文章编号: 1001-0920 (2012) 12-1870-06

基于光束角思想的最大间隔学习机

刘忠宝[1,2], 王士同[1]

(1. 江南大学 数字媒体学院，江苏 无锡 214122；2. 中北大学 电子与计算机科学技术学院，太原 030051)

摘 要: 受空间几何知识和光学领域光束角的启发, 提出了基于光束角思想的最大间隔学习机 (BAMLM). 该方法试图在模式空间中找到一个“光源”分别照射两类样本, 然后根据照射区域的不同确定样本类属. 分析发现, BAMLM 的核化形式等价于核化中心受限最小包含球 (CCMEB), 通过引入核心向量机将 BAMLM 扩展为基于核心向量机的 BAMLM (BACVM), 有效地解决了大规模样本的分类问题. 标准数据集和人工数据集上的实验表明了 BAMLM 和 BACVM 的有效性.

关键词: 光束角；模式分类；分类点；大规模样本

中图分类号: TP391 **文献标志码:** A

Maximum margin learning machine based on beam angle

LIU Zhong-bao[1,2], *WANG Shi-tong*[1]

(1. School of Digital Media，Jiangnan Univerisity，Wuxi 214122，China；2. School of Electronics and Computer Science Technology，North University of China，Taiyuan 030051，China. Correspondent：LIU Zhong-bao，E-mail: liu_zhongbao@hotmail.com)

Abstract: Inspired by space geometry and beam angle, a maximum margin learning machine based on beam angle(BAMLM) is proposed in this paper. The basic idea of BAMLM is to find a classified point in pattern space to separate two classes. Meanwhile, the kernelized BAMLM is equivalent to the kernelized center-constrained minimum enclosing ball(CCMEB), and BAMLM can be extended to BACVM by introducing core vector machine(CVM) which can solve the classification for large-scale datasets. Experimental results obtained from synthetic and standard datasets show the effectiveness of the proposed algorithms.

Key words: beam angle；pattern recognition；classified point；large-scale datasets

1 引 言

模式分类是模式识别中的一项重要内容. 在众多模式分类方法中, 基于边界的方法[1-3]运用最为广泛. 该方法通过几何形状, 如超平面[1,4]、超 (椭) 球[2,5]等, 将目标数据中的高密度区域映射到一个正半空间或者封闭的超 (椭) 球里, 同时保证包含了大部分目标数据且上述几何形状体积最小, 以达到最佳分类效果. 经典的支持向量机 (SVM) 及其变种[1, 2, 6-12]的基本思想是: 在空间内寻找一个超平面将两类分开; 支持向量数据描述 (SVDD)[2]采用最小体积超球约束目标数据以达到剔除奇异点的目的. Wei 等[5]利用超椭球代替了 SVDD 中的超球以考虑数据的结构信息. 类似的椭球模型还有最小体积包含椭球 (MVEE)[13]以及核最小体积覆盖椭球 (KMVCE)[14], 它们均是通过优化椭球体积来寻找最小超椭球.

由上述分析可知, 空间几何中平面 (线)、球 (椭球) 等已被广泛用于模式分类中. 关于空间几何中另一重要组成部分 —— 点, 能否作为分类依据值得研究. 借鉴光束角的思想, 本文提出一种新颖的模式分类方法, 其包括两部分: 基于光束角思想的最大间隔学习机 (BAMLM) 和基于核心向量机[9]的 BAMLM 即 BACVM. 该方法试图在模式空间中找到一个分类点将两类样本分开. 标准数据集和人工数据集上的实验表明, 该方法可同时解决中小规模和大规模样本分类问题, 且分类效果良好.

2 背景知识

2.1 光束角[15]

光学领域中, 光束角是指过光源轴线的同一平面

收稿日期: 2011-04-11；**修回日期:** 2011-08-06.

基金项目: 国家863计划项目(2007AA1Z158, 2006AA10Z313)；国家自然科学基金项目(60773206, 60704047).

作者简介: 刘忠宝(1981−), 男, 博士生, 从事模式识别的研究；王士同(1964−), 男, 教授, 博士生导师, 从事人工智能与机器学习等研究.

