核最优变换与聚类中心的算法

赵 峰[1,2]，张军英[2]，刘 敬[2]
(1. 山东工商学院 信电学院，山东 烟台 264005；
2. 西安电子科技大学 计算机学院，陕西 西安 710071)

摘要：基于核化原理，提出核最优变换与聚类中心算法. 算法通过非线性变换，将数据映射到核空间，并在核空间中执行最优变换与聚类中心算法. 该算法可提取稳健的非线性鉴别特征，解决复杂分布数据的模式分类问题. 同时，基于训练样本在核空间所张成的子空间的一组基，提出一个快速提取鉴别特征的计算方法，解决了一般核方法面临的“大训练集”难题. 基于 IRIS，YEAST，GLASS 等数据的分类实验验证了该方法的有效性.

关键词：核方法；最优聚类中心；最优变换
中图分类号：TP301 **文献标识码**：A **文章编号**：1001-2400(2009)01-0127-07

Kernel optimal transformation and cluster centers algorithm

ZHAO Feng[1,2], *ZHANG Jun-ying*[2], *LIU Jing*[2]
(1. School of Info. and Electro. Eng., Shandong Inst. of Business and Tech., Yantai 264005, China;
2. School of Computer Science and Technology, Xidian Univ., Xi'an 710071, China)

Abstract: The kernel optimal transformation and cluster centers algorithm (KOT-CC) is presented by using kernel methods. In the KOT-CC, all data are mapped to a kernel space via some nonlinear mapping and the optimal transformation and cluster centers (OT-CC) is performed in the kernel space. KOT-CC is a powerful technique for extracting nonlinear discriminant features and is very effective in solving pattern recognition problems which have serious overlap between the patterns of different classes. A fast algorithm for KOT-CC is also proposed based on the basis of the sub-space which is spanned by the training samples mapped into the kernel space, which can improve the efficiency of the feature extraction process and tackle the "large sample size" problem which many kernel methods may suffer from. The experiments based on the data of IRIS, YEAST, GLASS and so on, demonstrate the validity of the proposed new algorithm.

Key Words: kernel methods; optimal cluster centers; optimal transformation

模式识别已在诸如生物学、生理学、药学、计算机视觉、人工智能、图像处理等各种工程和科学领域成为越来越重要的问题[1]. 如何抽取数据的有效鉴别特征则是模式识别的一个关键环节. 特征提取的方法众多，基于样本类间距与类内距的特征提取是一种常用的有效方法，基本思想是寻找一组投影坐标轴(或投影向量、投影方向)，使得训练样本集沿这些投影轴投影后，尽量使得同类样本的投影点聚集，而异类样本的投影点分散. Fisher 判别分析(FLDA)是该类方法的典范代表，其实质是寻找一组投影向量使得训练样本集沿这些投影轴投影后，类间散度与类内散度之比达到最大. 目前该方法已经得到广泛推广与应用[2]. 但 Fisher 准则函数是从整体上考虑类内聚集性与类间差异性，反映的是“平均信息”或“总体信息”，可能为了保证类内数据更加聚集而牺牲类间的可分性，造成某些类别出现混叠现象，使得原先比较靠近的不同类别的数据投影后变得更加难以区分，容错能力受到一定限制.

收稿日期：2008-01-19
基金项目：国家自然科学基金资助(60574039，60371044)；国家部委预研项目资助(413070501)
作者简介：赵 峰(1974-)，男，西安电子科技大学博士研究生，E-mail：zhaofeng1016@126.com.

最优变换和聚类中心算法(OT-CC)[3],也属于基于样本类间距与类内距的特征提取方法,但思路与FLDA 不同. 其基本思想是通过一个变换矩阵(实质上矩阵的每行代表一个投影方向),使得变换后的每类数据分别聚集于预先指定的聚类中心,同时保持聚类中心之间尽可能分离,即优先考虑异类之间的差异性. 但OT-CC 本质上属于线性算法,只能提取数据的线性特征,而不能提取数据的非线性特征. 同时求解过程涉及矩阵的求逆运算,计算复杂度为 $O(N^3)$,其中 N 为数据的维数,因此对于"高维小样本"数据,类内散布矩阵面临奇异性问题,更主要的是求逆运算困难甚至不能实现.

基于统计学习理论的核学习算法[4-5],近年来受到越来越多的关注. 基本思想是通过一个非线性变换,将原空间的非线性问题转化为核空间上的线性问题,然后在核空间采用一些线性算法处理问题,间接解决了原空间的非线性问题. 其技巧在于无需明确非线性变化的具体形式,而是借助非线性变换的内积,即核函数,将实际计算保留在原空间进行. 该方法为人们开辟了一种解决非线性问题的新思路,目前已经应用到数据分析中的很多线性算法,例如,支撑向量机[6](SVM)、核主分量分析[7](KPCA)、核 Fisher 判别分析[8](KFDA)等都是基于相应的线性算法,通过核化原理而提出的.

