

# 求解非半正定核Huber-支持向量回归问题的 序列最小最优化算法

周晓剑<sup>1</sup>, 马义中<sup>1</sup>, 朱嘉钢<sup>2</sup>, 刘利平<sup>1</sup>, 汪建均<sup>1</sup>

(1. 南京理工大学 管理科学与工程系, 江苏 南京 210094; 2. 江南大学 信息工程学院, 江苏 无锡 214122)

**摘要:** 序列最小最优化(SMO)算法是求解大型支持向量机(SVM)问题的有效算法. 已有的算法都要求核函数是正定的或半正定的, 从而使其应用受到限制. 针对这种缺点, 本文提出一种新的SMO算法, 可求解非半正定核Huber-SVR问题. 提出的算法在保证收敛的前提下可使非半正定Huber-SVR能够达到比较理想的回归精度, 因而具有一定的理论意义和实用价值.

**关键词:** 支持向量机; 非半正定核; 序列最小最优化算法; Huber-支持向量回归机

**中图分类号:** TP18 **文献标识码:** A

## Sequential-minimal-optimization algorithm for solving Huber-support-vector-regression with non-positive semi-definite kernels

ZHOU Xiao-jian<sup>1</sup>, MA Yi-zhong<sup>1</sup>, ZHU Jia-gang<sup>2</sup>, LIU Li-ping<sup>1</sup>, WANG Jian-jun<sup>1</sup>

(1. Department of Management Science and Technology, Nanjing University of Science and Technology, Nanjing Jiangsu 210094, China;  
2. School of Information Engineering, Jiangnan University, Wuxi Jiangsu 214122, China)

**Abstract:** Sequential-minimal-optimization(SMO) algorithm is effective in solving large-scale support-vector-machine(SVM) problems. However, the existing algorithms require the kernel functions to be positive definite(PD) or positive semi-definite(PSD), thus limiting their applications. Having considered their deficiencies, we propose a new algorithm for solving Huber-SVR problems with non-positive semi-definite(non-PSD) kernels. This algorithm provides desirable regression-accuracies while ensuring the convergence. Thus, it is of theoretical and practical significance.

**Key words:** support-vector-machine; non-positive semi-definite kernel; sequential-minimal-optimization algorithm; Huber-support vector regression

### 1 引言(Introduction)

在实际中, 由于SVM的应用所涉及的样本点越来越多, 所以利用通用的优化工具包(如MINOS, LOQO等)进行优化往往达不到理想的效果, 计算所花的时间过长. 针对这一问题文献[1]提出了最早的专用算法——选块算法. 它的基本思想是, 去掉非支持向量所对应的训练点, 而只对支持向量计算相应的Lagrange乘子, 由于支持向量是未知的, 选块算法需要通过某种迭代方式逐步排除非支持向量, 选出支持向量所对应的块. 这种方法的优点是当支持向量的数目远远小于训练样本的数目时, 能够大大提高运算效率. 然而, 如果支持向量的数目本身就较多时, 随着算法迭代的次数的增多, 所选的块也会越来越大, 算法就变得十分缓慢了. 对于选块算法所面

临的问题, 文献[2]针对支持向量分类机(SVC)提出了“分解”算法, 并由文献[3]将其扩展到SVR. “分解”算法克服了选块算法的不足, 该算法并不是要找出所有支持向量, 每次只是针对很小的训练子集来求解. 序列最小最优化(SMO)算法是分解算法中工作集的元素个数为2的一种特殊情形即每次迭代过程中只调整相应的两个样本点的Lagrange乘子, 它只求解一个具有两个变量的最优化问题. 实际上, 这时工作集的规模已经减到最小, 原因是原问题存在着等式约束. 只要变动一个变量, 就必须同时调整另一个变量来保证不违反该约束.

现有的SMO算法分为求解SVC的SMO算法和求解SVR的SMO算法. 求解SVC的SMO算法最早是由Platt在1998年提出的, 同年并由Smola将其扩展

到求解SVR. 以上这两类求解大型支持向量机的SMO算法都是假定支持向量机的核函数为正定的并且满足Mercer条件. 但是在某些特定的应用领域中, 所使用的核函数并不能保证为正定核函数, 如神经网络中的Sigmoid核, 在某些条件下就是非半正定(除正定和半正定以外的矩阵的), 并且在实际应用中, 除了已知的一些核函数外, 往往难以检验核函数的正定性或半正定性, 这样现有的SMO算法的应用就受到了限制. 文献[4]提出的用于解决非半正定SVC优化问题的方法, 可以保持SVC的泛化性能, 说明了研究非半正定SVC的必要性. 但有关非半正定Huber-SVR的求解方法目前还未见有重要的研究成果公开发表. 本文将文献[4]的思想引入到Huber-SVR模型求解中, 以解决非半正定核Huber-SVR求解问题.

