适合大样本快速训练的最大夹角间隔核心集向量机

胡文军[1,2],王士同[1],邓赵红[1]
(1. 江南大学信息工程学院,江苏无锡 214122;2. 湖州师范学院信息与工程学院,浙江湖州 313000)

摘 要: 许多核化形式的分类方法如 SVM,SVDD 等都是对应一个二次规划(QP)问题,而核矩阵计算需要 $O(m^2)$空间复杂度,求解 QP 需要 $O(m^3)$时间复杂度,限制了这类方法对大样本数据的训练.本文基于一种新的分类间隔概念提出最大向量夹角间隔分类器 MAMC,目标是在样本空间找到最优向量 $\boldsymbol{c}$,测试样本通过 $\boldsymbol{c}$ 与训练样本之间的最大化向量夹角间隔 ρ(称为 Margin)实现分类.同时,文中证明了该方法的核化形式等价于核化的最小包络球 MEB 问题,并通过引入核心集向量机 CVM 将 MAMC 扩展为 MAM-CVM,进而快速实现对大样本的训练和分类.人造和真实数据集实验表明了 MAMC 和 MAM-CVM 算法的有效性.

关键词: 向量夹角间隔; 核化方法; 核心集向量机; 最小包络球
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2011) 05-1178-07

Maximum Vector-Angular Margin Core Vector Machine Suitable for Fast Training for Large Datasets

HU Wen-jun[1,2], WANG Shi-tong[1], DENG Zhao-hong[1]
(1. *School of Information Engineering, Jiangnan University, Wuxi, Jiangsu* 214122, *China*;
2. *School of Information and Engineering, Huzhou Teachers College, Huzhou, Zhejiang* 313000, *China*)

Abstract: Many kernelized classification methods, such as SVM and SVDD, are formulated as quadratic programming (QP) problems, but computing kernel matrix would require $O(m^2)$ computation, and solving QP may take up to $O(m^3)$, which limits these methods to train on large datasets. In this paper, a new classification method called Maximum Vector-Angular Margin Classifier (MAMC) is proposed, based on a new concept of margin called vector-angular margin, to find an optimal vector $\boldsymbol{c}$ in patterns' feature space and all the testing points can be classified in terms of the maximum vector-angular margin ρ between the vector $\boldsymbol{c}$ and all the training points. Meanwhile, the kernelized MAMC can be equivalently formulated as the kernelized Minimum Enclosing Ball (MEB), and thus MAMC can be extended to Maximum Vector-Angular Margin Core Vector Machine (MAM-CVM) by introducing Core Vector Machine (CVM) method, to solve the training and classification for large datasets. Experimental results on artificial and real datasets are provided to validate the effectiveness of the proposed methods here.

Key words: vector-angular margin, kernel method, core vector machine (CVM), minimum enclosing ball (MEB)

1 引言

分类是模式识别和机器学习中非常重要的一个内容.较著名的分类器包括:支持向量机(Support Vector Machine,SVM)[1],v-SVC[2],多核 SVM[3],最小类内散度 SVM(Minimum Within-class Scatter Support Vector Machine,MCSVM)[4],模糊核超球感知器(Fuzzy Kernel Hyperball Perceptron, FKHP)[5],椭球核机(Ellipsoidal Kernel Machine,EKM)[6]等用于解决二类及多类问题;支持向量数据描述(Support Vector Data Description,SVDD)[7],小球体大间隔(Small Sphere and Large Margin,SSLM)[8]等方法用于一类或非正常数据检测.一般地,核化使得这些方法适应性更好并能解决线性不可分问题,而对应核化后的方法往往可以描述成二次规划(Quadratic Programming, QP)问题,若是一个严格凸优化问题,必会得到全局最优解.但是,对于大样本数据集,QP 求解的计算复杂度是相当可观,比如含 m 个样本的 SVM,至少需要 $O(m^3)$时间和 $O(m^2)$空间复杂度[9].因此,如何降低计算复杂度成了研究热点,较流行的方法是通过逼近替代 QP 问题中核矩阵的计算[10],如 Nyström 法[11],贪婪逼近法[12],采样法[13],矩阵分解法[14]等.

最近,Tsang 等人证明了一类硬 SVDD、一/二类 L_2-SVM、L_2-SVR 以及 Ranking SVM 等价于 MEB 或中心约束最小球(Center Constrain MEB, CC-MEB)[10,15]问题,并提出核心集向量机(Core Vector Machine,CVM)[10]和一般化的 CVM 算法[15]用于逼近 MEB 和 CC-MEB,其时间复杂

收稿日期:2010-06-13;修回日期:2010-10-18
基金项目:国家自然科学基金(No.60903100,No.60975027)

度与训练样本数成线性关系,而空间复杂度与训练样本数无关[10,15]. Deng 等人建立了子集密度估计器(Reduced Set Density Estimator, RSDE)和模糊推理系统(Fuzzy Inference Systems, FIS)与 MEB 之间的联系[16,17],并提出了快速子集密度估计器(Fast Reduced Set Density Estimator, FRSDE)[16]和快速 ML 模糊推理系统(Fast Mamdani-Larsen FIS, ML-FIS)[17]用于大样本训练.