笔者首先对 OT-CC 进行了深入分析,指出了文献[3]在算法推导上存在的疏漏,并就其计算复杂度、特征提取能力等方面进行了分析,然后基于核化原理,提出一个新的非线性鉴别特征提取算法:核最优变换和聚类中心算法(KOT-CC). 即通过一个非线性变换,将数据映射到高维核空间,然后采用 OT-CC 处理数据,实现原空间数据的非线性特征提取.

KOT-CC 的鉴别特征提取能力,特别是非线性鉴别特征的提取能力比 OT-CC 有显著提高,甚至优于经典的 KFDA. 其计算复杂度与样本的数目 n 相关,为 $O(n^3)$,而同原空间数据的维数 N 无关,比较适用于"高维小样本"数据集的特征提取.

同时也注意到,对于大训练集,即 n 很大时,KOT-CC 同许多其他核方法[6-8]一样,计算困难. 为此,笔者基于训练集在核空间的空间结构,用训练集在核空间中所张成的子空间的一组基来表达最优变换矩阵,给出KOT-CC 的一个快速算法(FKOT-CC),该算法计算复杂度仅为 $O(r^3)$(其中 r 表示基的个数,一般情况下,$r \ll n$),解决了目前许多核方法面临的"大训练集"问题,加快了鉴别特征提取速度.

1 核最优变化与聚类中心

OT-CC 只是一个线性变换[3],只能提取数据的线性鉴别特征,而对非线性可分数据以及分布较为复杂的实测数据分类效果很不理想. 这里基于核化原理,首先通过一个非线性映射将输入空间的数据映射到另一空间,即所谓的核空间,然后在核空间中执行 OT-CC 算法. 将该算法简记为 KOT-CC. 下面给出详细的推导过程.

1.1 KOT-CC 算法推导

令 ϕ 表示原空间 R^N 到核空间 F 的一个非线性变换,即

$$\boldsymbol{x} \in R^N \xrightarrow{\phi} \boldsymbol{\phi}(\boldsymbol{x}) \in F \quad . \tag{1}$$

设 $\boldsymbol{W} = (w_1, w_2, \cdots, w_{g-1})^{\mathrm{T}}$ 为 F 中的一些列矢量 $\boldsymbol{w}_l (l = 1, 2, \cdots, g-1)$ 所构成的矩阵,作如下变换

$$\boldsymbol{y}_{ij} = \boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{x}_{ij}) + \boldsymbol{b} \quad . \tag{2}$$

则最优变换的目标就是寻求一个矩阵 $\boldsymbol{W}$ 及列矢量 $\boldsymbol{b}$,满足

$$J(\boldsymbol{W}, \boldsymbol{b}) = \underset{\boldsymbol{W},\boldsymbol{b}}{\operatorname{argmin}} \left(\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{u}_i - \boldsymbol{y}_{ij})^{\mathrm{T}} (\boldsymbol{u}_i - \boldsymbol{y}_{ij}) \right) \quad , \tag{3}$$

其中 $\boldsymbol{u}_i$ 为 $g-1$ 维列向量. 称 $\boldsymbol{W}$ 为最小平方误差准则下的最优变换矩阵. 式(3) 表明,经最优变换后,第 i 类数据的子像聚集于 $\boldsymbol{u}_i$ 的周围,所以称 $\boldsymbol{u}_i$ 为第 i 类目标在子像空间的聚类中心.

由 Mercer 定理[9],式(3)的最优解 $\boldsymbol{W}$ 的每一个行矢量,即 $\boldsymbol{w}_l^{\mathrm{T}}$,必然落在所有训练样本 $\boldsymbol{\phi}(\boldsymbol{x}_i)$ 所张成的子空间中,即存在 $\boldsymbol{\alpha}_l = (a_{l1}, a_{l2}, \cdots, a_{\mathrm{ln}})^{\mathrm{T}}$,满足

$$\boldsymbol{w}_l = \sum_{i=1}^{n} a_{li} \boldsymbol{\phi}(\boldsymbol{x}_i) \quad , \tag{4}$$

其中 $\boldsymbol{x}_i$ 表示输入空间的第 i 个训练样本,n 为总训练样本个数.将式(4)带入式(2),并在运算中用核函数表示向量的内积,即 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=<\boldsymbol{\phi}(\boldsymbol{x}_i),\boldsymbol{\phi}(\boldsymbol{x}_j)>$,则得到

$$\boldsymbol{y}_{ij}=\boldsymbol{B}\boldsymbol{k}_{ij}+\boldsymbol{b}\quad, \tag{5}$$

其中 $\boldsymbol{B}=(\alpha_1,\alpha_2,\cdots,\alpha_{g-1})^{\mathrm{T}}$ 为 $(g-1)\times n$ 矩阵,$\boldsymbol{k}_{ij}=(k(\boldsymbol{x}_1,\boldsymbol{x}_{ij}),k(\boldsymbol{x}_2,\boldsymbol{x}_{ij}),\cdots,k(\boldsymbol{x}_n,\boldsymbol{x}_{ij}))^{\mathrm{T}}$.此时,式(3)的优化问题转化为

$$J(\boldsymbol{B},\boldsymbol{b})=\underset{\boldsymbol{B},\boldsymbol{b}}{\operatorname{argmin}}\Big(\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{u}_i-\boldsymbol{y}_{ij})^{\mathrm{T}}(\boldsymbol{u}_i-\boldsymbol{y}_{ij})\Big)\quad. \tag{6}$$

