目标识别中特征空间核矩阵收缩方法*

郭 雷** 肖怀铁 付 强
国防科技大学电子科学与工程学院ATR实验室，长沙 410073

摘要 线性分类与非线性分类是模式识别领域的基础性课题. 核方法处理非线性分类问题有其独特的优势，核矩阵反映了输入样本在特征空间的位置关系，决定了样本在特征空间的可分性. 针对特征空间线性不可分问题，提出了特征空间核矩阵收缩的新概念和新方法. 首先定义了特征空间中样本数据的收缩因子以及样本数据相对于各类类心的收缩方法；然后理论推导得到样本数据收缩后的核矩阵，并且证明收缩后的数据可分性能更优. 最后的实验从核矩阵的性能度量以及核矩阵的分类性能两个方面验证了收缩后的核矩阵性能比收缩前性能更优.

关键词 核方法 核矩阵 特征空间 收缩因子

核方法是近年来机器学习领域的热点，最初用来解决两类分类问题的核方法叫做支持矢量机(support vector machine，SVM)[1]. 为了求解非线性可分问题，核方法将原始非线性可分的数据通过核函数映射到一个高维空间，在这个高维空间(也称为特征空间)可以进行内积运算，转换为线性分类问题. 核方法最大的优势就是在其高维空间内求解非线性分类问题的能力[2]. 核矩阵是输入样本在特征空间的内积运算结果和核方法的主要运算对象，反映了输入样本在特征空间的相对位置信息. 要在数据集内发现数据结构，必须通过核矩阵展现[3].

通过预先定义的核函数进行运算可以得到核矩阵，那么核函数类型及其参数决定了核矩阵的形式，也决定了数据在特征空间的结构，包括数据在特征空间可分性的情况. 为了使数据分类达到比较理想的结果，很多文献研究了核函数类型及其参数的最优选择方法[4—6]. 对于特征空间中线性不可分的样本，SVM通过定义惩罚因子c来控制对错分样本惩罚程度，实现错分样本的比例与算法复杂度的折中. 其本质是在原核矩阵的对角线元素上引入一个常数$\frac{1}{c}$，使得核矩阵的分类性能能够达到最优. 文献[7，8]还研究了数据线性可分的常用准则以及线性可分的程度，但也仅限于原始数据空间. 这些研究工作的实质就是通过改进分类器来构造分类性能最优的核矩阵，使得分类结果最优. 但是很多情况下，如雷达目标高分辨距离像(high resolution range profile，HRRP)识别问题，用于目标分类识别的数据无论是在原始空间还是在特征空间其可分性能比较差，即使选择最优的核函数构造核矩阵，最后的分类结果也难以达到满意的识别率.

为了使得数据在特征空间的分类结果更优，上述工作的思路是在不改变数据自身可分性的基础上设计最优分类器使数据分类结果更优. 与上述工作思路不同，本文将特征空间中可分性较差或者不可分的数据收缩，提高数据自身的可分性能，提出了特征空间核矩阵收缩的新概念和新方法. 首先定义了特征空间中样本数据的收缩因子，以及样本数据相对于类心的收缩方法；然后理论推导得到收缩后的核矩阵，并且证明收缩后的数据可分性能更优.

2008-04-15收稿，2008-05-26收修改稿
* 国家自然科学基金资助项目(批准号：60572138)
** E-mail：thunderblastg@yahoo.com.cn

最后的实验从核矩阵的性能度量以及核矩阵的分类性能两个方面验证了收缩后的数据核矩阵性能比收缩前性能更优，即收缩后的数据能达到更好的可分性能．此外，本文方法另外一个优点是对于核函数不敏感，即核函数类型及其参数的改变基本不会影响到最后的分类识别结果．

1 非线性分类问题

假定训练样本的集合用 $\boldsymbol{S}$ 表示，其表示形式为 $\boldsymbol{S}=(\boldsymbol{x}_i, y_i)$，$i=1, 2, \cdots, N$，$\boldsymbol{x}_i \in \mathrm{R}^d$，$y_i \in \{+1, -1\}$，其中 x_i 为训练样本的特征矢量，维数为 d 维，y_i 是类别标号，用 $+1$ 表示正类，用 -1 表示负类，N 表示训练样本的个数．分类面方程为 $g(\boldsymbol{x})=\boldsymbol{w}\cdot\boldsymbol{x}+b=0$，我们将判别函数进行归一化，使两类所有样本都满足 $|g(\boldsymbol{x})|\geqslant 1$，使离分类面最近的样本的 $|g(\boldsymbol{x})|=1$，这样分类间隔就等于 $\frac{2}{\|\boldsymbol{w}\|}$，因此使间隔最大等价于使 $\|\boldsymbol{w}\|$（或者 $\min\left(\frac{1}{2}\|\boldsymbol{w}\|^2\right)$）最小；而要求分类线对所有样本正确分类，就是要求它满足

$$y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]\geqslant 1, \quad i=1, 2, \cdots, N \tag{1}$$

