training data set; covariance matrix; eigenvector

核主分量分析(kernel principal component analysis, KPCA)通过 Mercer 核函数实现非线性映射, 将主分量分析(principal component analysis, PCA)方法在特征空间中推广。由于 KPCA 算法具有较强的提取数据非线性特征的能力, 因此它在模式识别、故障分析等领域得到了广泛的应用和重视<sup>[1-5]</sup>。但标准 KPCA 算法需要对一个  $M \times M$  的核矩阵( $M$  为训练样本个数)进行特征值分解, 当训练大样本集时, 该方法面临计算代价巨大的问题。

为此, 文献[6]提出将训练集划分成若干子集, 并将每个子集在特征空间中的协方差矩阵用一些特征向量近似表示。基于这种近似表示, 使 KPCA 在求解过程中, 只需对一个阶数等于这些特征向量数目之和的核矩阵进行特征值分解即可。该方法简单有效, 但是文中没有讨论每个子集应选取多少个特征向量, 以及训练集划分分子集的不同方式是否会明显影响最后选择的特征向量个数, 从而影响分解最终核矩阵的时间复杂度以及求解测试样本的特征提取时间。本文将针对这 2 个问题进行探讨。

### 1 基于子集划分的 KPCA

由文献[6], 设训练样本集  $X = \{x_i\}_{i=1}^N$ , 划分的

收稿日期: 2009-12-19.

基金项目: 国家“863”计划资助项目(2007AA04Z423, 2006AA01Z106); 国家自然科学基金资助项目(60576033); 福建省自然科学基金资助项目(2008J04001); 厦门市科技计划资助项目(3502Z20083031).

通信作者: 陈 旭. E-mail: shendacx@163.com.

$M$  个子集为  $X_k = \{x_{ki}\}_{i=1}^{N_k}, k=1, 2, \dots, M, N_k$  为子集  $X_k$  的样本个数. 显然, 有  $\sum_{k=1}^M X_k = X$  以及  $\sum_{k=1}^M N_k = N$  成立. 经过非线性变换后, 子集  $X_k$  对应的特征点集为  $\varphi_k = \{\varphi(x_{ki})\}_{i=1}^{N_k}$ . 记  $\varphi_k$  的矩阵形式为  $\varphi(X_k) = [\varphi(x_{k1}) \varphi(x_{k2}) \cdots \varphi(x_{ki}) \cdots \varphi(x_{kN})],$   
 $k = 1, 2, \dots, M.$   
 记训练样本集  $X = [x_1 x_2 \cdots x_N]$ , 子集  $X_k = [x_{k1} x_{k2} \cdots x_{kN_k}] (k=1, 2, \dots, M).$

对应地, 把核矩阵  $K$  分成  $M \times M$  块, 即  $K = (K_{ij}), i = 1, 2, \dots, M, j = 1, 2, \dots, M.$  其中:  $K_{ij} = \varphi^T(X_i)\varphi(X_j)$ . 记  $R_k = \varphi(X_k)\varphi^T(X_k)$ , 则由 KPCA 算法原理知:

$$\begin{aligned} R &= \frac{1}{N} \sum_{i=1}^N \varphi(x_i)\varphi^T(x_i) = \\ &= \frac{1}{N} \sum_{k=1}^M \sum_{i=1}^{N_k} \varphi(x_{ki})\varphi^T(x_{ki}) = \\ &= \frac{1}{N} \sum_{k=1}^M \varphi(X_k)\varphi^T(X_k) = \frac{1}{N} \sum_{k=1}^M R_k. \end{aligned}$$

设  $\lambda_1^k \geq \lambda_2^k \geq \cdots \geq \lambda_{N_k}^k$  为  $R_k$  的特征值,  $v_1^k, v_2^k, \dots, v_{N_k}^k$  为对应的特征向量, 则

$$R_k = \sum_{j=1}^{N_k} \lambda_j^k v_j^k (v_j^k)^T = \sum_{j=1}^{N_k} \tilde{v}_j^k (\tilde{v}_j^k)^T. \quad (1)$$

式中:  $\tilde{v}_j^k = \sqrt{\lambda_j^k} v_j^k = \varphi(X_k) a_j^k, a_j^k$  为矩阵  $K_{kk}$  的单位特征向量, 其对应特征值为  $\lambda_j^k$ .