定义 $\boldsymbol{k}_{ij}=(k(\boldsymbol{x}_1,\boldsymbol{x}_{ij}),k(\boldsymbol{x}_2,\boldsymbol{x}_{ij}),\cdots,k(\boldsymbol{x}_n,\boldsymbol{x}_{ij}))^{\mathrm{T}}$ 为核样本,结合式(5)与式(6),有

$$\begin{cases}\boldsymbol{B}=\boldsymbol{K}_{VW}\boldsymbol{K}_{WW}^{-1}\quad,\\ \boldsymbol{b}=\bar{\boldsymbol{u}}-\boldsymbol{B}\bar{\boldsymbol{k}}\quad,\end{cases} \tag{7}$$

其中

$$\begin{cases}\bar{\boldsymbol{u}}=\dfrac{1}{n_1+n_2+\cdots+n_g}\displaystyle\sum_{i=1}^{g}n_i\boldsymbol{u}_i\quad, & \bar{\boldsymbol{k}}=\dfrac{1}{n_1+n_2+\cdots+n_g}\displaystyle\sum_{i=1}^{g}\sum_{j=1}^{n_i}\boldsymbol{k}_{ij}\quad,\\ \boldsymbol{K}_{VW}=\displaystyle\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{u}_i-\bar{\boldsymbol{u}})(\boldsymbol{k}_{ij}-\bar{\boldsymbol{k}})^{\mathrm{T}}\quad, & \boldsymbol{K}_{WW}=\displaystyle\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{k}_{ij}-\bar{\boldsymbol{k}})(\boldsymbol{k}_{ij}-\bar{\boldsymbol{k}})^{\mathrm{T}}\quad.\end{cases} \tag{8}$$

需要说明的是式(7)需要对 $\boldsymbol{K}_{WW}$ 求逆运算,为了保证可逆,实际计算中采用正则化技术,用 $\bar{\boldsymbol{K}}_{WW}=\boldsymbol{K}_{WW}+\mu\boldsymbol{I}$ 取代 $\boldsymbol{K}_{WW}$,其中 μ 为一个较小的正数,$\boldsymbol{I}$ 为对应的单位矩阵.在这里的实验中取 $\mu=0.001$.

而最优聚类中心的选取,则仍然同 OT-CC 的选取方法一致[3].下面给出最优聚类中心的具体求解步骤:① $\boldsymbol{u}_1=(1,0,\cdots,0)^{\mathrm{T}}$;② $\boldsymbol{u}_2=(a_{21},a_{22},0,\cdots,0)^{\mathrm{T}}$,其中 a_{21},a_{22} 由条件 $\|\boldsymbol{u}_2\|=1$,$<\boldsymbol{u}_1,\boldsymbol{u}_2>=-1/(g-1)$ 确定;③ $\boldsymbol{u}_3=(a_{31},a_{32},a_{33},0,\cdots,0)^{\mathrm{T}}$.其中 a_{31},a_{32},a_{33} 由 $\|\boldsymbol{u}_3\|=1$,$<\boldsymbol{u}_1,\boldsymbol{u}_3>=<\boldsymbol{u}_2,\boldsymbol{u}_3>=-1/(g-1)$ 确定;④ 重复以上过程,一直求得 $\boldsymbol{u}_{g-1}=(a_{g-1,1},a_{g-1,2},\cdots,a_{g-1,g-1})^{\mathrm{T}}$;⑤ $\boldsymbol{u}_g$ 由 $\sum_{i=1}^{g}\boldsymbol{u}_i=0$ 得到.

确定了最优聚类中心,进而就可以按照式(7)得到最优变换.这样,对于输入样本 $\boldsymbol{x}$,由式(5)得到其对应的核空间的子像 $\boldsymbol{y}$ 为

$$\boldsymbol{y}=\boldsymbol{B}\boldsymbol{k}_x+\boldsymbol{b}\quad, \tag{9}$$

其中 $\boldsymbol{k}_x=(k(\boldsymbol{x}_1,\boldsymbol{x}),k(\boldsymbol{x}_2,\boldsymbol{x}),\cdots,k(\boldsymbol{x}_n,\boldsymbol{x}))^{\mathrm{T}}$,进而可依据子像 $\boldsymbol{y}$ 为识别特征,采用一定的分类方法进行分类.由 KOT-CC 的推导过程知道,无需明确非线性变换 $\boldsymbol{\phi}(\cdot)$,因为整个的求解过程只用到变换的内积函数,即核函数,这正是核化原理的巧妙所在.

1.2　性能比较分析

由 1.1 节知,KOT-CC 实质上是通过非线性映射加线性算法——OT-CC,间接实现了数据的非线性鉴别特征提取,但基本思想都是将数据经过最优变换,使得变换后的各类子像尽可能的聚集到特定的聚类中心,保证同类间的"趋同性",而每类的聚类中心却提前指定并满足两两间的距离相等且尽可能的分离,目的是增大异类间的"差异性".而经典的 FLDA 以及相应的核化方法——KFDA 同样是寻求一些投影方向,或者说一个变换矩阵,使得数据投影后的子像,类间离散度与类内离散度之比最大,从整体上考虑类间差异性与类内"趋同性",因此对于多分类问题,可能造成这种局面:某些类别间的分离明显,而忽视了一些类别,特别是数据少的类别间的差异性.而 OT-CC 以及 KOT-CC,则优先考虑类间距,保持各类的间距相等.

