基于微分容量控制的学习机

张 莉, 周伟达, 焦李成
(西安电子科技大学雷达信号处理重点实验室, 陕西西安 710071)

摘 要: 本文通过对函数集容量的分析, 得出用函数的微分来控制函数集容量的学习方法. 该方法不仅能用支撑矢量核函数而且可以采用其他的函数作为基函数. 基于样本的机器学习, 要求学习机在容量控制和过拟合之间取一个折衷, 从而保证学习机的推广能力和误差精度. 本文通过在微分容量控制和最小化经验误差之间作一个折衷, 提出基于微分容量控制的学习机. 仿真实验验证了我们的学习机具有良好的推广能力.

关键词: 机器学习; 支撑矢量机; 容量控制; 模式识别; 回归估计
中图分类号: TP391. 4 文献标识码: A 文章编号: 0372-2112 (2003) 10-1526-06

Learning Machine Based on Differential Capacity Control

ZHANG Li, ZHOU Wei-da, JIAO Li-cheng
(National Key Lab. for Radar Signal Processing, Xidian University, Xi. an, Shaanxi 710071, China)

Abstract: A differential method for capacity control is presented based on the analysis of the capacity of learning machines. Our method that can be applicable to the set of nonlinear hypothesis functions as well as the set of linear ones generalizes the theory about the capacity control of SVMs. A new learning machine is proposed based on differential capacity control method. In our learning machine, a good generalization performance can be obtained by the right balance struck between the empirical risk and the differential of the set of hypothesis functions that controls the machine capacity. Simulation results show the feasibility of our learning machine.

Key words: learning machine; support vector machines; capacity control; pattern recognition; regression estimation

1 引言

基于有限样本的机器学习, 其推广能力和学习精度是一个两难问题[1~4]. 通常情况下, 过高的学习精度会导致推广能力的下降, 这就是所谓的过拟合(overfitting)问题. 支撑矢量机引入结构风险最小化准则的思想, 实现了学习精度和机器容量之间的折衷, 较好地解决了这一问题[1~3,5,6]. 然而, 支撑矢量机的容量控制是对 VC 维数进行控制, 其推广能力的界是建立在 VC 维数的基础上(即是构造性的与分布无关的界).

本文提出了一种基于微分容量控制的学习机, 其推广能力的界是依赖于分布的界, 容量控制体现在对函数微分的控制上. 该学习机可以用任意一阶可微假设函数集作为待估计函数的目标函数集, 若函数集取线性函数集, 此时的学习机就是线性支撑矢量机. 我们从函数集容量概念上定性地说明了用假设函数集的微分来控制学习机的容量的可行性. 仿真实验的结果说明了该学习机无论采用 Mercer 核函数还是采用其他的假设函数集都能够得到较好的推广能力.

2 推广能力及微分容量控制

2.1 学习机的推广能力

基于样本的机器学习, 由于样本分布未知, 一般采用最小化经验风险的方法[1~3]. 由于实际中样本数有限, 纯粹的经验风险最小并不能保证良好的推广能力, 这就是过拟合问题. 小的经验风险要求大的假设空间, 小的假设空间又会导致经验风险增大. 这是一个两难问题. 由此可知, 学习机的学习精度与其容量之间的矛盾是不可调和的, 所能做的只是在两者之间作一个较好的折衷, 以获得较高的推广能力. Vapnik 等人所提出的结构风险最小化(SRM)归纳原则能够体现这一思想, 而支撑矢量机则能够实现这一思想.

下面我们先简要描述本文将要涉及到的统计学习理论中的一些概念和符号[1~3]. 设 $Q(z, \alpha)$, $\alpha \in \Lambda$ 是一个损失函数集, 考虑 l 个独立同分布样本 $\{z_1, \cdots, z_l\}$. 相应地定义如下的向量集,

$$q(\Lambda) = (Q(z_1, \alpha), \cdots, Q(z_l, \alpha)), \alpha \in \Lambda \quad (1)$$

当 $Q(z, \alpha)$ 为指示函数(函数值为 0 或 1)时, 则函数集的多样性 $N = N^{\Lambda}(z_1, \cdots, z_l)$ 的值是向量 $q(\Lambda)$ 取不同值的个数; 当 $Q(z, \alpha)$ 为实函数时, 情况稍有不同, 须先对向量集空间以某种度量(如 L_p 度量)进行 ε-网格划分, $N = N^{\Lambda}(\varepsilon; z_1, \cdots, z_l)$ 是向量集 $q(\Lambda)$, $\alpha \in \Lambda$ 的最小 ε-网格的元素个数, 对向量集合 $q(\Lambda)$, $\alpha \in \Lambda$, 如果

(1)存在 $N = N^{\Lambda}(\varepsilon; z_1, \cdots, z_l)$ 个向量 $q(\alpha_1), \cdots, q(\alpha_N)$, 使得对任意向量 $q(\alpha^*)$, $\alpha^* \in \Lambda$, 我们可以在这 N 个向量中

收稿日期: 2001-12-13; 修回日期: 2002-06-09
基金项目: 国家自然科学基金(No. 60073053, 60133010, 69831040)

找到一个 $q(A_r)$, 它以 E 靠近向量 $q(A^*)$ (在某个给定的度量下). 如在 L_2 度量下意味着

$$Q_{L_2}(q(A^*), q(A_r)) = \sqrt{\sum_{i=1}^{l} | Q(z_i, A^*) - Q(z_i, A_r) |^2} < E$$

(2) N 是具有这一特性的向量的最小数目, 则向量集合 $q(A)$, $A \in \Lambda$ 有一个最小的 ε2网格.