实际上,上述等价于 MEB 或 CC-MEB 的 L_2-SVM 方法是原始 L_2-SVM 的特殊化,而本文从一种新的分类间隔概念(称为向量夹角间隔)出发,提出一种新的分类算法,并能更直接的通过等价于 CC-MEB 问题解决大样本.思想是在样本空间中找到最优向量 $\boldsymbol{c}$,测试样本可以通过 $\boldsymbol{c}$ 与训练样本之间的向量夹角间隔进行分类,并尽可能使得分类的向量夹角间隔最大化,如图 1 所示,两类样本分别被蓝色和黑色线条夹角分开,故将本文方法称为最大向量夹角间隔分类器(Maximum Vector-Angular Margin Classifier, MAMC). MAMC 通过等价 CC-MEB 后,利用 CVM 技巧将其拓展为最大向量夹角间隔核心集向量机(Maximum Vector-Angular Margin Core Vector Machine, MAM-CVM)用于处理大样本.

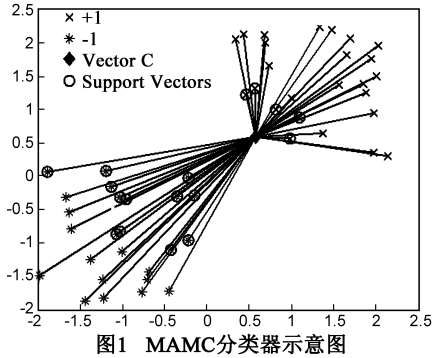

图1　MAMC分类器示意图

2　最大向量夹角间隔分类器

定义:含两类的样本集 $X=\{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in R^d\}$,$\boldsymbol{x}_i$ 为列向量,类标签 $y_i \in \{+1,-1\}$,当 $i=1,2,\cdots,m_1$ 时,$y_i=+1$,当 $i=m_1+1,\cdots,m_1+m_2(m_1+m_2=m)$时,$y_i=-1$.

2.1　原始形式

根据第 1 节思想,在样本特征空间找最优向量 $\boldsymbol{c}$,当采用$\dfrac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{c}}{\|\boldsymbol{c}\|.\|\boldsymbol{x}_i\|}$反映样本点 $\boldsymbol{x}_i$ 与 $\boldsymbol{c}$ 间向量夹角距离,则该算法的优化问题描述为:

$$\min_{\rho,\boldsymbol{c}} -v\rho+\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{c}\|^2 \tag{1}$$

$$\text{s.t.}\quad y_i\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{c}}{\|\boldsymbol{c}\|\cdot\|\boldsymbol{x}_i\|}\right)\geqslant\rho,1\leqslant i\leqslant m \tag{2}$$

其中,ρ 称为向量夹角距离(Angular Distance)意义上的间隔 Margin,式(1)中第 2 项的目的是为了让 $\boldsymbol{c}$ 尽可能地接近样本中心,以便使对应的 VC 维尽量小,$v>0$为调节参数.为方便推导,式(2)改成式(3)中的约束项,这也可方便地通过对样本正则化实现.由此得到优化问题的新模型:

$$\min_{\rho,\boldsymbol{c}} -v\rho+\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{c}\|^2 \tag{3}$$
$$\text{s.t.}\quad y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{c})\geqslant\rho,1\leqslant i\leqslant m$$

构造上述模型的拉格朗日方程:

$$L(\rho,\boldsymbol{c},\boldsymbol{\alpha})=-v\rho+\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{c}\|^2+\sum_{i=1}^{m}\alpha_i(\rho-y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{c}) \tag{4}$$

其中,$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_m)^{\mathrm{T}}\geqslant 0$ 是拉格朗日乘子,方程 $L(\rho,\boldsymbol{c},\boldsymbol{\alpha})$分别对原始问题变量 ρ 和 $\boldsymbol{c}$ 求偏导数:

$$\frac{\partial L}{\partial\rho}=0\Rightarrow\sum_{i=1}^{m}\alpha_i=v \tag{5}$$

$$\frac{\partial L}{\partial\boldsymbol{c}}=0\Rightarrow\boldsymbol{c}=\sum_{i=1}^{m}\left(\frac{1}{m}+\frac{\alpha_i y_i}{2}\right)\boldsymbol{x}_i \tag{6}$$

将式(5)、(6)代入式(4),可得原始问题的对偶形式:

$$\max_{\boldsymbol{\alpha}}\sum_{i=1}^{m}\alpha_i y_i\left(\frac{-4}{m}\sum_{j=1}^{m}\boldsymbol{x}_j^{\mathrm{T}}\right)\boldsymbol{x}_i-\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$$
$$\text{s.t.}\quad\sum_{i=1}^{m}\alpha_i=v,\alpha_i\geqslant 0,1\leqslant i\leqslant m \tag{7}$$

显然对偶问题是一个 QP 问题.为求解 ρ,构造集合 U:

$$U=\{\boldsymbol{x}_i\mid 0<\alpha_i<v,1\leqslant i\leqslant m\} \tag{8}$$

根据 K.K.T.定理,当 α_i 满足条件 $0<\alpha_i<v$ 时,式(3)中的约束不等式等号成立,因此:

$$\rho=P/|U| \tag{9}$$

其中,$P=\sum_{\boldsymbol{x}_i\in U}y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{c}=\sum_{\boldsymbol{x}_i\in U}y_i\boldsymbol{x}_i^{\mathrm{T}}\sum_{k=1}^{m}\left(\frac{1}{m}+\frac{\alpha_k y_k}{2}\right)\boldsymbol{x}_k$