满足上述条件且使最小 $\|\boldsymbol{w}\|^2$ 的分类面就是最优分类面．最优分类面参数 $\boldsymbol{w}$ 和 b 的求解可以转化为求解下面的 Lagrange 方程[1]

$$L(\boldsymbol{w}, b, \alpha)=\frac{1}{2}\langle \boldsymbol{w}, \boldsymbol{w}\rangle-\sum_{i=1}^{N}\alpha_i[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1] \tag{2}$$

其中 $\alpha_i\geqslant 0$ 为 Lagrange 乘子．根据 Wolf 对偶准则，得到在约束条件 $\sum_{i=1}^{N}\alpha_i y_i=0$ 和 $\alpha_i\geqslant 0$ 下对 α_i 求解下面泛函的最小值

$$\min_{\alpha}\frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}^i\cdot\boldsymbol{x}^j)-\sum_{i=1}^{N}\alpha_i \tag{3}$$

求解上述问题后，我们就可以得到最优分类函数是

$$f(\boldsymbol{x})=\mathrm{sgn}\{(\boldsymbol{w}\cdot\boldsymbol{x})+b\}=\mathrm{sgn}\left\{\sum_{i=1}^{N}\alpha_i y_i(\boldsymbol{x}_i\cdot\boldsymbol{x})+b\right\} \tag{4}$$

实际情况中一般遇到的问题大多是非线性分类问题．针对这类问题，SVM 首先通过非线性映射 $\Phi(\cdot)$ 将原始数据 x 映射到某一高维空间（称为特征空间），得到 $\Phi(\boldsymbol{x})$，然后在这一高维空间构造线性分类超平面 $g(\boldsymbol{x})=\boldsymbol{w}\cdot\Phi(\boldsymbol{x})+b=0$，如图 1 所示．

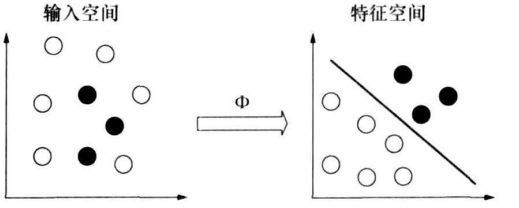

图 1 原始数据非线性映射到特征空间

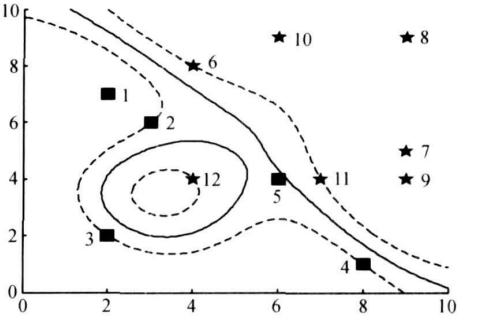

图 2 非线性可分的最优分类面

对于不可分的样本，(1)式的约束条件可以放宽为 $y_i[(\boldsymbol{w}\cdot\Phi(\boldsymbol{x}_i))+b]\geqslant 1-\xi_i$，$i=1, 2, \cdots, N$，$\xi_i>0$，则优化问题变为：$\min\left(\frac{1}{2}\|\boldsymbol{w}\|^2+c\sum_{i=1}^{N}\xi_i\right)$，其中 c 为惩罚因子．该优化问题相应的 Lagrange 方程通过 Wolf 对偶准则可以转化为：$\min_{\alpha}\frac{1}{2}\left[\Lambda^{\mathrm{T}}\boldsymbol{G}\Lambda+\frac{1}{c}\Lambda^{\mathrm{T}}\Lambda\right]-\Lambda^{\mathrm{T}}\boldsymbol{1}$，即

$$\min_{\alpha}\frac{1}{2}\left[\Lambda^{\mathrm{T}}\left(\boldsymbol{G}+\frac{1}{c}\boldsymbol{I}\right)\Lambda\right]-\Lambda^{\mathrm{T}}\boldsymbol{1} \tag{5}$$

约束条件为 $0\leqslant\alpha_i\leqslant c$. 其中 $\Lambda=[\alpha_1, \alpha_2, \cdots, \alpha_N]^{\mathrm{T}}$, $\boldsymbol{G}_{ij}=y_iy_j\boldsymbol{K}_{ij}$, $\boldsymbol{K}_{ij}$ 为核矩阵元素. 此时最优分类函数称为广义分类超平面, 它是平面上直线、R^3 中平面的推广, 其表达式是

$$f(\boldsymbol{x})=\mathrm{sgn}\left(\left(\boldsymbol{w}\cdot\Phi(\boldsymbol{x})\right)+b\right)=\mathrm{sgn}\left(\sum_{i=1}^{N}\alpha_iy_i\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x})+b\right) \tag{6}$$

该最优分类函数映射回原始数据空间形成的分类面如图 2 所示.

