文章编号:1001 - 9081(2010)01 - 0143 - 03

多分类簇支持向量机方法

王之怡,杨一帆
(西南财经大学 经济信息工程学院,成都 610074)
(wangzy_t@ swufe. edu. cn)

摘　要:针对支持向量机的多分类问题,提出一种新颖的基于非平行超平面的多分类簇支持向量机。它针对 k 模式分类问题分别训练产生 k 个分割超平面,每个超平面尽量靠近自身类模式而远离剩余类模式;决策时,新样本的类别由它距离最近的超平面所属的类决定,克服了一对一(OAO)和一对多(OAA)等传统方法存在的"决策盲区"和"类别不平衡"等缺陷。基于 UCI 和 HCL2000 数据集的实验表明,新方法在处理多分类问题时,识别精度显著优于传统多分类支持向量机方法。

关键词:支持向量机;超平面;核函数;手写体汉字识别
中图分类号: TP391.4　　**文献标志码**:A

Multi-class cluster support vector machines

WANG Zhi-yi, YANG Yi-fan
(*School of Economics Information Engineering, Southwest University of Finance and Economics, Chengdu Sichuan* 610074, *China*)

Abstract: Based on the idea of nonparallel hyperplanes, a novel multi-class cluster support vector machine method was proposed to settle the multi-class classification problem of support vector machines. For a k-class classification problem, it trained k-hyperplanes respectively, and each one lay as close as possible to self-class while being apart from the rest classes as far as possible. Then, labels of new samples were determined by the class of their nearest hyperplane, thus the inherent limitations of One-Against-One (OAO) and One-Against-All (OAA) methods can be avoided, such as "decision blind-area" and "unbalanced classes". Finally, experiments on UCI and HCL2000 datasets show that the proposed method significantly outperforms traditional OAO and OAA in terms of recognition accuracy.

Key words: Support Vector Machine (SVM); hyperplane; kernel function; handwritten Chinese character recognition

0　引言

支持向量机(Support Vector Machine, SVM)是一种新型的基于统计学习理论的机器学习方法,具有推广能力强、维数不敏感等优点[1]。大量研究表明[2-3],它在处理文本分类、手写体汉字识别、图像检索等高维小样本数据时,识别精度显著优于神经网络、贝叶斯网络、隐马尔可夫模型等传统机器学习方法。然而,SVM 在本质上解决的是二值模式分类问题,虽然已提出 OAA(One-Against-All)、OAO(One-Agianst-One)、ECOC 等方法能够将它扩展到多分类问题[4],但是,这些方法尚存在"决策盲区"、"不平衡类"等缺陷,分类器的推广能力受到一定影响。多分类支持向量机方法仍然是机器学习的研究热点之一。

本文将广义特征值支持向量机(Proximal Support Vector Machine via Generalized Eigenvalues, GEPSVM)的非平行超平面的思想引入到多分类问题[5],提出一种新颖的多分类簇支持向量机方法,并推导了它的非线性形式。基于 UCI 和 HCL2000 的实验表明,新方法能够取得非常优异的识别精度。

1　支持向量机及其多分类问题

考虑 n 维实空间 $\mathbf{R}^n$ 中的模式二分类问题[1]:已知 m 个训练样本 $A_i(i = 1,2,\cdots,m)$ 和它们的类别标记 $y_i \in \{1, -1\}$,寻找一个分类函数使得 $\mathbf{R}^n$ 划分为两个子空间,两类样本分别属于不同的子空间。这里,$A_i = (A_{i1}, A_{i2}, \cdots, A_{in}) \in \mathbf{R}^n$。在两类样本线性可分的情况下,超平面(1)可将它们区分开:

$$\boldsymbol{\varpi} \cdot x + b = 0 \tag{1}$$

其中,$\boldsymbol{\varpi} \in \mathbf{R}^n$ 且 $b \in \mathbf{R}$。考虑到两类样本到超平面都应该有一定的距离,超平面还应该满足如下的不等式约束:

$$y_i(A_i\boldsymbol{\varpi} + b) \geqslant 1;\ i = 1,2,\cdots,m \tag{2}$$

显然,满足要求的分类超平面不止一个。在结构风险最小化准则下,支持向量机寻找最优分类超平面,使得两类之间的分类间隔 $\frac{2}{\|\boldsymbol{\varpi}\|_2}$ 最大,等价于求解如下的二次优化:

