核协方差成分分析方法及其在聚类中的应用

闫晓波 王士同 郭慧玲
(江南大学数字媒体学院 无锡 214122)

摘 要 以降维前后密度总和与Renyi熵的差(Densities-vs-Entropy,D-vs-E)尽量靠近为准则,得到了一种新的特征降维方法,而D-vs-E是由核特征空间的协方差矩阵导出的,因此称为核协方差成分分析(Kernel Covariance Component Analysis,KCCA)。将D-vs-E发展为广义D-vs-E(generalized D-vs-E)。KCCA通过将数据投影在使D-vs-E最大的KPCA轴方向得到转换后的低维数据,但是所选取的KPCA轴不一定对应于核矩阵最大的几个特征值。与基于Renyi熵的KECA相比,KCCA是基于D-vs-E的。基于广义D-vs-E的KCCA数据转换方法应用于聚类的结果显示,它在对高斯核参数的选择上具有更强的鲁棒性。

关键词 核熵成分分析,核协方差成分分析,聚类,协方差矩阵,高斯核参数,雷尼熵
中图法分类号 TP181 **文献标识码** A

Kernel Covariance Component Analysis and its Application in Clustering

YAN Xiao-bo WANG Shi-tong GUO Hui-ling
(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract A new feature dimensionality reduction method called kernel covariance component analysis (KCCA) was put forward on the criterion that the transformed data best preserves the concept of the difference (denoted as D-vs-E) between total densities and Renyi entropy of the input data space, induced from kernel covariance matrix. The generalized version of D-vs-E was also developed here. KCCA achieves its goal by projections onto a subset of D-vs-E preserving kernel principal component analysis (KPCA) axes and this subset does not generally need to correspond to the top eigenvalues of the corresponding kernel matrix, in contrast to KPCA. However, KCCA is rooted at the new concept of D-vs-E rather than Renyi entropy. Experimental results also show that KCCA is more robust to the choice of Gaussian kernel bandwidth when it is used in clustering.

Keywords KECA, KCCA, Clustering, Covariance matrix, Gaussian kernel bandwidth, Renyi entropy

特征提取与选择[1-5]在模式分析和机器智能领域是非常关键的,它的目的是将维数较高的数据转换为典型的低维数据,以便揭示出数据的内在结构。用映射(或变换)的方法把原始特征变换为新特征,称为特征提取(Feature Extraction),常用的变换有傅立叶变换、小波变换、PCA变换、ICA变换、Gabor变换等。希望通过变换,用较少的特征来近似表示原来的对象,而且误差尽量小。在所有正交线性变换中,这种最优变换是Karhunen-Loeve(KL)变换,相应的特征提取方法被称为主成分分析(Principle Component Analysis,PCA)。从原始特征中挑选出一些最有代表性、可分性能最好的特征,称为特征选择(Feature Selection)。谱方法(spectral methods)是数据转换领域的一个主导研究方向。谱方法是基于特定矩阵的最大或最小的一些特征值或谱(spectrum)和相应的特征向量来实现数据转换。其中最为著名的是主成分分析法,它需要构造数据的相关系数矩阵。主成分分析法是线性方法,转换后的数据是不相关的,并且最大程度地保留了原数据的方差。还有一种线性方法称为度量多维标度(Metric Multidimensional Scaling,Metric MDS)[6],它需要构造数据的内积矩阵,转换后的数据在最小二乘的意义上最大程度地保留了原数据的内积。我们可以证明Metric MDS与PCA是等价的[7]。

KPCA[8,9]是一种非常有影响力的非线性的数据转换谱方法,这种方法需要先把原数据非线性映射到高维的核特征空间,核特征空间中的内积可以用一个半正定的核函数计算得出,这样就可以构造出核特征空间的内积矩阵(核矩阵),然后在这个核矩阵上进行metric MDS操作。也就是说KPCA就是运用在线性可分的高维的特征空间中的PCA方法。KPCA有很多应用,文献[10]中的谱聚类方法就是C-mean在以KPCA的主轴为基坐标下聚类的。KPCA还可以用于模式降噪和分类。还有很多其它的谱方法,区别在于矩阵的构造方法不同。例如ISOMAP(isometric mapping)[13]算法用点之间的geodesic距离代替点之间的欧氏距离,然后用Metric MDS算法进行数据降维。最近提出的还有MVU(Maximum Variance Unfolding)[14]算法、LLE(Locally Linear Em-

到稿日期:2011-11-29 返修日期:2012-02-17 本文受国家自然科学基金(90820002),江苏省自然科学基金(BK2009067)资助。
闫晓波(1987-),女,硕士生,主要研究方向为人工智能、模式识别,E-mail:hnpyyxb@163.com;王士同(1964-),男,教授,博士生导师,主要研究方向为模式识别、人工智能、生物信息学;郭慧玲(1989-),女,硕士生,主要研究方向为人工智能、模式识别、图像处理。

bedding)[15]算法等。Robert Jenssen 提出的 KECA[11]与其它谱方法有很大不同，主要体现在两个方面：第一，这种数据转换方法揭示了输入空间数据集关于 Renyi 熵的结构；第二，这种方法不一定选择核矩阵最大的那些特征值以及相应的特征向量。

本文提出的 KCCA 方法是基于协方差矩阵的特征降维方法，这种方法要求降维前后的 D-vs-E 尽量接近，而 D-vs-E 是由核空间的协方差矩阵引出的。这种方法的特点是：第一，散度为 σ 时，降维前后 Renyi 熵尽量接近；第二，散度 σ 缩小时，降维前后数据集形状的变化尽量小。KCCA 方法是通过对输入空间数据集的 D-vs-E 来估计有贡献的 KPCA 轴方向上的投影，以达到数据转换和降维，而这些轴不一定对应于核矩阵最大的那些特征值和特征向量。如果能使得降维前后的 Renyi 熵的变化尽可能地小，并且当散度 σ 变小时，降维后的数据集也能较好地保持降维前的形状，就能使聚类的结果更稳定，σ 的选取也就更容易了。