令 $H^+(l)$ 是对应函数集的 VC 熵, $H^+_{ann}(l)$ 是其退火熵, 它们都可以表征函数集的容量并可以用函数集的多样性来描述

$$H^+(l) = E\ln N \tag{2}$$

$$H^+_{ann}(l) = \ln EN \tag{3}$$

$G^+(l)$ 表示函数集的生长函数, 这三个量存在着如下的关系[1~3]:

$$H^+(l) \leqslant H^+_{ann}(l) \leqslant G^+(l) \tag{4}$$

不失一般性, 令损失函数集为完全有界非负的, 即 $0 \leqslant Q(z, A) \leqslant B, A \in \Lambda$. 则学习机的推广能力的界以至少 $1-\eta$ 的概率成立[2]:

$$R(A) \leqslant R_{emp}(A) + \frac{B\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(A)}{B\varepsilon}}\right) \tag{5}$$

其中 $R(A)$ 表示实际风险和 $R_{emp}(A)$ 表示经验风险. 对不同的 ε 产生不同的界, 对依赖于分布的推广能力的界(即基于 VC 退火熵函数的界)有

$$\varepsilon = 4\,\frac{H^+_{ann}(2l) - \ln(\eta/4)}{l} \tag{6}$$

对与分布无关的推广能力的界有

$$\varepsilon = 4\,\frac{G^+(2l) - \ln(\eta/4)}{l} \tag{7}$$

对构造性的与分布无关的推广能力的界有

$$\varepsilon = 4\,\frac{h(\ln(2l/h) + 1) - \ln(\eta/4)}{l} \tag{8}$$

其中 h 表示函数集的 VC 维数, 其大小与函数集的容量大小成正比. 由于式(4)和

$$G^+(l) \leqslant h\left(\ln\left(\frac{l}{h}\right) + 1\right) \tag{9}$$

可知, 式(8)表示的界最松和式(6)表示的界最紧.

支撑矢量机采用式(8)来表示其推广能力的界, 其容量控制就是控制 VC 维数的大小. 支撑矢量机是在经验风险和 VC 维数之间取折衷, 从而得到好的推广能力. 这里, 我们要构造一个学习机, 其推广能力的界是依赖与样本的分布, 即式(6).

2.2　微分容量控制

我们定义函数集序列, $S_k = \{Q(z, A), A \in \Lambda_k\}$, $k = 1, 2, \cdots$, 满足

$$P\{A \in \Lambda_k, \left|\frac{\partial Q(z, A)}{\partial z}\right| \leqslant A_k\}, \text{且 } A_1 \leqslant A_2 \leqslant \cdots \leqslant A_k \leqslant \cdots$$

由上述定义, 显然 $S_1 \subset S_2 \subset \cdots \subset S_k \subset \cdots$.

对于函数集序列 S_k, 如果其容量随着序号 k 的增加而增大, 那么在 S_k 中的经验风险最小值将随之减小, 但是不等式(5)右边第二项或等式(6)的值会增加. 我们类似地引入结构风险, 来选择函数集 S_k 和最小化经验风险, 从而得到实际风险的最好的界. 这样, 我们就能构造一个新的具有容量控制的学习机.