$$\underset{\boldsymbol{\varpi},b}{\mathrm{Min}} \frac{1}{2}\boldsymbol{\varpi}^{\mathrm{T}}\boldsymbol{\varpi} \tag{3}$$

$$\text{s.t. } y_i(A_i\boldsymbol{\varpi} + b) \geqslant 1;\ i = 1,2,\cdots,m$$

然而,完全满足优化问题(3)的精确分类超平面是不经常存在的。因此,考虑到某些训练样本可能不满足约束条件(2),引入松弛变量:

$$\zeta_i \geqslant 0;\ i = 1,2,\cdots,m \tag{4}$$

根据结构风险最小化准则,分类超平面不仅应该使分类间隔最大,同时应保持训练样本分类误差尽可能小。因此,优化问题(3)变为:

$$\underset{\boldsymbol{\varpi},b,\zeta}{\mathrm{Min}} \frac{1}{2}\boldsymbol{\varpi}^{\mathrm{T}}\boldsymbol{\varpi} + C\sum_{i=1}^{m}\zeta_i \tag{5}$$

收稿日期:2009 - 07 - 21;**修回日期**:2009 - 08 - 31。
基金项目:国家自然科学基金资助项目(69732010);西南财经大学科学研究基金资助项目(QN0806)。
作者简介:王之怡(1964 -),男,上海人,副教授,主要研究方向:图像处理、电子支付;　杨一帆(1970 -),男,四川乐山人,博士研究生,主要研究方向:电子支付。

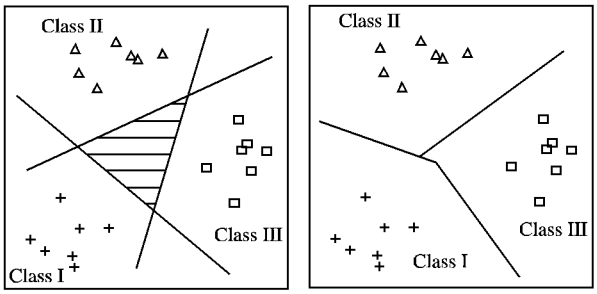
s.t. $y_i(A_i\boldsymbol{\varpi}+b)\geqslant 1-\zeta_i;\zeta_i\geqslant 0,i=1,2,\cdots,m$

其中，C 称为“罚参数”，在分类器的推广能力和误分类率之间进行折中。求解优化(3)或(5)的对偶问题，得到支持向量机的决策函数，如下：

$$f(x)=\mathrm{sgn}(\boldsymbol{\varpi}\cdot\boldsymbol{x}+b) \tag{6}$$

然而，支持向量机是二值分类器。为解决 k 模式分类问题($k\geqslant 2$)，OAA 和 OAO 等方法通过组合多个二值分类器进行集体决策来实现多分类[4,6]。

OAA 方法总共训练 k 个二值 SVM 分类器，其中每个分类器的决策超平面将一类模式与剩余所有类的模式分割开，如图 1(a)所示。OAO 方法在每一对模式之间都训练一个二值 SVM 分类器，总共训练 $\frac{k(k-1)}{2}$ 个分类器，如图 1(b)所示。进行决策时，二值 SVM 分类器分别对新样本 $\boldsymbol{x}$ 进行判别，并按照多数投票的方法决策出 $\boldsymbol{x}$ 的最终类别。



(a) OAA (b) OAO

图 1 OAA 和 OAO 多分类方法

然而，OAA 方法存在以下明显缺陷：

1) 决策“盲区”难以避免(图 1 的阴影区域)。如果输入样本位于该区域，则 OAA 方法不容易对它的类别进行正确判别。

2) 容易引起不平衡类的问题。当 k 取值较大时，任意一类样本与其剩余样本的规模差异很大，使得分类器的推广能力受到影响。

对于 OAO 方法，它训练的部分二值 SVM 分类器需要被迫进行错误决策，将致使推广能力下降。例如，图 1 中在“Ⅰ”和“Ⅲ”类样本上训练的二值分类器为 $H_{\mathrm{Ⅰ,Ⅲ}}$；如果输入样本 $\boldsymbol{x}$ 属于“Ⅱ”类，显然，$H_{\mathrm{Ⅰ,Ⅲ}}$ 无论怎样决策，都会做出错误判断。