由以上分析可以看出, 对于不可分的样本, SVM 通过加入惩罚因子 c 来控制分类错误最小, 这实际上是使得不满秩的矩阵 $\boldsymbol{G}$ 加入 $\frac{1}{c}\boldsymbol{I}$ 之后变为满秩矩阵, 从而能够实现线性分类. SVM 具有清晰的几何意义, 其优化过程等价于求解特征空间中两类样本形成的闭凸包或者增大闭凸包之间的距离[9].

文献[4—6] 研究了核函数 $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y})$ 及其参数的最优选取方法, 本质是将数据映射到不同的特征空间(由于核函数类型及其参数不同, 那么映射后的特征空间也不同)来构造不同的核矩阵, 以使得两类样本之间的闭凸包最大, 构造最优分类器以达到较好的分类识别结果.

但是对于无论在原始空间还是特征空间可分性能很差或者根本无法分类的数据, 即使选择最优的核函数及其参数或者选择最优惩罚因子, 都难以做到核矩阵满秩, 所以类似于(3)式的优化过程也就无法收敛到最优的唯一解. 基于这种优化结果构造的分类器显然不是最优分类器, 也就无法达到最优的分类识别性能. 此外, 在实际应用中, 通过训练样本反复训练而得到的最优核函数模型反而有可能造成对测试数据泛化能力的下降.

与上述的设计最优分类器的思路不同, 本文提出特征空间数据收缩的新方法, 在特征空间对数据进行收缩, 改变数据自身的可分性, 使得不可分的数据变得线性可分, 构造分类性能更优的核矩阵.

2 特征空间核矩阵收缩方法

2.1 特征空间收缩因子

令 $\boldsymbol{S}^+=(\boldsymbol{x}_i^+, y_i)$, $i=1, 2, \cdots, N_+$, $y_i=+1$ 和 $\boldsymbol{S}^-=(\boldsymbol{x}_i^-, y_i)$, $i=1, 2, \cdots, N_-$, $y_i=-1$ 分别表示+1 类和−1 类训练样本集, 其中 N_+ 和 N_- 分别表示+1 类和−1 类训练样本数目, $N=N_++N_-$ 表示总的训练样本个数. 样本 $\boldsymbol{x}_i$ 映射到特征空间为 $\Phi(\boldsymbol{x}_i)$ (其中 $\Phi(\boldsymbol{x}_i)=[\phi_1(\boldsymbol{x}_i), \phi_2(\boldsymbol{x}_i), \cdots, \phi_l(\boldsymbol{x}_i)]^{\mathrm{T}}$, l 为映射后特征空间维数, l 可能为有限维, 也可能为无限维). 在特征空间中, $\Phi(\boldsymbol{x}_i)$ 分别位于不同的特征球内部[10], 则+1 类样本的特征球球心可表示为 $\boldsymbol{m}_+=\frac{1}{N_+}\sum_{p=1}^{N_+}\Phi(\boldsymbol{x}_p^+)$, −1 类样本的特征球球心可表示为 $\boldsymbol{m}_-=\frac{1}{N_-}\sum_{p=1}^{N_-}\Phi(\boldsymbol{x}_p^-)$. 本节定义样本数据 $\Phi(\boldsymbol{x}_i)$ 相对于各个特征球球心的收缩因子, 并且根据该收缩因子将数据向球心方向收缩, 以此构造分类性能更优的数据结构.

定义 1 令特征空间中样本数据 $\Phi(\boldsymbol{x}_i)$ 相对于球心的收缩因子为 $\mu(0\leqslant\mu\leqslant1)$, 收缩后的训练样本数据分别为

$$\overline{\Phi}(\boldsymbol{x}_i^+)=(1-\mu)\boldsymbol{m}_++\mu\Phi(\boldsymbol{x}_i^+) \tag{7}$$

$$\overline{\Phi}(\boldsymbol{x}_i^-)=(1-\mu)\boldsymbol{m}_-+\mu\Phi(\boldsymbol{x}_i^-) \tag{8}$$

文献[11] 定义了原始数据空间的数据收缩因子, 本文将其推广到核映射后的特征空间. 数据收缩过程如图 3 所示.

由于核映射之后的特征空间维数不确定, 那么

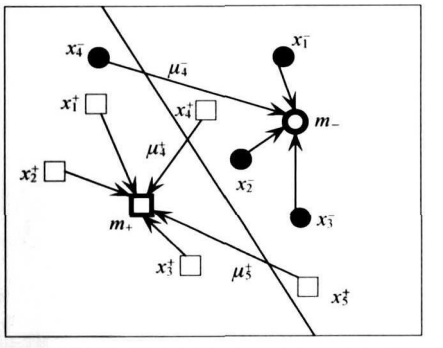

图 3 数据收缩示意图

原始空间的分类面的直接求解方法不同于特征空间中的求解方法.

下面给出闭凸集的严格定义并且详细证明数据在特征空间收缩后的可分性更优.

定义 **2**[12] 设 A 是线性空间 E 的子集，若对 A 中任意两点 x 与 y，有

$$\lambda x+(1-\lambda)y\in A,\quad \lambda\in[0,1],$$

则称 A 是 E 中的凸集.