下面, 我们来说明函数集微分的范数能够控制其容量. 不失一般性, 我们假设损失函数集中所有函数均为有界的一致连续函数 $0 \leqslant Q(z, A) \leqslant B, A \in \Lambda$. 这个假设在实际中也是合理的, 虽然函数集的支撑可能无界, 但我们仅关心样本分布区域的函数值. 设 Z 表示样本的原空间, 给定一个 $A \in \Lambda$, 损失函数将样本映射到某个有界空间 T: $Z \xrightarrow{Q(z, A)} T$, 其中 $T = (Q(z_1, A), \cdots, Q(z_l, A))$. 为了简化, 现假设样本集仅为一个样本 $z = (x, y)$, 其分布未知, 则 T 空间的维数为 1. 我们对空间 T 进行 ε2网格划分, 则函数集的多样性 N, 其可能的最大值等于网格总数. 考虑某一网格 NET^t_{ε}, t_0 表示该网格的中心, E 表示网格的半径. 如果 $t \in NET^t_{\varepsilon}$, 则 $|t - t_0| \leqslant E$. 这样由于函数的一致连续性, 对应于空间 T 中的任一网格, 在样本空间 Z 至少有一个或多个邻域 D^z_{δ} 与之对应. 在此我们假设映射为一一映射, 关于多对一的映射可以对样本空间 Z 中每一个邻域逐个考虑. 那么对空间 T 中一个网格 NET^t_{ε}, 由于 $Q(z, A)$ 一致连续, 所以在 Z 空间中有且只有一个邻域 D^z_{δ}, 使得 $D^z_{\delta} \xrightarrow{Q(z, A)} NET^t_{\varepsilon}$. 显然如果 $z \in D^z_{\delta}$, 则 $|Q(z, A) - Q(z_0, A)| \leqslant E$. 其中 $Q(z_0, A) = t_0$, 邻域 D^z_{δ} 的半径定义为 $D = \underset{D}{\text{average}} + (z - z_0) +$, 其中 $\underset{D}{\text{average}}$ 表示在集合 D 上取平均, $D = \{z | Q(z) - Q(z_0) = E\}$. 由于函数 $Q(z, A)$ 在 Z 上一致连续, 在集合 D^z_{δ} 中, 如果 $z \neq z_0$, 则 $Q(z, A) \neq Q(z_0, A)$; 反之当 z 远离 z_0 时, $Q(z, A)$ 也远离 $Q(z_0, A)$, 所以当 E 取较小的正数时, 集合 $D = \{z | Q(z) - Q(z_0) = E\}$ 应该分布在集合 D^z_{δ} 的边界上, 如图 1 所示.

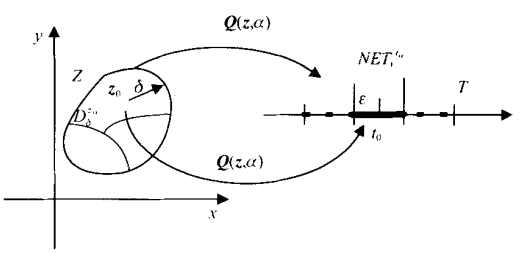

图 1　一维一个样本情形下, 样本空间 Z 与输出空间 T 由损失函数建立的映射关系

现在我们在集合 D^z_{δ} 的边沿上取一点使得 $Q(z, A) - Q(z_0, A) = E$. 在邻域 D^z_{δ} 上使用微分中值定理, 我们有

$$Q(z, A) - Q(z_0, A) = (z - z_0)^T Q'_z(H, A) \tag{10}$$

其中 H 取介于 z 和 z_0 之间的向量, $Q'_z(z, A)$ 表示 $Q(z, A)$ 对变量求变化率. 式(10)可以变成下面的不等式:

$$|Q(z, A) - Q(z_0, A)| = |(z - z_0)^T Q'_z(H, A)| \leqslant |(z - z_0)| \cdot |Q'_z(H, A)|$$

即

$$E \leqslant |(z - z_0)| \cdot |Q'_z(H, A)|$$

因此有

$$|(z - z_0)| \geqslant \frac{E}{|Q'_z(H, A)|} \tag{11}$$

由上式可知, 即损失函数在邻域 D^z_{δ} 中微分的范数减小

时, 在样本空间 Z 中, 与输出 T 中的网格 NET_{ε}^{l} 对应的集合 $D_{\delta}^{z_0}$ 半径将随之增大. 由于样本空间 Z 的大小不变, 这样在样本空间总的网格数将减少. 这意味着在输出空间 T 中的总网格数也将减小, 其实是输出空间 T 在减小, 如图 2 所示. 总网格数的减少必将导致函数集多样性 N 的减少, 因此我们可以粗略地说, 对于上面我们定义的空间序列 S_k, 有 N_k [N_{k+1}, k = 1, 2, , 成立.

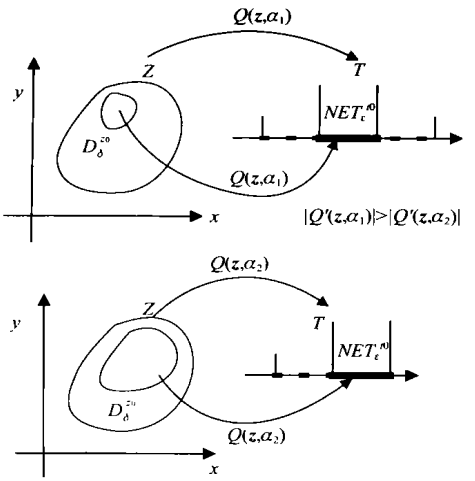

图 2 当| Qc(H, A_1) | > | Qc(H, A_2) | 时, 在样本空间 Z 中的网格增大了, 网格的总数相应减小; 而在输出空间 T 中, T 减小了, 从而导致网格数减小.

对于多个样本的情况. 设样本数为 l, 则样本空间 Z 的维数为 2l 维, 输出空间 T 为l 维. 我们同样对输出空间进行 E2 网格划分, 在样本空间 Z 的每一个邻域上使用 l 次微分中值定理, 我们将得到与上述完全一致的结果. 同样如果样本维数为 d 维, 那么推导中的损失函数对 z 的微分将是一个d+ 1 维的偏微分, 最终得到结果也是一致的.

