

# 基于矩阵模式的最小类内散度支持向量机

皋 军<sup>1,2,3</sup>, 王士同<sup>1,3</sup>

(1. 江南大学信息工程学院, 江苏无锡 214122; 2. 盐城工学院信息工程学院, 江苏盐城 224001;  
3. 浙江大学 CAD&CG 国家重点实验室, 浙江杭州 310027)

**摘 要:** 基于最小类内散度支持向量机 (MCSVMs) 提出一种新的矩阵模式的最小类内散度支持向量机 (MCSVMs<sup>matrix</sup>)。同时为了更好地解决非线性分类问题, 将 Mercer 核函数引入到 MCSVMs<sup>matrix</sup> 方法中, 并提出基于矩阵模式的非线性支持向量机: Ker-MCSVMs<sup>matrix</sup>。上述两种方法不但继承了 MCSVMs 的优点, 而且由于将矩阵模式的类内散度矩阵引入到支持向量机中, 从而在理论上可以较好地解决了 MCSVMs 方法在处理小样本高维数据集时类内散度矩阵奇异性问题, 同时降低了求解类内散度矩阵及其逆矩阵和权重矢量的时间、空间复杂度。因此, 在一定程度上提高了分类精度。实验结果也表明 MCSVMs<sup>matrix</sup>、Ker-MCSVMs<sup>matrix</sup> 具有上述优势。

**关键词:** 支持向量机; 矩阵模式; 类内散度矩阵; 人脸识别

**中图分类号:** TP391 **文献标识码:** A **文章编号:** 0372-2112 (2009) 05-1051-07

## Matrix Pattern Based Minimum Within-Class Scatter Support Vector Machines

GAO Jun<sup>1,2,3</sup>, WANG Shi-tong<sup>1,3</sup>

(1. School of Information Engineering, Jiangnan University, Wuxi, Jiangsu 214122, China;  
2. School of Information Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu 224001, China;  
3. State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, Zhejiang 310027, China)

**Abstract:** Based on minimum within-class scatter support vector machines (MCSVMs), a new matrix pattern based MCSVMs (MCSVMs<sup>matrix</sup>) is presented. Accordingly, it is extended by introducing Mercer's kernels in order to solve the problem of nonlinear decision boundaries, which presents a significant matrix pattern based nonlinear support vector machines: Ker-MCSVMs<sup>matrix</sup>. The above-mentioned approaches not only keep the merits of MCSVMs, but, owing to introducing matrix pattern based within-class scatter matrix into support vector machines, theoretically better solve the singular problem of within-class scatter matrix when small sample size problems are dealt with, reduce the time/place complexity when within-class scatter matrix, its invertible matrix and coefficient vector omega are calculated. Hence, the classification accuracy is improved to certain extent. Experimental results indicate the above advantages of the proposed methods: both MCSVMs<sup>matrix</sup> and Ker-MCSVMs<sup>matrix</sup>.

**Key words:** SVMs; matrix pattern; within-class scatter matrix; face recognition

### 1 引言

支持向量机 (Support Vector Machine, SVM) 是数据挖掘中一项新技术<sup>[1,2]</sup>, 它借助于统计学习理论和最优化方法解决机器学习问题。该方法已经在众多的模式识别领域得到了成功地运用 (比如对象识别<sup>[3]</sup>、生物信息学<sup>[4,15]</sup>、人脸识别<sup>[5,6]</sup>)。在文献<sup>[5]</sup>中, A. Tefas 等人在充分分析了 Fisher 判别准则后, 将 Fisher 判别准则<sup>[4]</sup>重新写成具有新约束条件的二次规划的形式, 从而提出了最小类内散度支持向量机 (Minimum Within-class Scatter Support Vector Machines, MCSVMs), 该方法充分结合了

Fisher 判别准则和支持向量机的各自优势, 同时从理论上说明了 Fisher 判别优化的实质是具有约束的最小二乘优化问题<sup>[5]</sup>。然而通过分析发现, 在理论上并没有很好地解决类内散度矩阵奇异性问题, 特别是在处理具有高维特征的小样本数据集时表现的适应性不强。文献<sup>[6]</sup>通过构造总散度矩阵, 根据 PCA 方法并结合 Hilbert-Schmidt 理论在一定程度上克服最小类内散度支持向量机在处理小样本数据时类内散度矩阵奇异性问题。但是通过分析得知, 该方法和原方法相比在求解散度矩阵、逆矩阵和权矢量所对应的时间、空间复杂度并没有降低, 而且在该方法中所得到的  $\tilde{S}_w$  并不能确保非奇

收稿日期: 2008-06-27; 修回日期: 2008-10-23

基金项目: 国家 863 高技术研究发展计划 (No. 2007AA1Z158, No. 2006AA10Z313); 国家自然科学基金 (No. 60773206, No. 90820002); 2007 年教育部高等学校创新创业工程重大培育项目; 浙江大学 CAD&CG 国家重点实验室开放课题 (No. A0802); 国防应用基础研究基金 (No. A1420461266)

异<sup>[6]</sup>.

近来文献[7~10]提出并使用矩阵模式来表示训练样本而不是按照传统的矢量方式来表示.特别是文献[10]将矩阵模式引入到最小二乘支持向量机,对传统的SVM方法进行了改造,形成了新的基于矩阵模式的最小二乘支持向量机(matLSSVM),提高了分类效果.然而该方法只是对标准的SVM方法进行了改造,没有考虑类内散度对分类的影响,而且没有形成真正意义上的非线性支持向量机,从而限制了该方法在非线形分类中的运用.