内光强为最大光强 1/2 的两束光, 如图 1 所示. 光导管系统的光源与照明区域密切相关. 从光束角角度看, 模式分类的目标是在模式空间中找到一个“光源”分别照射两类样本, 根据照射区域的不同确定样本类属.

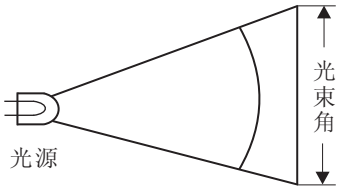


图 1　光束角示意图

2.2　相关算法

规定: 对于一个包含 N 个模式二类划分问题, 设给定训练集合 $T=(x_1,y_1),\cdots,(x_N,y_N)$. 其中: $x_i\in R^d(1\leqslant i\leqslant N_1+N_2=N)$ 为输入数据, $y_i\in\{1,-1\}$ 为类别标签. 当 $1\leqslant i\leqslant N_1$ 时, $y_i=1$; 当 $N_1+1\leqslant i\leqslant N$ 时, $y_i=-1$. 第 1 类含有 N_1 个模式 $\{x_i,y_i\}_{i=1}^{N_1}$, 第 2 类含有 N_2 个模式 $\{x_j,y_j\}_{i=N_1+1}^{N}$.

2.2.1　支持向量机

支持向量机是一种基于统计学习理论的机器学习方法, 其几何原理是使用两个带有最大间隔的平行超平面, 将两类样本尽可能地分开. 设超平面方程为 $w^{\mathrm{T}}x+b=0$, 分类间隔为 $2/\|w\|$, 则该最优化问题可描述为如下 2 种形式.

1) 线性形式

$$\min_{w,b,\xi_i}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i; \tag{1}$$

$$\text{s.t. }\ y_i(w^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i,$$
$$\xi_i\geqslant 0,\ i=1,2,\cdots,N. \tag{2}$$

其中: C 为惩罚因子, 它控制对错分样本的惩罚程度, $C=0$ 时表示线性可分, $C>0$ 时表示线性不可分; 对于线性不可分或事先未知是否线性可分的情况, 通过引入松弛因子 ξ_i 允许错分样本的存在.

2) 非线性形式

$$\min_{w,b,\xi_i}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i;$$

$$\text{s.t. }\ y_i(w^{\mathrm{T}}\varphi(x_i)+b)\geqslant 1-\xi_i,$$
$$\xi_i\geqslant 0,\ i=1,2,\cdots,N. \tag{3}$$

其中 $\varphi(x_i)$ 为从原始样本空间到高维特征空间的映射.

由 Lagrangian 定理[16]可将原问题转化为如下对偶形式:

$$\max_{\alpha}\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}; \tag{4}$$

$$\text{s.t. }\ \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}=0,\ \boldsymbol{\alpha}\geqslant\mathbf{0}. \tag{5}$$

其中: $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$, $\mathbf{1}=[1,1,\cdots,1]^{\mathrm{T}}$, 核函数 $\boldsymbol{K}=[y_iy_jk(x_i,x_j)]=[y_iy_j\varphi(x_i)^{\mathrm{T}}\varphi(x_j)]$, $\boldsymbol{Y}=[y_1,y_2,\cdots,y_N]^{\mathrm{T}}$, $\mathbf{0}=[0,0,\cdots,0]^{\mathrm{T}}$.

2.2.2　SVDD 及 MEB 问题

SVDD 是受 SVM 的启发提出的, 用于单类分类或数据描述问题. SVDD 设法寻找一个以 c 为中心、R 为半径能够包含所有样本的最小超球体. SVDD 分为硬边界 SVDD 和软边界 SVDD, 本文重点关注硬边界 SVDD. 求最小超球的半径就是求解以下的二次规划问题.

1) 线性形式

$$\min\ R^2; \tag{6}$$

$$\text{s.t. }\ \|c-x_i\|^2\leqslant R^2,\ i=1,2,\cdots,N. \tag{7}$$

其中: c 为超球体球心, R 为超球体半径.

2) 非线性形式

$$\min\ R^2;$$

$$\text{s.t. }\ \|c-\varphi(x_i)\|^2\leqslant R^2,\ i=1,2,\cdots,N. \tag{8}$$

其中 $\varphi(x_i)$ 为从原始样本空间到高维特征空间的映射.