这样我们说明了不管样本在样本空间 Z 中如何分布, 随着损失函数在样本空间 Z 上微分范数的减小, 函数集的多样性 N 将随之减小, 从而使得 VC 退火熵 $H_{ann}^{+}(l)$ 小. 这样保证了不等式(5) 右边的第二项减小, 或者式(6)的减小, 实现了对函数集容量的控制.

3 基于微分容量控制的学习机

具体对于各类问题而言, 损失函数 L(y, f(x, A)) 可以取不同的形式, 在此取如下通用的损失函数:

$$Q(z, A) = L(y, f(x, A)) = |f(x, A) - y| \qquad (12)$$

其中 z= (x, y). 因此最小化假设函数集的微分和最小化损失函数集的微分是等价的. 对于给定的学习样本对{(x_i, y_i) | x_i I R^d, y_i I R, i= 1, , , l}, 我们构造如下的风险泛函: 由此, 我们可以构造这样的学习机:

$$\text{最小化} \qquad C\#R_{emp}(A) + \frac{1}{2}\sum_{i=1}^{l} + fc(x, A) + ^2 \qquad (13)$$

基于微分容量控制的学习机的学习过程就是最小化式(13), 其中 C> 0 是在容量和学习精度之间的折衷因子. 式(13) 中第二项为:

$$+ fcx_i, A) + ^2 = + \frac{5fc(x_i, A)}{5x} + ^2$$

其中 $\frac{5fc(x_i, A)}{5x} = \left[\frac{5fc(x_i, A)}{5x^1}, \frac{5fc(x_i, A)}{5x^d}\right]^T$, 上标^{T}表示向量或矩阵的转置以及上标^{i}表示样本x 的第 i 个分量. 不失一般性, 我们采用任意一阶可微函数 g(x) 来作为函数集 f(x, A)的基函数:

$$f(x, A) = \sum_{i=0}^{p} A_i g_i(x) \qquad (14)$$

其中 g(x)= [g_0(x) g_1(x) g_p(x)]^T, p> 1 是自定义的参数. g_0(x) 值只取 0 或 1, 当 g_0(x) = 1, 表示估计函数有阈值; 反之则没有. 下面我们给出常用的几种 g(x) 的表达式:

p 线性支撑矢量机的情况

$$g_i(x) = x^{(i)}, \ i = 1, , , d$$

其中 x= [$x^{(1)}$ $x^{(1)}$, $x^{(d)}$] I R^d 和p= d.

p g(x) (x I R^d) 可以取核函数[7], 如高斯核函数和多项式核函数, 此时 p= l.

$$g_i(x) = K(x, x_i), \quad i = 1, , , l \qquad (15)$$

p g(x) (x I R) 可以取多项式函数(与多项式核函数是不同的) .

$$g(x) = [1 \quad x \quad x^2 \quad , \quad x^p]^T \quad \text{上标表示指数.} \qquad (16)$$

p g(x) (x I R) 可以取子波函数, a_i, b_i 分别表示子波的伸缩和平移因子[8, 9].

$$g_i(x) = h\left(\frac{x - b_i}{a_i}\right), \ i = 1, , , l \qquad (17)$$

式(16) 和(17) 也可以扩展到高维的情况, 这里我们不加以讨论.

对式(14) 求对样本 x 的变化率, 有

$$+ fc(x, A) + ^2 = + \sum_{i=1}^{p} {}_i \frac{5g_i(x)}{5x} + ^2 \qquad (18)$$

令

$$H_{ij} = \sum_{k=1}^{l}\left(\left(\frac{5g_i(x)}{5x}\Big|_{x=x_k}\right)^T\left(\frac{5g_j(x)}{5x}\Big|_{x=x_k}\right)\right) \qquad (19)$$

其中 $\frac{5g_i(x)}{5x} = \left[\frac{5g_i(x)}{5x^1}, \frac{5g_i(x)}{5x^d}\right]^T$ 和 x= [x^1 x^2 , x^d]^T. 可知 H 是对称的、半正定的矩阵. 由式(19), 可把式(16) 改写为:

$$+ fc(x, A) + ^2 = \sum_{i,j=1}^{p} A_i A_j H_{ij} = A^T H A \qquad (20)$$

至此, 风险泛函式(12) 可以重写为:

$$C\#R_{emp}(A) + \frac{1}{2}A^T H A \qquad (21)$$

其中 A= [A_1 A_2 , A_p]^T.

在训练样本点上, 令

$$G = [g(x_1) \quad g(x_2) \quad , \quad g(x_l)]^T_{l@p} \qquad (22)$$

矩阵 G 中忽略了 g_0(x) . 则式(11) 可以改写为矩阵形式

$$f(x) = GA + A_0 \qquad (23)$$

当采用的线性估计函数 f(x, A) = A# x + A_0 时, H 为单位矩阵, 则风险泛函为:

$$C\#R_{emp}(A) + \frac{1}{2} + A + ^2 \qquad (24)$$

这与线性支撑矢量机的风险泛函是一样的.