因此,本文将矩阵模式的类内散度矩阵<sup>[9]</sup>引入到最小化类内散度支持向量机(MCSVMs)中,并形成矩阵模式的约束条件,提出基于矩阵模式的最小类内散度支持向量机(MCSVMs<sup>matrix</sup>).MCSVMs<sup>matrix</sup>方法和原方法相比不但继承了原方法的特点,同时具有如下优势:从理论上较好地解决原方法类内散度矩阵奇异性问题;降低求解散度矩阵逆矩阵的复杂度,提高分类效果;首次提出基于矩阵模式的非线性类内散度支持向量机.

### 2 最小类内散度支持向量机:MCSVMs

**定义 1<sup>[4]</sup>** 假设有  $n$  个样本组成的样本集  $D = \{x_1, \dots, x_n\}$ ,  $x_i \in R^M$ , 它们分别属于 2 个不同的类 ( $C_+$ ,  $C_-$ ), 其中大小为  $n_k$  样本子集  $D_k$  属于第  $k$  类, 给定分类决策平面的法向量  $\mu$ , 则类内散度为:  $S_w$ . 其中:

$$S_w = \sum_{k=1}^2 \sum_{x \in D_k} (x - u_k)(x - u_k)^T \quad (1)$$

称为类内散度矩阵

$$u_k = \frac{1}{n_{kx}} \sum_{x \in D_k} x, (k=1, 2) \quad (2)$$

称为均值

**定义 2<sup>[5]</sup>** 设一训练样本集  $TS = \{(x_i, y_i) | i=1, \dots, n\} \subset R^M \times \{1, -1\}$  则 MCSVMs 的原始优化问题为:

$$\min_{\mu, b} \frac{1}{2} \mu^T S_w \mu + C \sum_{i=1}^n \xi_i \quad (3)$$

$$\text{subject to } y_i(\mu^T x_j + b) - 1 - \xi_i, i=1, \dots, n \quad (4)$$

其中:  $\xi_i$  为松弛因子,  $C > 0$  为惩罚参数根据广义的拉格朗日函数和相应的 Karush-Kuhn-Tucker (KKT) 条件<sup>[12]</sup>可以得到式(3)、(4)对应的对偶问题:

$$\min \frac{1}{2} H^T H - 1^T \quad (5)$$

$$\text{subject to } y_i \xi_i = 0, 0 \leq \xi_i \leq C$$

其中:  $H = (h_{ij})_{n \times n}$ ,  $h_{ij} = y_i y_j x_i^T S_w^{-1} x_j$ ,  $\mathbf{1} = (1, \dots, 1)^T \in R^n$ , 根据式(5)的解并按照传统 SVM 解法可以得到 MCSVMs 方法的偏移量  $b$ 、分类函数  $g(x)$ :

$$b = \frac{1}{n_{TS}} \sum_{i \in TS} (y_i - \sum_{j=1}^n y_j x_j^T S_w^{-1} x_i) \quad (6)$$

$$g(x) = \mu^T x + b \begin{cases} > 0 \Rightarrow x \in C_+ \\ < 0 \Rightarrow x \in C_- \end{cases} \quad (7)$$

从以上分析结果来看,在 MCSVMs 方法中由于必须要计算矩阵  $S_w^{-1}$ , 因此在处理小样本高维数据时确实存在类内散度矩阵奇异性的问题, 因此本文使用适当的策略将矩阵模式引入到 MCSVMs 方法中, 从而使得此类问题在一定程度上得以解决.

### 3 基于矩阵模式的最小类内散度支持向量机

在 MCSVMs 方法中当类内散度矩阵  $S_w$  奇异时, 该方法根本不可能得到决策超平面法向量, 因此本文引入矩阵模式的类内散度矩阵, 提出矩阵模式的最小类内散度支持向量机(MCSVMs<sup>matrix</sup>)和非线性矩阵模式的最小类内散度支持向量机(Ker-MCSVMs<sup>matrix</sup>).

#### 3.1 线性矩阵模式的类内散度支持向量机: MCSVMs<sup>matrix</sup>

**定义 3<sup>[9]</sup>** 假设有  $n$  个样本组成的集合  $D = \{A_1, \dots, A_n\}$ ,  $A_i \in R^{d_1 \times d_2}$  它们分别属于 2 个不同的类 ( $C_+$ ,  $C_-$ ), 其中大小为  $n_k$  样本子集  $D_k$  属于第  $k$  类, 给定分类决策平面的法向量  $\mu$ , 则矩阵模式类内散度为:  $\mu^T S_w^{mat} \mu$ .

其中:

$$S_w^{mat} = \sum_{k=1}^2 \sum_{A \in D_k} (A - U_k)(A - U_k)^T \quad (8)$$

称为矩阵模式的类内散度矩阵

$$U_k = \frac{1}{n_{kA}} \sum_{A \in D_k} A, (k=1, 2) \quad (9)$$

称为均值

**定理 1<sup>[8]</sup>** 在 MCSVMs<sup>matrix</sup> 方法中当选取的训练样本集为  $D = \{A_1, \dots, A_n\}$ ,  $A_i \in R^{d_1 \times d_2}$ , 如果当  $n \geq 2 + \frac{d_1}{\min(d_1, d_2)}$  成立时, 则定义 3 中的矩阵模式的类内散度矩阵  $S_w^{mat}$  非奇异.

定理 1 可以充分说明 MCSVMs 方法不能很好地解决小样本高维数据集的根本原因.