对于 KECA 与 KCCA 方法转换后的数据集，在核特征空间中不同的簇关于原点都保持着一定的角度，这在某种意义上揭示了数据潜在的分类信息，但是 KECA 的这种结构对核参数 σ 的变化比较敏感，核参数 σ 很小的变化就能导致转换后的数据结构发生很大的改变，这就使得最优或接近最优的核参数值区间小，难以捕捉，算法的鲁棒性不好，对 KECA 的应用造成不便。而 KCCA 解决了这个问题，我们采用一种基于角度的谱聚类算法分别对经过 KCCA 和 KECA 转换得到的数据集进行聚类。实验结果表明，KCCA 在核参数的选择上具有更好的鲁棒性。在核方法中，核参数的选择至关重要。径向基核函数中核参数 σ 的选取对核函数的性能影响很大，合适的 σ 是核函数具有良好性能的保证。过大的 σ，样本的“势力范围”会过大，以致于一些毫无关系的训练样本会干扰对新的测试样本做出正确判断；过小的 σ 则会导致核函数只有记忆功能而无法对新的样本进行判断。所以，选择合适的核宽度需要在两者之间进行权衡。为了获得高性能的核方法，核参数选择的鲁棒性就显得尤为重要。

本文第 1 节简单介绍谱数据转换方法 KECA；第 2 节给出 KCCA 的原理、与现有谱数据转换方法的比较以及算法步骤，并以美国邮政区号手写数字识别数据集(USPS)中的数字 0 和 9 为例做特征降维实验，从实验结果更直观地看到 KPCA，KECA 和 KCCA 这几种谱数据转换方法的异同；第 3 节介绍 KCCA 在聚类中的应用。

1 核熵成分分析(KECA)

KECA 是由 R. Jenssen[11]提出的一种数据转换和降维方法，为了便于阅读，以下采用与文献[11]相同的数学符号。KECA 的提出基于两个概念，一个是 Renyi 熵[16]

$$H(p) = -\log V(p) = -\log\int p^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{1}$$

一个是 Parzen 窗密度估计[17]

$$\hat{p}(\boldsymbol{x})=\frac{1}{N}\sum_{\boldsymbol{x}_t\in D}k_\sigma(\boldsymbol{x},\boldsymbol{x}_t) \tag{2}$$

式中，$D=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$。

用 $\hat{p}(\boldsymbol{x})$ 的期望近似地表示 $V(p)$ 为

$$\hat{V}(p)=\frac{1}{N}\sum_{\boldsymbol{x}_t\in D}\hat{p}(\boldsymbol{x}_t)$$

$$=\frac{1}{N^2}\sum_{\boldsymbol{x}_t\in D}\sum_{\boldsymbol{x}_{t'}\in D}k_\sigma(\boldsymbol{x}_t,\boldsymbol{x}_{t'})=\frac{1}{N^2}\mathbf{1}^{\mathrm{T}}\boldsymbol{K}\mathbf{1} \tag{3}$$

式中，$(N\times N)$ 的矩阵 $\boldsymbol{K}$ 中标号为 (t,t') 的元素等于 $k_\sigma(\boldsymbol{x}_t,\boldsymbol{x}_{t'})$，$\mathbf{1}$ 是元素全为 1 的 $(N\times 1)$ 的列向量。由此，KECA 数据转换可以表示为

$$\boldsymbol{\Phi}_{eca}=\boldsymbol{D}_l^{\frac{1}{2}}\boldsymbol{E}_l^{\mathrm{T}}:\min_{\lambda_1e_1,\cdots,\lambda_{N'}e_N}\hat{V}(p)-\hat{V}_l(p)$$

$$=\boldsymbol{D}_l^{\frac{1}{2}}\boldsymbol{E}_l^{\mathrm{T}}:\min_{\lambda_1e_1,\cdots,\lambda_{N'}e_N}\frac{1}{N^2}\mathbf{1}^{\mathrm{T}}(\boldsymbol{K}-\boldsymbol{K}_{eca})\mathbf{1} \tag{4}$$

式中，$\hat{V}_l(p)=\frac{1}{N^2}\mathbf{1}^{\mathrm{T}}\boldsymbol{E}_l\boldsymbol{D}_l\boldsymbol{E}_l^{\mathrm{T}}\mathbf{1}=\frac{1}{N^2}\mathbf{1}^{\mathrm{T}}\boldsymbol{K}_{eca}\mathbf{1}$，$K_{eca}=\boldsymbol{E}_l\boldsymbol{D}_l\boldsymbol{E}_l^{\mathrm{T}}$。

与 PCA 或 KPCA 相比，KECA 并不一定选择核矩阵最大的那些特征值和对应的特征向量，因此得出的数据转换结果也是截然不同的。KECA 转换后的数据的不同类之间关于原点成一定的角度，揭示了数据集的类结构，因此一种基于角度的谱聚类方法被提出，它能够很好地对 KECA 转换后的数据集进行聚类，一些实验结果也证明了在某些情况下，基于 KECA 的谱聚类算法能够得出比基于 Laplacian 矩阵的谱聚类算法相当甚至更好的结果。

2 核协方差成分分析(KCCA)

2.1 D-vs-E 的定义

KECA 是基于 Parzen 窗密度估计的，其中，$k_\sigma(\boldsymbol{x},\cdot)$ 的形状对 Parzen 窗密度估计的影响并不太。为了便于分析和计算，其中的 $k_\sigma(\boldsymbol{x},\cdot)$ 可以采用核参数为 σ 的高斯核函数。因为式(5)，KECA 也可以看作是降维前后核特征空间的数据的平均向量的欧几里德长度变化的最小化问题。

$$\hat{V}(p_\sigma)=\frac{1}{N^2}\mathbf{1}^{\mathrm{T}}\boldsymbol{K}\mathbf{1}=\frac{1}{N^2}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{1}$$

$$=\langle\frac{1}{N}\sum_{\boldsymbol{x}_i\in D}\phi(\boldsymbol{x}_i),\frac{1}{N}\sum_{\boldsymbol{x}_{i'}\in D}\phi(\boldsymbol{x}_{i'})\rangle=\|\boldsymbol{m}\|^2 \tag{5}$$