2.2　核化形式

一般地,真实样本空间很难做到准确划分,为此需要进行核化,其实质是找到一个合适的映射 $\varphi:\boldsymbol{x}_i\in R^d\rightarrow\varphi(\boldsymbol{x}_i)\in R^D(d\ll D)$,并用核函数 $k(\boldsymbol{u},\boldsymbol{v})$表示映射后的内积 $\varphi(\boldsymbol{u})^{\mathrm{T}}\varphi(\boldsymbol{v})$,式(3)核化后的对偶问题为:

$$\max_{\boldsymbol{\alpha}}\sum_{i=1}^{m}\alpha_i y_i\left(\frac{-4}{m}\sum_{j=1}^{m}\varphi(\boldsymbol{x}_j)^{\mathrm{T}}\right)\varphi(\boldsymbol{x}_i)-\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)$$
$$\text{s.t.}\quad\sum_{i=1}^{m}\alpha_i=v,\alpha_i\geqslant 0,1\leqslant i\leqslant m$$

用 $k(\cdot,\cdot)$替换上式内积:

$$\max_{\boldsymbol{\alpha}}\sum_{i=1}^{m}\alpha_i y_i\left(\frac{-4}{m}\sum_{j=1}^{m}k(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)-\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{10}$$
$$\text{s.t.}\quad\sum_{i=1}^{m}\alpha_i=v,\alpha_i\geqslant 0,1\leqslant i\leqslant m$$

对应 $\boldsymbol{c}$ 和 P 的核化形式：

$$\boldsymbol{c}=\sum_{i=1}^{m}\left(\frac{1}{m}+\frac{\alpha_i y_i}{2}\right)\varphi(\boldsymbol{x}_i) \tag{11}$$

$$\begin{aligned}P&=\sum_{\boldsymbol{x}_i\in U}\alpha_i\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{c}\\&=\sum_{\boldsymbol{x}_i\in U}y_i\sum_{k=1}^{m}\left(\left(\frac{1}{m}+\frac{\alpha_k y_k}{2}\right)k(\boldsymbol{x}_k,\boldsymbol{x}_i)\right)\end{aligned} \tag{12}$$

2.3 决策函数

为测试一个新样本 $\boldsymbol{x}\in R^d$ 的类别，只需判断是否在对应的夹角中，其决策函数为：

$$\begin{aligned}f(\boldsymbol{x})&=\mathrm{sgn}(\varphi(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{c})\\&=\mathrm{sgn}\left(\sum_{i=1}^{m}\left(\frac{1}{m}+\frac{\alpha_i y_i}{2}\right)k(\boldsymbol{x}_i,\boldsymbol{x})\right)\end{aligned} \tag{13}$$

2.4 MAMC 的优势与计算复杂度分析

根据 Vapnik-Chervonenkis 维(VC 维)理论可知，更小的 VC 维可以获得更优的分类界面[5,6]，并且 VC 维大小满足条件：$\mathrm{VC}\leqslant\min\{r^2/M^2,d\}+1$，这里 r 是包络所有样本的球半径，M 是分类间隔，d 是样本维数. 可知 SVM 追求间隔 M(SVM 的分类间隔为 $2/\|\boldsymbol{w}\|$)最大，而 SVDD 追求最小包络球(即 r 最小)使得 VC 更小. 对于 MAMC 的数学模型，要求最优向量 $\boldsymbol{c}$ 尽可能接近样本中心进而使得样本包络球半径 r 尽可能小，同时通过最大化夹角间隔 ρ(分类间隔为 2ρ)强化 VC 维更小. 可见，MAMC 从这两方面获得更小 VC 维从而获得更优的分类界面，第 5 节 MAMC 实验结果也验证了这一结论.

但是，MAMC 的核化模型跟其他核化方法(如 SVM，SVDD)一样，实质上都是求解 QP 问题. 式(10)中的一次和二次项复杂度均为 $O(m^2)$，但一次项中 $\sum_{j=1}^{m}k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 实际上是核矩阵的第 i 行元素之和，因此 MAMC 的计算复杂度仍为 $O(m^2)$，求解 QP 需要 $O(m^3)$，所以 MAMC 同 SVM、SVDD 一样很难实现大样本训练，那么 MAMC 是否也像一类硬 SVDD、一/二类 L_2-SVM 等价于 MEB 问题，并实现对大样本训练呢?

3 MEB 和 CC-MEB 最小球问题

最小包络球 MEB 问题在模式识别中得到了广泛研究和应用[14~20]，特别是利用 MEB 的核心集 Core Set 解决大样本问题. 定义球心为 $\boldsymbol{c}$，半径为 R 的球 $B(\boldsymbol{c},R)$，给定 m 个数据的样本集 $X=\{\boldsymbol{x}_i\mid\boldsymbol{x}_i\in R^d\}$，其中 $\boldsymbol{x}_i$ 为列向量，$1\leqslant i\leqslant m$，原始样本空间通过函数 $\varphi:\boldsymbol{x}_i\to\varphi(\boldsymbol{x}_i)$ 映射到高维特征空间 Γ.