分析可知，支持向量机强制它的决策超平面 $\boldsymbol{\varpi}\cdot\boldsymbol{x}+b=0$ 平分两个边界超平面 $\boldsymbol{\varpi}\cdot\boldsymbol{x}+b=\pm 1$ 之间的最大间隔。这对于平衡类情况下的二模式分类问题非常适合，但对多模式分类问题，反而是分类器推广能力下降的主要原因之一。本文将 GEPSVM 的非平行超平面的思想引入到多分类问题，提出一种新颖的多分类方法(MC-SVM)。

2 多类簇支持向量机

假设 $A_i^{(1)}(i=1,2,\cdots,m_1)$ 和 $A_i^{(2)}(i=1,2,\cdots,m_2)$ 分别表示属于“+1”类和“-1”类的训练样本。GEPSVM 方法通过求解两个不平行的超平面 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\varpi}^{(1)}+b^{(1)}=0$ 和 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\varpi}^{(2)}+b^{(2)}=0$ 共同对未知样本的类别进行决策。每一个超平面尽量靠近一类样本而尽量远离另外一类样本，并可由以下的优化问题求解[5]：

$$\underset{\boldsymbol{\varpi},b\neq 0}{\mathrm{Min}}\ \frac{(\|\boldsymbol{A}^{(1)}\boldsymbol{\varpi}+\boldsymbol{e}b\|_2^2+\delta\|[\boldsymbol{\varpi},b]^{\mathrm{T}}\|_2^2)}{\|\boldsymbol{A}^{(2)}\boldsymbol{\varpi}+\boldsymbol{e}b\|_2^2} \tag{7}$$

其中：$\boldsymbol{A}^{(1)},\boldsymbol{A}^{(2)}$ 分别是 $m_1\times n$ 和 $m_2\times n$ 维矩阵，是两类样本的矩阵表示；$\delta>0$ 是控制超平面复杂度的正则参数。文献[5]指出，去掉超平面的平行限制条件后，分类器的推广能力并不会受到显著影响；相反，由于式(7)是一个典型的广义特征值问题，能够大幅度地提高分类器的训练速度。

本文将非平行超平面的思想扩展到 SVM 的多分类问题，提出一种新颖的多分类簇支持向量机——MC-SVM。它针对 k 模式的分类问题，同时求解 k 个超平面，其中每个超平面尽量靠近自身类模式，而远离剩余类的模式(如图 2 所示)。

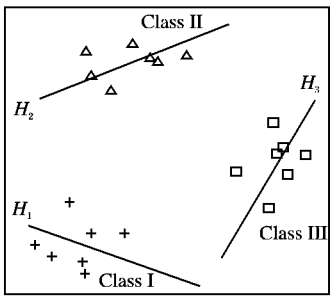


图 2 多分类簇支持向量机的超平面

设属于第 k 类的样本分别用矩阵 $\boldsymbol{A}^{(1)},\boldsymbol{A}^{(2)},\cdots,\boldsymbol{A}^{(k)}$ 表示，数量分别为 $m_1,m_2,\cdots,m_k$(且满足 $m_1+m_2+\cdots+m_k=m$)。此外，令 k 个超平面分别表示为 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\varpi}^{(i)}+b^{(i)}=0,i=1,2,\cdots,k$，则它们可以通过求解如下的一组二次优化问题得到：

$$\underset{\boldsymbol{\varpi}^{(1)},b^{(1)},\boldsymbol{\zeta}}{\mathrm{Min}}\ \frac{1}{2}\|\boldsymbol{A}^{(1)}\boldsymbol{\varpi}^{(1)}+\boldsymbol{e}_1b^{(1)}\|_2^2+C_1\bar{\boldsymbol{e}}_1\boldsymbol{\zeta} \tag{8}$$

$$\text{s.t. } \bar{\boldsymbol{Y}}^{(1)}(\bar{\boldsymbol{A}}^{(1)}\boldsymbol{\varpi}^{(1)}+\bar{\boldsymbol{e}}_1b^{(1)})\geqslant\bar{\boldsymbol{e}}_1-\boldsymbol{\zeta},\boldsymbol{\zeta}\geqslant 0$$