就计算复杂度而言,OT-CC,FLDA 都要对类内散布矩阵求逆运算[3],计算复杂度为 $O(N^3)$,其中 N 为样本维数,因此对于"高维小样本"数据,类内散布矩阵面临奇异性问题,更主要的是求逆运算困难甚至不能实现.而 KOT-CC,KFDA 的计算复杂度为 $O(n^3)$,(见式(7)),与训练样本数目 n 有关,与维数 N 无关,所以不存在"小样本问题",但同时却陷入"大样本问题",即训练数据多时计算困难.

2　KOT-CC 的快速算法

由 1.2 节知,KOT-CC 的计算复杂度为 $O(n^3)$,对于大训练集数据计算困难.基于训练集在核空间的空

间结构,给出一种 KOT-CC 的快速算法,方便描述起见,简称 FKOT-CC.

2.1 FKOT-CC 的推导

由式(4)可以看出,在核空间中,所有的训练样本 $\boldsymbol{\phi}(\boldsymbol{x}_i)(1\leqslant i\leqslant n)$ 都参与变换矩阵 $\boldsymbol{W}$ 的表达. 事实上,训练样本 $\boldsymbol{\phi}(\boldsymbol{x}_i)$ $(1\leqslant i\leqslant n)$ 在核空间所张成的子空间$\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$,其空间结构可由该子空间的一组基来捕获[10]. 基的个数 r,一般情况下满足$r\ll n$. 设 $\boldsymbol{\phi}(\boldsymbol{x}_{b1}),\boldsymbol{\phi}(\boldsymbol{x}_{b2}),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_{br})$ 为子空间$\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$ 的一组基,则最优变换矩阵 $\boldsymbol{W}$ 的每一个行矢量 $\boldsymbol{w}_l^{\mathrm{T}}$(见式(2)) 可由 $\boldsymbol{\phi}(\boldsymbol{x}_{b1}),\boldsymbol{\phi}(\boldsymbol{x}_{b2}),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_{br})$ 线性表示

$$\boldsymbol{w}_l=\sum_{i=1}^{r}t_{li}\boldsymbol{\phi}(\boldsymbol{x}_{bi}) \quad . \tag{10}$$

将式(10)代入式(2),并结合式(3),容易推导出:对任意输入数据 $\boldsymbol{x}$, 最优变换为

$$\boldsymbol{y}=\widetilde{\boldsymbol{B}}\tilde{\boldsymbol{k}}_x+\tilde{\boldsymbol{b}} \quad , \tag{11}$$

其中 $\tilde{\boldsymbol{k}}_x=(k(\boldsymbol{x}_{b1},\boldsymbol{x}),k(\boldsymbol{x}_{b2},\boldsymbol{x}),\cdots,k(\boldsymbol{x}_{br},\boldsymbol{x}))^{\mathrm{T}}$,而 $\widetilde{\boldsymbol{B}},\tilde{\boldsymbol{b}}$ 满足

$$\begin{cases}\widetilde{\boldsymbol{B}}=\widetilde{\boldsymbol{K}}_{VW}\widetilde{\boldsymbol{K}}_{WW}^{-1} & ,\\ \tilde{\boldsymbol{b}}=\bar{\boldsymbol{u}}-\widetilde{\boldsymbol{B}}\,\bar{\tilde{\boldsymbol{k}}} & ,\end{cases} \tag{12}$$

其中

$$\begin{cases}\bar{\boldsymbol{u}}=\dfrac{1}{n_1+n_2+\cdots+n_g}\displaystyle\sum_{i=1}^{g}n_i\boldsymbol{u}_i \quad , \qquad \bar{\tilde{\boldsymbol{k}}}=\dfrac{1}{n_1+n_2+\cdots+n_g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\tilde{\boldsymbol{k}}_{ij} \quad ,\\ \widetilde{\boldsymbol{K}}_{VW}=\displaystyle\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{u}_i-\bar{\boldsymbol{u}})(\tilde{\boldsymbol{k}}_{ij}-\bar{\tilde{\boldsymbol{k}}})^{\mathrm{T}} \quad , \qquad \widetilde{\boldsymbol{K}}_{WW}=\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\tilde{\boldsymbol{k}}_{ij}-\bar{\tilde{\boldsymbol{k}}})(\tilde{\boldsymbol{k}}_{ij}-\bar{\tilde{\boldsymbol{k}}})^{\mathrm{T}} \quad ,\\ \tilde{\boldsymbol{k}}_{ij}=(k(\boldsymbol{x}_{b1},\boldsymbol{x}_{ij}),k(\boldsymbol{x}_{b2},\boldsymbol{x}_{ij}),\cdots,k(\boldsymbol{x}_{br},\boldsymbol{x}_{ij}))^{\mathrm{T}} \quad .\end{cases} \tag{13}$$

由式(12)知,快速算法的计算复杂度为 $O(r^3)$,仅与子空间$\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$ 的基的个数有关,大大降低了 KOT-CC 的计算复杂度,更重要的是测试阶段的特征提取速度也显著改善,这是因为 FKOT-CC 进行特征提取,只需计算 r 个核函数(见式(11)),而 KOT-CC 则需要计算 n 个核函数(见式(9)).