为了构造矩阵模式的决策函数, 本文的 MCSVMs<sup>matrix</sup> 方法采用和文献[10]相同的方法来定义:

$$g(A) = \mu^T A + b \quad (10)$$

其中:  $A \in R^{d_1 \times d_2}$ ,  $\mu = (\mu_1, \dots, \mu_{d_1})^T \in R^{d_1}$ ,  $b = (b_1, \dots, b_{d_2})^T \in R^{d_2}$

**定义 4** 设训练样本集  $matTS = \{(A_i, y_i) | i=1, \dots, n\} \subset R^{d_1 \times d_2} \times \{1, -1\}$ , 则 MCSVMs<sup>matrix</sup> 的原始优化问题为:

$$\min_{\mu, b} \frac{1}{2} \mu^T S_w^{mat} \mu + C \sum_{i=1}^n \xi_i \quad (11)$$

$$\text{subject to } y_i(\mu^T A_i + b) - 1 - \epsilon_i, i = 1, \dots, n \quad (12)$$

其中:  $\epsilon_i$  为松弛因子,  $C > 0$  为惩罚参数

**定理 2** MCSVMs<sup>matrix</sup> 原始优化问题式(11)、(12)的对偶问题为:

$$\min \frac{1}{2} \mu^T H^{mat} \mu - 1^T \mu \quad (13)$$

$$\text{subject to } y_i \mu_i = 0, 0 \leq \mu_i \leq C$$

其中:  $H^{mat} = (h_{ij}^{mat})_{n \times n}$ ,  $h_{ij}^{mat} = y_i y_j (A_j^T S_w^{mat-1} A_i)$ ,  $\mathbf{1} = (1, \dots, 1)^T$ ,  $R^n$

**证明** 原始优化问题对应的拉格朗日函数为:

$$L(\mu, b, \epsilon, \lambda) = \frac{1}{2} \mu^T S_w^{mat} \mu + C \sum_{i=1}^n \epsilon_i - \sum_{i=1}^n \lambda_i (y_i(\mu^T A_i + b) - 1 + \epsilon_i) - \sum_{i=1}^n \mu_i (\mu_i - C) \quad (14)$$

其中:  $\lambda = (\lambda_1, \dots, \lambda_n)^T$ ,  $\mu = (\mu_1, \dots, \mu_n)^T$  是拉格朗日系数, 根据 KKT 条件:

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \mu = S_w^{mat-1} \sum_{i=1}^n y_i A_i \lambda_i \quad (15)$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n y_i \lambda_i = 0 \quad (16)$$

$$\frac{\partial L}{\partial \epsilon_i} = 0 \Rightarrow \epsilon_i = C - \mu_i \quad (17)$$

$$\frac{\partial L}{\partial \lambda_i} = 0 \Rightarrow y_i(\mu^T A_i + b) - 1 + \epsilon_i = 0 \quad (18)$$

将式(15)、(16)、(17)结果代入式(14)定理成立。

同时,为了保证式(11)最小化,根据优化理论,梯度下降迭代法具有收敛到局部最优解的特性<sup>[11]</sup>,因此在本文的 MCSVMs<sup>matrix</sup>方法中采用和文献[10]相似的方法得到关于矢量  $\mu$  的迭代公式:

$$\mu^{(k+1)} = \mu^{(k)} - \frac{\partial L}{\partial \mu} = \mu^{(k)} + \sum_{i=1}^n y_i A_i \lambda_i \mu^{(k)} \quad (19)$$

其中:  $\eta > 0$  是学习率,  $k$  是迭代次数

当矢量  $\mu$  经过有限次迭代收敛到  $\mu^*$  后,将  $\mu^*$  代入到式(15)得到  $\mu$  的当前值  $\mu^*$ ,则  $\mu^*$  就是在固定矢量  $\lambda^*$  前提下得到的凸优化问题的局部最优解<sup>[13]</sup>。因此,可以依据当前的  $\mu^*$ ,  $\lambda^*$  来构造决策函数式(10)中的偏移量  $b^*$ 。

$$b^* = \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^n y_j \lambda_j^* (A_j^T \mu^*))^T S_w^{mat-1} (A_i^T \mu^*) \quad (20)$$

由此,可以得到 MCSVMs<sup>matrix</sup> 算法:

Step1:初始化矢量  $\mu_1$ , 惩罚参数  $C$ , 学习率  $\eta$ , 迭代次数  $maxIter$ , 误差  $\epsilon$ ,  $k = 1$ ;

Step2:对矢量表示的样本根据文献[9]的方法转换成矩阵模式,并计算  $S_w^{mat-1}$ ;如果训练样本使用矩阵模式表示,直接计算  $S_w^{mat-1}$ ;

Step3:使用二次规划方法求解对偶问题式(13);

Step4:使用式(15)计算  $\mu^{(k)} = S_w^{mat-1} \sum_{i=1}^n y_i A_i \lambda_i^{(k)}$ ;

Step5:使用式(19)计算  $\mu^{(k+1)} = \mu^{(k)} + \eta \sum_{i=1}^n y_i A_i \lambda_i^{(k)} \mu^{(k)}$ ;

Step6:如果  $\|\mu^{(k+1)} - \mu^{(k)}\| > \epsilon$  &&  $k < maxIter$ , 并令  $k = k + 1$ , 转到 Step3, 否则输出  $\mu^*$ ,  $\lambda^*$ , 转到 Step7;

Step7:使用式(20)计算决策函数偏移量  $b^*$ 。

根据上述方法所得结果就可以使用(10)对任意矩阵模式的测试样本  $A$  进行分类:

$$g(A) = \mu^T A \mu^* + b^* \begin{cases} > 0 \Rightarrow A \in C_+ \\ < 0 \Rightarrow A \in C_- \end{cases} \quad (21)$$