$$\vdots$$

$$\underset{\boldsymbol{\varpi}^{(k)},b^{(k)},\boldsymbol{\zeta}}{\mathrm{Min}}\ \frac{1}{2}\|\boldsymbol{A}^{(k)}\boldsymbol{\varpi}^{(k)}+\boldsymbol{e}_kb^{(k)}\|_2^2+C_k\bar{\boldsymbol{e}}_k\boldsymbol{\zeta} \tag{9}$$

$$\text{s.t. } \bar{\boldsymbol{Y}}^{(k)}(\bar{\boldsymbol{A}}^{(k)}\boldsymbol{\varpi}^{(k)}+\bar{\boldsymbol{e}}_kb^{(k)})\geqslant\bar{\boldsymbol{e}}_k-\boldsymbol{\zeta},\boldsymbol{\zeta}\geqslant 0$$

其中，$\bar{\boldsymbol{A}}^{(1)}$ 表示由第Ⅰ类之外的所有训练样本构成的数据矩阵，维数为 $(m-m_1)\times n$；$\bar{\boldsymbol{Y}}^{(1)}$ 是其类标记(均视为“-1”类)组成的对角矩阵。$\boldsymbol{\zeta}$ 是松弛向量，$\boldsymbol{e}_1$ 和 $\bar{\boldsymbol{e}}_1$ 是全 1 向量。

上述 k 个优化问题显然是同构的，并且和 SVM 的优化问题(5)在形式上一致，这是称之为“多类簇支持向量机”的原因。以式(8)为例对它的特点进行分析。

显然，目标函数的第一项表示所有属于第Ⅰ类的样本距离超平面的欧氏距离之和。约束条件要求所有不属于第Ⅰ类的样本与超平面的距离都至少为 1。与 SVM 类似，引入松弛向量 $\boldsymbol{\zeta}$ 对约束条件进行松弛。目标函数的第二项表示对所有靠近超平面的非第Ⅰ类的样本进行惩罚，以使得它们尽量远离超平面。因此，优化问题(8)求解得到的是靠近第 1 类样本，而远离剩余所有样本的最优超平面。

注意到，优化问题(8)中只对第Ⅰ类之外的样本进行约束，这种方式避免了多类 SVM 方法中常见的“类别不平衡”问题。

优化问题(8)的拉格朗日对偶问题为[1]：

$$L(\boldsymbol{\varpi}^{(1)},b^{(1)},\boldsymbol{\zeta},\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{2}\|\boldsymbol{A}^{(1)}\boldsymbol{\varpi}^{(1)}+\boldsymbol{e}_1b^{(1)}\|_2^2+$$

$$C_1\bar{\boldsymbol{e}}_1\boldsymbol{\zeta}-\boldsymbol{\alpha}^{\mathrm{T}}(\bar{\boldsymbol{Y}}^{(1)}(\bar{\boldsymbol{A}}^{(1)}\boldsymbol{\varpi}^{(1)}+\bar{\boldsymbol{e}}_1b^{(1)})-\bar{\boldsymbol{e}}_1+\boldsymbol{\zeta})-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\zeta} \tag{10}$$

其中，$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_{(m-m_1)})^{\mathrm{T}}$ 和 $\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_{(m-m_1)})^{\mathrm{T}}$ 是大于 0 的拉格朗日乘子。根据约束优化问题取极值时的 KKT 条件，函数 L 应满足以下条件：

$$\frac{\partial L}{\partial \boldsymbol{\varpi}^{(1)}} = \boldsymbol{A}^{(1)\mathrm{T}}(\boldsymbol{A}^{(1)}\boldsymbol{\varpi}^{(1)} + \boldsymbol{e}_1 b^{(1)}) - \bar{\boldsymbol{A}}^{(1)\mathrm{T}}\bar{\boldsymbol{Y}}^{(1)}\boldsymbol{\alpha} = 0 \tag{11}$$

$$\frac{\partial L}{\partial b^{(1)}} = \boldsymbol{e}_1^{\mathrm{T}}(\boldsymbol{A}^{(1)}\boldsymbol{\varpi}^{(1)} + \boldsymbol{e}_1 b^{(1)}) - \bar{\boldsymbol{e}}_1^{\mathrm{T}}\bar{\boldsymbol{Y}}^{(1)}\boldsymbol{\alpha} = 0 \tag{12}$$

$$\frac{\partial L}{\partial \boldsymbol{\zeta}} = C_1\bar{\boldsymbol{e}}_1 - \boldsymbol{\alpha} - \boldsymbol{\beta} = 0 \tag{13}$$