从  $R_k$  的  $N_k$  个特征值中选择最大的前  $p_k$  个特征值  $\lambda_1^k, \lambda_2^k, \dots, \lambda_{p_k}^k$ , 使之满足:

$$\sum_{k=1}^M \sum_{i=p_k+1}^{N_k} \lambda_i^k \ll \sum_{k=1}^M \sum_{i=1}^{N_k} \lambda_i^k.$$

令

$$\tilde{R}_k = \sum_{i=1}^{p_k} \tilde{v}_i^k (\tilde{v}_i^k)^T, \quad (2)$$

$$\tilde{R} = \frac{1}{N} \sum_{k=1}^M \tilde{R}_k = \frac{1}{N} \sum_{k=1}^M \sum_{i=1}^{p_k} \tilde{v}_i^k (\tilde{v}_i^k)^T, \quad (3)$$

显然有  $\tilde{R}_k \approx R_k, \tilde{R} \approx R$ , 这样对  $R$  的特征值分解, 就可以近似地转换为对  $\tilde{R}$  的特征值分解. 设  $\tilde{\lambda}, \tilde{v}$  为  $\tilde{R}$  的特征值和对应的特征向量, 即有

$$\tilde{R} \tilde{v} = \tilde{\lambda} \tilde{v}. \quad (4)$$

注意到, 存在系数  $\tilde{a}_{kj} (k=1, 2, \dots, M, j=1, 2, \dots, p_k)$ , 使得

$$\tilde{v} = \sum_{k=1}^M \sum_{j=1}^{p_k} \tilde{v}_j^k \tilde{a}_{kj} = \sum_{k=1}^M \sum_{j=1}^{p_k} \varphi(X_k) a_j^k \tilde{a}_{kj}. \quad (5)$$

将式(2)、(5)代入式(4), 可得

$$\tilde{K} \tilde{a} = \tilde{\lambda} \tilde{a}.$$

式中:  $\tilde{K} = (\tilde{K}_{ij}) (i=1, 2, \dots, M, j=1, 2, \dots, M)$  为分

块矩阵,  $\tilde{K}_{ij} = A_i^T K_{ij} A_j$  为  $p_i \times p_j$  的矩阵,

$$A_i = [a_1^i a_2^i \cdots a_{p_i}^i],$$

$$\tilde{a} = [\tilde{a}_{11} \cdots \tilde{a}_{1p_1} \tilde{a}_{21} \cdots \tilde{a}_{2p_2} \cdots \tilde{a}_{M1} \cdots \tilde{a}_{Mp_M}].$$

这样利用近似核矩阵  $\tilde{K}$  就可以得到特征向量  $\tilde{a}$ .

对于测试样本  $z$ , 它在特征向量  $\tilde{v}$  的投影为

$$\begin{aligned} (\tilde{v})^T \varphi(z) &= \sum_{k=1}^M \sum_{j=1}^{p_k} \tilde{a}_{kj} (a_j^k)^T \varphi^T(X_k) \varphi(z) = \\ &= \sum_{k=1}^M \sum_{j=1}^{p_k} \tilde{a}_{kj} (a_j^k)^T \tau_k. \end{aligned} \quad (6)$$

式中:  $\tau_k = [k(x_{k1} z) k(x_{k2} z) \cdots k(x_{kN_k} z)]$ .

## 2 基于核聚类分析的 KPCA 改进算法

由基于子集划分的 KPCA 算法<sup>[6]</sup>可知, 其关键性能指标是  $\tilde{N} = \sum_{k=1}^M p_k$ .  $\tilde{N}$  越小, 既可使对  $\tilde{K}$  特征分解的复杂度  $O(\tilde{N}^3)$  降低; 同时由于  $\tilde{K}$  的特征向量  $\tilde{a}$  的维数为  $\tilde{N}$ , 若能减小  $\tilde{N}$ , 则根据式(6)可知, 可能改善测试样本的特征提取速度.

显然, 影响  $\tilde{N}$  值大小的因素有子集个数  $M$ 、每个子集选取的特征向量个数  $p_k$ . 另外, 子集的不同划分方式对  $\tilde{N}$  也有比较明显的影响.