由 Lagrangian 定理, 可将原问题转化为如下对偶形式:

$$\max_{\alpha}\ \boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}; \tag{9}$$

$$\text{s.t. }\ \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ \boldsymbol{\alpha}\geqslant 0. \tag{10}$$

其中: $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$, $\mathbf{1}=[1,1,\cdots,1]^{\mathrm{T}}$, 核函数 $\boldsymbol{K}=[k(x_i,x_j)]=[\varphi(x_i)^{\mathrm{T}}\varphi(x_j)]$, $\mathbf{0}=[0,0,\cdots,0]^{\mathrm{T}}$.

文献 [9] 指出, 硬边界 SVDD 等价于最小包含球 (MEB) 问题, 该结论对于本文研究具有重要意义.

3　基于光束角思想的最大间隔学习机

光学领域中, 光导管系统的光源与照明区域密切相关. 实际应用要求光源尽可能照射整个目标区域. 基于此思想, 本文提出了基于光束角思想的最大间隔学习机 (BAMLM). 从光学角度, BAMLM 可理解为在样本空间中寻找一个“光源”分别照射两类样本,

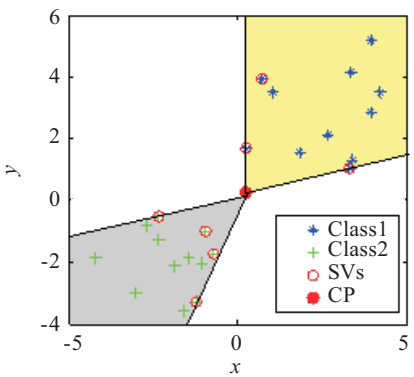


图 2　BAMLM 工作原理示意图

根据照射区域的不同对样本进行分类; 从空间几何角度, BAMLM 可理解为在样本空间内寻找一个分类点, 通过计算样本与分类点间的夹角来判断样本类属. BAMLM 工作原理如图 2 所示. 图中: Class 1 和 Class 2 分别表示两类样本, SVs 表示支持向量, CP 表示分类点.

3.1 线性形式

基于上述分析, BAMLM 目标是在样本空间中寻找分类点 c, 保证两类分开且两类间隔最大. 采用 $x_i^{\mathrm{T}}c$ 反映样本点 x_i 与 c 间的夹角, 则最优化问题可描述为

$$\min_{c,\rho,\xi_i}\ \frac{1}{N}\sum_{i=1}^{N}\|x_i-c\|^2-\upsilon\rho+C\sum_{i=1}^{N}\xi_i^2; \tag{11}$$

$$\text{s.t. }\ y_i(x_i^{\mathrm{T}}c+\xi_i)>\rho,\ i=1,2,\cdots,N. \tag{12}$$

其中: c 为分类点; ρ 为类间夹角间隔; υ 为可调参数; C 为惩罚因子, 用于惩罚错分样本; ξ_i 为松弛因子.

上述目标函数中, $\dfrac{1}{N}\sum\limits_{i=1}^{\mathrm{T}}\|x_i-c\|^2$ 使得样本距离“光源”最近, 在一定程度上避免了奇异点对分类的影响; $-\upsilon\rho$ 保证类间夹角间隔最大, 其中可调参数 υ 与支持样本数密切相关(详见 5.2.1 节); 松弛因子使样本与分类点间的夹角“软化”, 即松弛因子使 BAMLM 对分类边界附近的样本点具有一定的容错性. 此外, 为了后面数学推导的需要, 采用松弛因子的平方 ξ_i^2 有利于将 BAMLM 转化为 CCMEB 形式.

根据 Lagrangian 定理, 上述原问题的 Lagrangian 方程为

$$L(c,\rho,\alpha,\xi_i)=\frac{1}{N}\sum_{i=1}^{N}\|x_i-c\|^2-\upsilon\rho+C\sum_{i=1}^{N}\xi_i^2+\sum_{i=1}^{N}\alpha_i(\rho-y_ix_i^{\mathrm{T}}c-y_i\xi_i), \tag{13}$$

其中 Lagrangian 乘子 $\alpha_i\geqslant 0$.