下面讨论 MCSVMs<sup>matrix</sup>方法、MCSVMs 方法对应的运算量:在 MCSVMs 方法中,要想  $S_w$  非奇异,根据定义 1 对应的假设前提并结合定理 1,则 MCSVMs 方法在计算散度矩阵  $S_w$  的时间复杂度为  $O(M^3)$ ,其中  $M$  表示每个样本的维数。然而在 MCSVMs<sup>matrix</sup>算法中,根据定义 3 的假设前提,则计算矩阵模式散度矩阵  $S_w^{mat}$  运算量为  $O(\frac{n}{d^2} M^2)$ ,其中  $n$  为样本数,当该算法处理高维数据集时,根据定理 1,当  $n \ll M$  时就能保证  $S_w^{mat}$  非奇异,则计算  $S_w^{mat}$  的时间复杂度可以认为  $O(M^2)$ ;同样在处理高维数据集时,在该方法的迭代过程中多次计算矢量  $\mu$ 、所需要的运算量为  $O(k(2nM + \frac{M^2}{d^2}))$ ,其中  $k$  为迭代次数,结合算法 Step6 则可以认为  $k$  为有界量,则迭代过程中多次计算矢量  $\mu$ 、所需要的运算量为  $O(M^2)$ 。因此,本文的 MCSVMs<sup>matrix</sup>算法对应的时间复杂度为  $O(M^2)$ 。

诚然,本文 MCSVMs<sup>matrix</sup>算法和一些基本的 SVM 方法相比,由于在实现过程中需要计算矩阵模式的散度矩阵  $S_w^{mat}$ ,以及通过迭代产生矢量  $\mu$ 、,因此具有相对较高的时间复杂度。当然,由于以上讨论的各种方法都要经过二次规划求解过程,因此在讨论时间复杂度时我们没有考虑二次规划求解的运算量。

### 3.2 非线性的 MCSVMs<sup>matrix</sup>方法:Ker-MCSVMs<sup>matrix</sup>

通过研究式(13)、(20)和(21)发现,在三式中都有相似的二次式  $(A_i^T S_w^{mat-1} A_j)$  出现,因此可以采用与 MCSVMs 相似的方法提出非线性的 MCSVMs<sup>matrix</sup>: Ker-MCSVMs<sup>matrix</sup>。

由于  $S_w^{mat-1}$  是实正定矩阵,则可以将二次式  $(A_i^T S_w^{mat-1} A_j)$  定义成内积的形式:  $(S_w^{mat-1/2} A_i)^T (S_w^{mat-1/2} A_j) = (S_w^{mat-1/2} A_i, S_w^{mat-1/2} A_j)$ ,由此,可以寻找一非线性函数  $\phi$  将矢量  $S_w^{mat-1/2} A_i$  映射到高维特征空间:

$$\phi: S_w^{mat^{-1/2}} A_i \quad \phi( S_w^{mat^{-1/2}} A_j ) \quad (22)$$

$$(\phi( S_w^{mat^{-1/2}} A_i ), \phi( S_w^{mat^{-1/2}} A_j )) = K( S_w^{mat^{-1/2}} A_i , S_w^{mat^{-1/2}} A_j ) \quad (23)$$

其中  $K(,)$  一般是 Mercer 核函数, 根据式 (23) 可以分别得到 Ker-MCSVMs<sup>matrix</sup> 方法的对偶问题、决策函数的偏移量和决策函数. Ker-MCSVMs<sup>matrix</sup> 的对偶问题为:

$$\min \frac{1}{2} \quad \tau H_{ker}^{mat} - 1^T \quad (24)$$

subject to  $y_i \quad i=0, 0 \quad i \quad C,$

其中:  $H_{ker}^{mat} = ( H_{ij, ker}^{mat} )_{n \times n} \quad H_{ij, ker}^{mat} = y_i y_j K( S_w^{mat^{-1/2}} A_j^* , S_w^{mat^{-1/2}} A_i^* ), \mathbf{1} = (1, \dots, 1)^T \quad R^n$

Ker-MCSVMs<sup>matrix</sup> 的决策函数偏移量为:

$$b_{ker}^* = \frac{1}{n} \sum_{i=1}^n (y_i - y_j) K( S_w^{mat^{-1/2}} A_i^* , S_w^{mat^{-1/2}} A_j^* ) \quad (25)$$

Ker-MCSVMs<sup>matrix</sup> 的决策分类函数为:

$$g(A) = \sum_{i=1}^n y_i K( S_w^{mat^{-1/2}} A^* , S_w^{mat^{-1/2}} A_i^* ) + b_{ker}^* \begin{cases} > 0 \Rightarrow A \in C_+ \\ < 0 \Rightarrow A \in C_- \end{cases} \quad (26)$$