### 2.1 特征向量的选取

由基于子集划分的 KPCA 算法知, 算法的关键是用  $\tilde{R}, \tilde{R}_k$  来近似  $R, R_k$ . 对比式(1)和(2)可知, 组成  $\tilde{R}_k$  的  $p_k$  个特征向量与  $R_k$  的  $N_k$  个特征值中最大的前  $p_k$  个特征值  $\lambda_1^k, \lambda_2^k, \dots, \lambda_{p_k}^k$  相对应. 提出采用特征值的累计贡献率来确定  $p_k$  的值:

设  $R_k$  的  $N_k$  个特征值为  $\lambda_1^k, \lambda_2^k, \dots, \lambda_{N_k}^k$ . 记

$$S_k = \sum_{i=1}^{N_k} \lambda_i^k, S_k(t) = \sum_{i=1}^t \lambda_i^k.$$

令  $R_k$  前  $t$  个特征值的累积贡献率为

$$\gamma(t) = \frac{S_k(t)}{S_k}.$$

给定一个比率  $r_a$ , 逐渐增加  $t$  直到某一个  $t_s$ , 有  $\gamma(t_s) \geq r_a$ , 则  $p_k = t_s$ , 即选择  $R_k$  的最大的前  $t_s$  个特征向量, 然后利用式(2)和(3)可求得  $\tilde{R}$  和  $\tilde{R}_k$ .

### 2.2 KKM 划分子集

基于子集划分的 KPCA 算法需要将训练样本集划分为若干子集. 子集的不同划分方式会导致算法不同的性能, 但文献[6]中并未讨论子集的划分方法. 本文提出从聚类的角度出发, 以距离为相似度准则把相似样本聚在一类, 把每一类作为一个子集, 以此对训练样本集进行划分. 又由于 KPCA 算法实际上是对训练样本经过核函数变换到特征空间中的特征点进行 PCA 分析, 因此本文对这些特征点而非训

训练样本聚类,即采用核聚类方法.典型的是核 K-均值(kernel K-means, KKM)聚类<sup>[7]</sup>:

设训练样本集为  $X = \{x_i\}_{i=1}^N$ ,通过某种非线性映射  $\phi$  映射到某一特征空间  $H$  得到  $\phi(x_1), \phi(x_2), \dots, \phi(x_N)$ ,  $\{\phi(x_i)\}_{i=1}^N$  就是待聚类的样本.核 K-均值聚类算法实际上就是最小化目标函数

$$J = \sum_i \sum_j \|\phi(x_{ij}) - c_i\|^2.$$

式中:  $x_{ij}$  为第  $i$  类的第  $j$  个样本,  $\phi(x_{ij})$  是其在特征空间中的投影,  $c_i$  为第  $i$  类的中心.

由  $\frac{\partial J}{\partial c_i} = 0$ , 可得

$$c_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \phi(x_{ij}),$$

则对于某个给定的样本  $x$ ,它在特征空间中与某类中心  $c_i$  的距离为

$$\begin{aligned} D = \|\phi(x) - c_i\|^2 &= \|\phi(x) - \frac{1}{N_i} \sum_{j=1}^{N_i} \phi(x_{ij})\|^2 = \\ &< \phi(x), \phi(x) > - \frac{2}{N_i} < \phi(x), \sum_{j=1}^{N_i} \phi(x_{ij}) > + \\ &\frac{1}{N_i^2} \sum_{l=1}^{N_i} \sum_{m=1}^{N_i} < \phi(x_{il}), \phi(x_{im}) >. \end{aligned}$$

由核函数定义:

$$\begin{aligned} K(x, x) &= < \phi(x), \phi(x) >, \\ D &= K(x, x) - \frac{2}{N_i} \sum_j K(x, x_{ij}) + \\ &\frac{1}{N_i^2} \sum_l \sum_m K(x_{il}, x_{im}), \end{aligned}$$

这样,就可以根据最近邻原则将每个样本分配给相应类别.

### 3 实验结果与分析

如上节所述,影响  $\tilde{N}$  值大小的因素包括子集个数  $M$ 、每个子集选取的特征向量个数  $p_k$  以及子集的不同划分方式.若对子集的特征向量按照本文提出的累积贡献率选择选取,则  $p_k$  的值由给定的比率  $r_a$  决定.实验部分 3.1 节比较在相同的子集个数  $M$ 、比率  $r_a$  的条件下,3 种不同的划分子集方式(“一般”划分、平分和核聚类划分)对  $\tilde{N}$  值的影响;实验 3.2 节给出以上 3 种划分子集的方式所对应的 KPCA 方法提取数据主要特征向量的实验结果.实验采用著名的机器学习平台怀卡托智能分析环境(Waikato environment for knowledge analysis, WEKA)<sup>[8]</sup>,以及常用软件如 Matlab 7.3 等.