$$\bar{\boldsymbol{Y}}^{(1)\mathrm{T}}(\bar{\boldsymbol{A}}^{(1)}\boldsymbol{\varpi}^{(1)} + \bar{\boldsymbol{e}}_1 b^{(1)}) \geqslant \bar{\boldsymbol{e}}_1 - \boldsymbol{\zeta}, \boldsymbol{\zeta} \geqslant 0 \tag{14}$$

$$\boldsymbol{\alpha}^{\mathrm{T}}(\bar{\boldsymbol{Y}}^{(1)\mathrm{T}}(\bar{\boldsymbol{A}}^{(1)}\boldsymbol{\varpi}^{(1)} + \bar{\boldsymbol{e}}_1 b^{(1)}) - \bar{\boldsymbol{e}}_1 + \boldsymbol{\zeta}) = 0, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\zeta} = 0 \tag{15}$$

$$\boldsymbol{\alpha} \geqslant 0, \boldsymbol{\beta} \geqslant 0 \tag{16}$$

定义新矩阵 $\boldsymbol{P} = [\boldsymbol{A}^{(1)} \quad \boldsymbol{e}_1]$ 和 $\boldsymbol{Q} = [\bar{\boldsymbol{A}}^{(1)} \quad \bar{\boldsymbol{e}}_1]$,以及向量 $\boldsymbol{u}^{\mathrm{T}} = [\boldsymbol{\varpi}^{(1)\mathrm{T}} \quad b^{(1)}]$,合并式(11)和(12),得到:

$$\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{u} - \boldsymbol{Q}^{\mathrm{T}}\bar{\boldsymbol{Y}}^{(1)}\boldsymbol{\alpha} = 0 \Rightarrow \boldsymbol{u} = (\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P})^{-1}\boldsymbol{Q}^{\mathrm{T}}\bar{\boldsymbol{Y}}^{(1)}\boldsymbol{\alpha} \tag{17}$$

由式(13)和(16)容易推得:

$$0 \leqslant \boldsymbol{\alpha} \leqslant C_1 \tag{18}$$

将式(13)和(17)代入函数 L,经过整理得到优化问题(8)的 Wolfe 对偶问题:

$$\underset{\boldsymbol{\alpha}}{\mathrm{Max}}\ \bar{\boldsymbol{e}}_1^{\mathrm{T}}\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P})^{-1}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\alpha} \tag{19}$$

s.t. $0 \leqslant \boldsymbol{\alpha} \leqslant C_1$

实矩阵 $\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}$ 是 $(n+1)\times(n+1)$ 阶半正定,$|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}| > 0$ 时则逆矩阵存在;若出现 $|\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}| = 0$ 的情况,可以采用岭回归的方法,用 $(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P} + \varepsilon\boldsymbol{I})$ 的逆替代[7]。其中,$\varepsilon > 0$ 是非常小的实数。

显然,优化(19)与标准 SVM 的对偶问题在形式上完全一致,许多成熟的算法可以对它进行求解[1]。

求解出 $\boldsymbol{\alpha}$ 的值后,代入式(17),立即可以第 1 类样本的超平面方程 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\varpi}^{(1)} + b^{(1)} = 0$。对于剩余的 $k-1$ 个超平面,可以按照类似的方法求出它们的超平面方程,即:

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\varpi}^{(i)} + b^{(i)} = 0, i = 1,2,\cdots,k \tag{20}$$

此后,多类簇支持向量机的决策函数为:

$$f(x) = \underset{i=1,2,\cdots,k}{\mathrm{argMin}} |\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\varpi}^{(i)} + b^{(i)}| \tag{21}$$

也即,MC-SVM 将新样本 $\boldsymbol{x}$ 的类别判断为其最靠近的超平面所在的类别。相比于 OAA 和 OAO 方法通过多数投票来决定 x 所属的类别,MC-SVM 能够有效避免“决策盲区”的发生(见图 1)。

3　非线性核函数的情况

核函数在现代机器学习中起着重要作用。满足 Mercer 条件的核函数能够将 n 维输入空间 $\mathbf{R}^n$ 中的样本映射到某个高维特征空间 F,即非线性映射 $\Phi: \boldsymbol{x} \mapsto \varphi(\boldsymbol{x})$ 使得样本在 F 中更容易被分类[1]。通过引入核函数 $K(\boldsymbol{x},\boldsymbol{z}) = \varphi(\boldsymbol{x}) \cdot \varphi(\boldsymbol{z})$,对多类簇支持向量机进行扩展。