$L(c,\rho,\alpha,\xi_i)$ 分别对 ρ,c,ξ_i 等变量求偏导, 并令各偏导方程等于零, 可得

$$\frac{\partial L}{\partial\rho}=-\upsilon+\sum_{i=1}^{N}\alpha_i=0, \tag{14}$$

$$\frac{\partial L}{\partial\rho}=-\frac{2}{N}\sum_{i=1}^{N}(x_i-c)-\sum_{i=1}^{N}\alpha_iy_ix_i=0, \tag{15}$$

$$c=\sum_{i=1}^{N}\Big(\frac{1}{N}+\frac{1}{2}\alpha_iy_i\Big)x_i,$$

$$\frac{\partial L}{\partial\xi_i}=2C\xi_i-\alpha_iy_i=0,\ \xi_i=\frac{\alpha_iy_i}{2C}. \tag{16}$$

将式 (14)~(16) 代入目标函数 (13) 中, 求得原问题的对偶形式为

$$\max_{\alpha}\ -\frac{4}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_iy_ix_i^{\mathrm{T}}x_j-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_jx_i^{\mathrm{T}}x_j-\frac{1}{C}\sum_{i=1}^{N}\alpha_i^2; \tag{17}$$

$$\text{s.t. }\ \sum_{i=1}^{n}\alpha_i=\upsilon,\ \alpha_i\geqslant 0,\ i=1,2,\cdots,N. \tag{18}$$

3.2 非线性形式

在非线性情况下, 通过一个满足 Mercer 条件的核函数对输入样本进行高维映射, 并在高维空间中进行模式分类. BAMLM 的核化形式如下:

$$\min_{c,\rho,\xi_i}\ \frac{1}{N}\sum_{i=1}^{N}\|\varphi(x_i)-c\|^2-\upsilon\rho+C\sum_{i=1}^{N}\xi_i^2; \tag{19}$$

$$\text{s.t. }\ y_i(\varphi(x_i)^{\mathrm{T}}c+\xi_i)\geqslant\rho,\ i=1,2,\cdots,N. \tag{20}$$

其中映射函数 $\varphi:R^d\to R^D(D\gg d)$ 将原始样本空间映射到高维特征空间.

BAMLM 的核化对偶式为

$$\max_{\alpha}\ -\frac{4}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_iy_ik(x_i,x_j)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jy_iy_jk(x_i,x_j)-\frac{1}{C}\sum_{i=1}^{N}\alpha_i^2; \tag{21}$$

$$\text{s.t. }\ \sum_{i=1}^{N}\alpha_i=\upsilon,\ \alpha_i\geqslant 0,\ i=1,2,\cdots,N. \tag{22}$$

其中核函数 $k(x_i,x_j)=\varphi(x_i)^{\mathrm{T}}\varphi(x_j)$.

3.3 类间夹角间隔 ρ 的求解

为了达到最佳的模式分类效果, BAMLM 要求类间夹角间隔 ρ 最大. 由 KKT 条件可知, 对于支持向量, 式 (20) 的等号成立, 即

$$\rho=y_i(\varphi(x_i)^{\mathrm{T}}c+\xi_i). \tag{23}$$

BAMLM 类间夹角间隔 ρ 取各支持向量对应 ρ 的平均值. 设支持向量集为 $S=\{x_i|\alpha_i>0,\ i=1,2,\cdots,N\}$. 将每个 $x_i\in S$ 代入式 (23) 中并求平均可得如下表达式:

$$\rho=\frac{1}{|S|}\sum_{x_i\in S}y_i\Big(\sum_{i=1}^{N}\Big(\frac{1}{N}+\frac{1}{2}\alpha_iy_i\Big)k(x_i,x_j)+\xi_i\Big), \tag{24}$$

其中 ξ_i 可由式 (16) 求得.

3.4 判别函数

为了判断新模式 $x\in R^d$ 的类别, BAMLM 通过比较该模式与分类点间的夹角以及类间夹角间隔的关系来确定该模式的类属. BAMLM 的决策函数如下:

$$f(x)=\mathrm{sgn}(\varphi(x)^{\mathrm{T}}c-\rho)=$$

$$\mathrm{sgn}\Big(\sum_{i=1}^{N}\Big(\frac{1}{N}+\frac{1}{2}\alpha_i y_i\Big)k(x_i,x)-\rho\Big). \tag{25}$$

若 $f(x)>0$, 则 x 属于第 1 类; 若 $f(x)<0$, 则 x 属于第 2 类. 将上述决策函数称为"夹角差决策函数".