式中，$\boldsymbol{m}=\frac{1}{N}\sum_{\boldsymbol{x}_i\in D}\phi(\boldsymbol{x}_i)$ 是核特征空间数据集的平均向量，设降维后的数据集为 $\boldsymbol{\Phi}_{eca}=[\boldsymbol{\Phi}_{eca}(\boldsymbol{x}_1),\cdots,\boldsymbol{\Phi}_{eca}(\boldsymbol{x}_N)]$，降维后的熵表示为

$$\hat{V}_k(p_\sigma)=\frac{1}{N^2}\mathbf{1}^{\mathrm{T}}\boldsymbol{K}_{eca}\mathbf{1}=\|\boldsymbol{m}_{eca}\|^2 \tag{6}$$

式中，$\boldsymbol{m}_{eca}=\frac{1}{N}\sum_{\boldsymbol{x}_i\in D}\phi_{eca}(\boldsymbol{x}_i)$ 是转换后的数据 $\boldsymbol{\Phi}_{eca}$ 的平均向量. 所以 KECA 数据转换还可以表示为

$$\boldsymbol{\Phi}_{eca}=\boldsymbol{D}_k^{\frac{1}{2}}\boldsymbol{E}_k^{\mathrm{T}}:\hat{V}(p_\sigma)-\hat{V}_k(p_\sigma)$$

$$=\boldsymbol{D}_k^{\frac{1}{2}}\boldsymbol{E}_k^{\mathrm{T}}:\min_{\lambda_1e_1,\cdots,\lambda_{N'}e_N}\|\boldsymbol{m}\|^2-\|\boldsymbol{m}_{eca}\|^2 \tag{7}$$

观察矩阵 $\boldsymbol{m}\boldsymbol{m}^{\mathrm{T}}$，由式(8)建立了 $\hat{V}(p_\sigma)$ 与 $\boldsymbol{m}\boldsymbol{m}^{\mathrm{T}}$ 的等价关系。

$$\int_{\boldsymbol{x}\in D}\phi^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{m}\boldsymbol{m}^{\mathrm{T}}\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\int_{\boldsymbol{x}\in D}\hat{p}_\sigma^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\hat{V}(p_\sigma) \tag{8}$$

式(8)引起了我们对核特征空间中数据集的协方差矩阵 $\frac{1}{N}\sum_{i=1}^{N}(\phi(\boldsymbol{x}_i)-\boldsymbol{m})(\phi(\boldsymbol{x}_i)-\boldsymbol{m})^{\mathrm{T}}$ 的思考，将 $\boldsymbol{m}\boldsymbol{m}^{\mathrm{T}}$ 替换成协方差矩阵，得到

$$\int_{\boldsymbol{x}\in D}\phi^{\mathrm{T}}(\boldsymbol{x})(\frac{1}{N}\sum_{i=1}^{N}(\phi(\boldsymbol{x}_i)-\boldsymbol{m})(\phi(\boldsymbol{x}_i)-\boldsymbol{m})^{\mathrm{T}})\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$=\int_{\boldsymbol{x}\in D}\frac{1}{N}\sum_{i=1}^{N}\phi^{\mathrm{T}}(\boldsymbol{x})\phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-$$

$$\int_{x\in D}\phi^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{m}\boldsymbol{m}^{\mathrm{T}}\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$
$$=\int_{x\in D}\frac{1}{N}\sum_{i=1}^{N}k_\sigma(\boldsymbol{x},\boldsymbol{x}_i)k_\sigma(\boldsymbol{x},\boldsymbol{x}_i)\mathrm{d}\boldsymbol{x}-\int_{x\in D}\hat{p}_\sigma^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{9}$$

对于高斯核函数，由于 $k_\sigma(\boldsymbol{x},\boldsymbol{x}_i)k_\sigma(\boldsymbol{x},\boldsymbol{x}_i)=k_{\frac{\sigma}{\sqrt{2}}}(\boldsymbol{x},\boldsymbol{x}_i)$，因此

$$Eq.(15)=\int_{x\in D}\frac{1}{N}\sum_{i=1}^{N}k_{\frac{\sigma}{\sqrt{2}}}(\boldsymbol{x},\boldsymbol{x}_i)\mathrm{d}\boldsymbol{x}-\int_{x\in D}\hat{p}_\sigma^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$
$$=\int_{x\in D}\hat{p}_{\frac{\sigma}{\sqrt{2}}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}-\int_{x\in D}\hat{p}_\sigma^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{10}$$

式(10)中的第二项对应于$\hat{V}(p_\sigma)$，进而对应于 $H(p_\sigma)$。式(10)中的第一项可以近似地表示为

$$\int_{x\in D}\hat{p}_{\frac{\sigma}{\sqrt{2}}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\approx\sum_{i=1}^{N}\hat{p}_{\frac{\sigma}{\sqrt{2}}}(\boldsymbol{x}_i)=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}k_{\frac{\sigma}{\sqrt{2}}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

因此，式(10)可以近似地表示为

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}k_{\frac{\sigma}{\sqrt{2}}}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}k_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{11}$$

于是式(11)导出了 D-vs-E(Densities-vs-Entropy)的概念，D-vs-E 就是核参数为 $\sigma/\sqrt{2}$ 时的密度总和与熵$\hat{V}(p_\sigma)$的差。将式(11)中的$\frac{1}{N}$约去，然后重写为 $\mathbf{1}^{\mathrm{T}}\tilde{\boldsymbol{K}}\mathbf{1}$，其中 $\tilde{\boldsymbol{K}}$ 是$N\times N$的矩阵，下标为(i,j)的元素为 $k_{\frac{\sigma}{\sqrt{2}}}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\frac{1}{N}k_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_j)$，$\mathbf{1}$ 是元素全为 1 的 $N\times 1$ 的列向量。设转换后的数据为 $\boldsymbol{\Phi}_{\alpha\sigma}=[\boldsymbol{\Phi}_{\alpha\sigma}(\boldsymbol{x}_1),\cdots,\boldsymbol{\Phi}_{\alpha\sigma}(\boldsymbol{x}_N)]$，KCCA 数据转换表示为

$$\boldsymbol{\Phi}_{\alpha\sigma}=\boldsymbol{D}_k^{\frac{1}{2}}\boldsymbol{E}_k^{\mathrm{T}}:\min_{\lambda_1\boldsymbol{e}_1,\cdots,\lambda_N\boldsymbol{e}_N}\mathbf{1}^{\mathrm{T}}(\tilde{\boldsymbol{K}}-\tilde{\boldsymbol{K}}_{\alpha\sigma})\mathbf{1} \tag{12}$$