假设 MC-SVM 在高维空间 F 中的 k 个超平面为:

$$K(\boldsymbol{x},\boldsymbol{B}^{\mathrm{T}})\boldsymbol{\psi}^{(i)} + b^{(i)} = 0;\ i = 1,2,\cdots,k \tag{22}$$

其中,$\boldsymbol{B}^{\mathrm{T}} = [\boldsymbol{A}^{(1)\mathrm{T}} \quad \bar{\boldsymbol{A}}^{(1)\mathrm{T}}]$ 是全部样本的数据矩阵。显然,当核函数为线性函数 $K(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}$ 时,令 $\boldsymbol{\varpi}^{(i)} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\psi}^{(i)}$ 便可恢复出公式(20)的线性超平面。

容易得出非线性核函数情况下的多类簇支持向量机的原始优化问题:

$$\underset{v^{(1)},b^{(1)},\boldsymbol{\zeta}}{\mathrm{Min}}\ \frac{1}{2}\|K(\boldsymbol{A}^{(1)\mathrm{T}},\boldsymbol{B}^{\mathrm{T}})\boldsymbol{\psi}^{(1)} + \boldsymbol{e}_1 b^{(1)}\|_2^2 + C_1\bar{\boldsymbol{e}}_1\boldsymbol{\zeta} \tag{23}$$

s.t. $\bar{\boldsymbol{Y}}^{(1)}(K(\bar{\boldsymbol{A}}^{(1)\mathrm{T}},\boldsymbol{B}^{\mathrm{T}})\boldsymbol{\psi}^{(1)} + \bar{\boldsymbol{e}}_1 b^{(1)}) \geqslant \bar{\boldsymbol{e}}_1 - \boldsymbol{\zeta}, \boldsymbol{\zeta} \geqslant 0$

$$\vdots$$

$$\underset{v^{(k)},b^{(k)},\boldsymbol{\zeta}}{\mathrm{Min}}\ \frac{1}{2}\|K(\boldsymbol{A}^{(k)\mathrm{T}},\boldsymbol{B}^{\mathrm{T}})\boldsymbol{\psi}^{(k)} + \boldsymbol{e}_k b^{(k)}\|_2^2 + C_k\bar{\boldsymbol{e}}_k\boldsymbol{\zeta} \tag{24}$$

s.t. $\bar{\boldsymbol{Y}}^{(k)}(K(\bar{\boldsymbol{A}}^{(k)\mathrm{T}},\boldsymbol{B}^{\mathrm{T}})\boldsymbol{\psi}^{(k)} + \bar{\boldsymbol{e}}_k b^{(k)}) \geqslant \bar{\boldsymbol{e}}_k - \boldsymbol{\zeta}, \boldsymbol{\zeta} \geqslant 0$

与前面类似,通过拉格朗日对偶方法对(23)进行求解。经过整理,得到其对偶问题:

$$\underset{\boldsymbol{\alpha}}{\mathrm{Max}}\ \bar{\boldsymbol{e}}_1^{\mathrm{T}}\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{\alpha} \tag{25}$$

s.t. $0 \leqslant \boldsymbol{\alpha} \leqslant C_1$

其中,$\boldsymbol{H} = [K(\boldsymbol{A}^{(1)\mathrm{T}},\boldsymbol{B}^{\mathrm{T}}) \quad \boldsymbol{e}_1]$ 和 $\boldsymbol{G} = [K(\bar{\boldsymbol{A}}^{(1)\mathrm{T}},\boldsymbol{B}^{\mathrm{T}}) \quad \bar{\boldsymbol{e}}_1]$;而向量 $\boldsymbol{v} = [\boldsymbol{\psi}^{(1)\mathrm{T}} \quad b^{(1)}]^{\mathrm{T}}$ 满足:

$$\boldsymbol{v} = [\boldsymbol{\psi}^{(1)\mathrm{T}} \quad b^{(1)}]^{\mathrm{T}} = (\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{G}^{\mathrm{T}}\bar{\boldsymbol{Y}}^{(1)}\boldsymbol{\alpha} \tag{26}$$