4 CCMEB 及 BACVM

Tsang 等[9]提出了核心向量机 (CVM), 该方法把 QP 的求解转化为最小包含球问题, 并使用一个逼近率为 $(1+\varepsilon)$ 的近似算法得到核心集. 该核心集规模远小于原始样本规模, 通过对该核心集训练可得到理想的分类效果. 此外, 核心集规模仅与参数 ε 有关, 而与样本数及样本维数无关, 该结论从理论上保证了 CVM 适用于大规模样本分类问题.

4.1 中心受限最小包含球

中心受限最小包含球 (CCMEB) 是 MEB 问题的扩展. 设 $\delta_i\in R$, 将原核空间的样本点扩展为 $\begin{bmatrix}\varphi(x_i)\\ \delta_i\end{bmatrix}$, 将原球心扩展为 $\begin{bmatrix}c\\ 0\end{bmatrix}$, 则式 (8) 变为 (26). 结合式 (6) 可得如下的 CCMEB:

$$\begin{aligned}&\min\ R^2;\\ &\text{s.t. } \|c-\varphi(x_i)\|^2+\delta_i^2\leqslant R^2,\ i=1,2,\cdots,N.\end{aligned} \tag{26}$$

由 Lagrangian 定理, 易得上述问题的对偶形式为

$$\max_{\alpha}\ \boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K}+\Delta)-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}; \tag{27}$$

$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ \boldsymbol{\alpha}\geqslant\mathbf{0}. \tag{28}$$

其中: $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$, $\boldsymbol{K}=[k(x_i,x_j)]=[\varphi(x_i)^{\mathrm{T}}\varphi(x_j)]$, $\Delta=[\delta_1^2,\delta_2^2,\cdots,\delta_N^2]^{\mathrm{T}}\geqslant 0$, $\mathbf{0}=[0,0,\cdots,0]^{\mathrm{T}}$, $\mathbf{1}=[1,1,\cdots,1]^{\mathrm{T}}$.

对于任意的常数 $\eta\in R$, 有

$$\begin{aligned}&\max_{\alpha}\ \boldsymbol{\alpha}^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K}+\Delta-\eta\mathbf{1})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha};\\ &\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ \boldsymbol{\alpha}\geqslant\mathbf{0}.\end{aligned} \tag{29}$$

因 η 与 α 无关, 易知式 (27) 与 (29) 同解. 任何形如式 (29) 且 $\Delta\geqslant 0$ 者均可视为 MEB 问题[9].

4.2 BAMLM 与 CCMEB 的关系

令 $\beta_i=\dfrac{1}{\nu}\alpha_i$, 将其代入式 (21) 和 (22), 有

$$\begin{aligned}\max_{\beta}\ &-\frac{4\nu}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\beta_i y_i k(x_i,x_j)-\\ &\sum_{i=1}^{N}\sum_{j=1}^{N}\beta_i\beta_j y_i y_j k(x_i,x_j)-\frac{1}{C}\sum_{i=1}^{N}\beta_i^2;\end{aligned} \tag{30}$$

$$\text{s.t } \sum_{i=1}^{N}\beta_i=1,\ \beta_i\geqslant 0,\ i=1,2,\cdots,N. \tag{31}$$

式 (31) 和 (32) 等价于

$$\max_{\alpha}\ \boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(\boldsymbol{K})+\Delta-\eta\mathbf{1})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}; \tag{32}$$

$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}=1,\ \boldsymbol{\alpha}\geqslant\mathbf{0}. \tag{33}$$

其中

$$\boldsymbol{K}=[y_i y_j k(x_i,x_j)+\mu_{ij}],$$

$$\mu_{ij}=\begin{cases}1/C,\ i=j;\\ 0,\ i\neq j;\end{cases}$$

$$\Delta=-\mathrm{diag}(\boldsymbol{K})-\frac{4}{N\nu}y_i\sum_{j=1}^{N}k(x_i,x_j)+\eta\mathbf{1}.$$

当 η 取值足够大时, 总能保证 $\Delta\geqslant 0$, 则 BAMLM 等价于 CCMEB 问题.