由于 $\boldsymbol{S}^+$ 和 $\boldsymbol{S}^-$ 在特征空间中分别位于不同的特征球内[10]，该特征球即是特征空间的凸集，由上述的凸集定义可知，由(7)和(8)式定义的+1 类和−1 类样本收缩后仍然是原凸集的内点，即样本收缩后仍然位于本类特征球内.

令收缩后的特征球球心分别为：$\overline{\boldsymbol{m}}_+=\frac{1}{N_+}\sum_{p=1}^{N_+}\overline{\Phi}(\boldsymbol{x}_p^+)$ 和 $\overline{\boldsymbol{m}}_-=\frac{1}{N_-}\sum_{p=1}^{N_-}\overline{\Phi}(\boldsymbol{x}_p^-)$，收缩后的样本数据能保持样本中心不变[11]，即收缩后的正类样本特征球球心 $\overline{\boldsymbol{m}}_+$ 与负类样本特征球球心 $\overline{\boldsymbol{m}}_-$ 仍与收缩前相同，那么两个特征球球心之间的距离保持不变，所以 $d(\boldsymbol{m}_+,\boldsymbol{m}_-)=d(\overline{\boldsymbol{m}}_+,\overline{\boldsymbol{m}}_-)$. 类间、类内距离比值准则是分类过程一个常用的准则函数，这个比值越大，说明两类数据之间的可分性越好[13]. 由于数据收缩后特征球球心保持不变，故类间距离保持不变，那么只要收缩后的数据类内距离之和变小，则可以证明收缩后的数据可分性比收缩前更优.

下面将证明数据在特征空间收缩后类内距离和变小.

定理 **1** 数据在特征空间收缩后的数据比收缩前的数据类内距离减小，即

$$\begin{aligned}&\sum_{i=1}^{N_+}\|\overline{\Phi}(\boldsymbol{x}_i^+)-\overline{\boldsymbol{m}}_+\|^2\leqslant\sum_{i=1}^{N_+}\|\Phi(\boldsymbol{x}_i^+)-\boldsymbol{m}_+\|^2\\&\sum_{i=1}^{N_-}\|\overline{\Phi}(\boldsymbol{x}_i^-)-\overline{\boldsymbol{m}}_-\|^2\leqslant\sum_{i=1}^{N_-}\|\Phi(x_i^-)-\boldsymbol{m}_-\|^2\end{aligned}\tag{9}$$

证明 收缩前类内距离和为 $\sum_{i=1}^{N_+}\|\Phi(\boldsymbol{x}_i^+)-\boldsymbol{m}_+\|^2$，由于收缩因子 $0\leqslant\mu\leqslant1$，那么

$$\sum_{i=1}^{N_+}\|\Phi(\boldsymbol{x}_i^+)-\boldsymbol{m}_+\|^2\geqslant\sum_{i=1}^{N_+}\|(1-\mu)\boldsymbol{m}_++\mu\Phi(\boldsymbol{x}_i^+)-\boldsymbol{m}_+\|^2,$$

即 $\sum_{i=1}^{N_+}\|\Phi(\boldsymbol{x}_i^+)-\boldsymbol{m}_+\|^2\geqslant\sum_{i=1}^{N_+}\|\overline{\Phi}(\boldsymbol{x}_i^+)-\boldsymbol{m}_+\|^2$.

假设 ξ 为 $\overline{\Phi}(\boldsymbol{x}_i^+)$ 特征空间中的任意一点，令 $f(\xi)=\sum_{i=1}^{N_+}\|\overline{\Phi}(\boldsymbol{x}_i^+)-\xi\|^2$，对 $f(\xi)$ 求导，令导数为 0，得到

$$2\sum_{i=1}^{N_+}\xi-2\sum_{i=1}^{N_+}\xi\overline{\Phi}(\boldsymbol{x}_i^+)=0\Rightarrow\xi=\frac{1}{N_+}\sum_{i=1}^{N_+}\overline{\Phi}(\boldsymbol{x}_i^+)$$

即只有当 ξ 为收缩后的类心时，$f(\xi)$ 达到最小值，令 $\overline{\boldsymbol{m}}_+=\frac{1}{N_+}\sum_{i=1}^{N_+}\overline{\Phi}(\boldsymbol{x}_i^+)$，得到

$$\sum_{i=1}^{N_+}\|\Phi(\boldsymbol{x}_i^+)-\boldsymbol{m}_+\|^2\geqslant\sum_{i=1}^{N_+}\|\overline{\Phi}(\boldsymbol{x}_i^+)-\overline{\boldsymbol{m}}_+\|^2,$$

证毕.

由以上证明可以看出，数据收缩后类间距离保持不变，而类内距离和变小，使得不同类别数据之间的交集变小，数据之间的可分性能更优.