3.1 MEB 最小球

最小包络球 MEB 问题可用下列优化问题描述：

$$\begin{aligned}&\min_{R,\boldsymbol{c}}R^2\\&\text{s.t.}\quad\|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2\leqslant R^2\end{aligned} \tag{14}$$

引入核函数核化后，MEB 问题可描述为：

$$\begin{aligned}&\min_{R,\boldsymbol{c}}R^2\\&\text{s.t.}\quad\|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2\leqslant R^2\end{aligned} \tag{15}$$

通过构造拉格朗日函数，可以得到该优化问题的对偶形式：

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}\\&\text{s.t.}\quad\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1\end{aligned} \tag{16}$$

其中，$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_m)^{\mathrm{T}}_{1\times m}\geqslant 0$ 是拉格朗日乘子，$\boldsymbol{1}=(1,1,\cdots,1)^{\mathrm{T}}_{1\times m}$，$\boldsymbol{K}=[k(\boldsymbol{x}_i,\boldsymbol{x}_j)]_{m\times m}$ 为核矩阵.

Tsang 等人在文献[10,15]中指出，如果引入的核函数满足条件 $k(\boldsymbol{x}_i,\boldsymbol{x}_i)=k$($k$ 为某一常数)，那么核化的对偶形式如同式(16)的优化问题均可视为 MEB 问题，如硬边界一类 SVDD 和一/二类 L_2-SVM 核化形式，这样一些求解 MEB 问题的优势方法可以使用，如利用 MEB 核心集的快速逼近法等.

3.2 中心约束 CC-MEB 最小球

文献[15]Tsang 等人将 MEB 拓展到 CC-MEB 后，建立了 L_2-SVR，Ranking SVM 等算法与 MEB 之间的联系，从而利用一般化的 CVM 算法实现对大样本的训练.

在 CC-MEB 中，将 Γ 空间中每个样本点 $\varphi(\boldsymbol{x}_i)$ 引入一属性项 $\delta_i\in R$，构建一个拓展的 MEB 高 1 维特征空间 Γ'，其样本为 $\begin{bmatrix}\varphi(\boldsymbol{x}_i)\\ \delta_i\end{bmatrix}$，且拓展的 MEB 中心点约束为 $\begin{bmatrix}\boldsymbol{c}\\ 0\end{bmatrix}$，其中 $\boldsymbol{c}$ 是 MEB 在 Γ 特征空间中的中心点. 因而 MEB 在 Γ' 空间中优化问题可表述为式(17)，即为 CC-MEB 问题.

$$\begin{aligned}&\min_{R,\boldsymbol{c}}R^2\\&\text{s.t.}\quad\|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}^2+\delta_i\leqslant R^2\end{aligned} \tag{17}$$

其对偶形式如下：

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K})+\boldsymbol{\Delta})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}\\&\text{s.t.}\quad\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1\end{aligned} \tag{18}$$

其中，

$$\boldsymbol{\Delta}=[\delta_1^2,\delta_2^2,\cdots,\delta_m^2]^T\geqslant 0 \tag{19}$$

此时，有

$$R=\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K})+\boldsymbol{\Delta})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}} \tag{20}$$

$$\boldsymbol{c}=\sum_{i=1}^{m}\alpha_i\varphi(\boldsymbol{x}_i) \tag{21}$$

Γ' 空间中任一点到球心距离的平方：

$$\left\|\begin{bmatrix}\varphi(\boldsymbol{x}_i)\\ \delta_i\end{bmatrix}-\begin{bmatrix}\boldsymbol{c}\\ 0\end{bmatrix}\right\|^2=\|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2+\delta_i^2 \tag{22}$$

由于约束条件 $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1$，式(18)可改写成：

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K})+\Delta-\eta\boldsymbol{1})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}\\&\text{s.t.}\quad\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1\end{aligned} \tag{23}$$

其中 $\eta \in R$ 为某一常数.

结论:形如式(23)且满足条件式(19)的优化问题就可视为 MEB 问题,只是此时为 CC-MEB 问题.

4 最大向量夹角间隔核向量机

4.1 MAMC 与 CC-MEB 关系

MAMC 同样可以达到 SVM 和 SVDD 分类精度,甚至更高(见第 5 节实验),但样本较大时,MAMC 能否适用,为此我们给出下面的定理.

定理 1　MAMC 等价于 CC-MEB 问题.

证明:令 $v\lambda_i = \alpha_i$,并代入式(10)可得:

$$\begin{aligned}\max_{\lambda_i, v} & \sum_{i=1}^{m} v\lambda_i y_i \left(\frac{-4}{m} \sum_{j=1}^{m} k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right) \\ & - \sum_{i=1}^{m}\sum_{j=1}^{m} v\lambda_i v\lambda_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{s.t.} \quad & \sum_{i=1}^{m} v\lambda_i = v, 1 \leqslant i \leqslant m\end{aligned} \tag{24}$$

整理后得:

$$\begin{aligned}\max_{\lambda_i, v} & \sum_{i=1}^{m} \lambda_i y_i \left(\frac{-4}{mv} \sum_{j=1}^{m} k(x_i, x_j)\right) \\ & - \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{s.t.} \quad & \sum_{i=1}^{m} \lambda_i = 1, 1 \leqslant i \leqslant m\end{aligned} \tag{25}$$

并重新令 $\lambda_i = \alpha_i$,则:

$$\begin{aligned}\max_{\alpha} & \boldsymbol{\alpha}^{\mathrm{T}}(\operatorname{diag}(\widetilde{\boldsymbol{K}}) + \boldsymbol{\Delta} - \eta\mathbf{1}) - \boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\boldsymbol{K}}\boldsymbol{\alpha} \\ \text{s.t.} \quad & \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} = 1\end{aligned} \tag{26}$$

其中,

$$\widetilde{\boldsymbol{K}} = [\tilde{k}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{m\times m} = [y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{m\times m} \tag{27}$$

$$\boldsymbol{\Delta} = -\operatorname{diag}(\widetilde{\boldsymbol{K}}) + \eta\mathbf{1} + \boldsymbol{T}_{linearterm} \tag{28}$$

$$\boldsymbol{T}_{linearterm} = [T_i]_{m\times 1} = \left[y_i \frac{-4}{mv} \sum_{j}^{m} k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right]_{m\times 1} \tag{29}$$

当取 $\eta = \max(\operatorname{diag}(\widetilde{\boldsymbol{K}}) - \boldsymbol{T}_{linearterm})$时,条件式(19)必定满足.证毕.