对于实际使用 Ker-MCSVMs<sup>matrix</sup> 方法的两点说明: 在式 (24)、(25)、(26) 中矢量  $\tau$  都使用的是  $\mathbf{1}$ , 而  $\mathbf{1}$  是 MCSVMs<sup>matrix</sup> 方法经过梯度下降法迭代收敛而成的, 这种选取策略是合适的. 这是因为: 一方面在 Ker-MCSVMs<sup>matrix</sup> 方法中当矢量  $\tau$  固定时, 确实可以得到非线性方法相对于矢量  $\tau$  的局部最优解, 此时就设定固定矢量  $\tau$  为  $\mathbf{1}$ ; 另一方面根据式 (22), 在 Ker-MCSVMs<sup>matrix</sup> 方法中是将  $S_w^{mat^{-1/2}} A^*$  整体映射到高维特征空间, 由于  $\mathbf{1}$  是收敛值, 这样在一定程度上保证单元  $S_w^{mat^{-1/2}} A^*$  是所有相似输入单元中保留信息最丰富、性质最稳定的, 因此, 在 Hilbert 空间中所对应的内积更能正确反映输入单元之间的相似性. 诚然, 就 Ker-MCSVMs<sup>matrix</sup> 方法本身来说, 一般可以认为在使用该方法前已经得到最优矢量  $\mathbf{1}$ , 所以, 在本文的实验部分测试 Ker-MCSVMs<sup>matrix</sup> 方法时, 我们认为在测试前已经得到了最优矢量  $\mathbf{1}$ .

可以根据求解问题来具体设计相应的核函数, 这是因为  $S_w^{mat^{-1}}$  是实正定矩阵, 则  $S_w^{mat^{-1/2}} R^{d_1 \times d_1} [6]$ , 则式 (23) 是一个普遍意义上的核函数, 因此, 在解决实际问题时可以采取和标准核函数

相似的构造方法得到 Ker-MCSVMs<sup>matrix</sup> 方法中对应的核函数. 然而, 通过研究常用核函数发现, 在构造这些常用核函数时并不需要单独计算  $S_w^{mat^{-1/2}} A^*$  这个分量. 表 1 提供了几种常用核函数在 Ker-MCSVMs<sup>matrix</sup> 方法中实际使用的形式.

### 4 实验

通过以上理论分析, 本文方法不管在处理小样本高维数据集, 还是在算法的复杂度方面都比 MCSVMs 方法具有明显的优势. 下面通过测试两种不同数据表示形式的小样本高维数据集来说明上述优势. 一种是高维的矢量表示的基因数据集: 结肠癌基因数据集 (Colon Tumor) [14~16] (<http://www.molbio.princeton.edu/Colondata>), 另一种为矩阵模式表示的图像数据集: Yale、ORL 人脸识别数据集 [9,10] (<http://www.cs.uiuc.edu/homes/dengcai2/>).

#### 4.1 测试 Colon Tumor 基因数据集

结肠癌数据集 (Colon Tumor) 是二分类数据集, 该数据集包括 62 个样本, 每一个样本包括 2000 维特征, 其中有 40 个正类样本, 22 个负类样本. 在测试本文的 MCSVMs<sup>matrix</sup> 和非线性的 Ker-MCSVMs<sup>matrix</sup> 方法时, 任意选取正类 30 个、负类 20 个作为训练样本, 剩余的 12 作为测试样本, 并令误差控制量  $\epsilon = 1e-3$ , 最大迭代次数  $maxIter = 1000$ ,  $\mathbf{1} = (1, \dots, 1)^T$ . 通过测试 Colon Tumor 基因数据集来说明本文方法在将矢量数据转换成矩阵模式后的基本分类能力 (表 2).

实验具体设计过程为: 根据文献 [9] 的方法将每个矢量数据分别转换为  $15 \times 80, 40 \times 50, 80 \times 25$  三种矩阵模式来分别测试. 根据文献 [14~16] 对 Colon Tumor 数据集的分析, 在测试过程中核化的 C-SVM 和 Ker-MCSVMs<sup>matrix</sup> 都采用 Gaussian RBF 核分类器. 为了更好地选择 MCSVMs<sup>matrix</sup>、Ker-MCSVMs<sup>matrix</sup> 中惩罚参数  $C$ 、学习率参数  $\eta$  和核宽度参数  $\gamma$ , 以及 C-SVM、核化的 C-SVM 方法中的相应参数, 本实验使用常用的 10-折交叉验证 [17] 方法对上述参数进行选取. 其中上述参数测试取值分别为:  $\eta = [0.5, 1, 10]$ ,

表 1 Ker-MCSVMs<sup>matrix</sup> 方法中几种常用的核函数形式

| Type of Classifiers      | Kernel Functions  |
|--------------------------|---|
| Linear                   | $K( S_w^{mat^{-1/2}} A_i^* , S_w^{mat^{-1/2}} A_j^* ) = (A_j^*)^T S_w^{mat^{-1}} (A_i^*)$                               |
| Gaussian RBF             | $K( S_w^{mat^{-1/2}} A_i^* , S_w^{mat^{-1/2}} A_j^* ) = \exp( - (A_j^* - A_i^*)^T S_w^{mat^{-1}} (A_j^* - A_i^*) ) / 2$ |
| Polynomial of degree $q$ | $K( S_w^{mat^{-1/2}} A_i^* , S_w^{mat^{-1/2}} A_j^* ) = ((A_j^*)^T S_w^{mat^{-1}} (A_i^*) + c)^q$                       |
| Sigmoid                  | $K( S_w^{mat^{-1/2}} A_i^* , S_w^{mat^{-1/2}} A_j^* ) = \tanh( k (A_j^*)^T S_w^{mat^{-1}} (A_i^*) + d )$                |

其中  $c > 0, c = 0, k > 0, d < 0$

100, 1000],  $C = [0.5, 1, 2^4, 2^8, 2^{10}]$ ,  $\gamma = [0.5, 1, 2^4, 2^8, 2^{10}]$ , 则交叉验证的结果和相应的分类精度、测试 CPU 时间见表 2.