#### 3.1 不同划分子集方式对 $\tilde{N}$ 值的影响

设有  $M$  个子集,一般的划分是子集元素个数  $N_k$  ( $k=1, 2, \dots, M$ ) 为满足  $\sum_{k=1}^M N_k = N$  的任意一组数,本文称之为“一般”划分(general dividing),与之对应的 KPCA 方法称为 GDKPCA;如果划分到每个子集的元素个数  $N_k$  ( $k=1, 2, \dots, M$ ) 基本相同(约为  $N/M$ ),本文称之为平分(equal dividing),对应的 KPCA 方法称为 EDKPCA;本文提出的采用核聚类方法(kernel clustering)对训练集划分,聚类数目设为子集个数  $M$ ,核聚类划分对应的 KPCA 方法称为 KCKPCA.将 GDKPCA、EDKPCA 和 KCKPCA 应用于人工数据和实际数据,比较在相同的累积贡献率  $r_a$  和核参数条件下,这 3 种划分方式对最后特殊向量总数  $\tilde{N}$  值的影响.对每个数据绘制出不同子集划分对应 KPCA 方法的  $\tilde{N}$  值随子集个数  $M$  变化的曲线,如图 1 所示.

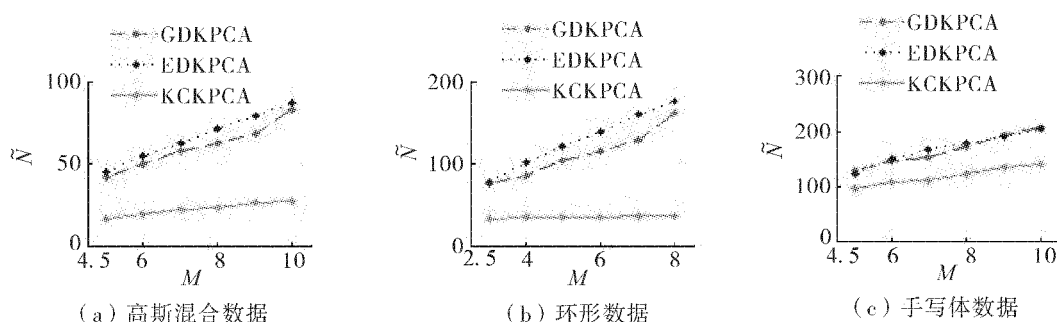


图1 不同的子集划分方式对特殊向量总数目的影响

Fig. 1 The influences of different partitions of a subset on the size of eigenvector

图 1(a) 中的高斯混合数据(mixture data)由 3 个二维高斯分布数据点生成,各类的均值为  $(-0.5, -0.2)$ 、 $(0.0, 0.6)$ 、 $(0.5, 0.0)$ ,方差均为 0.01,每一类产生 500 个点,共 1 500 个样本.

累积贡献率  $r_a = 0.95$ ,选用高斯核函数  $K_1(x, y) = e^{-\rho_1 \|x-y\|^2}$ ,  $\rho_1 = 2.5$ .

图 1(b) 中的环形数据(ring data)由两类圆形数据产生,每类由  $N = 251$  个样本组成,共有样本

502 个. 每一类均加上均值  $\mu=0$ , 方差  $\sigma^2=0.001$  的高斯噪声. 累积贡献率  $r_a=0.95$ , 选用高斯核函数  $K_2(\mathbf{x}, \mathbf{y}) = e^{-\rho_2 \|\mathbf{x}-\mathbf{y}\|^2}$ ,  $\rho_2=7.5$ .

图 1(c) 中的手写体数据 (handwritten digits) 来自 UCI 标准数据库<sup>[9]</sup>, 实验从数据库中产生 5 个实例 (instances), 每个实例包含 2 000 个数据, 其中 1 500 个用于训练, 500 个用于测试, 每个数据包含 64 个维度. 实验以这 5 个实例实验结果的均值作为最终数据结果. 累积贡献率  $r_a=0.95$ , 选用多项式核

函数  $K_3(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + b)^d$ , 其中  $b=0.5, d=0.3$ .