因此,可以利用 CC-MEB 的逼近算法(一般化的 CVM)快速找出样本核心集 Core Set[15~21],然后利用 MAMC 对核心集训练获得判决函数,从而获得了利用 MEB 逼近方法来解决大样本问题的另外一个版本 MAM-CVM.

4.2 MAM-CVM 算法实现

实际上,MAMC 中的每个样本 $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$在 MAM-CVM 中可以看成$\begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_i)\\ \delta_i\end{bmatrix} = \begin{bmatrix}y_i\varphi(\boldsymbol{x}_i)\\ \delta_i\end{bmatrix}$.此时,采用一般化的 CVM 算法需要计算最小球的半径和样本点到球心的距离,半径采用式(20)计算,而样本点到球心的距离需要将式(21)、(22)改成式(30)、(31),并通过式(31)计算.

$$\boldsymbol{c} = \sum_{i=1}^{m} \alpha_i \tilde{\varphi}(\boldsymbol{z}_i) = \sum_{i=1}^{m} \alpha_i y_i \varphi(\boldsymbol{x}_i) \tag{30}$$

$$\begin{aligned}\left\| \begin{bmatrix}\tilde{\varphi}(\boldsymbol{z}_l)\\ \delta_l\end{bmatrix} - \begin{bmatrix}\boldsymbol{c}\\ 0\end{bmatrix} \right\|^2 &= \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &- 2\sum_{i=1}^{m} \alpha_i y_i y_l k(\boldsymbol{x}_i, \boldsymbol{x}_l) + y_l y_l k(\boldsymbol{x}_l, \boldsymbol{x}_l) + \delta_l^2\end{aligned} \tag{31}$$

MAM-CVM 算法分两个阶段,第一阶段采用 CC-MEB 的一般化的 CVM 算法找出核心集 S(请参考文献[10,15]),第二阶段利用 MAMC 对核心集进行训练,并输出判决函数.具体如下:

算法 1 MAM-CVM

阶段 1	通过一般化 – CVM 获得 Core Set
Step 1	初始化 $t=0, S_t, \boldsymbol{c}_t$ 和 R_t
Step 2	所有样本点均被 $B(\boldsymbol{c}_t, (1+\varepsilon)R_t)$包络,则 $S = S_t$ 转入阶段 2,否则转入 Step 3
Step 3	找出离 $\boldsymbol{c}_t$ 最远的点 z,加入核心集中,即 $S_{t+1} = S_t \cup \{z\}$
Step 4	找出新的 MEB: $B(\boldsymbol{c}_{t+1}, R_{t+1})$
Step 5	转 Step 1
阶段 2	获得决策函数
Step 6	采用 MAMC 对核心集 S 训练
Step 7	根据式(13)输出决策函数

算法中的 Step3 需要计算 $B(\boldsymbol{c}_t, (1+\varepsilon)R_t)$球外的每个点(设 n 个)到 $\boldsymbol{c}_t$ 的距离,并找出最远点 z,由式(31)可知其复杂度为 $O(|S_t|^2 + n|S_t|)$,当 n 很大时计算很可观.为此,本算法采用文献[12]提出的概率加速方法,通过随机采样一个子集,当子集 Size = 59 时,所有最远点的 5% 在该子集中的概率将会达到 95%,这样在保证性能的基础上使计算复杂度降为 $O(|S_t|^2)$,而一般的 $|S_t| \ll n$,从而提高算法的效率.

5 实验研究

为评价本文算法,实验包含两部分:其一是 MAMC 实验,本部分实验使用 8 种中小量样本数据集,其中 1 种是人造香蕉形数据,其他 7 种真实数据集可以从网站 http://www.ics.uci.edu/~mlearn/MLRepository.html 和 http://ict.ewi.tudelft.nl/~davidt/index.html 下载;其二是 MAM-CVM 实验,使用 5 种大样本数据集,1 种是 4×4 CheckBoard 人造数据,其他 4 种真实数据集可以从网站 http://www.ics.uci.edu/~mlearn/MLRepository.html 下载,实验均在 3.0GHz 主频,2G 内存,XP 系统的计算机及 Matlab 2009a 平台上实现.考虑到训练数据的不平衡,两部分实验均记录了正负类的分类精度 a^+ 和 a^-,并用几何平均度量方法式(32)评价最终结果,简称 g 为几何精度,该方法常用于不平衡数据集[8,22].

$$g = \sqrt{a^+ \cdot a^-} \tag{32}$$

其中 a^+ 和 a^- 分别指正类和负类分类精度,采用

$$a^+ = \frac{\#\ \text{positive samples correctly classified}}{\#\ \text{total positive samples classified}} \times 100\%$$

$$a^- = \frac{\#\ \text{negative samples correctly classified}}{\#\ \text{total negative samples classified}} \times 100\%$$

计算.所有实验均采用径向基核函数实现:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\boldsymbol{x}_i - \boldsymbol{x}_j^2}{h}) \tag{33}$$

5.1 MAMC 实验

此节实验参数选择如下:核函数带宽参数 h 以训练样本的平均 2 范数的平方 s 为基准,并在网格{$s/128, s/64, s/32, s/16, s/8, s/4, s/2, s, 2s, 4s, 8s, 16s, 32s$}中搜索至最优,调节参数 v 在网格{2,5,7,10,20,50,70,90}搜索至最优.