式中，$\tilde{\boldsymbol{K}}_{\alpha\sigma}=\boldsymbol{\Phi}_{\alpha\sigma}^{\mathrm{T}}\boldsymbol{\Phi}_{\alpha\sigma}=\boldsymbol{E}_k\boldsymbol{D}_k\boldsymbol{E}_k^{\mathrm{T}}$。需要注意的一点是 $\tilde{\boldsymbol{K}}$ 可能不是半正定的矩阵，我们可以把式(11)与式(3)结合起来，这也就把 D-vs-E 与 Renyi 熵结合了起来。

2.2 广义 D-vs-E

一个合理的数据转换方法应该保证数据的散度在降维前后尽可能地保持不变。理论和实践上，协方差矩阵是数据集散度的一个比较有效的度量。从前面的分析可以看出，D-vs-E 的概念是由核特征空间的协方差矩阵导出的，所以，D-vs-E 也是一个对数据集散度合理科学的度量。除此之外，D-vs-E 的原理还可以从另外的角度来看。

首先，在统计学上，当数据降维后，反映数据集散度的核半径 σ 会有变小的趋势，这有违我们的愿望。所以只考虑有固定 σ 值的$\hat{V}(p_\sigma)$是远远不够的。观察式(11)，其中第一项采用了核半径为 $\sigma/\sqrt{2}$ 的高斯核函数，这就反映了以上所说的趋势。由于我们不能预测数据集散度缩小的速度，为了更好地体现这样一个缩小的趋势，使用多重的速率组合，并与 KECA 中的 Renyi 熵结合，得到了广义 D-vs-E。

$$\mu\sum_{i=1}^{N}\sum_{j=1}^{N}(k_{\frac{\sigma}{\alpha_1}}(\boldsymbol{x}_i,\boldsymbol{x}_j)+\cdots+k_{\frac{\sigma}{\alpha_n}}(\boldsymbol{x}_i,\boldsymbol{x}_j))+\sum_{i=1}^{N}\sum_{j=1}^{N}k_\sigma(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{13}$$

式中，$\mu>0$ 是一个调整系数，$\alpha_1,\cdots,\alpha_n>1$，$n$ 是一个正整数。根据经验得出 $n=3,\alpha_1,\cdots,\alpha_n\in[1.5,4]$时，就能得到较好的结果和可以接受的复杂度。如果没有特别的说明，KCCA 指的都是基于广义 D-vs-E 的 KCCA。

2.3 KCCA 与现有数据转换方法的比较

本质上，KCCA 和很多现有的非线性谱数据转换方法是一样的，只是核矩阵和投影轴的选择有所不同。和 KECA 相比，KCCA 具有以下特点。

- KCCA 数据转换揭示了输入空间数据集与 D-vs-E 或广义 D-vs-E 有关的数据结构，而不是与 Renyi 熵有关的数据结构；
- KCCA 并不一定选取核矩阵最大的几个特征值和对应的特征向量。这就又为我们提供了一种核成分分析方法。

后面的数据转换实验将显示出 KCCA 转换后的数据集分布像 KECA 一样具有明显的角度结构，于是同样可以用于基于角度的谱聚类算法。

从多核学习的角度来看，D-vs-E 和广义 D-vs-E 是基于不同核半径的多个核的组合。在 Parzen 窗密度估计中，核参数的选择远比核函数的形状重要[18,19]。实验也显示了 KECA 对核半径的选择非常敏感，尤其是相对来说较小的数据集。为了增强对核半径选择的鲁棒性，一种方法是采用多核学习框架，例如用多核的非负线性组合，我们可以设计多核的 KECA 如下：

$$\boldsymbol{\Phi}_{MECA}=\boldsymbol{E}_k^{\frac{1}{2}}\boldsymbol{E}_k^{\mathrm{T}}:\min_{\alpha_1,\alpha_2,\cdots,\alpha_m}\min_{\lambda_1\boldsymbol{e}_1,\cdots,\lambda_N\boldsymbol{e}_N}\mathbf{1}^{\mathrm{T}}(\boldsymbol{K}^m-\boldsymbol{K}_{\alpha\sigma}^m)\mathbf{1} \tag{14}$$

式中，$\boldsymbol{K}^m=\sum_{i=1}^{m}\alpha_i\boldsymbol{K}_{\sigma_i}$，$\boldsymbol{K}_{\sigma_i}$ 表示原 KECA 中核半径为 σ_i 时的核矩阵 $\boldsymbol{K}$，$\sum_{i=1}^{m}\alpha_i=1,\alpha_i>0,\sigma_i$ 可以随机选取。为了得到式(14)的解，需要一个两级的最小化过程，这个过程会非常繁琐，并且还多了 m 个参数。特别是，我们往往对 m 没有先前的了解。通常的处理是 m 必须是一个比较大的整数：8，10 或者更大。

把 KCCA 看作是多核数据转化方法的特例。但是，与式(14)不同，因为 KCCA 有比较明确的统计学上的解释，于是就提供了一个天然的多核学习框架，而且不需要在约束条件 $\sum_{i=1}^{m}\alpha_i=1$ 下进行两级最小化的繁琐过程。

KCCA 还有一个与众不同的特点，D-vs-E 和广义 D-vs-E 内独特的多核组合中，其它的 σ_i 必须比 Renyi 熵中的核半径小，而不是随机选取。实验结果证明 3 个核的组合是比较好的选择，这显然比 m 小得多。