2.2 子空间基的确定

基于子空间 $\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$ 的一组基$\boldsymbol{\phi}(\boldsymbol{x}_{b1}),\boldsymbol{\phi}(\boldsymbol{x}_{b2}),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_{br})$, 2.1 节给出了 KOT-CC 的一个快速算法. 下面需要解决的问题是: 如何寻找 $\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$ 的一组基. 因为非线性变换 $\boldsymbol{\phi}$ 只是以隐式形式出现,无法直接明确知道 $\boldsymbol{\phi}(\boldsymbol{x}_i)$,因此$\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$ 的基不能显式给出. 从向量间的线性相关性理论考虑,给出子空间 $\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$ 一组基的确定办法. 假设利用 t $(t\leqslant N)$ 个样本$\{\boldsymbol{x}_i\}_{i=1}^{t}$完成训练,得到子空间$\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{i=1}^{t}$的一组基 $\boldsymbol{\phi}(\boldsymbol{x}_{b1}),\boldsymbol{\phi}(\boldsymbol{x}_{b2}),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_{bs})$,对于新的样本 $\boldsymbol{x}$,判定 $\boldsymbol{\phi}(\boldsymbol{x}_{b1}),\boldsymbol{\phi}(\boldsymbol{x}_{b2}),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_{bs}),\boldsymbol{\phi}(\boldsymbol{x})$ 是否线性无关?如果无关,则令$\boldsymbol{\phi}(\boldsymbol{x}_{b,s+1})=\boldsymbol{\phi}(\boldsymbol{x})$,构成一个新的线性无关组 $\boldsymbol{\phi}(\boldsymbol{x}_{b1}),\boldsymbol{\phi}(\boldsymbol{x}_{b2}),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_{bs}),\boldsymbol{\phi}(\boldsymbol{x}_{b,s+1})$. 当遍历所有训练样本时,所求的线性无关组 $\boldsymbol{\phi}(\boldsymbol{x}_{b1}),\boldsymbol{\phi}(\boldsymbol{x}_{b2}),\cdots,\boldsymbol{\phi}(\boldsymbol{x}_{br})$ 即为子空间$\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$ 的一组基. 下面给出具体算法步骤如下:

(i) 初始化 在训练集 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ 任选一样本 $\boldsymbol{x}$,满足 $k(\boldsymbol{x},\boldsymbol{x})\neq 0$,令$S=\{\boldsymbol{x}\}$,$D=\{\boldsymbol{x}\}$,$G=1/k(\boldsymbol{x},\boldsymbol{x})$,$t=1$.

(ii) 如果 $t=N$,则输出 D, 终止程序. 否则下一步.

(iii) 在 $\bar{S}=X-S$ 中任选一样本 $\boldsymbol{x}^*$,令 $S=S\cup\{\boldsymbol{x}^*\}$,$t=t+1$; 并验证下式是否成立

$$\boldsymbol{k}_{tt}-\boldsymbol{k}_{st}^{\mathrm{T}}G\boldsymbol{k}_{st}=0 \quad , \tag{14}$$

其中 $\bar{\boldsymbol{x}}_i\in D$,$(\boldsymbol{k}_{st})_i=k(\bar{\boldsymbol{x}}_i,\boldsymbol{x}^*)$,$\boldsymbol{k}_{tt}=k(\boldsymbol{x}^*,\boldsymbol{x}^*)$.

(iv) 如果式(14)成立,返回(ii); 否则令

$$D=D\cup\{\boldsymbol{x}^*\},\quad G=\frac{1}{\boldsymbol{k}_{tt}-\boldsymbol{k}_{st}^{\mathrm{T}}G\boldsymbol{k}_{st}}\begin{bmatrix}(\boldsymbol{k}_{tt}-\boldsymbol{k}_{st}^{\mathrm{T}}G\boldsymbol{k}_{st})G+G\boldsymbol{k}_{st}\boldsymbol{k}_{st}^{\mathrm{T}}G & -G\boldsymbol{k}_{st}\\ -\boldsymbol{k}_{st}^{\mathrm{T}}G & 1\end{bmatrix} \quad ,$$

返回(ii).

程序终止时,集合 D 中的向量经非线性变换$\boldsymbol{\phi}$后,即为子空间$\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant M}$ 的一组基. 需要说明的是考虑到实际的计算误差,采用式(14) 进行线性无关性判别式,采用$\boldsymbol{k}_{tt}-\boldsymbol{k}_{st}^{\mathrm{T}}G\boldsymbol{k}_{st}\leqslant\varepsilon$ 取代式(14),其中 ε 是一个小的正数,在这里的实验中,$\varepsilon=0.01$.