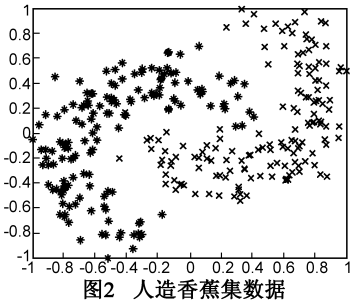

图2 人造香蕉集数据

人造香蕉形数据的标准差为 1,正负类样本各 150 个,如图 2 所示,表 1 给出了 8 种数据集的特征属性.实验前数据进行了归一化处理,从全体样本中随机取 70%作为训练,剩余样本用于测试,每个数据集进行 5 次随机交叉验证,并以平均后的 g 作为最终分类精度.为与 SVM 和 SVDD 在同等条件相比,数据同样进行归一化处理,SVM 和 SVDD 算法中的核函数及其参数搜索均与 MAMC 一致,交叉验证方法也相同,SVM 实验使用 SVM tool box 软件包,并采用硬划分;SVDD 实验中两个参数 C_1 和 C_2 均取为 1.测试精度采用几何精度 g 评价,并通过训练时间 train-time(单位:s)和测试样本的分类时间 class-time(单位:s)比较了 3 种算法的运行效率,train-time(单位:s)和 class-time(单位:s)采用平均值给出,即分别累加各次训练和分类时间,最终以平均值给出.表 2 给出了实验结果.

表 1 MAMC 实验数据集

数据集	维数	样本数	+1 类	-1 类
Banana	2	300	150	150
SpectfHeart	44	267	212	55
Ionosphere	34	351	225	126
Biomed	5	194	127	67
Hepatitis	19	155	123	32
Liver	6	345	200	145
Iris	4	150	50	100
Wine	13	178	59	119

表 2 MAMC 测试精度比较

数据集	MAMC			SVM			SVDD		
	g(%)	train-time(s)	class-time(s)	g(%)	train-time(s)	class-time(s)	g(%)	train-time(s)	class-time(s)
Banana	99.12	3.5288	0.0983	98.71	20.9180	0.3522	95.56	4.7224	0.6925
SpectfHeart	77.94	1.1041	0.0767	72.07	14.7868	0.2832	70.29	1.6214	0.6736
Ionosphere	87.16	3.1875	0.1287	88.75	31.9469	0.4772	93.12	4.0839	1.0192
Biomed	85.93	0.9266	0.0391	85.74	5.6572	0.1483	83.34	0.9827	0.3577
Hepatitis	84.45	0.2869	0.0265	74.67	3.3440	0.0947	71.09	0.3837	0.2913
Liver	65.83	12.3148	0.1154	66.19	28.2404	0.4558	63.37	8.7712	0.9103
Iris	95.96	1.0667	0.0234	95.21	3.0325	0.0868	89.62	0.5964	0.2471
Wine	100.00	0.4059	0.0346	100.00	4.9873	0.1245	98.97	0.7601	0.3034

从表 2 可以看出,除 Liver 和 Ionosphere 数据集外 MAMC 分类几何精度均优于 SVM 和 SVDD,这是因为 MAMC 模型从球半径和分类间隔 ρ 两方面同时减小 VC 维从而获得更优分类界面的结果.训练(除 Liver 和 Iris 外)和分类时间均少于 SVM 和 SVDD,效率较高,这是由于 MAMC 只需要获得最优向量 $\boldsymbol{c}$ 而不像 SVM(或 SVDD)需要同时获得 $\boldsymbol{w}$ 和偏移项 b(或球心 $\boldsymbol{c}$ 和半径 R).可见,本文的 MAMC 算法相比而言具有较好的优势.

5.2 MAM-CVM 实验

此节分别对 4×4 CheckBoard 数据集,如图 3 所示,和真实的 UCI 大样本数据集进行测试研究,UCI 数据集进行了一定的预处理,如多类数据只取其中两类,或多类数据合并成 2 类,或补充丢失的属性等,表 3 给出了数据集的属性.核函数带宽 h 在网格{$s/32, s/16, s/8, s/4, s/2, s, 2s$}中搜索至最优,调节参数 v 在网格{10,20,50,70,90}搜索至最优.实验精度和效率分别采用几何精度 g 和训练时间 train-time(单位:s)评价.

实验 1 利用 4×4 CheckBoard 数据集分析逼近参数 ε 对 MAM-CVM 算法的影响.表 4 给出了实验结果,从表 4可以看出 ε 越小,精度越高,但训练时间越长,因此实际应用时,应当考虑两者之间的协调,一般地 ε 取 1e-6 将满足绝大部分情况.