2.4 KCCA 数据转换算法步骤

将数据集 $\boldsymbol{\Phi}$ 转换成 k 维的低维数据，KCCA 数据转换步骤如下：

1. 对给定的数据集 $\boldsymbol{\Phi}$，根据式(13)算出其基于广义 D-vs-E 的核矩阵 $\tilde{\boldsymbol{K}}$，其中 $\alpha_i>1,n=3$；

2. 对 $\tilde{\boldsymbol{K}}$ 进行特征分解 $\tilde{\boldsymbol{K}}=\boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^{\mathrm{T}}$，其中 $\boldsymbol{D}=diag([\lambda_1,\cdots,\lambda_N])$，$\boldsymbol{E}=[\boldsymbol{e}_1,\cdots,\boldsymbol{e}_N]$，那么 $\boldsymbol{U}=\boldsymbol{\Phi}\boldsymbol{E}\boldsymbol{D}^{-\frac{1}{2}}$ 的列向量即为 KPCA 轴；

3. 计算每个 KPCA 轴对 Renyi 熵估计的贡献：$(\sqrt{\lambda_i}\boldsymbol{e}_i^{\mathrm{T}}\mathbf{1})^2$ $(i=1,2,\cdots,N)$，选取前 k 个对 Renyi 熵估计贡献最大的 KPCA 轴组成 $\boldsymbol{U}_k$，对应的特征值构成对角阵 $\boldsymbol{D}_k$，对应的特征向量构成 $\boldsymbol{E}_k$；

4. 将数据集 $\boldsymbol{\Phi}$ 在 $\boldsymbol{U}_k$ 上投影为 $\boldsymbol{\Phi}_{\alpha\sigma}=P_{U_k}\boldsymbol{\Phi}=\boldsymbol{D}_k^{\frac{1}{2}}\boldsymbol{E}_k^{\mathrm{T}}$，$\boldsymbol{\Phi}_{\alpha\sigma}$ 即为转换得到的数据。

2.5 数据转换实验

使用 USPS，即美国邮政区号手写数字识别数据集中的数字 0 和 9 来做特征降维实验，把 256 个特征降至 2 个特征，在二维空间表示出降维后的样本点，核参数选取 3 到 10.8 的

以 0.2 为步长的一系列值，比较中心化的 KPCA、未中心化的 KPCA、KECA 和 KCCA 的数据转换结果有何异同。

图 2 中显示了未中心化的 KPCA 降维的结果。当 $\sigma \leqslant 6$ 时，一个类集中在(0,0)点，另一个类在一个方向分散开来。但是 $\sigma \geqslant 6.2$ 时，未中心化的 KPCA 和 KECA 是相同的，它们都是在前两个最大的特征值所对应的特征向量方向进行投影而达到降维的。为了实验的完整性，图 1 列出了中心化的 KPCA 的降维结果，可以看出与未中心化的 KPCA 和 KECA 的结果都不同。

图 3 中，当 $\sigma \geqslant 3.4$ 时，两个类出现了 KECA 中典型的角度结构。随着 σ 的增加，两个类的分布逐渐变得松散。

图 4 显示了 KCCA 的降维结果，与 KECA 相比，随着 σ 值的增加，样本分布的形状结构变化变得更加缓慢了。σ 值变得较大时，两个类依然有明显的角度结构，并且样本点在两个角度的方向上保持紧凑、不分散。KCCA 的降维结果也有着典型的角度结构。

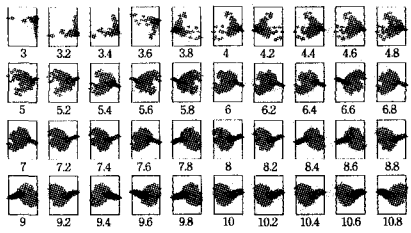

图 1　中心化的 KPCA

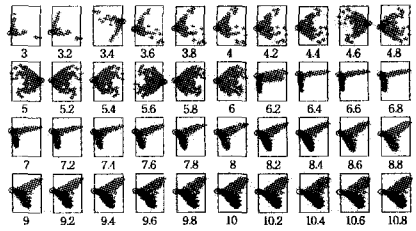

图 2　未中心化的 KPCA

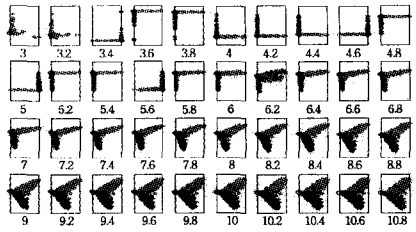

图 3　KECA

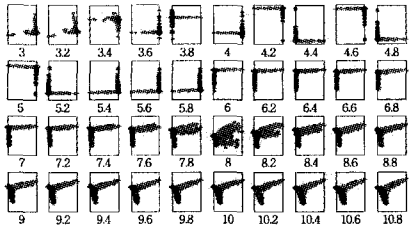

图 4　KCCA

3　KCCA 在聚类中的应用

3.1　KCCA 谱聚类

聚类属于非监督模式识别问题，它是按照某种相似程度的度量，使相似的样本归为一类，而将不相似的样本归于不同的类。聚类的过程完全依赖于样本之间的特征差别，如果能够用 KCCA 方法增加对样本特征的优化，聚类将会取得更好的结果。

数据经过 KCCA 转换后，不同的簇关于核特征空间的原点分布在不同的角度，因此聚类的价值函数必须能够捕捉到这种角度结构。概率密度函数之间的 Cauchy-Schwarz(CS)散度与核特征空间中的平均向量所成角度的余弦是一致的[12]。已知第 i 个簇 $p_i(\boldsymbol{x})$ 的概率密度函数与总体 $p(\boldsymbol{x})$ 的概率密度函数的 CS 散度为 $D_{CS}(p_i,p)=-\log V_{CS}(p_i,p)$，其中

$$V_{CS}(p_i,p)=\frac{\int p_i(\boldsymbol{x})\ p(\boldsymbol{x})\mathrm{d}x}{\sqrt{\int p_i^2(\boldsymbol{x})\mathrm{d}x\int p^2(\boldsymbol{x})\mathrm{d}x}} \tag{15}$$