利用累积贡献率标准选取特征向量, 衡量的是将全体特征向量压缩为部分向量的压缩率. 由图 1 可知, 采用同样的特征向量压缩率, 只要给定子集个数  $M$ , 用 KCKPCA 方法总能得到比其他 2 种方法更少的特征向量个数  $\tilde{N}$ . 减小  $\tilde{N}$  值有助于降低核矩阵  $\tilde{\mathbf{K}}$  特征分解的复杂度  $O(\tilde{N}^3)$ , 并能改善测试样本的特征提取速度 (如表 1).

表 1 不同的子集划分方式的性能比较

Table 1 The performance comparison of different partitions of the subset

算法	mixture 数据		ring 数据		handwritten digits 数据	
	特征分解	测试样本	特征分解	测试样本	特征分解	测试样本
	核矩阵时间比	投影时间比	核矩阵时间比	投影时间比	核矩阵时间比	投影时间比
GDKPCA	12.60	1.38	38.54	3.57	1.60	1.19
EDKPCA	3.30	1.57	36.70	3.87	1.68	1.22
KCKPCA	1.00	1.00	1.00	1.00	1.00	1.00

表 1 中的数据时间比是在子集数目为 10 时, 用 GDKPCA、EDKPCA 算法的时间比上 KCKPCA 算法的时间, 可以看出 KCKPCA 算法的特征分解核矩阵时间以及测试样本投影时间最少, 子集数目为其他值时, 上述结论仍然成立.

### 3.2 不同划分的 KPCA 提取数据特征的效果

#### 3.2.1 绘制投影 ring 数据主分量的等高线

为对比 KPCA、GDKPCA、EDKPCA、KCKPCA 提取数据集特征的效果, 绘制投影 ring 数据主分量的等高线, 如图 2 所示.

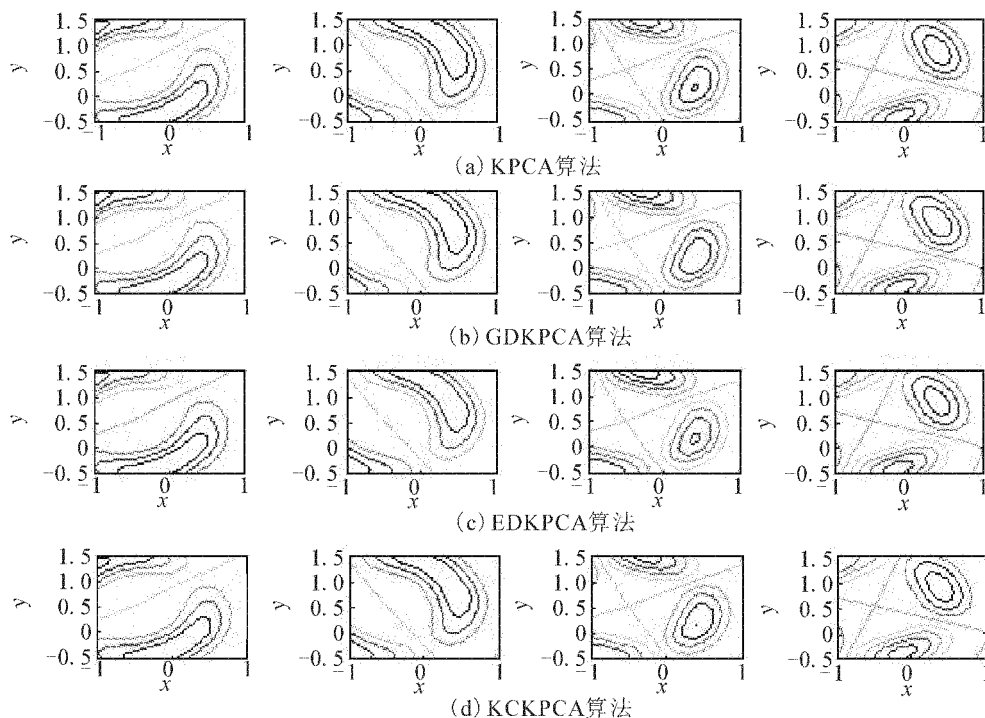


图 2 用于提取数据特征的 4 种 KPCA 算法的效果对比

Fig. 2 The comparison of four kinds of KPCA algorithms for feature extraction

由图 2 可见,图中从上到下每一横排对应的 KPCA 算法分别为标准 KPCA、GDKPCA(一般划分)、EDKPCA(平分)、KCKPCA(核聚类划分);从左到右每一纵排分别对应按从大到小排列的前面 4 个特征向量的投影图.图 2 中同一曲线上的点表示:在特征空间中,这些点在主轴方向的投影值相同,该曲线称为等高线.子集的特征向量选取按照累积贡献率  $r_a = 0.95$  来选择.由图 2 可见,不同样本子集选择方法都能取得同标准 KPCA 算法基本相同的等高线,说明它们提取的数据集特征基本相同.