4.3 BACVM 算法

基于上述分析, 本文提出 BACVM 算法. 算法具体步骤如下.

Step 1: 初始化 $c_t, R_t, S_t, \varepsilon, t=0$.

Step 2: 对于 $\forall z$, 如果 $\varphi(z)\in B(c_t,(1+\epsilon)R)$, 则转 Step 5; 否则转 Step 3.

Step 3: 如果 $\varphi(z)$ 距离球心 c_t 最远, 则 $S_{t+1}=S_t\bigcup\{\varphi(z)\}$.

Step 4: 寻找最新最小包含球 $B(S_{t+1})$, 并设置 $c_t=c_{B(S_{t+1})}, R_t=R_{B(S_{t+1})}$.

Step 5: $t=t+1$, 转 Step 2.

Step 6: BAMLM 对核心集 S_t 进行训练并得到如式 (25) 的决策函数.

上述算法中: $B(c,R)$ 是球心为 c、半径为 R 的最小包含球, S_t 为核心集, t 为迭代次数, ε 为终止参数.

5 实验分析

实验目的是考察 BAMLM 和 BACVM 分别在中小规模和大规模数据集上的有效性. 实验环境为 3 GHz Pentium4 CPU, 256M RAM, Windows XP 及 Matlab7.0. 实验选取的核函数为高斯核函数

$$k(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\delta^2). \tag{34}$$

5.1 实验参数设置

BAMLM 的分类精度与参数选择密切相关. 目前参数选择的主要方法有: 单一验证估计、留一法、k 倍交叉验证法以及基于样本相似度的方法等. 本文采用 5 倍交叉验证法.

参数通过网格搜索策略[17]加以选择. 高斯核函数的方差 δ 在网格 $\{\bar{x}/2\sqrt{2},\bar{x}/2,\bar{x}/\sqrt{2},\bar{x},\sqrt{2}\bar{x},2\bar{x},2\sqrt{2}\bar{x}\}$ 中搜索选取, 其中 $\bar{x}$ 为训练样本平均范数的平方根; C-SVC 中, 惩罚因子 C 在网格 {0.01, 0.05, 0.1, 0.5, 1, 5, 10} 中搜索选取; ν-SVC 中, 参数 ν 在网格 {0.1, 0.5, 1, 5, 10} 中搜索选取; BAMLM 中, 可调参数 ν 在网格 {0.1, 0.5, 1, 5, 10} 中搜索选取, 惩罚因子 C 在网格 {0.1, 0.5, 1, 5, 10} 中搜索选取.

5.2 中小规模数据集

5.2.1 参数对 BAMLM 的影响

由式(11)可知, BAMLM 有3个重要参数: 高斯核函数的方差δ、可调参数ν以及惩罚因子C, 其中可调参数ν影响支持向量数. 实验数据集选择 Wine, Iris, Heart 和 Spectf, 实验结果如图3所示.

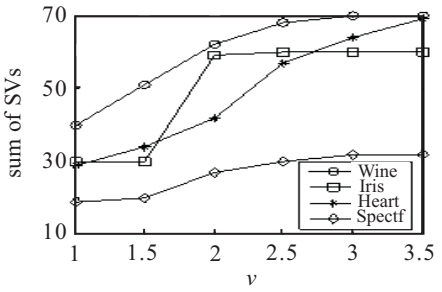


图 3 可调参数ν与支持向量数的关系

由图3可以看出, 支持向量数随着ν值增大而增加. 实验表明, 该结论对于其他 UCI 数据集亦成立.

5.2.2 UCI 数据集

实验数据集见表1. 其中: Total 表示样本总数, Class 1 表示第1类的样本数, Class 2 表示第2类的样本数, Dim 表示样本维数. 高斯核函数方差δ, C-SVC 的惩罚因子C, ν-SVC 的参数ν, BAMLM 的可调参数ν以及惩罚因子C均通过5倍交叉验证法获得.

在取得最佳参数后, 依次在实验数据集上运行 C-SVC, ν-SVC 以及 BAMLM, 实验参数和实验结果记录于表2, 其中 mean 为训练样本平均范数.