2.2 特征空间数据收缩后的核矩阵

将数据在特征空间收缩后，改变了数据原来的分布形式，而核矩阵反映了样本在特征空间的相对位置信息，所以样本在特征空间收缩后必定会改变样本原来的核矩阵. 这里需要说明的是：样本通过非线性映射到特征空间后的维数可能为有限维，也可能为无穷维，所以无法直接求解分类超平面. 而核矩阵是通过样本在特征空间的内积运算得到的，避免了无穷维带来的求解分类超平面的维数灾难. 本节主要推导得到数据在特征空间收缩后的核矩阵形式.

假设原核矩阵表示为 $\boldsymbol{K}=[\boldsymbol{K}_{ij}]_{N\times N}$，核矩阵还满足 Mercer 定理，即 $\boldsymbol{K}_{ij}=\langle\Phi(\boldsymbol{x}_i),\Phi(\boldsymbol{x}_j)\rangle$. 对于样本集 $\boldsymbol{S}$，其核矩阵可以用分块矩阵表示为：

$$\boldsymbol{K}=\begin{bmatrix}\boldsymbol{K}^{++}_{N_+\times N_+} & \boldsymbol{K}^{+-}_{N_+\times N_-}\\ \boldsymbol{K}^{-+}_{N_-\times N_+} & \boldsymbol{K}^{--}_{N_-\times N_-}\end{bmatrix}_{N\times N} \tag{10}$$

其中

$$\begin{aligned}\boldsymbol{K}^{++}_{N_+\times N_+}&=\left[\langle\Phi\left(\boldsymbol{x}_i^+\right),\Phi\left(\boldsymbol{x}_j^+\right)\rangle\right]_{N_+\times N_+}\\ \boldsymbol{K}^{--}_{N_-\times N_-}&=\left[\langle\Phi\left(\boldsymbol{x}_i^-\right),\Phi\left(\boldsymbol{x}_j^-\right)\rangle\right]_{N_-\times N_-}\\ \boldsymbol{K}^{+-}_{N_+\times N_-}&=\boldsymbol{K}^{-+}_{N_-\times N_+}{}^{\mathrm{T}}=\left[\langle\Phi\left(\boldsymbol{x}_i^+\right),\Phi\left(\boldsymbol{x}_j^-\right)\rangle\right]_{N_+\times N_-}\end{aligned} \tag{11}$$

(11)式中 $\boldsymbol{K}^{++}_{N_+\times N_+}$ 和 $\boldsymbol{K}^{--}_{N_-\times N_-}$ 分别表示+1 类样本和−1 类样本类内数据核矩阵，$\boldsymbol{K}^{+-}_{N_+\times N_-}$ 表示+1 类样本和−1 类样本数据之间的核矩阵.

令收缩后的核矩阵表示为$\overline{\boldsymbol{K}}=[\overline{K}_{ij}]_{N\times N}$，同样可以用分块矩阵来表示

$$\overline{\boldsymbol{K}}=\begin{bmatrix}\overline{\boldsymbol{K}}^{++}_{N_+\times N_+} & \overline{\boldsymbol{K}}^{+-}_{N_+\times N_-}\\ \overline{\boldsymbol{K}}^{-+}_{N_-\times N_+} & \overline{\boldsymbol{K}}^{--}_{N_-\times N_-}\end{bmatrix}_{N\times N} \tag{12}$$

数据收缩后的核矩阵$\overline{\boldsymbol{K}}$中的各个元素可以用原核矩阵$\boldsymbol{K}$ 的元素推导得到

$$\begin{aligned}\overline{\boldsymbol{K}}^{++}_{ij}&=\langle\overline{\Phi}\left(\boldsymbol{x}_i^+\right),\overline{\Phi}\left(\boldsymbol{x}_j^+\right)\rangle=\\ &\left\langle\left(1-\mu\right)\boldsymbol{m}_+ +\mu\Phi\left(\boldsymbol{x}_i^+\right),\right.\\ &\left.\left(1-\mu\right)\boldsymbol{m}_+ +\mu\Phi\left(\boldsymbol{x}_j^+\right)\right\rangle=\\ &\left\langle\left(1-\mu\right)\left(\frac{1}{N_+}\sum_{i=1}^{N_+}\Phi\left(\boldsymbol{x}_i^+\right)\right)+\mu\Phi\left(\boldsymbol{x}_i^+\right),\right.\\ &\left.\left(1-\mu\right)\left(\frac{1}{N_+}\sum_{j=1}^{N_+}\Phi\left(\boldsymbol{x}_j^+\right)\right)+\mu\Phi\left(\boldsymbol{x}_j^+\right)\right\rangle=\\ &\left(1-\mu\right)\left(1-\mu\right)\frac{1}{N_+}\sum_{i=1}^{N_+}\Phi\left(\boldsymbol{x}_i^+\right)\cdot\\ &\frac{1}{N_+}\sum_{j=1}^{N_+}\Phi\left(\boldsymbol{x}_j^+\right)+\frac{1}{N_+}\left(1-\mu\right)\mu\Phi\left(x_j^+\right)\cdot\\ &\sum_{i=1}^{N_+}\Phi\left(\boldsymbol{x}_i^+\right)+\frac{1}{N_+}\mu\left(1-\mu\right)\Phi\left(\boldsymbol{x}_i^+\right)\cdot\\ &\sum_{j=1}^{N_+}\Phi\left(\boldsymbol{x}_j^+\right)+\mu\mu\Phi\left(\boldsymbol{x}_i^+\right)\Phi\boldsymbol{x}\left(x_j^+\right)=\\ &\frac{1}{N_+^2}\left(1-\mu\right)^2\sum_{p,q=1}^{N_+}\boldsymbol{K}_{pq}+\frac{1}{N_+}\mu\left(1-\mu\right)\sum_{p=1}^{N_+}\boldsymbol{K}_{pj}+\\ &\frac{1}{N_+}\mu\left(1-\mu\right)\sum_{q=1}^{N_+}\boldsymbol{K}_{qi}+\mu^2\boldsymbol{K}_{ij}\end{aligned} \tag{13}$$