假设 N_i 个数据点 $\boldsymbol{x}_n \in D_i$ 由对应簇 C_i 的 $p_i(\boldsymbol{x})$ 生成。这些数据点是 N 个由 $p(\boldsymbol{x})$ 生成的数据点 $\boldsymbol{x}_t \in D$ 的子集。通过 Parzen 窗得到

$$\hat{V}_{CS}(p_i,p)=\cos\angle(\boldsymbol{m}_i,\boldsymbol{m}) \tag{16}$$

式中，$\boldsymbol{m}_i=\frac{1}{N_i}\sum_{\boldsymbol{x}_n\in D_i}\phi(\boldsymbol{x}_n)$ 是核特征空间簇 C_i 的平均向量，$\boldsymbol{m}=\frac{1}{N}\sum_{\boldsymbol{x}_t\in D}\phi(\boldsymbol{x}_t)$ 是数据总体在核空间的平均向量。CS 散度主要是度量了这些平均向量所成角度的余弦值。

聚类的基于角度的价值函数可以表示为

$$J(C_1,\cdots,C_C)=\sum_{i=1}^{C}N_i\cos\angle(\boldsymbol{m}_i,\boldsymbol{m}) \tag{17}$$

用 KCCA 转换后的数据 $\boldsymbol{\Phi}_{\alpha a}$ 代表 $\boldsymbol{\Phi}$ 来优化基于角度的价值函数。优化的过程就是用角度距离代替欧几里德距离，在 KCCA 空间中运用 C-mean 算法，因此这种方法一定能够最终收敛到一个局部最优值。KCCA 谱聚类算法步骤如下：

1. 用 KCCA 方法得到 $\boldsymbol{\Phi}_{\alpha a}$；
2. 初始化平均向量 $\boldsymbol{m}_i$，$i=1,\cdots,C$；
3. 对所有的 t：$\boldsymbol{x}_t \to C_i$：$\max_i\cos(\phi_{\alpha a}(\boldsymbol{x}_t),\boldsymbol{m}_i)$；
4. 更新平均向量；
5. 重复步骤 3、步骤 4，直到收敛。

在步骤 2 中，初始化平均向量时，选择两个数据点 $\phi_{\alpha a}(\boldsymbol{x}_t)$ 与 $\phi_{\alpha a}(\boldsymbol{x}_{t'})$ 满足 $\min_{t,t'}\cos(\phi_{\alpha a}(\boldsymbol{x}_t),\phi_{\alpha a}(\boldsymbol{x}_{t'}))$ 来代表 $\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$。当 $i>2$ 时，选择 $\phi_{\alpha a}(\boldsymbol{x}_{t'})$ 满足 $\min_{t'}\sum_{j=1}^{i-1}\cos(\phi_{\alpha a}(\boldsymbol{x}_{t'}),\boldsymbol{m}_j)$ 来代表 $\boldsymbol{m}_i$。经验证明这是一个鲁棒性非常强的初始化方法，能在几次迭代后收敛。判断是否收敛的方法是看价值函数在前后两次迭代中的变化是否小于某一个较小的值。

3.2　KPCA、KECA 与 KCCA 在聚类中的对比实验

使用几个常用的数据集，分别用基于 KPCA、KECA 与 KCCA 的聚类算法对这些数据集进行聚类。图 5－图 10 中纵轴表示错误率，横轴表示高斯核半径 σ。KCCA 聚类时，令 $n=3$，$\alpha_1=1.5$，$\alpha_2=2.0$，$\alpha_3=4.0$，不同的数据集中 μ 的取值不同。

图 5 显示了美国邮政区号手写数字识别数据集(USPS)中的 3 个类：6，8 和 9 的聚类结果。KCCA 聚类方法中 $\mu=16$。KPCA 在 $\sigma=8.0$ 时，取得的最小错误率为 24.7%；KECA 在 $\sigma=4$ 时，取得的最小错误率为 11.6%；KCCA 在 $\sigma=9.6$ 时，取得的最小错误率为 12.6%。当 $\sigma>9.6$ 时，相当长的一段区间内 KCCA 错误率一直保持最低，KECA 其次，KPCA 波动较大且错误率较高。

图 6 是对数据集 image 的聚类结果，KCCA 聚类中 μ 取

1。KPCA 在 $\sigma=4.1$ 时，取得的最小错误率为 38.3%；KECA 在 $\sigma=0.8$ 时，取得的最小错误率为 36.3%；KCCA 在 $\sigma=0.9$ 时，取得的最小错误率为 35.85%。KCCA 在区间[5.3,13.3]的错误率不大于 39%，错误率比较低的区间明显大于 KECA 和 KPCA。

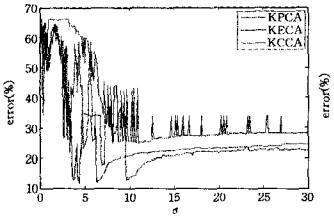

图 5 USPS

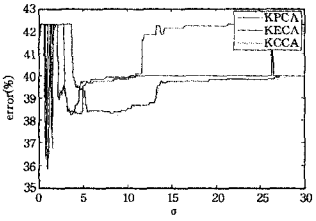

图 6 image

图 7 是对数据集 ringnorm 的聚类结果，KCCA 方法中 $\mu=10$。KPCA 在 $\sigma=2.9$ 时，取得的最小错误率为 1.3%；KECA 在 $\sigma=2.0$ 时，取得的最小错误率为 10.25%；KCCA 在 $\sigma=2.8$ 时，错误率为 1.75%，达到最小。在区间(0,2)，KCCA 的错误率最高，但在 $\sigma>8.0$ 的相当长的区间内，KCCA 的错误率最低。

图 8 是对数据集 heart 的聚类结果，KCCA 方法中 $\mu=13$。KPCA 在 $\sigma=14.9$ 时，取得的最小错误率为 14.1%；KECA 在 $\sigma=2.6$ 时，取得的最小错误率为 12.4%；KCCA 在 $\sigma=4.6$ 时，取得的最小错误率同样为 12.4%。相比之下，KPCA 的波动最大，KECA 次之，KCCA 相对来说最为稳定，并且较小错误率的区间能保持较长的宽度。