对于式(21),不同的学习问题只是损失函数不同,也就是经验风险 $R_{emp}(A)$ 不同. 下面我们将分别对模式识别和回归估计给出基于微分容量控制学习机的风险泛函.

3.1.1　用于模式识别

已知训练样本对 $(x_1, y_1), \cdots, (x_l, y_l)$,其中 $x \in R^d$ 和 $y \in \{-1, +1\}$. 我们的目的是要构造一个超平面把不同类别的样本点分开. 对模式识别问题,损失函数可以变形为:

$$L(y, f(x, A)) = \begin{cases} 0, & yf(x, A) \geqslant 1 \\ 1 - yf(x, A), & yf(x, A) < 1 \end{cases} \tag{25}$$

令 $\xi_i = L(y_i, f(x_i, A))$,则经验风险函数表示为:

$$R_{emp}(A) = \frac{1}{l}\sum_{i=1}^{l}\xi_i \tag{26}$$

其决策函数为

$$f^*(x) = \mathrm{sgn}\left(\sum_{i=1}^{p}A_i g_i(x) + A_0\right) \tag{27}$$

其中 sgn 表示符号函数. 把式(26)代入风险泛函式(21),有下面的规划:

$$\text{P1: 最小化} \quad \frac{1}{2}\sum_{i,j=1}^{p}A_iA_jH_{ij} + C\sum_{i=1}^{l}\xi_i \tag{28}$$

$$\text{约束} \quad y_i\left(\sum_{j=1}^{p}A_j g_j(x_i) + A_0\right) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 0,\ i = 1, \cdots, l$$

我们对原规划 P1 求其 Wolf 对偶得到对偶规划 D1:

$$\text{D1: 最大化} \quad \sum_{i=1}^{l}K_i - \frac{1}{2}\sum_{i,j=1}^{l}K_iK_j(HG)_{ij} \tag{29}$$

$$\sum_{i=1}^{l}K_iy_i = 0, \quad 0 \leqslant K_i \leqslant C,\ i = 1, \cdots, l$$

其中 K_i 是原规划 P1 对应的 Lagrange 乘子, $HG = \overline{G}H^+\overline{G}^T$, $\overline{G}_{ij} = y_iG_{ij}$ 和 H^+ 表示 H 矩阵的广义逆矩阵. 在从原规划 P1 到对偶规划 D1 的转换中,我们有

$$A = H^+\overline{G}^TK \tag{30}$$

阈值 A_0 不能在对偶规划中求出,因此要利用 KKT 条件,即

$$K_i\left[y_i\left(\sum_{j=1}^{p}A_jg_j(x_i) + A_0\right) + 1 - \xi_i\right] = 0 \tag{31}$$

当 $0 < K_i < C$ 时,有 $\xi_i = 0$,因此可得到

$$A_0 = \frac{1}{|I|}\sum_{i\in I}\left(y_i - \sum_{j=1}^{p}A_jg_j(x_i)\right) \tag{32}$$

其中 $I = \{i \mid 0 < K_i < C\}$.

把式(30)得到的 $A_i(i = 1, \cdots, l)$ 和式(32)得到的阈值 A_0 代入到决策函数式(27),就可以完成模式识别问题.

3.1.2　用于回归分析和函数逼近

已知训练样本对 $(x_1, y_1), \cdots, (x_l, y_l)$,其中 $x \in R^d$ 和 $y \in R$. 我们的目的是要逼近由样本点确定的目标函数. 如果样本点是有噪的,即 $y = f(x) + n$,其中 n 的分布独立于 x. 那么这时表示的问题是回归估计问题. 如果样本点不含噪声,则是函数逼近问题. 对于这两种问题,我们可以用同样的方法来求解.

令 E 表示给定的学习精度,把式(2)的损失函数变形为 E 不敏感损失函数[6]:

$$L(y, f(x, A)) = |(y - f(x, A)|_E = \begin{cases} |(y - f(x, A)| - E, & |y - f(x, A)| > E \\ 0, & |y - f(x, A)| \leqslant E \end{cases} \tag{33}$$