同理，可求得−1 类样本收缩后的核矩阵元素为

$$\begin{aligned}\overline{\boldsymbol{K}}^{--}_{ij}&=\langle\overline{\Phi}\left(\boldsymbol{x}_i^-\right),\overline{\Phi}\left(\boldsymbol{x}_j^-\right)\rangle=\\ &\left\langle\left(1-\mu\right)\boldsymbol{m}_- +\mu\Phi\left(\boldsymbol{x}_i^-\right),\left(1-\mu\right)\boldsymbol{m}_- +\mu\Phi\left(\boldsymbol{x}_j^-\right)\right\rangle=\\ &\frac{1}{N_-^2}\left(1-\mu\right)^2\sum_{p,q=1}^{N_-}\boldsymbol{K}_{pq}+\frac{1}{N_-}\mu\left(1-\mu\right)\sum_{p=1}^{N_-}\boldsymbol{K}_{pj}+\\ &\frac{1}{N_-}\mu\left(1-\mu\right)\sum_{q=1}^{N_-}\boldsymbol{K}_{qi}+\mu^2\boldsymbol{K}_{ij}\end{aligned} \tag{14}$$

则+1 类样本与−1 类样本之间的核矩阵元素为

$$\begin{aligned}\overline{\boldsymbol{K}}^{+-}_{ij}&=\langle\overline{\Phi}\left(\boldsymbol{x}_i^+\right),\overline{\Phi}\left(\boldsymbol{x}_j^-\right)\rangle=\\ &\left\langle\left(1-\mu\right)\boldsymbol{m}_+ +\mu\Phi\left(\boldsymbol{x}_i^+\right),\left(1-\mu\right)\boldsymbol{m}_- +\mu\Phi\left(x_j^-\right)\right\rangle=\\ &\frac{1}{N_+}\frac{1}{N_-}\left(1-\mu\right)^2\sum_{p=1}^{N_+}\sum_{q=1}^{N_-}\boldsymbol{K}_{pq}+\frac{1}{N_+}\mu\left(1-\mu\right)\cdot\\ &\sum_{p=1}^{N_+}\boldsymbol{K}_{pj}+\frac{1}{N_-}\mu\left(1-\mu\right)\sum_{q=1}^{N_-}\boldsymbol{K}_{qi}+\mu^2\boldsymbol{K}_{ij}\end{aligned} \tag{15}$$

由(13)—(15)式可以看出，(12)式确定的收缩后的核矩阵可以根据收缩因子与原核矩阵求得.

确定数据收缩后的核矩阵之后，也就确定了样本在特征空间收缩之后的位置关系，那么通过求解收缩后核矩阵的 Lagrange 方程即可得到分类超平面，同时也可以直接利用核矩阵相似度求得收缩后的核矩阵的性能. 本文第 3 节给出了详细的实验过程与结果分析.

2.3 特征空间数据收缩以及训练识别步骤

(1) 选定训练数据集与测试数据集；

(2) 确定合适的收缩因子 μ；

(3) 在特征空间对训练数据进行收缩，并且根据(13)—(15)式计算训练数据收缩后的核矩阵；

(4) 采用二次优化技术优化(5)式，得到训练数据的最优分类超平面，超平面的形式如(6)式；

(5) 将每一个测试数据 $\boldsymbol{x}_i$ 相对于每一类特征球球心均做收缩，令得到收缩后的数据分别表示为 $\boldsymbol{x}_i^{(+)}$ 和 $\boldsymbol{x}_i^{(-)}$；

(6) 将 $\boldsymbol{x}_i^{(+)}$ 和 $\boldsymbol{x}_i^{(-)}$ 代入由训练数据确定的最优分类超平面，根据得到的数值大小判定 $\boldsymbol{x}_i$ 应属的类别.

3 实验与分析

3.1 二维数据收缩实验

为了验证本文提出方法的有效性，首先利用随机产生的二维数据和标准数据集中的 Banana 数据[14] 进行实验，原始数据和收缩后的数据分别如图 4—7 所示.