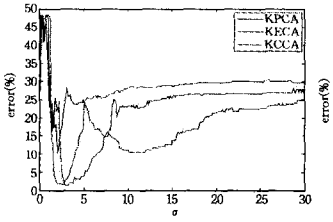

图 7 ringnorm

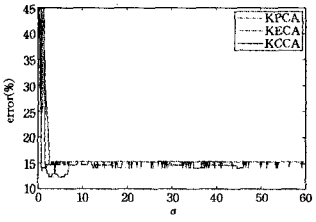

图 8 heart

图 9 是对数据集 banana 的聚类结果，KCCA 方法中 $\mu=5$。KPCA 在 $\sigma=0.4$ 时，取得的最小错误率为 40.5%；KECA 在 $\sigma=3.6$ 时，取得的最小错误率为 39.3%；KCCA 在 $\sigma=20.3$ 时，错误率为 39.5%，达到最小。虽然 KCCA 在(2,8)的区间内错误率在三者中最高，但是在 $\sigma>13$ 的相当长一段区间内，KCCA 的错误率比 KPCA 和 KECA 都小。

图 10 是对数据集 thyroid 的聚类结果，KCCA 方法中 $\mu=7$。KPCA 在 $\sigma=1.4$ 时，取得的最小错误率为 5.7%；KECA 在 $\sigma=1.1$ 时，取得的最小错误率为 5.7%；KCCA 在 $\sigma=2.4$ 时，错误率为 7.1%，达到最小。KCCA 在相当宽的区间(2.4,25)内错误率在三者中最低。

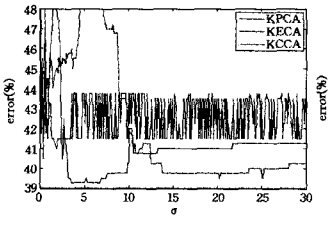

图 9 banana

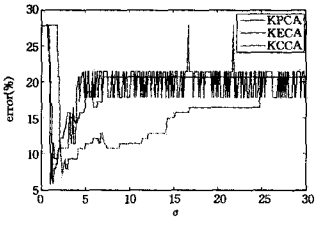

图 10 thyroid

可以很清楚地看到，KCCA 在比较宽的区间上错误率比 KPCA 和 KECA 都低。也就是说，KCCA 更稳定，这就意味着 σ 值更容易选取。以上实验说明，与 KPCA 和 KECA 相比，KCCA 对高斯核半径的选择具有更强的鲁棒性。

结束语 本文由核特征空间的协方差矩阵导出了 D-vs-E 的概念，基于 D-vs-E，提出了一种新的特征降方法 KCCA。这种新的数据转换方法具有 KECA 的优点，同时弥补了 KECA 的不足。KCCA 不仅考虑到了输入空间的 Renyi 熵，而且确保了数据集的散度在降维前后变化最小。由于它天然的多核学习框架，因此 KCCA 对高斯核半径的选择有更强的鲁棒性，基于 KCCA 的谱聚类结果也就更稳定。从时间复杂度来看，KCCA 只是在计算核矩阵时增加了时间复杂度，在性能更好的同时并没有增大核矩阵的规模。

参 考 文 献

[1] Jain A K, Mao J C, Duin R P W, et al. Statistical pattern recognition: A review [Review][J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37

[2] Hu Q, Pedrycz W, Yu D, et al. Selecting Discrete and Continuous Features Based on Neighborhood Decision Error Minimization [J]. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics, 2010, 40(1): 137-150

[3] Vasconcelos M, Vasconcelos N. Natural Image Statistics and Low-Complexity Feature Selection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 228-244

[4] Ghanty P, Pal N R. Prediction of Protein Folds: Extraction of New Features, Dimensionality Reduction, and Fusion of Heterogeneous Classifiers[J]. IEEE transactions on nanobioscience, 2009, 8(1): 100-110

[5] Doi J, Yamanaka M. Discrete Finger and Palmar Feature Extraction for Personal Authentication[J]. IEEE Transactions on Instrumentation and Measurement, 2005, 54(6): 2213-2219

[6] Cox T, Cox M. Multidimensional scaling(Second edition)[M]// Cox D R, Isham V, Keiding N, et al. Monographs on Statistics and Applied Probability 88. London: Chapman & Hall/CRC, 2001

[7] Williams C. On a connection between kernel PCA and metric multidimensional scaling[M]// Leen T, Dietterich T, Tresp V. Advances in Neural Information Processing Systems 13. Cambridge, MA: MIT Press, 2001: 675-681

[8] Scholkopf B, Smola A, Muller K. Nonlinear Component Analysis as a Kernel Eigenvalue Problem[J]. Neural Computation, 1998, 10(5): 1299-1319

[9] Scholkopf B, Smola A, Muller K. Kernel Principal Component analysis [M]. Advances in Kernel Methods-Support Vector Learning. Cambridge MA: MIT Press, 1999: 327-352

[10] Zha H, He X, Ding C, et al. Spectral Relaxation for K-means Clustering[J]. Advances in Neural Information Processing System, 2002, 2(14): 1057-1064

[11] Jenssen R. Kernel Entropy Component Analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(5): 847-880

[12] Jenssen R, Eltoft T. A New Information Theoretic Analysis of Sum-of-Squared-Error Kernel Clustering[J]. Neurocomputing, 2009, 72(1): 23-31

[13] Tenenbaun J, Silva V, Langford J. A Global Geometric Framework for Nonlinear Dimensionality Reduction[J]. Science, 2000, 290(5500): 2319-2323

[14] Weinberger K, Saul L. Unsupervised Learning of Image Manifolds by Semidefinite Programming[J]. International Journal of Computer Vision, 2006, 70(1): 77-90

[15] Roweis S, Saul L. Nonlinear Dimensionality Reduction by Locally Linear Embedding[J]. Science, 2000, 290(5500): 2323-2326

[16] Renyi A. On Measures of Entropy and Information[J]. Selected Papers of Alfred Renyi, 1976, 2: 565-580

[17] Parzen E. On the Estimation of a Probability Density Function and the Mode[J]. The Annals of Math. Statistics, 1962, 32: 1065-1076