从表 2 可以看出当矢量表示的数据被转换成矩阵模式时, 本文的 MCSVMs<sup>matrix</sup> 和 Ker-MCSVMs<sup>matrix</sup> 方法同样表现出较强的分类能力, 这说明在将矢量型数据转换成矩阵模式时, 本文方法在一定程度上并没有降低原始数据的信息量. 根据定理 1, 由于在测试过程中选取训练样本数为 50, 这样就会充分保障本文方法在处理不同矩阵模式时得到的类内散度的非奇异性, 而 MCSVMs 方法根本没有办法处理该训练数据集 (因为对应的类内散度矩阵则是奇异的), 这就在一定程度上说明本文方法在处理小样本高维数据集具有较强的分类能力. 同时根据表 2 三种不同矩阵模式的测试效果来看, 对于不同的矩阵模式精度存在着差异, 这说明本文的 MCSVMs<sup>matrix</sup>、Ker-MCSVMs<sup>matrix</sup> 方法所得到的分类精度在一定程度上依赖于转换的矩阵模式.

还可以从表 2 中还可以看出本文的方法 MCSVMs<sup>matrix</sup> 在 CPU 所消耗的时间比 C-SVM 方法要多, 这是由于本文的 MCSVMs<sup>matrix</sup> 方法在分类过程中需要进行迭代优化运算, 诚然当选用模式适当, 还是有比较满意的效果的 (比如模式取  $\gamma = 1$  &  $C = 1, 40 \times 50$  时, CPU 时间只要 0.312 秒). 而本文的核化方法 Ker-MCSVMs<sup>matrix</sup> 方法 (假

表 2 MCSVMs<sup>matrix</sup>、Ker-MCSVMs<sup>matrix</sup>和 C-SVMs 分类精度、CPU 时间比较

| Algorithm   | Accuracy  | CPU Time (s) |
|---|---|--------------|
| C-SVMs  | 0.83333 ( $C = 1$ )   | 0.124        |
| C-SVMs with Gaussian RBF Kernel                       | 0.83333 ( $C = 1$ & $\gamma = 1$ )                                | 0.203        |
| MCSVMs <sup>matrix</sup>                              | 0.75 ( $\gamma = 1$ & $C = 1, 25 \times 80$ )                     | 1.2792       |
|   | 0.75 ( $\gamma = 1$ & $C = 1, 40 \times 50$ )                     | 0.312        |
|   | 0.83333 ( $\gamma = 1$ & $C = 0.5, 80 \times 25$ )                | 12.8914      |
| Ker-MCSVMs <sup>matrix</sup> with Gaussian RBF Kernel | 0.83333 ( $\gamma = 1$ & $C = 1$ & $\gamma = 10, 25 \times 80$ )  | 0.3432       |
|   | 0.83333 ( $\gamma = 1$ & $C = 1$ & $\gamma = 0.5, 40 \times 50$ ) | 0.3588       |
|   | 0.91667 ( $\gamma = 1$ & $C = 0.5$ & $\gamma = 1, 80 \times 25$ ) | 0.47392      |

设在测试前已经得到最优矢量  $\gamma^*$  ) 和核化的 C-SVM 在消耗的 CPU 时间上具有相同数量级, 因此, 本文的 Ker-MCSVMs<sup>matrix</sup> 不但在精度上有一定的优势, 而且具有一定的运算效率.

### 4.2 测试图像数据

常规的人脸图像数据是使用矩阵模式进行存储的, 因此在实验过程中分别使用 2D-FLDA + 1NN<sup>[8]</sup>、mat-FLSSVM<sup>[10]</sup> 和本文的 MCSVMs<sup>matrix</sup>、Ker-MCSVMs<sup>matrix</sup> 方法测试 Yale、ORL 人脸识别数据集来表明本文方法在处理以矩阵模式存储的二分类问题时的分类能力. Yale 数据集 (32 × 32) 是包含 15 个类别的人脸数据, 同一类中有 11 种不同表情的人脸数据 (见图 1); ORL 数据集 (32 × 32) 包含 40 个类别的人脸数据, 同一类中有 10 种不同表情的人脸数据 (见图 2).