### 3.2.2 UCI 数据库 handwritten digits 识别

本实验的目的是比较 3 种基于样本子集方法

(GDKPCA、EDKPCA 和 KCKPCA)的效果和性能.首先用 GDKPCA、EDKPCA、KCKPCA 提取训练样本集的主要特征向量,然后将测试样本集投影到这些特征向量方向上,得到特征提取后的测试样本集.在著名的机器学习平台 WEKA 上利用 SMO(sequential minimal optimization)分类器对特征提取后的测试样本集进行识别.实验从数据库中产生 5 个实例(instances),每个实例包含 2 000 个数据,其中 1 500 个用于训练,500 个测试,每个数据包含 64 个维度.设定各子集选择特征向量个数为 10,需满足累积贡献率  $r_a = 0.95$ ,实验结果如表 2 所示.

表 2 3 种方法(GDKPCA、EDKPCA 和 KCKPCA)的性能比较

Table 2 The performance comparison of three methods (GDKPCA, EDKPCA and KCKPCA)

方法	核矩阵尺寸	提取特征个数	测试误差率/%	样本投影时间比
GDKPCA	205	65	4.8	1.40
EDKPCA	203	64	5.0	1.42
KCKPCA	140	50	4.8	1.00

由表 2 可知,通过 GDKPCA、EDKPCA 和 KCKPCA 3 种变换后得到的新的测试集用于识别都能得到比较准确的结果(测试误差率分别为 4.8%, 5.0% 和 4.8%).但同时注意到核聚类划分 KPCA 算法(KCKPCA)得到的核矩阵尺寸、提取特征个数和样本的投影时间均为最小,显示出一定的优越性.

### 3.3 实验结果分析

由以上实验结果可知,在实际应用中确定了划分子集的个数  $M$  以及选择子集特征向量的累积贡献率  $r_a$  后,用核聚类划分子集的 KPCA 方法(KCKPCA)能够得到与 GDKPCA、EDKPCA 基本相同的特征提取效果,但比 GDKPCA、EDKPCA 更能降低核矩阵  $\tilde{K}$  特征分解的复杂度  $O(\tilde{N}^3)$  以及改善测试样本的特征提取速度.

为了验证算法的稳定性,实验首先把数据产生多个实例,然后对每个实例分别测试,最后对这些实例的实验结果求统计平均作为该数据的实验结果.实验结果表明,KCKPCA 方法相对于 GDKPCA、EDKPCA 方法的优越性在测试数据集上是鲁棒的.

## 4 结束语

标准 KPCA 算法作为一种非常有效的非线性特征处理方法被广泛研究和应用,但是,当训练大样本集时该方法面临的计算代价大的问题限制了它的进一步应用.因此,研究如何改善大样本集的计算代价

具有十分重要的意义.本文在文献[6]中提出的基于子集划分的 KPCA 算法的基础上作了进一步的工作,具体地,提出了用核聚类方法来划分子集和采用特征值累积贡献率来选择特征向量.实验验证了核聚类方法的有效性、优越性和累积贡献率原则的有效性.这种划分子集的方法和选择特征向量的方式为应用 KPCA 方法进行非线性特征处理给出了新思路.

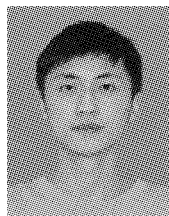
但是,核聚类划分子集的 KPCA 方法仍有一些问题需要进一步研究.首先,需要在更多的、大量的数据集上测试 KCKPCA 方法,以验证它是否具有普遍适用性.如果不能普遍适用,就需要分析它具有适用性的条件以及最适合应用它的条件.其次,划分子集的个数  $M$  直接影响  $\tilde{N}$  值大小,是否能根据子集的性质来估计较为合适的子集个数  $M$ ,这个合适值又应该怎样估计,另外,如何进一步改善测试阶段的特征提取速度等问题都需要进一步研究.