[18] Deng Zhao-hong, Chung Fu-lai, Wang Shi-tong. FRSDE: Fast reduced set density estimator using minimal enclosing ball approximation[J]. Pattern Recognition, 2008, 41: 1363-1372

[19] Kollios G, Gunopulos D. Efficient biased sampling for approximate clustering and outlier detection in large datasets[J]. IEEE Trans. Knowledge and Date Engineering, 2003, 15(5): 1170-1187

(上接第 210 页)

由表 2 数据可以看出，每个极化方式的识别正确率都在 80%以上，说明提取的特征向量可从不同方面表征目标属性，保留较多的目标信息，有利于解决文中研究的分类问题。通过多极化融合的分类结果可知，在单极化特征向量表征目标属性的基础上，多极化融合又加入了目标的极化属性，更全面地保留了目标信息，所以目标识别正确率比单极化目标识别提高了 4%～11%。多分类器动态组合在每个极化方式的识别结果都等于或优于其它同一基分类器组合，对 4 个极化方式分类结果进行融合后，目标识别正确率提高了 2%～5%，验证了文中给出的多分类器动态组合方法的有效性。

2. 对不同方位角范围的多极化数据集进行分类，验证所提取特征向量和本文给出的 HRRP 目标识别方法对不同姿态角目标的分类性能。选择不同方位数据集进行分类，分类结果统计于表 3。

表 3 不同方位数据集的分类正确率(%)

	HH	HV	VH	VV	多极化融合
0～10°	95.32 ±1.25	99.17 ±0.56	93.95 ±1.65	98.68 ±0.39	99.81 ±0.08
20～30°	95.68 ±0.86	98.98 ±0.69	94.85 ±1.38	98.47 ±0.64	99.84 ±0.09
50～60°	97.91 ±2.26	100.00	99.21 ±0.13	100.00	100.00
80～90°	98.92 ±0.67	98.84 ±0.58	99.47 ±0.13	99.02 ±0.32	99.97 ±0.05
110～120°	96.32 ±0.50	100.00	100.00	98.55 ±0.60	100.00
140～150°	93.65 ±0.78	97.76 ±0.51	96.85 ±1.17	97.87 ±0.62	99.35 ±0.23
160～170°	94.32 ±1.25	99.37 ±0.31	94.96 ±1.28	97.35 ±0.54	99.72 ±0.11

由表 3 中的数据可知，不同方位角的分类正确率变化较小，分类效果稳定，验证了(SK_{PQ}, CV_{PQ}, L_{PQ})特征向量和本文给出的 HRRP 目标识别方法良好的分类性能。

表 4 不同特征向量的分类正确率(%)

	HH	HV	VH	VV	多极化融合
文献[4]提取特征向量	69.93 ±2.91	84.23 ±3.06	72.64 ±5.27	69.80 ±3.10	88.82 ±3.15
文献[5]提取特征向量	75.29 ±1.92	91.18 ±2.85	91.38 ±1.78	75.12 ±1.83	94.61 ±3.54
本文提取特征向量	93.03 ±0.62	94.50 ±2.38	88.04 ±2.32	90.37 ±2.20	99.07 ±0.69

3. 用文献[4]提出的目标强散射中心维数、能量聚集区长度、目标散射中心分布熵 3 种平移不变特征以及文献[5]提出的能量聚集区长度、强散射中心数目、HRRP 序列方差和一维像的“信源熵” 4 种平移不变特征分别组成的特征向量进行分类，并与本文给出的特征向量的分类结果进行比较。表 4 记录了不同特征向量的分类结果，可以看出，本文提取的(SK_{PQ}, CV_{PQ}, L_{PQ})特征向量分类正确率较高。

结束语 多极化多特征融合是提高雷达系统目标识别能力的有效途径，本文给出的基于低维平移不变特征向量和多分类器动态组合的多特征多极化融合方法，有效地降低了样本维数和计算复杂度，并通过实验验证了该方法的有效性和可行性。但是由于其提取的平移不变特征是一维的数字特征，只能表征部分目标属性，虽然多特征多极化融合能在一定程度上弥补目标信息的损失，但仍然会造成部分目标信息丢失，不利于对目标的正确识别。因此，提取更有效的目标极化特征向量，提出更合理的多极化多特征融合方法是下一步研究的方向。

参 考 文 献

[1] 吴顺君，梅晓春. 雷达信号处理和数据处理技术[M]. 北京：电子工业出版社，2008

[2] Jacobs S P. Automatic target recognition using high-resolution radar range profiles[D]. Dissertation, Washington University, 1999

[3] 代大海. 极化雷达成像及目标特征提取研究[D]. 长沙：国防科学技术大学，2008

[4] 许人灿，姜卫东，陈曾平. 目标一维距离像特征提取方法研究[J]. 系统工程与电子技术，2005，27(7)：1173-1174，1191

[5] 徐庆，王秀春，李青. 基于高分辨一维像的目标特征提取方法[J]. 现代雷达，2009，31(6)：60-63

[6] 曹向海，刘宏伟，吴顺君. 多极化多特征融合的雷达目标识别研究[J]. 系统工程与电子技术，2008，30(2)：261-264

[7] 李丽亚，刘宏伟，纠博，等. 基于核函数的多极化 HRRP 识别[J]. 西安电子科技大学学报：自然科学版，2010，37(1)：49-55

[8] 杜兰. 雷达高分辨率距离像目标识别方法研究[D]. 西安：西安电子科技大学，2007

[9] 胡可云，田凤占，黄宽厚. 数据挖掘理论与应用[M]. 北京：清华大学出版社，北京交通大学出版社，2008

[10] 陈冰，张化祥. 集成学习的多分类器动态组合方法[J]. 计算机工程，2008，34(24)：218-220

[11] Sun J, Li H. Listed companies' financial distress prediction based on weighted majority voting combination of multiple classifiers[J]. Expert Systems with Applications, 2008, 35(3): 818-827

[12] 边肇祺，张学工，等. 模式识别(第二版)[M]. 北京：清华大学出版社，2000