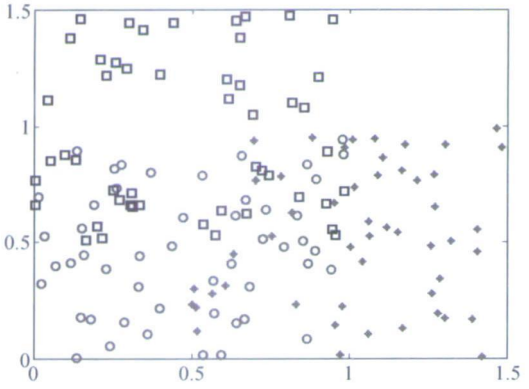

图 4 随机产生的二维原始数据

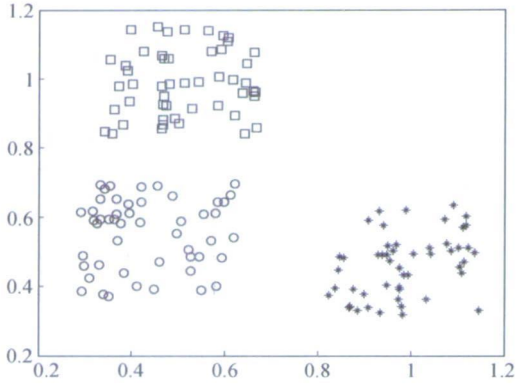

图 5 收缩后的二维数据(μ=1/3)

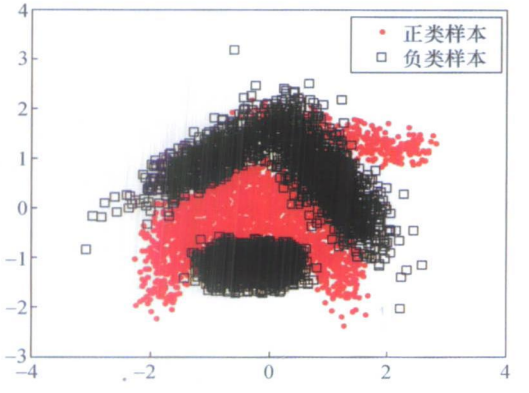

图 6 原始 Banana 训练数据

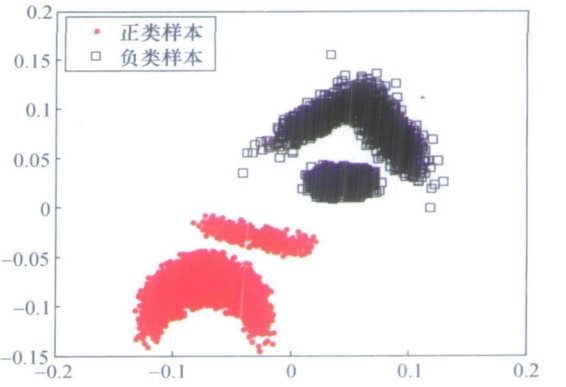

图 7 收缩后的 Banana 训练数据(μ= 0.9)

从图 4—7 可以明显的看出，采用本文提出的数据收缩方法之后，原来两类线性不可分的数据变的线性可分，而且可以达到无错误分类.

3.2 特征空间数据核矩阵收缩实验

3.2.1 核矩阵相似性测度 核矩阵包含所有可提供给学习机的关于输入在特征空间中相对位置的信息. 很自然的，如果要在数据集内发现结构，数据必须通过核矩阵展现这一结构. 由前面的证明可知，数据收缩后，类内距离变小，数据更靠近本类类心，那么收缩后的核矩阵使得数据的可分性更好，能够达到更优的分类性能. 核矩阵的优劣可以用核的相似性的概念来衡量，即定义核排列：

$$A(\boldsymbol{K}_1, \boldsymbol{K}_2) = \frac{\langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F}{\sqrt{\langle \boldsymbol{K}_1, \boldsymbol{K}_1 \rangle_F \langle \boldsymbol{K}_2, \boldsymbol{K}_2 \rangle_F}}$$，表示归一化的 $\boldsymbol{K}_1$，$\boldsymbol{K}_2$ 的 Frobenious 范数. 令目标矩阵 $\boldsymbol{K}_2 = \boldsymbol{Y}\boldsymbol{Y}^T$，$\boldsymbol{Y} = [y_1, y_2, \cdots, y_N]^T$，$A(\boldsymbol{K}_1, \boldsymbol{K}_2)$ 大小反映了矩阵 $\boldsymbol{K}_1$ 的分类性能，它满足 $0 \leqslant A(\boldsymbol{K}_1, \boldsymbol{K}_2) \leqslant 1$. $A(\boldsymbol{K}_1, \boldsymbol{K}_2)$ 值越大，说明 $\boldsymbol{K}_1$ 分类性能越好[3].