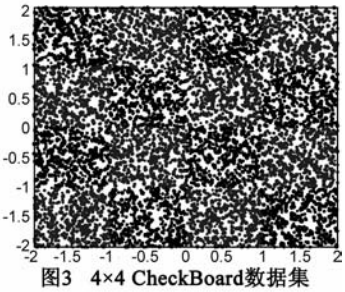

图3 4×4 CheckBoard数据集

表 3　MAM-CVM 实验数据集

数据集	维数	样本数	+1 类	-1 类
4×4CheckBoard	2	500,000	250,000	250,000
Waveform	21	3,304	1,657	1,647
Waveform + noise	40	3,345	1,692	1,653
Letter Recognition	16	20,000	8,387	11,613
Statlog_Shuttle	9	43,500	34,108	9,392

表 4　逼近参数 ε 对 MAM-CVM 的影响($h=s, v=20$)

ε	几何精度 $g(\%)$	Core Set 子集样本数 CVs	train-time(s)
1e-2	61.64	7	1.0625
1e-3	65.46	25	3.3594
1e-4	73.66	34	5.0313
1e-5	84.67	50	8.3438
1e-6	92.63	67	12.9531
1e-7	92.28	92	23.4219
1e-8	96.56	138	61.8438

实验 2　MAM-CVM 全局性实验.取全部样本为训练样本,测试样本如下选取:如果训练样本数目少于5000,则从训练样本中随机抽取 50%,否则随机抽取2000 个样本进行测试.实验参数及结果在表 5 中给出.从表 5 可以看出,几何精度除 Letter Recognition 外均达到了 90% 以上,并且此时 Core Set 子集样本数 CVs 远远小于训练样本数.如果训练样本不是全体样本,MAM-CVM 其适应性(或称泛化性)如何,为此我们进行局部性实验.

表 5　MAM-CVM 全局性效率及测试精度

数据集	ε	几何精度 $g(\%)$	Core Set 子集样本数 CVs	train-time(s)
4×4CheckBoard	1e-6	94.38	106	41.4531
Waveform	1e-5	95.53	464	554.8750
Waveform + noise	1e-6	97.67	443	642.9531
Letter Recognition	1e-5	87.02	514	1701.719
Statlog_Shuttle	1e-8	99.39	46	2.6406

实验 3　MAM-CVM 局部性实验.在全体样本中随机抽取 10%,30%,50%,70% 和 90% 用于训练,并从全体样本中随机抽取 500 个进行测试,注意此实验和实验 2 是不同的.表 6 给出了比较结果,表中训练样本数以全体样本数的百分数表示,从表 6 可以看出,对于大样本而言,训练样本越多,精度并不一定越高,反而可能增加不必要的运行时间.

表 6　MAM-CVM 局部性效率及测试精度

训练样本数	4×4CheckBoard		Waveform		Waveform + noise		Letter Recognition		Statlog_Shuttle	
	$g(\%)$	train-time(s)	$g(\%)$	train-time(s)	$g(\%)$	train-time(s)	$g(\%)$	train-time(s)	$g(\%)$	train-time(s)
10%	93.31	5.7656	90.38	4.1563	88.79	5.2969	86.57	994.9844	98.77	2.1094
30%	95.12	17.7188	94.01	50.4844	93.60	40.6563	86.16	151.016	99.56	1.7813
50%	91.65	9.7968	94.34	96.9531	93.60	96.0938	89.70	1548.438	100	0.9844
70%	93.79	13.8750	94.32	376.0156	97.85	460.0625	88.30	1765.188	98.87	1.0781
90%	94.18	16.3594	93.33	304.0000	97.00	888.1563	81.32	2035.063	99.14	1.3594

实验 4　MAM-CVM 和 SVM、MAMC 精度和效率比较实验.利用 4×4 CheckBoard 数据集,从全体样本中随机抽取 1e+2,3e+2,5e+2,1e+3,3e+3,5e+3,1e+4,3e+4 作为训练样本,从全体样本中随机抽取 500 个作为测试样本,实际上这里采用了局部性实验方案,图 4 和图 5 给出了比较结果.

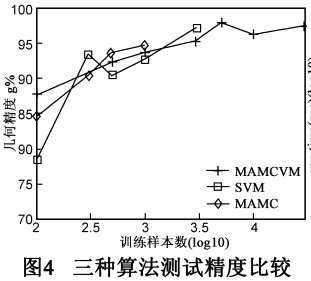

图4　三种算法测试精度比较

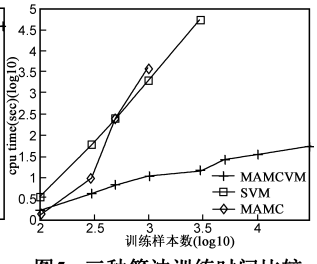

图5　三种算法训练时间比较

注意,图 4 横轴和图 5 横、纵轴采用 log10 坐标绘制,从图 4 可知训练样本在 300 以上,精度均能保证 90%以上,但从图 5 可知,MAM-CVM 的训练时间远远少于 SVM 和 MAMC,当样本超过 1000 时,MAMC 和 SVM 的训练时间较长,如当为 1e+3 时,MAMC 训练时间超过 4000 秒,当为 3e+3 时,SVM 训练时间超过 50000 秒,从图 5 还可知,样本增加 SVM 和 MAMC 训练时间快速递增,当超过 5e+3 时,由于时间过长未能进行.而当样本为 3e+4 时,MAM-CVM 训练时间 54 秒,由此说明,在样本较大时,当选择合适的逼近参数 ε 后 MAM-CVM 比其他两种方法更加有优势.