3　实验分析

对一些仿真数据及实测数据进行鉴别特征提取,然后采用最小距离法进行识别,以讨论文中所提方法的性能.其中最小距离法是:以每类训练样本的鉴别特征的平均作为识别模板,比如 KCT-OO,其识别模板为每类训练样本经最优变换后的子像的平均,则各类目标的模板库为 $(\overline{y_1},\overline{y_2},\cdots,\overline{y_g})$,对于待测样本 $\boldsymbol{x}$,设其对应的鉴别特征,即子像为 $\boldsymbol{y}$,若 $i=\underset{k=1,2,\cdots,g}{\operatorname{argmin}}\|\boldsymbol{y}-\overline{y_k}\|$,则 $\boldsymbol{y}$ 属于第 i 类.

3.1　鉴别特征提取能力分析

对于复杂的非线性分类问题,比如双螺旋问题,由于 OT-CC 是线性算法,因此无法对双螺线进行分类,而 KOT-CC 却可以很好的提取相应的鉴别特征.

图 1 显示的是 KOT-CC 及 FKOT-CC 对双螺旋问题所获得的分类边界示意图.其中双螺旋数据的获取采用如下方法获取.第 1 类样本集为:$\{\boldsymbol{x}_i^1\}_{i=1,2,\cdots,108}$,其中 $\boldsymbol{x}_i=\{\rho_i\cos\theta_i,\rho_i\sin\theta_i\}$ 且 $\rho_i=\exp(\alpha\theta_i)$,$\theta_i=i\Delta\theta$,$\alpha=0.1$,$\Delta\theta=-\pi/18$.第 2 类样本集是第 1 类的样本关于原点的对称点的集合,即 $\boldsymbol{x}_i^2=-\boldsymbol{x}_i^1$.两类的数据图像见图 1.所选用的核函数为参数 $\sigma=0.04$ 的高斯核函数(RBF).由图 1 可以看出,KOT-CC 与 FKOT-CC 所获取的分类边界几乎相同,它们都可以对样本进行正确分类而且几乎处于两类模式样本的中间,分类曲线也比较光滑,说明了 KOT-CC 与 FKOT-CC 能够较好的提取稳健的非线性鉴别特征.

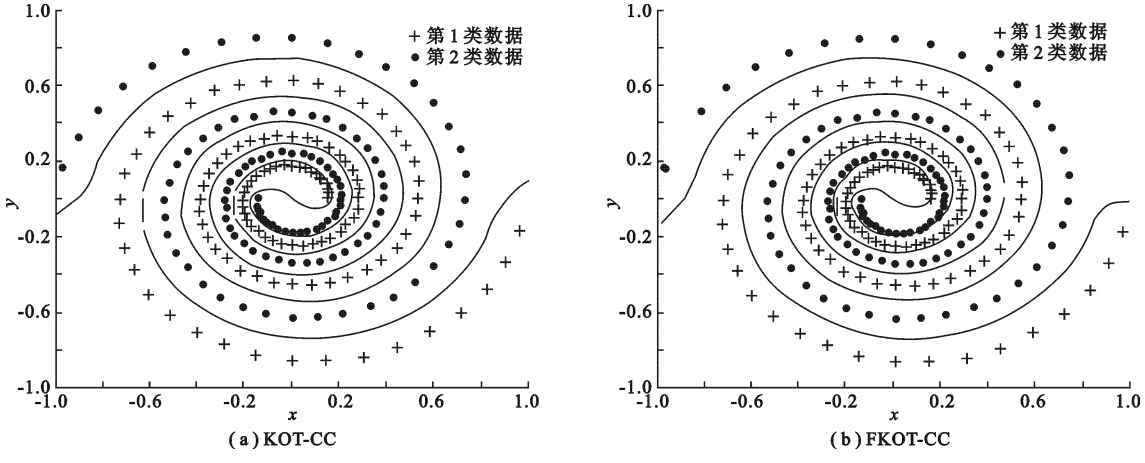

图 1　KOT-CC 与 FKOT-CC 在双螺旋数据上的分类曲线

3.2　计算性能分析

3.2.1　高维小样本实验

人脸识别是目前模式识别研究领域中极具挑战性的研究问题,受到人们的普遍重视和广泛研究[11],它是一个典型的"小样本问题".从 Yale Face Database B 数据库中选取 4 个典型人脸,所选对象包括了不同肤色,不同性别,比较具有代表性.每个人物的人脸包裹 64 幅正面位置的图像,图像的光照强度、拍摄的俯仰角等均不相同.每幅图像的大小为192(w)×164(h),具有 256 个灰度级.实验中,采用等间隔从每类中选取一半作为训练数据,其余作为测试数据.

由前面的分析可知,OT-CC 的计算复杂度为 $O(N^3)$,其中 N 为样本维数.由于人脸数据维数高达192×164 维,因此 OT-CC 需要处理一个192×164 阶矩阵的逆运算问题,其计算量是惊人的甚至不可能求解,而且如此高维的数据压缩到一个维数与类别数相当的空间,所提取的鉴别特征也不理想.限于计算机的内存空间,无法给出 OT-CC 的计算结果.