表 1 实验数据集

Datasets	Total	Class1	Class2	Dim
Wine	125	55	70	13
Iris	100	50	50	4
Liver	345	145	200	4
Heart	190	145	45	13
Spectf	225	190	35	44
Glass	145	70	75	9
Pima	765	265	500	8

表 2 SVC, ν-SVC, BAMLM 分类结果

Datasets	SVC	ν-SVC	BAMLM
Wine	$C=0.01,\ \delta^2=$mean/8 91.7%	$\nu=0.1,\ \delta^2=$mean/8 93.3%	$C=1,\ \delta^2=2*$mean **96.7**%
Iris	$C=0.01,\ \delta^2=$mean/8 **100**%	$\nu=0.1,\ \delta^2=$mean/8 **100**%	$C=0.1,\ \delta^2=$mean/8 **100**%
Liver	$C=0.01,\ \delta^2=2*$mean 63.5%	$\nu=0.5,\ \delta^2=2*$ mean 65.9%	$C=10,\ \delta^2=$mean/8 **76.5**%
Heart	$C=0.01,\ \delta^2=8*$mean 78.1%	$\nu=0.1,\ \delta^2=$mean/8 75.6%	$C=5,\ \delta^2=$mean/4 **81.7**%
Spectf	$C=1,\ \delta^2=$mean/4 66.0%	$\nu=0.1,\ \delta^2=$mean/4 66.3%	$C=1,\ \delta^2=4*$mean **73.8**%
Glass	$C=0.5,\ \delta^2=$mean/2 **63.2**%	$\nu=0.1,\ \delta^2=$mean/4 61.8%	$C=0.1,\ \delta^2=$mean/8 59.2%
Pima	$C=0.01,\ \delta^2=$mean/8 66.2%	$\nu=0.1,\ \delta^2=$mean/4 **67.3**%	$C=1,\ \delta^2=$mean/8 66.5%

由表2可见, 与C-SVC 和ν-SVC 相比, BAMLM 在UCI 数据集上具有较好的分类效果. 在Wine, Liver, Heart, Spectf 数据集上BAMLM 的分类精度好于C-SVC 和ν-SVC; 在 Iris, Pima 数据集上三者的分类精度相当或基本相当; 在 Glass 数据集上 BAMLM 分类精度略逊于其他两种算法, 但分类精度基本可以接受.

综上所述, BAMLM 在人工数据集和 UCI 数据集均可达到较理想的分类效果. 实验发现, BAMLM 分类精度与样本的分布情况密切相关. 样本分布较密, 则 BAMLM 分类精度较高; 反之, 分类精度较低. 主要原因与 BAMLM 工作原理有关: 样本分布越密集, 则“光源”越容易照射到该样本区域; 反之亦然.

5.3 中大规模数据集

实验数据集见表3. 数据集 Chess, Contraceptive,

表 3 中大规模数据集

Dataset	Total	Class1	Class2	Dim
Chess	3 196	2 294	904	37
Contraceptive	1 140	629	511	10
Magic	19 020	12 332	6 688	11
Checkboard	450 000	250 000	200 000	2
Forest	58 102	28 329	29 773	54

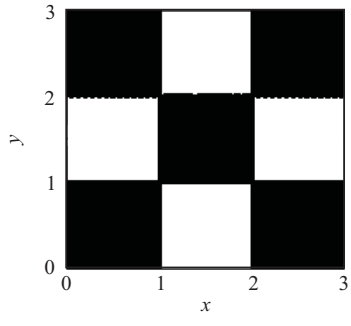


图 4 Checkboard 数据集

Magic 下载于 www.ics.uci.edu/mlearn/MLRepository.html; 数据集 Forest 下载于 www/cse.ust.hk/ivor/cvm.html; Checkboard 为人工数据集 (如图 4 所示).

5.3.1　参数对 BACVM 的影响

1) 可调参数 ν 对支持向量数的影响.

由 5.2.1 节可知, MAMLM 的支持向量数随着可调参数 ν 的增大而增加. 研究发现, BACVM 支持向量数也遵循同样规律.