实验选用国际通用的标准数据集[14]，特征空间中数据收缩前与收缩后的核矩阵性能如表 1 所示. 核函数均选用 Guass 核函数.

表 1 特征空间中数据收缩前后的核矩阵性能

数据集	收缩前 $A(\boldsymbol{K}_1, \boldsymbol{K}_2)$	收缩后 $A(\boldsymbol{K}_1, \boldsymbol{K}_2)$
Heart	0.36	0.98
Thyroid	0.51	0.85
Breast cancer	0.21	0.90
Diabetis	0.19	0.95

3.2.2 识别率的衡量

(1) 基准数据集识别实验

表 2 给出了基准数据集在特征空间收缩前后的识别率变化情况. 从表 1 和表 2 两个衡量标准来看, 收缩后的基准数据集比收缩前的性能有了很大的提高, 达到了满意的识别率.

表 2 特征空间数据收缩前后的识别率

数据集	收缩因子 μ	收缩前平均识别率/%	收缩后平均识别率/%
Banana	0.9	88.47	91.23
Heart	0.95	83.00	100
Thyroid	0.9	98.67	100
Breastcancer	0.9	74.03	96.71
Diabetis	0.9	77.00	100
German	0.9	80.67	100
Waveform	0.9	89.48	100
Splice	0.9	89.93	100

(2) 雷达目标 HRRP 数据集识别实验

采用某电磁特性计算软件计算得到不同弹道点的 4 种飞机目标 HRRP, 分别是 B737, B747, F16, F18, 每一种飞机有 855 个弹道点, 即 855 幅 HRRP, 每幅 HRRP 特征维数为 150. 实验过程抽取每一类数据的 1/3 做训练数据, 全部的 HRRP 做测试数据. 将 HRRP 映射到特征空间后数据收缩前后得到的识别结果如表 3 所示.

表 3 雷达目标 **HRRP** 数据特征空间收缩前后的识别率

目标类型	收缩前识别率/%	收缩后识别率/%
B737	96.86	99.50
B747	96.49	99.30
F16	78.22	96.52
F18	72.13	96.73

3.3 讨论

本节实验可以看出, 从相似性和识别率两个角度来衡量数据在特征空间收缩后得到的核矩阵的性能都非常优异, 几乎能够得到无错误分类, 而且实验过程结果对于核参数的选择并不敏感, 主要原因是数据在特征空间经过收缩后, 其可分性已经变得很好, 所以核函数类型及其参数对于分类结果的影响不大. 此外, 雷达目标 HRRP 实验中, F16 和 F18 两类目标的识别率得到了非常大的提高.

4 结论

为了求解非线性可分问题, 核方法将原始非线性可分的数据通过核函数映射到一个高维空间, 在这个高维空间(也称为特征空间)可以进行内积运算, 来求解线性可分问题. 对于特征空间中线性不可分的样本, 可以引入惩罚因子和寻找最优的核函数, 其实质是设计最优的分类器来提高数据的正确分类识别性能. 本文思路与上述思路不同, 本文将数据在特征空间进行收缩, 使得线性不可分问题变为线性可分问题, 提出了特征空间核矩阵收缩的新方法, 得到性能更优的核矩阵, 使得目标数据分类性能更优. 实验验证了本文方法的可行性与适用性.

参 考 文 献

1 Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer, 2000

2 Cristianini N, Shawe-Taylor J 著. 支持向量机导论. 李国正, 王 猛, 曾华军, 译. 北京: 电子工业出版社, 2006

3 Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. UK: Cambridge University Press, 2004

4 Chapelle O, Vapnik VN, Bousquet O, et al. Choosing multiple parameters for support vector machines. Machine Learning, 2002, 16(1): 131—159

5 Keerthi SS. Efficient tuning of SVM hyper parameters using radius/margin bound and iterative algorithms. IEEE Trans Neural Networks, 2002, 13(5): 1225—1229

6 Peng XY, Wu HX, Peng Y. Parameter selection method for SVM with PSO. Chinese Journal of Electronics, 2006, 15(4): 638—642

7 Elizondo D. The linear separability problem: Some testing methods. IEEE Transactions on Neural Networks, 2006, 17(2): 330—344

8 Ben-Israel A, Levin Y. The geometry of linear separability in data sets. Linear Algebra and its Applications, 2006, 416(1): 75—87

9 Bennett K, Champbell O. Support vector machines: Hype or hallelujah?. SIGKDD Explorations, 2000, 2(2): 1—13

10 Tax DMJ, Duin RPW. Support vector data description. Machine Learning, 2004, 54(1): 45—66

11 陶 卿, 孙德敏, 范劲松. 基于闭凸包收缩的最大边缘线性分类器. 软件学报, 2002, 13(3): 404—409

12 吴 翊, 屈田兴. 应用泛函分析. 长沙: 国防科技大学出版社, 2002

13 孙即祥. 现代模式识别. 长沙: 国防科技大学出版社, 2002

14 Rätsch G. Benchmarks data sets. Available. http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm [2007-07-16]