而 KOT-CC 却可以很好的解决这个问题,其求解过程只需求解一个阶数等于训练样本数的逆矩阵,计算复杂度为 $O(n^3)$(见式(7)),与训练样本数目 n 有关,而与维数 N 无关,不存在"小样本问题".在笔者的实验中,训练样本数仅为 128 个.同样,FKOT-CC 对于"小样本问题",其计算复杂度更小,因为 FKOT-CC 的

计算法杂度为 $O(r^3)$,仅与子空间$\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$ 基的个数有关,一般情况下$r\ll n$. 在这里的实验中,基的个数为 92 个. KOT-CC 与 FKOT-CC 的训练样本识别率与测试样本识别率相同,分别为 100%与 96.25%. 实验中所选用的核函数为参数 $\sigma=3\times10^7$ 的 RBF.

3.2.2 大训练数据实验

表 1 是在相同实验平台下,KOT-CC/FKOT-CC 在四高斯分布分类问题上的识别情况比较. 四类高斯分布数据分别由 $N(0,4;1)$,$N(4,0;1)$;$N(0,0;1)$,$N(4,4;0.25)$ 所生成的样本的分类问题. 其中 $N(a,b;c)$ 表示均值为(a,b),方差为 c 的高斯分布. 选取不同数目的训练数据进行训练,而测试数据则不变,均为 100.

表 1 KOT-CC,FKOT-CC 在高斯分布数据上的识别情况比较

训练样本数 n	基的个数 r	测试时间/s		训练识别率/%		测试识别率/%		核参数
		KOT-CC	FKOT-CC	KOT-CC	FKOT-CC	KOT-CC	FKOT-CC	
400	192	2.01	0.61	99	99	99	98	0.1
800	259	2.63	0.74	99	99	97	96	0.1
1 200	281	5.50	0.83	99	98	96	97	0.1
2 000	324	9.22	0.95	99	99	100	98	0.1

由表 1 可以看出:(1)二者的识别率基本一致,小的差别主要由计算误差引起的.(2)随着训练数据 n 的增多,基的个数 r 逐渐增加,但增长速度明显减慢,如当 n 由1 200 增加到2 000 时,r 由 281 增加到 324,仅仅增加了 43 个. 这是因为随着数据的增加,数据间的相关性增强. 反映到计算量上,FKOT-CC 的计算量较 KOT-CC 明显改善. 比如当$n=2\,000$ 时,KOT-CC 需要对一个阶数为2 000的矩阵求逆,而 FKOT-CC 则只需对一个阶数为 324 的矩阵求逆.(3)就测试速度而言,相比 KOT-CC,FKOT-CC 的测试速度显著提高,特别是 n 较大时,优势更加明显. 比如当$n=2\,000$ 时,FKOT-CC 的测试速度几乎是 KOT-CC 测试速度的 10 倍($9.22/0.95\approx9.71$). 这是因为提取待测数据的鉴别特征,KOT-CC 需要计算 n 个核函数(见式(9)). 而 FKOT-CC 则只需计算 r 个核函数(见式(11)).

3.3 与核 Fisher 鉴别分析的比较

为了进一步验证 KOT-CC/FKOT-CC 的性能,选择了一些不同领域的实测数据,分别采用 OT-CC,KOT-CC,FKOT-CC 以及经典的特征提取方法 KFDA 进行特征提取,然后采用最小距离法进行识别.

3.3.1 实测数据的描述

Iris 数据是从 3 种类型的鸢尾属植物每种 50 个样本所获得的植物萼片长度、宽度以及花瓣长度、宽度数据,数据特征为 4 维,共 150 个样本,3 类. Yesat 是 3 类真实的真核生物酵母菌基因数据,共 277 个样本,每个样本含有 17 个基因的表达水平,各类分别有 67,135,75 个样本. Glass 数据含有 214 个样本,每个样本含有 9 个特征,分别代表玻璃碎片的折射率及如 Na,Mg,Al 等 8 种物质的氧化物的百分含量,共 6 类,分别代表窗用浮法玻璃、窗用普通玻璃、车用玻璃、容器玻璃、餐具用玻璃和车灯用玻璃,各类样本数分别为 70,76,17,13,9,29 个样本. 第 4 组数据是 B-52、歼-6 和歼-7 3 种飞机的缩比模型微波暗室转台数据(记为 Darkroom data). 目标的方位角变化范围为 0°~155°,俯仰角恒为 5°,平均方位角采样间隔为 0.43°. 该数据是 101 维的,样本数分别为 322,311,451. 实验中,采用等间隔从每类中选取一半作为训练数据,其余作为测试数据.

3.3.2 实验结果分析

表 2 给出了各种方法在不同数据集上的实验结果,其中 n/r 分别表示训练样本个数与基的个数. 各种核方法所采用的核函数均为 RBF,核参数 σ 见表 2,核函数的形式及参数是依据训练识别率最好而选取的. 表 2 中 σ 的取值,第 1 个是 KOT-CC/FKOT-CC 所对应的核参数,第 2 个是 KFDA 所对应的核参数. 其中测试速度是指对于同样的测试集,同样的实验平台下的 10 次实验的平均时间.