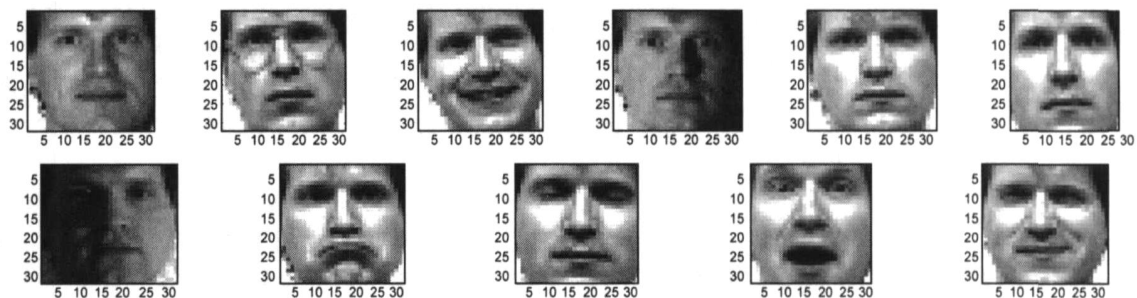


图1 Yale数据集中某一类中所有图像

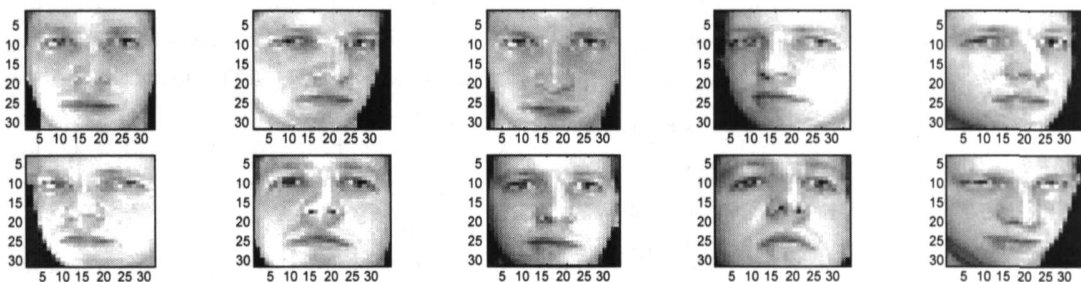


图2 ORL数据集中某一类所有图像

实验设计: 在测试本文 MCSVMs<sup>matrix</sup>、Ker-MCSVMs<sup>matrix</sup> 方法处理图像数据分类能力时, 采用随机抽取 Yale、ORL 数据集中的任意两类作为测试数据集, 其中在 Yale

数据集中随机抽取三个数据子集为: Yale. 1、Yale. 2、Yale. 3, 同时在每一个数据子集中随机选取 10 个图像样本作为训练样本, 剩下的 12 个样本作为测试样本; 在

ORL 数据集中随机抽取的三个数据子集为:ORL\_1、ORL\_2、ORL\_3,同时在每一个数据子集中随机选取 8 个样本作为训练样本,剩下的 12 个样本作为测试样本.令误差控制量  $\epsilon = 1e - 8$ ,最大迭代次数  $maxIter = 500$ ,  $\mathbf{1} = (1, \dots, 1)^T$ ,学习率参数  $\eta = 1$ 、惩罚参数  $C = 0.5$ ,  $d = 1$  而在 2D-FLDA + 1NN 方法中令  $d = 10$ ,matFLSSVM 参数选取和本文方法相同.分类效果、测试消耗的 CPU 时间见表 3.

表 3 四种分类方法的精度、CPU 时间比较

| Datasets | Algorithm                    | Accuracy       | CPU Time (s)   |
|----------|------------------------------|----------------|----------------|
| Yale_1   | 2D-FLDA                      | 0.75           | 0.0347         |
|          | matFLSSVM                    | 0.75           | 11.622         |
|          | MCSVMs <sup>matrix</sup>     | 0.83333        | 19.126         |
|          | Ker-MCSVMs <sup>matrix</sup> | <b>0.91667</b> | <b>0.0312</b>  |
| Yale_2   | 2D-FLDA                      | 0.8333         | <b>0.0539</b>  |
|          | matFLSSVM                    | 0.8333         | 18.97          |
|          | MCSVMs <sup>matrix</sup>     | 0.8333         | 19.921         |
|          | Ker-MCSVMs <sup>matrix</sup> | <b>0.91667</b> | 0.0624         |
| Yale_3   | 2D-FLDA                      | 0.91667        | <b>0.0453</b>  |
|          | matFLSSVM                    | 0.8333         | 18.892         |
|          | MCSVMs <sup>matrix</sup>     | 0.91667        | 20.412         |
|          | Ker-MCSVMs <sup>matrix</sup> | <b>1</b>       | 0.078          |
| ORL_1    | 2D-FLDA                      | 0.91667        | <b>0.01563</b> |
|          | matFLSSVM                    | <b>1</b>       | 17.363         |
|          | MCSVMs <sup>matrix</sup>     | 0.91667        | 16.599         |
|          | Ker-MCSVMs <sup>matrix</sup> | <b>1</b>       | 0.0624         |
| ORL_2    | 2D-FLDA                      | 0.8333         | <b>0.0429</b>  |
|          | matFLSSVM                    | 0.91667        | 13.806         |
|          | MCSVMs <sup>matrix</sup>     | <b>1</b>       | 16.645         |
|          | Ker-MCSVMs <sup>matrix</sup> | <b>1</b>       | 0.0468         |
| ORL_3    | 2D-FLDA                      | 0.8333         | 0.0276         |
|          | matFLSSVM                    | 0.8333         | 13.354         |
|          | MCSVMs <sup>matrix</sup>     | 0.91667        | 15.382         |
|          | Ker-MCSVMs <sup>matrix</sup> | <b>0.91667</b> | <b>0.016</b>   |

从表 3 可以看出,在处理人脸数据集时本文 MCSVMs<sup>matrix</sup>、Ker-MCSVMs<sup>matrix</sup> 方法具有较强的分类能力.其中 MCSVMs<sup>matrix</sup> 方法和 matFLSSVM 方法相比具有较高的分类精度,这说明本文方法中引入类内散度对提高人脸识别效果起到一定的作用.特别是非线性方法 Ker-MCSVMs<sup>matrix</sup> 在测试以上 6 个数据子集都具有最高的分类精度(其中 3 个数据子集时精度达到 100%).从消耗的 CPU 时间来看,由于 matFLSSVM 方法和本文的 MCSVMs<sup>matrix</sup> 方法由于都需要进行迭代优化,则所消耗的时间远远高于 2D-FLDA 方法.然而,测试本文的 Ker-MCSVMs<sup>matrix</sup> 方法(假设在测试前已经得到最优矢量  $\mathbf{w}^*$ )和 2D-FLDA 方法相比具有较相似的 CPU 时间,这说明如果给本文 Ker-MCSVMs<sup>matrix</sup> 方法赋予适当的矢量  $\mathbf{w}^*$ ,则可以保证较好的分类精度和分类效率.