令 $\xi_i = (y_i - f(x_i, A))_E$ 和 $\xi_i^* = (f(x_i, A) - y_i)_E$,则经验风险函数表示为:

$$R_{emp}(A) = \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{34}$$

把式(34)代入到风险泛函式(21),可以得到下面的规划 P2:

$$\text{P2: 最小化} \quad \frac{1}{2}\sum_{i,j=1}^{p}A_iA_jH_{ij} + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{35}$$

$$\text{约束} \quad y_i - \left(\sum_{j=1}^{p}A_jg_j(x_i) + A_0\right) \leqslant E + \xi_i$$

$$\left(\sum_{j=1}^{p}A_jg_j(x_i) + A_0\right) - y_i \leqslant E + \xi_i^*$$

$$\xi_i^*, \xi_i \geqslant 0,\ i = 1, \cdots, l$$

我们同样可以利用 Lagrange 乘子方法把原规划 P2 变成其对偶规划 D2:

$$\text{D2: 最大化} \quad -E\sum_{i=1}^{l}(K_i + K_i^*) + \sum_{i=1}^{l}(K_i^* - K_i)y_i - \frac{1}{2}\sum_{i,j=1}^{l}(K_i^* - K_i)(K_j^* - K_j)(HG)_{ij} \tag{36}$$

$$\sum_{i=1}^{l}(K_i^* - K_i) = 0$$

$$K_i^*, K_i \in [0, C],\ i = 1, \cdots, l$$

其中 K_i 和 K_i^* 是规划 P2 的 Lagrange 乘子, $HG = \overline{G}H^+\overline{G}^T$. 在从原规划 P2 到对偶规划 D2 的转换中,我们有

$$A = H^+K^T(K^* - K) \tag{37}$$

利用 KKT 条件可求出阈值 A_0,即

$$K_i\left[E + y_i - \left(\sum_{j=1}^{p}A_jg_j(x_i) + A_0\right) + \xi_i\right] = 0 \tag{38}$$

$$K_i^*\left[E - y_i + \left(\sum_{j=1}^{p}A_jg_j(x_i) + A_0\right) + \xi_i^*\right] = 0 \tag{39}$$

当时 $0 < K_i < C$,有 $\xi_i = 0$,因此可得到

$$A_0 = \frac{1}{|I|}\sum_{i\in I}\left(y_i - \sum_{j=1}^{p}A_jg_j(x_i) + E\right) \tag{40}$$

其中 $I = \{i \mid 0 < K_i < C\}$. 同理有:

$$A_0 = \frac{1}{|I^*|}\sum_{i\in I^*}\left(y_i - \sum_{j=1}^{p}A_jg_j(x_i) - E\right) \tag{41}$$

其中 $I^* = \{i \mid 0 < K_i^* < C\}$. 最终的阈值是取式(40)和式(41)的平均.

4　仿真实验

4.1.1　识别问题

例 1　双螺旋问题的识别. 双螺旋问题可谓是模式识别方法的试金石[10]. 其数据集 $(x_1, y_1), \cdots, (x_l, y_l)$ 由下面的表达式所定义:

$$x_i = [(k_iH_i + A_j)\cos H_i \quad (k_jH_i + A_j)\sin H_i], \ y_i = j,\ j = 1, 2$$

其中 k_j 和 A_j 都是常量,分别代表速度和起始距离, H 是以弧

度为单位的相角. 这里令 $k_1=k_2=2$, $A_1=1$, $A_2=3$ 和 HI [0, 4P]. 我们在区间[0, 4P]上等间隔采样, 取 126 个点, 再等间隔取其中的 26 个作为训练样本, 共 52 个训练样本, 200 个检测样本. 这里用高斯核函数 $K(x_1, x_j)=\exp(-\ +x_i-\ x_j+^2/2B^2)$ 作为假设函数集的基函数. 图 3 显示了两种不同方法对双螺旋的分类情况, 图中的实线表示分界线, 即函数 $f(x)=0$, 虚线表示函数 $f(x)=?\ 1$ 的曲线. 从图 3 中可以看出基于微分容量控制学习机的性能和支撑矢量机的性能是一样的, 而且它们检测样本集上的识别率来都是 100%.

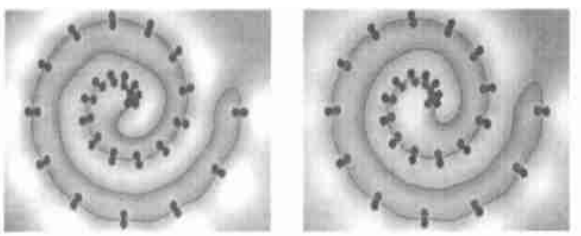

图 3 两类学习机对双螺旋的识别比较

(a) 基于微分容量控制的学习机对双螺旋的识别

(b) 支撑矢量机双螺旋的识别

例 2 一维人工数据的识别. 数据集$(x_1, y_1, \ , \ , x_l, y_l)$由下面的表达式所定义 $y=\begin{cases}1, & \text{当 } 1[\ x<3 \text{ 和 } 5[\ x<7\\ -1, & \text{当 } 3[\ x<5 \text{ 和 } 7[\ x<9\end{cases}$. 在每个小区间内随机取 55 个点, 随机取 5 个作为训练样本, 则共有 20 个训练样本, 200 个检测样本. 我们采用多项式函数式(13)作为决策函数的基函数, 令 $p=5$, 即取 5 阶多项式. 对 200 次实验取平均, 得到平均识别率为 92160% 和 91172%. 图 4 画出了两类样本的分界线. 图中星号和圆圈分别代表两类样本, 实线是分界线.