## 参考文献:

- [1] 谢永华,陈伏兵,张生亮,杨静宇.融合小波变换与 KPCA 的分块人脸特征抽取与识别算法[J].中国图象图形学报,2007,12(4):666-671.
- XIE Yonghua, CHEN Fubing, ZHANG Shengliang, YANG Jingyu. Features extraction and recognition of intersected human face based on wavelet transform and KPCA[J]. Journal of Image and Graphics, 2007, 12(4): 666-671.

- [2] 曾庆虎, 邱静, 刘冠军, 谭晓栋. 基于 KPCA-HSMM 设备退化状态识别与故障预测方法研究[J]. 仪器仪表学报, 2009, 30(7): 1341-1346.  
ZENG Qinghu, QIU Jing, LIU Guanjun, TAN Xiaodong. Research on equipment degradation state recognition and fault prognostics method based on KPCA-hidden semi-Markov model[J]. Chinese Journal of Scientific Instrument, 2009, 30(7): 1341-1346.
- [3] LI Ying, LEI Xiaogang, BAI Bendu, ZHANG Yanning. Information compression and speckle reduction for multifrequency polarimetric SAR images based on kernel PCA[J]. Journal of Systems Engineering and Electronics, 2008, 19(3): 493-498.
- [4] KIM K I, PARK S H, KIM H J. Kernel principal component analysis for texture classification[J]. IEEE Signal Processing Letters, 2001, 8(2): 39-41.
- [5] ROSIPAL R, GIROLAMI M, TREJO L J, CICHOCKI A. Kernel PCA for feature extraction and denoising in non-linear regression[J]. Neural Computing & Applications, 2001, 10(3): 231-243.
- [6] ZHENG Wenming, ZOU Cairong, ZHAO Li. An improved algorithm for kernel principal component analysis[J]. Neural Processing Letters, 2005, 22(1): 49-56.
- [7] 孔锐, 张国宣, 施泽生, 等. 基于核的 K 均值聚类[J]. 计算机工程, 2004, 30(11): 12-14.  
KONG Rui, ZHANG Guoxuan, SHI Zesheng, et al. Kernel-based K-means clustering[J]. Computer Engineering, 2004, 30(11): 12-14.
- [8] The University of Waikato. WEKA[EB/OL]. [2009-12-02]. <http://www.cs.waikato.ac.nz/ml/weka>.
- [9] UCI. Semeion handwritten digit data set[EB/OL]. [2009-12-02]. <http://archive.ics.uci.edu/ml/datasets/Semeion+Handwritten+Digit>.

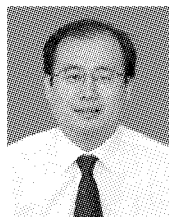
## 作者简介:



邓 貌,男,1986 年生,硕士研究生,主要研究方向为优化算法、神经网络、模式识别等。



陈 旭,女,1978 年生,副研究员,主要研究方向为图像处理、模式识别、神经网络等。



陈天翔,男,1966 年生,副研究员,主要研究方向为神经网络、智能计算等。

## 全国图像图形学学术会议

## The 15th National Conference on Image and Graphics(NCIG 2010)

第 15 届全国图像图形学学术会议(NCIG2010)将于 2010 年 12 月在广州召开. NCIG 2010 旨在为从事图像图形相关领域的基础研究和应用推广的广大专家学者和工程技术人员提供一个相互交流的平台. 会议将邀请国内外著名的图像图形学领域的专家与会并做大会报告,邀请企业参加产品展示会. 同时,将召开学会理事会.

NCIG 涵盖了计算机图形学、图像处理、视频通讯、虚拟现实、三维可视化、医学影像、数字艺术和游戏设计、机器学习、信息安全等广泛领域.

会议将由清华大学出版社正式出版“会议论文集”,并且除在参会宣读的青年论文中评选优秀论文外,还将推荐数十篇优秀论文到有关杂志上发表. 欢迎踊跃投稿和参会交流.

会议网站:<http://ncig2010.gdut.edu.cn>

电子邮件:[ncig2010@gdut.edu.cn](mailto:ncig2010@gdut.edu.cn); [ncig2010@163.com](mailto:ncig2010@163.com)