## 5 总结和未来工作

本文通过分析文献[5]的 MCSVMs 方法发现该方法在处理小样本高维数据集时会引起类内散度矩阵奇异性问题,因此将矩阵模式的类内散度引入到支持向量机中,提出基于矩阵模式的类内散度支持向量机: MCSVMs<sup>matrix</sup>,该方法不但继承了原方法的优点,而且在理论上讨论了矩阵模式类内散度矩阵非奇异性的条件,从而为该方法在处理小样本高维数据集时提供了理论和应用上的支持.为了更好地处理非线性分类问题,本文通过将 Mercer 核函数引入到 MCSVMs<sup>matrix</sup> 方法中首次提出真正意义上基于矩阵模式的非线性支持向量机: Ker-MCSVMs<sup>matrix</sup>,通过实验表明该方法具有较强的非线性分类能力.诚然,本文方法在求解矢量  $\mathbf{w}^*$  时使用了经典的梯度下降迭代方法,从而在一定程度上增加了算法的时间复杂度,因此,如何选择更好的优化方法去求解矢量  $\mathbf{w}^*$ 、提高方法的分类效率将是我们以后研究的目标.

### 参考文献:

- [1] Vanpanik V. Statistical Learning Theory [M]. New York: Wiley Press, 1998.
- [2] Scholkopf B, Smola A. Learning with Kernels-Support Vector Machine, Regularization, Optimization, and Beyond [M]. Cambridge, MA: MIT Press, 2002.
- [3] Pontil M, Verri A. Support vector machine for 3D object recognition [J]. IEEE Trans on PAMI, 1998, 20(6): 673-646.
- [4] 吴小俊, 杨静宇, 王士同, 等. 广义统计不相关最优鉴别矢量集的一个理论结果 [J]. 电子学报, 2004, 32(10): 1720-1722.  
Wu Xiaojun, Yang Jingyu, Wang Shitong, et al. A theoretical result on the generalized optimal set of statistically uncorrelated discriminant vectors [J]. Acta Electronica Sinica, 2004, 32(10): 1720-1722. (in Chinese)
- [5] Tefas A, Kotropoulos C, Pitas I. Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication [J]. IEEE Transactims on PAMI, 2001, 23(7): 735-746.
- [6] Zafeiriou S F, Tefas A, Pitas I. Minimum class variance support vector machines [J]. IEEE Transactims on Image Processing, 2007, 16(10): 2551-2564.
- [7] Yang J, Zhang D, Frangi A F, Yang J U. Two-dimension PCA: a new approach to appearance-based face representation and recognition [J]. Pattern Anal. Mach. Intell., 2004, 26(1): 131-137.
- [8] Li M, Yuan B Z. 2D-LDA: a statistical linear discriminant analysis for image matrix [J]. Pattern Recognition Letters, 2004

- (26) :527-532.
- [9] Chen S, Zhu Y, Zhang D, Yang J. Feature extraction approaches based on matrix pattern: matPCA and matFLDA[J]. Pattern Recognition Letters, 2005, (26) :1157-1167.
- [10] Wang Z, Chen S. New least squares support vector machines based on matrix patterns[J]. Neural Processing Letters, 2007, (26) :41-56.
- [11] 边肇祺, 张学工. 模式识别[M]. 北京:清华大学出版社, 2001.
- [12] 邓乃阳, 田英杰. 数据挖掘中的新方法-支持向量机[M]. 北京:科学出版社, 2004.
- [13] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods [M]. Cambridge:Cambridge University Press, 2000.
- [14] 李建中, 杨昆, 高宏. 考虑样本不平衡模型无关的基因选择方法[J]. 软件学报, 2006, 17(7) :1485-1493.  
Li Jianzhong, Yang Kun, Gao Hong. Model-Free gene selection method by considering unbalanced samples[J]. Journal of Software, 2006, 17(7) :1485 - 1493. (in Chinese)
- [15] 邓林, 马尽文, 裴健. 秩和基因选取方法及其肿瘤诊断中的应用[J]. 科学通报, 2004, 49(15) :1652 - 1657.  
Deng Lin, Ma Jinwen, Pei Jian. Rank sum method for related gene selection and its application to tumor diagnosis[J]. Chinese Science Bulletin, 2004, 49(15) :1652 - 1657. (in Chinese)
- [16] 王树林, 王戟, 等. 肿瘤信息基因启发式宽度优先搜索算法研究[J]. 计算机学报, 2008, 31(4) :636 - 649.

Wang Shulin, Wang Ji, et al. Heuristic breadth-first search algorithm for information gene selection based on gene expression profiles[J]. Chinese Journal of Computers, 2008, 31(4) : 636-649. (in Chinese)

- [17] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性简化[J]. 软件学报, 2008, 19(3) :640-649.

Hu Qinghua, Yu Daren, Xie Zongxia. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. Journal of Software, 2008, 19(3) : 640-649. (in Chinese)

#### 作者简介:



皋 军 男. 副教授. 1971 年生于江苏阜宁. 1996 年、2004 年分别在四川大学、南京航空航天大学获得理学学士、工学硕士学位. 2007 年进入江南大学信息工程学院攻读博士学位, 从事数据挖掘、人工智能、模式识别和模糊系统方面的有关研究工作.

E-mail : gjxlin@yahoo.cn



王士同 男. 教授、博士生导师、中国计算机学会高级会员, 1964 年生于江苏邗江. 1984 年、1987 年在南京航空航天大学获得工学学士、硕士学位. 现为江南大学信息工程学院院长, 从事人工智能、模式识别、模糊系统、医学图像处理和生物信息学方面的研究工作.