由表 2 可以看出:(1)就识别率而言,核方法较线性算法 OT-CC 的识别率有明显提高,这是因为通过核化原理,将原空间的非线性问题转化为核空间的线性问题,实现了数据的非线性鉴别特征的提取.(2)笔者所提方法 KOT-CC/FKOT-CC 同 KFDA 相比,识别性能相当,甚至优于 KFDA,如类别数较多的 Glass 数据. 原因在于 KOT-CC 所构造的最优变换,优先考虑类间差异性,而且不同类别的差异性同等对待,即各类聚类

中心的距离均等.但 KFDA 则从整体上考虑类内聚集性与类间差异性,是类内距与类间距的一个折中,或者说可能为了保证类内数据更加聚集而牺牲类间的可分性,这就有可能造成某些类别出现混叠现象.(3)就测试识别速度而言,KOT-CC 与 KFDA 相当,因为二者都需要计算 n 个核函数,而 FKOT-CC 只需计算 r 个核函数,因此识别速度明显改善.如对线性相关性较强的 Iirs 数据,$n/r = 75/12$,差别较为悬殊,FKOT-CC 的识别速度是 KOT-CC/KFDA 的 3.7 倍之多.而对于线性相关性较弱的数据 Darkroom data,$n/r = 542/380$,FKOT-CC 的识别速度提高也不是特别显著.

表 2　各种方法在不同数据集上的识别结果

	n/r	σ	测试识别率/%				测试速度/s			
			OT-CC	KOT-CC	FKOT-CC	KFDA	OT-CC	KOT-CC	FKOT-CC	KFDA
Iris	75/12	2.0/1.0	94.0	97.3	98.6	97.3	0.002 5	0.18	0.048	0.176
Yeast	140/53	50.0/9.0	81.0	86.0	85.0	84.6	0.004 0	0.61	0.210	0.620
Glass	107/60	3.0/0.8	59.0	75.0	75.0	71.0	0.003 8	0.31	0.190	0.310
Darkroom	542/380	0.1/0.1	93.5	99.6	99.0	99.0	0.023 0	10.02	6.720	10.380

4　结 束 语

分析 OT-CC 算法的基础上,基于核化理论,提出一个新的非线性算法——KOT-CC,同时基于子空间 $\{\boldsymbol{\phi}(\boldsymbol{x}_i)\}_{1\leqslant i\leqslant n}$ 的一组基,给出一种快速算法 FKOT-CC.理论与实验分析表明:(1)KOT-CC/FKOT-CC 能够提取稳健的鉴别特征,特别是数据的非线性鉴别特征,特征提取能力较 OT-CC 有明显改善,甚至优于经典的 KFDA.(2)由于 KOT-CC/FKOT-CC 的计算复杂度分别为 $O(n^3)$ 与 $O(r^3)$,与训练数据的维数无关,克服了 OT-CC 所面临的"高维小样本"难题;同时,一般情况下 $r \ll n$,因此快速算法 FKOT-CC 不仅计算量较小,特征提取速度加快,对大训练集识别问题同样适用.(3)特别需要强调的是笔者所提出的快速算法 FKOT-CC 设计思路,具有一般性,可适用于其他核方法的简化计算,如 KFDA,KPCA 等.

有待进一步研究的问题是由于 KOT-CC/FKOT-CC 算法中各类所对应的聚类中心间的夹角相等,所以对于类别数较大的识别问题,各聚类中心的夹角变小,识别效果不太理想,可以考虑"一对多"、"一对一"的解决方案;另外一个是 KOT-CC/FKOT-CC 与 KFDA 的具体区别,有待进一步从理论上进行分析探讨.

参考文献:

[1] Anil K J, Robert P W D, Mao Jianchang. Statistical Pattern Recognition: a Review [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37.

[2] 杨健, 杨静宇, 叶晖. Fisher 线形鉴别分析的理论研究及其应用[J]. 自动化学报, 2003, 29(4): 481-493.
Yang Jian, Yang Jingyu, Ye Hui. Theory of Fisher Linear Discriminant Analysis and Its Application[J]. Acta Automatica Sinica, 2003,29 (4): 481-493.

[3] 周代英, 沈晓峰, 杨万麟. 最优聚类中心雷达目标一维距离像识别[J]. 系统工程与电子技术, 2002, 24(4): 44-46.
Zhou Daiying, Shen Xiaofeng, Yang Wanling. Recognition of Radar Target Based on Optimal Cluster Centers Using Range Profile[J]. Systems Engineering and Electronics, 2002, 24(4): 44-46.

[4] Muller K B, Mika S, Ratsch G, et al. An Introduction to Kernel-based Learning Algorithms[J]. IEEE Trans on Neural Networks, 2001, 12(2): 181-201.

[5] 田盛丰. 基于核函数的学习算法[J]. 北方交通大学学报, 2003, 27(4): 1-8.
Tian Shengfeng. Kernel-Based Learning Algorithms[J]. Chinese Journal of Northern Jiaotong University, 2003, 27(4): 1-8.

[6] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.

[7] Schölkopf B, Smola A J, Müller K R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem [J]. Neural Computation, 1998, 10(6): 1299-1319.

(下转第 161 页)