6　结论

MAMC 算法是在特征空间中找到最优点,通过该点和训练样本之间的向量夹角距离实现分类,方法独特.并从 CC-MEB 角度建立了 MAMC 与 MEB 等价关系,在此基础上提出 MAM-CVM 算法用于解决大样本问题.虽然实验中仅使用径向基核函数,但其他核函数同样适用于本文算法,并且本文算法涉及的 QP 求解均可以采用 SMO 求解子 QP 来完成,可以进一步提高算法的效率.在 MAM-CVM 算法第 3 步,本文直接采用了概率加速方法 $Size=59$,并未深入地进行理论分析和探讨;同

时 MAM-CVM 算法第 1 步的初始化也会影响到算法的效率和测试精度,如何初始化本文也没有进行深入研究.对于这些问题我们将会继续深入,此外,在将来的研究工作中,将继续跟踪和研究最小包络球的相关理论及其方法在其他模式识别领域中的应用.

参考文献

[1] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995, 20(3): 273 – 297.

[2] Schölkopf B, Smola A, Williamson RC, Bartlett PL. New support vector algorithms[J]. Neural Computation, 2000, 12: 1207 – 1245.

[3] Hu M, Chen Y, Kwok J T. Building sparse multi-kernel SVM classifiers[J]. IEEE Trans on Neural Networks, 2009, 20(5): 827 – 839.

[4] 皋军,王士同.基于矩阵模式的最小类内散度支持向量机[J].电子学报,2009,37(5):1051 – 1057.
Gao Jun, Wang Shitong. Matrix pattern based minimum within-class scatter support vector machines[J]. Acta Electronica Sinica, 2009, 37(5): 1051 – 1057. (in Chinese)

[5] Chung FL, Wang S T, Deng Z. H., et al. Fuzzy kernel hyperball perceptron[J]. ASC, 2004, 5: 67 – 74.

[6] Shivaswamy P, Jebara T. Ellipsoidal kernel machines[J]. Artificial Intelligence and Statistics, AISTATS, 2007.

[7] David M J Tax, et al. Support vector data description [J]. Machine Learning, 2004, 54(1): 45 – 66.

[8] Wu M. R, Ye J P. A small sphere and large margin approach for novelty detection using training data with outliers [J]. PAMI, 2009, 31: 1 – 5.

[9] Collobert R, Bengio S., Bengio Y. A parallel mixture of SVMs for very large scale problems[J]. Neural Computation, 2002, 14(5): 1105 – 1114.

[10] Tsang I W, Kwok J T, et al. Core vector machines: Fast SVM training on very large data sets [J]. Journal of Machine Learning Research, 2005, 6: 363 – 392.

[11] Williams C, Seeger M. Using the Nyström method to speed up kernel machines[A]. Advances in Neural Information Processing Systems[C]. Cambridge, MA: MIT Press, 2001. 13: 682 – 688.

[12] Smola A, Schölkopf B. Sparse greedy matrix approximation for machine learning[A]. Proc. 17th ICML[C]. Stanford, CA, 2000. 911 – 918.

[13] Achlioptas D, McSherry F, Schölkopf B. Sampling techniques for kernel methods[A]. Advances in Neural Information Processing Systems[C]. Dietterich T, Becker S, Ghahramani Z. Eds. Cambridge, MA: MIT Press, 2002. 14: 335 – 342.

[14] Fine S, Scheinberg K. Efficient SVM training using low-rank kernel representations[J]. Journal of Machine Learning Research, 2001, 2: 243-264.

[15] Tsang I W, Kwok, J T, Zurada J M. Generalized core vector machines[J]. IEEE Trans on Neural Networks, 2006, 17(5): 1126 – 1140.

[16] Deng Z H, Chung F L, Wang S. T. FRSDE: fast reduced set density estimator using minimal enclosing ball approximation [J]. Pattern Recognition, 2008, 41: 1363 – 1372.

[17] Chung F L, Deng Z H, Wang S. T. From minimum enclosing ball to fast fuzzy inference system training on large datasets [J]. IEEE Trans on Fuzzy Systems, 2009, 17(1): 173 – 184.

[18] Bădoiu M, Clarkson K L. Optimal core-sets for balls[J]. Computational Geometry: Theory and Applications, 2008, 40(1): 14 – 22.

[19] Bădoiu M, Har-Peled S, Indyk P. Approximate clustering via core sets[A]. Proc 34th Annu ACM Symp. Theory Comput. [C]. Montreal, QC, Canada, 2002, 250 – 257.

[20] Asharaf S, Murty M N, Shevade S K. Multiclass core vector machine[A]. Proc 24th ICML[C]. Corvalis, OR. 2007, 41 – 48.

[21] Tsang I W, Kocsor A, Kwok J T. Simpler core vector machines with enclosing balls[A]. Proc 24th ICML[C]. Corvalis, OR, 2007. 911 – 918.

[22] Kubat M., Matwin S. Addressing the curse of imbalanced training sets: one-sided selection[A]. Proc 14th ICML[C]. Nashville, Morgan Kaufmann Pubishers, 1997. 179 – 186.

作者简介

胡文军 男,1977 年生于安徽绩溪.2000 年、2003 年分别在安徽工程大学、山东理工大学获得工学学士、硕士学位,2009 年进入江南大学信息工程学院攻读博士学位,主要从事模式识别、人工智能等方面的研究.
E-mail: hoowenjun@yahoo.com.cn

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

邓赵红 男,1981 年生于安徽蒙城.副教授,博士,主要研究方向模糊建模与计算智能.