因此,求解出二次优化问题(25)后,即可得到第Ⅰ类样本在高维空间 F 中的超平面方程。对于剩余的 $k-1$ 个超平面,可以按相同的方法得到。

4　实验与结果分析

为了验证本文提出的多类簇支持向量机方法的性能,首先在 7 个 UCI 评测数据集上分别采用 MC-SVM、OAA、OAO 和 ECOC[4] 方法训练分类器,并比较它们取得的识别精度。其中,ORHD 和 PRHD 分别表示数据集“optical recognition of handwritten digits”和“pen-Based recognition of handwritten digits”。

公平起见,实验中所有分类器均采用高斯核函数,支持向量机的罚参数和核参数通过 5-fold 交叉验证方法选取。表 1 给出了 4 种多分类方法获得的测试精度。

表 1　UCI 数据集上获得的测试精度比较　%

数据集	MC-SVM	OAA	OAO	ECOC
segment	97.81	96.17	96.29	96.32
heart	86.68	84.07	83.81	85.92
vehicle	89.30	87.05	87.47	88.66
letter	88.89	86.21	86.34	88.04
waveform	90.55	88.54	88.76	88.60
ORHD	99.13	98.26	98.20	98.73
PRHD	99.20	98.03	98.31	99.02

实验结果表明,在所有数据集上,MC-SVM 训练分类器获得的测试精度显著优于 OAA 和 OAO 方法,平均提升了约 1.90% 和 1.77%。而 ECOC 方法相对不稳定,在 segment 和 waveform 数据集上的精度明显偏低。

进一步,将 MC-SVM 方法应用在金融手写体汉字自动识别系统中,检验其训练多分类器时的性能。训练数据取自北京邮电大学手写汉字样本库 HCL2000 的 21 个金融汉字(如图 3 所示)。

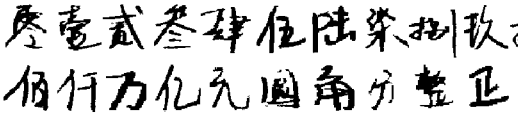


图 3　常用的 21 个金融手写体汉字

每个汉字随机抽取 800 个图像样本(前 500 个用于训练,后 300 个用于测试)。图像经预处理后,分别采用弹性网格骨架方向特征(EM_TDF)、弹性网格轮廓方向特征(EM_CDF)、弹性网格边缘方向特征(EM_EDF)和弹性网格轮廓方向角特征(EM_CDAF)、弹性网格笔画方向特征(EM_SDF)5种特征

(下转第 149 页)

4　结语

本文针对多序列比对问题,设计了一个基于 GC-GM 的穷举遗传算法,把传统遗传算法对多序列比对矩阵的处理转变为对空位矩阵的处理,无须存储每个残基,只需存储比对中空位位置,极大地节省了计算机的存储空间;针对空位矩阵设计了新的 GC 算子和 GM 算子,有力地拓展了搜索空间,避免了早熟收敛。模拟数值实验表明所提出的算法比较有效地解决多序列比对问题,为多序列比对问题提供了一条新的思路。

参考文献:

[1] 刘立芳,霍红卫,王宝树. PHGA-COFFEE: 多序列比对问题的并行混合遗传算法求解[J]. 计算机学报, 2006,29(5):727 -733.

[2] 胡桂武,郑启轮,彭宏. 基于遗传算法与星比对的多序列比对混合算法[J]. 计算机应用, 2004,24(5):90 -91.

[3] 张敏,方伟武,张俊华,等. 基于迭代渐进多序列比对算法研究[J]. 计算机工程, 2005,31(17):32 -33.

[4] 王小平,曹立明. 遗传算法—理论、应用与软件实现[M]. 西安:西安交通大学出版社,2002.

[5] 葛宏伟,梁艳春. 基于隐马尔可夫模型和免疫粒子群优化的多序列比对算法[J]. 计算机研究与发展, 2006,43(8):1330 -1336.

[6] HORNG J T, WU LICHENG, LIN CHINGMEI, *et al*. A genetic algorithm for multiple sequence alignment[J]. Soft Computing, 2005, 9(6):407 -420.

[7] NEEDLEMAN S B, WUNSCH C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins[J]. Journal of Molecular Biology, 1970, 48(3): 443 -453.