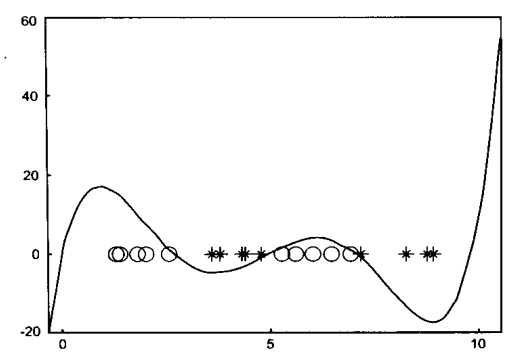

图 4 多项式函数对非线性可分的两类样本构成的分类边界

412 逼近问题

设数据集$(x_1, y_1), \ , \ , (x_l, y_l)$是由下面的一维非线性函数所定义的[11]:

$$y_i=\sin(x_i)+\frac{1}{3}\sin(3x_i)-\ 2\sin\left(\frac{x_i}{2}\right),\ i=1, \ , \ , l$$

其中 $x\,I\ [0, 2P]$. 我们在区间[0, 2P]上等间隔采样 158 个点, 等间隔取其中的 27 个点作为训练样本, 其余的作为检测样本. 我们用如下的子波函数来作为估计函数的基函数

$$h\left(\frac{x-\ b}{a}\right)=\cos\left(1.75\,\frac{x-\ b}{a}\right)\exp\left(-\ \frac{(x-\ b)^2}{2a^2}\right)$$

对所有的 i, 令 $b_i=x_i$ 以及 $a_i=a$ 为某一常数. 图 5 显示了在 E= 0l1 精度下, 基于微分容量控制的学习机对一维非线性函数的逼近结果. 表 3 给出了不同精度下, 学习机所得到的训练误差和检测误差. 误差是用真实目标函数和所得的逼近之间的标准差来定义的, 即 $\sqrt{\frac{1}{l}\sum_{i=1}^{l}(y_i-\hat{y}_i)^2}$, 其中 $\hat{y}$ 表示对 y 的逼近. 从表 1 中我们可以看出, 训练误差和检测误差在同一数量级上, 这说明学习机的推广能力是好的. 在学习精度小于某个值时, 误差不会随着精度的减小而减小, 而是趋于某个常数, 这保证了学习机不会产生过拟合现象.

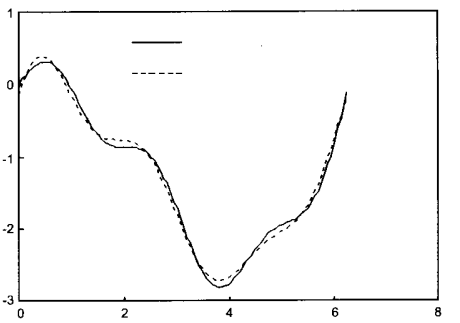

图 5 一维非线性函数及基于微分容量控制的学习机对它的逼近

表 1 对一维非线性函数在不同精度下的逼近结果

	E= 0.1	E= 0.01	E= 0.001	E= 0
训练误差	0.0735	0.0086	0.0045	0.0045
检测误差	0.0709	0.0087	0.0047	0.0047

413 回归估计问题

设数据集$(x_1, y_1, \ , \ , x_l, y_l)$是由区间[- 10, 10]上的一维 sinc 函数定义的; y 值受到正态分布的噪声的影响:

$$y_i=\sin(x_i)/x_i+\ n_i,\ i=1, \ , \ , l$$

其中 $n\sim N(0, R^2)$. 在区间[- 10, 10]内等间隔采样 101 个数据, 我们要从这些有噪的数据出发估计 sinc 函数 $\sin(x)/x$. 这里采用高斯核函数作为估计函数的基函数. 表 2 给出了基于微分容量控制的学习机和支撑矢量机在同等条件下的 20 次实验的平均结果. 表中的性能指标是训练集和测试集的真实目标函数与所得估计之间的标准差. 图 6 是在 E= 0l1 和 R= 0l1 下, 本文提出的学习机和支撑矢量机的逼近情况. 图 7 是在同等噪声方差、不同精度下, 基于微分容量控制的学习机的逼近情况.