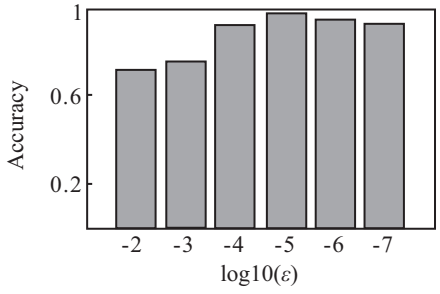
2) 终止参数 ε 对精度及训练时间的影响.

由 BACVM 算法可知, 终止参数 ε 越小, 算法迭代次数越多, 样本训练时间越长. 因此选择恰当的终止参数 ε 至关重要.

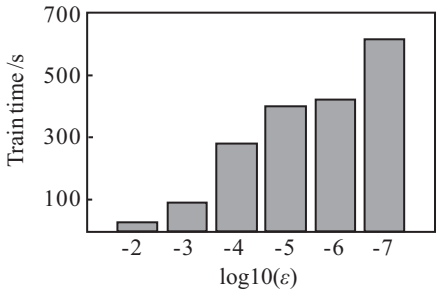
实验选取 60% 的 Chess 数据集作为训练样本, 剩下的作为测试样本. 终止参数 ε 在网格 $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ 中搜索选取. 实验结果如图 5 所示. 由图 5 可以看出, 终止参数 ε 不仅影响到算法的分类精度, 而且影响到样本的训练时间. 不失一般性, 选取 $\varepsilon = 10^{-6}$.

5.3.2　BACVM 分类性能

分别取表 3 中数据集的 20%, 40%, 60%, 80%作为训练样本, 从剩余样本中任取 500 个作为测试样本, 实验结果记录于表 4.



(a) ε 与分类精度的关系



(b) ε 与样本训练时间的关系

图 5　终止参数 ε 对 BACVM 的影响

表 4　BACVM 分类结果

Train	Chess		Contraceptive		Magic		Checkborad		Forest	
Size/%	Accurancy/%	Time/s	Accurancy/%	Time/s	Accurancy/%	Time/s	Accurancy/%	Time/s	Accurancy/%	Time/s
20	81.0	138.4	88.0	15.7	75.6	262.3	88.4	257.9	92.2	332.8
40	95.1	193.0	89.6	25.3	93.2	407.8	93.2	311.4	99.2	286.4
60	100.0	419.8	90.4	38.9	94.6	721.9	96.2	556.8	98.6	323.4
80	100.0	477.6	91.0	82.1	97.6	765.7	97.4	846.3	99.0	479.2

由表 4 可以看出, 随着训练样本规模的增大, BACVM 分类精度和训练时间未必一定增加, 因此, 训练样本规模直接影响 BACVM 的分类性能. 尽管 BACVM 的分类效率依赖于参数选取, 但其在解决大规模样本分类问题上的有效性是传统方法所不具备的.

6　结　论

基于边界的分类方法中, 超平面、超(椭) 球等几何形状运用较为广泛. 空间几何另一重要组成部分 —— 点能否作为分类依据值得研究. 本文受空间几何知识和光学领域光束角启发, 提出了基于光束角思想的最大间隔学习机 BAMLM. 从光学角度, BAMLM 可理解为在样本空间中寻找一个"光源"分别照射两类样本, 根据照射区域的不同对样本进行分类; 从空间几何角度, BAMLM 可理解为在样本空间内寻找一个分类点, 通过计算样本与分类点间的夹角来判断样本类属. 分析发现 BAMLM 的核化形式等价于核化 CCMEB, 通过引入核心向量机将 BAMLM 扩展为 BACVM, 可有效地解决大规模样本的分类问题. 标准数据集和人工数据集上的实验结果均表明了本文方法的有效性.

参考文献(References)

[1] Scholkopf B, Platt J, Shawe-Taylor J, et a1. Estimating the support of high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471.

[2] Tax D M J, Duin R P W. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.

[3] 冯爱民, 薛晖, 刘学军, 等. 增强型单类支持向量机[J]. 计算机研究与发展, 2008, 45(11): 1858-1864.
(Feng A M, Xue H, Liu X J, et a1. Enhanced one-class SVM[J]. J of Computer Research and Development, 2008, 45(11): 1858-1864.)

[4] Lauckriet G R G, Ghaoui L E, Jordan M. Robust novelty detection with single-class MPM[M]. Cambridge: MIT Press, 2002: 905-912.

(下转第1880页)