[8] FENG D F, DOOLITTLE R F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees[J]. Journal of Molecular Evolution, 1987, 25(4): 351 -360.

[9] NOTREDAME C, HIGGINS D G. SAGA : Sequence alignment by genetic algorithm[J]. Nucleic Acids Research, 1996, 24(8):1515 -1524.

[10] WANG YI, LI G B. Multiple sequence alignment using an exhaustive and greedy algorithm[J]. Journal of Bioinformatics and Computational Biology, 2005,3(2): 243 -255.

[11] THOMPSON J D, PLEWNIAK F, POCH O. BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs[J]. Bioinformatics 1999,15(1):87 -88.

[12] ISOKAWA M, WAYAMA M, SHIMIZU T. Multiple sequence alignment using a genetic algorithm[EB/OL]. [2009 -04 -20]. http://www.genome.jp/manuscripts/GIW96/Poster/GIW96P06.pdf.

[13] ZHANG C, WONG A K. A genetic algorithm for multiple molecular sequence alignment[J]. Computer Applications in the Biosciences, 1997, 13(6):565 -581.

[14] NOTREDAME C, O' BRIEN E A, HIGGINS D G. RAGA: RNA sequence alignment by genetic algorithm[J]. Nucleic Acids Research, 1997, 25(22):4570 -4580.

[15] WANG L, JIANG T. On the complexity of multiple sequence alignment[J]. Journal of Computational Biology, 1994, 1(4): 337 -348.

(上接第 145 页)

提取方法产生 5 组不同的数据集,每个样本含 256 维特征。

在生成的 5 个数据集上,分别采用 MC-SVM、OAA、OAO、ECOC 方法训练多类分类器,模型参数仍通过 5-fold 交叉验证方法选取。表 2 给出高斯核函数情况下,4 种多分类方法获得的测试精度。

显然,将 MC-SVM 应用于金融汉字识别问题能取得较好的识别精度,相比其他 3 种多类支持向量机方法,平均识别精度提高了 1.13% ~1.64%。其中,在"弹性网格轮廓方向角特征"数据集上,采用高斯核函数的 MC-SVM 取得99.68% 的最高测试精度,高于剩余 3 种方法中最高的98.76%。

表 2　HCL2000 数据集上获得的测试精度比较　%

数据集	MC-SVM	OAA	OAO	ECOC
EM_TDF	97.81	96.33	96.49	96.62
EM_CDF	97.29	95.56	95.43	95.97
EM_EDF	99.03	97.71	98.02	98.30
EM_CDAF	99.68	98.30	98.24	98.76
EM_SDF	98.46	97.37	97.50	98.03

基于 UCI 和 HCL2000 数据集上的实验结果表明,MC-SVM 由于采用了非平行超平面去逼近不同类样本的分布区域,避免了 OAA、OAO 等方法存在的"决策盲区"和"不平衡类"等不足,能够很好地解决传统支持向量机在处理多分类问题时的困难。

5　结语

本文针对 OAA 和 OAO 等多分类支持向量机方法存在的"决策盲区"和"类别不平衡"等问题,提出一种新颖的基于非平行超平面的多类簇支持向量机方法,并利用拉格朗日函数推导了它的对偶二次优化形式进行求解。基于 UCI 和 HCL2000 数据集的实验结果表明,新方法能够获得非常高的识别精度,明显优于传统的多类支持向量机方法。

参考文献:

[1] VAPNIK V N. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000.

[2] 高学,金连文,尹俊勋. 一种基于支持向量机的手写汉字识别方法[J]. 电子学报,2002,30(5):651 -654.

[3] DONG J X, KRZYZAK A, SUEN C Y. An improved handwritten Chinese character recognition system using support vector machine [J]. Pattern Recognition Letters, 2005,26(12):1849 -1856.

[4] HSU C W, LIN C J. A comparison of methods for multiclass support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415 -425.

[5] MANGASARIAN O L, WILD E W. Multisurface proximal support vector machine classification via generalized eigenvalues[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005,27 (12):1 -6.

[6] WANG Y C, CASASENT D. New support vector-based design method for binary hierarchical classifiers for multi-class classification problems[J]. Neural Networks, 2008,21(4):502 -510.

[7] SAUNDERS C, GAMMERMAN A, VOVK V. Ridge regression learning algorithm in dual variables[C]// Proceedings of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1998: 515 -521.