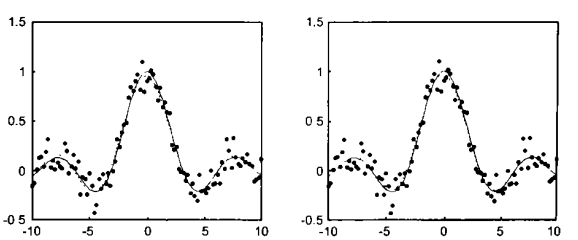

图 6 一维 sinc 函数和对它的逼近, R= 0.1, E= 0.1

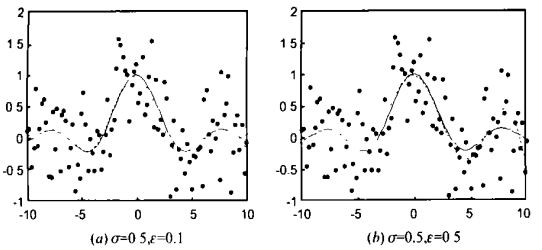

图 7　一维 sinc 函数及基于微分容量扩展的学习机对它的逼近

表 2　对一维 sinc 函数的回归性能比较

学习精度和噪声标准差	基于微分容量控制的学习机		支撑矢量机	
	训练误差	检测误差	训练误差	检测误差
E= 0.1, R= 0.1	0.0390	0.0382	0.0395	0.0386
E= 0.1, R= 0.5	0.1535	0.1474	0.1486	0.1435
E= 0.5, R= 0.5	0.1695	0.1651	0.1670	0.1638

5　结论和讨论

本文提出了一种基于微分容量控制的学习机. 该学习机可以选择任意一阶可微函数集来作为学习机的假设函数集. 其中线性微分容量控制学习机即相当于线性支撑矢量机. 我们根据统计学习理论中函数集容量概念说明了能够用假设函数集的微分范数来控制学习机的容量. 仿真实验的结果说明了该学习机无论采用 Mercer 核函数还是采用其他的假设函数集都能够得到较好的推广能力.

本文的结论在正则理论中也是非常合理的. 文献[12, 13]已经说明正则技术与支撑矢量机在统计学习理论的框架下是统一的. 正则范函中的稳定子, 在正则理论中的解释是限制函数的光滑度, 实际上就是通过限制函数的光滑性实现对假设函数集的容量控制. 本文提出的微分容量控制, 可以看成正则范函中的高通滤波器为 $1/G(s) = s^2$. 这样等价于本文也为高通滤波器采用 $1/G(s) = s^2$ 的正则技术给出了统计学习理论中的解释.

本文的仿真都是在训练样本较少的情况下得到的. 对于大训练样本集而言, 支撑矢量机已经有了解决大规模样本的快速算法[14~16]. 而基于微分容量控制学习机是否有快速算法, 需要在以后的工作中进一步研究.

参考文献:

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.

[2] Vapnik V. Statistical Learning Theory[M]. New York: John Wiley and Sons, Inc., 1998.

[3] Vapnik V. An overview of statistical learning theory[J]. IEEE TRANS. Neural Networks, 1999, 10(5): 988-999.

[4] 边肇祺, 张学工, 等. 模式识别[M]. 北京: 清华大学出版社, 2000 (第 2 版).

[5] C J C Burges. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 1-47.

[6] A Smola, B Sch&lkopf. A tutorial on support vector regression [EB/OL]. NeuroCOLT, Rep. 19, 1998. Available http://svm.first.gmd.de.

[7] C Saunders, et al. Support vector machine---reference manual [R]. Technical Report CSD-TR-98-03, Department of Computer Science, Royal Holloway, University of London, Egham, UK, 1998.

[8] 赵松年, 熊小芸. 子波变换与子波分析[M]. 北京: 电子工业出版社, 1997.

[9] 张贤达, 保铮. 非平稳信号分析与处理[M]. 北京: 国防工业出版社, 1998.

[10] 吴佑寿, 赵明生, 丁晓青. 一种激励函数可调的新人工神经网络及应用[J]. 中国科学(E 辑), 27(1): 55-60, 1997.

[11] 石卓尔, 焦李成, 保铮. 子波神经网络[A]. 中国神经网络 1993 学术大会论文集[C]. 西安: 1993 年中国神经网络学术会, 85-96, 1993.

[12] T Evgeniou, M Pontil, T Poggio. Regularization networks and support vector machines[A]. In A. J. Smola, P. L. Bartlett, B. Sch?lkopf, and D. Schuurmans, editors, Advances in Large Margin Classifiers[C]. Cambridge, MA: MIT Press, 2000.

[13] T Evgeniou, M Pontil, T Poggio. A unified framework of regularization networks and support vector machines[R]. A. I. Memo No. 1654, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1999.

[14] J Platt. Fast training of support vector machines using sequential minimal optimization[A]. In B. Sch&lkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in Kernel Methods Support Vector Learning[C]. Cambridge, MA: MIT Press, 1999. 185-208

[15] E Osuna, R Freund, G Girosi. Improved training algorithm for support vector machines[A]. Proc. IEEE NNSP. 97[C]. Amelia Island: IEEE, 1997.

[16] V Vapnik. Estimation of Dependences Based on Empirical Data[M]. New York: Springer-Verlag, 1982.

作者简介:

张　莉　女, 1975 年生于贵州省, 西安电子科技大学在读博士生, 主要研究方向有模式识别、人工神经网络和数据挖掘. Email: zhangli@rsp.xidian.edu.cn.

周伟达　男, 1974 年生于浙江省, 西安电子科技大学在读博士生, 主要研究方向有机器学习、统计学习理论和智能信号处理. Email: zhouwd@rsp.xidian.edu.cn.