### 2 Huber-SVR的基本形式(Basic form of Huber-SVR)

给定数据集 $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)\}$ (其中 $l$ 代表样本数目), 若SVR的损失函数取Huber函数<sup>[5]</sup>:

$$\tilde{c}(\xi) = \begin{cases} (\xi)^2 / 2\mu, & |\xi| < \mu, \\ |\xi| - \mu/2, & \text{其他.} \end{cases} \quad (1)$$

(其中:  $\mu$ 为损失函数的参数,  $\xi$ 为预测值与真实值的偏差), 则构成Huber-支持向量回归机(Huber-SVR), 其模型如下:

$$\begin{aligned} \min \Phi(\mathbf{w}, \boldsymbol{\xi}) &= \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l (\tilde{c}(\xi_i) + \tilde{c}(\xi_i^*)), \quad (2) \\ \text{s.t.} \quad &\begin{cases} w \cdot x_i + b - y_i \leq \xi_i, & i = 1, \dots, l, \\ y_i - w \cdot x_i - b \leq \xi_i^*, & i = 1, \dots, l, \\ \xi, \xi_i^* \geq 0, & i = 1, \dots, l. \end{cases} \end{aligned}$$

为便于问题的求解, 可将模型(2)写成Lagrange对偶模型:

$$\begin{aligned} \min_{\bar{\boldsymbol{\alpha}} \in \mathbb{R}^l} F(\bar{\boldsymbol{\alpha}}) &= \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \bar{\alpha}_i \bar{\alpha}_j K(x_i, x_j) - \\ &\quad \sum_{i=1}^l \bar{\alpha}_i y_i + \frac{\mu}{2C} \sum_{i=1}^l \bar{\alpha}_i^2, \quad (3) \\ \text{s.t.} \quad &\sum_{i=1}^l \bar{\alpha}_i = 0, -C \leq \bar{\alpha}_i \leq C, i = 1, \dots, l. \end{aligned}$$

其中 $\bar{\alpha}_i = \alpha_i^* - \alpha_i$ ,  $\alpha_i^{(*)}$ 为Lagrange乘子.

这样就把原问题转化为求解式(3)的最优化问题. 为便于推导, 可将(3)写成矩阵的形式:

$$\begin{aligned} \min_{\bar{\boldsymbol{\alpha}}} F(\bar{\boldsymbol{\alpha}}) &= \frac{1}{2} \bar{\boldsymbol{\alpha}}^T \mathbf{K}' \bar{\boldsymbol{\alpha}} - \mathbf{y}^T \bar{\boldsymbol{\alpha}}, \quad (4) \\ \text{s.t.} \quad &\sum_{i=1}^l \bar{\alpha}_i = 0, -C \leq \bar{\alpha}_i \leq C, i = 1, \dots, l. \end{aligned}$$

其中 $\mathbf{K}' = \mathbf{K} + \sigma \mathbf{I}$ ,  $\sigma = \sqrt{\mu/2C}$ .

### 3 求解非半正定核 Huber-SVR 的理论前提 (Theoretical premise for solving Huber-SVR with non-PSD kernels)

当核函数为非半正定时, 式(4)中的核矩阵 $\mathbf{K}'$ 就不能写成如下内积的形式:

$$\mathbf{K}'(x_i, x_j) \equiv \varphi(x_i)^T \varphi(x_j). \quad (5)$$

因此式(4)也就不能还原到式(2)那种形式. 从而当从式(4)中解出 $\boldsymbol{\alpha}$ 后, 还不能确定它就是式(2)的解, 或者说, 式(2)与式(4)的对偶关系不再成立. 也就无法得到有效的回归函数. 对此, 令

$$w(\boldsymbol{\alpha}) = \sum_i \alpha_i \varphi(x_i), \quad (6)$$

考虑如下新的二次规划问题:

$$\min \Phi(\boldsymbol{\alpha}, \boldsymbol{\xi}) = \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{K}' \boldsymbol{\alpha} + C \sum_{i=1}^l (\tilde{c}(\xi_i) + \tilde{c}(\xi_i^*)), \quad (7)$$

$$\text{s.t.} \quad \begin{cases} (\mathbf{K}' \boldsymbol{\alpha})_i + b - y_i \leq \xi_i, & i = 1, \dots, l, \\ y_i - (\mathbf{K}' \boldsymbol{\alpha})_i - b \leq \xi_i^*, & i = 1, \dots, l, \\ \xi, \xi_i^* \geq 0, & i = 1, \dots, l. \end{cases}$$

上式是把式(6)代入式(2)得到的. 这样做的目的是为了研究当用式(4)求出 $\boldsymbol{\alpha}$ 后, 用 $\boldsymbol{\alpha}$ 来构建回归函数时误差是否会很大. 为此, 有如下定理.

**定理 1** 如果 $\hat{\boldsymbol{\alpha}}$ 是(4)的稳定点, 则它是(7)的可行解.

**证** 注意到 $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^* - \boldsymbol{\alpha}$ , 为便于求证, 先将(4)还原为:

$$\min_{\bar{\boldsymbol{\alpha}}} F(\bar{\boldsymbol{\alpha}}) = \frac{1}{2} (\boldsymbol{\alpha}^* - \boldsymbol{\alpha})^T \mathbf{K}' (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}) - \mathbf{y}^T (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}).$$

假设 $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}^* - \hat{\boldsymbol{\alpha}}$ 是(4)的稳定点, 则它满足KKT条件. 即存在Lagrange乘子 $\hat{\gamma}^{(*)}, \hat{\lambda}^{(*)}$ 满足:

$$\mathbf{K}'(\hat{\boldsymbol{\alpha}}^* - \hat{\boldsymbol{\alpha}}) - \mathbf{y} + \hat{b}\mathbf{e} - \hat{\gamma}^* + \hat{\lambda}^* = 0, \quad (8a)$$

$$-\mathbf{K}'(\hat{\boldsymbol{\alpha}}^* - \hat{\boldsymbol{\alpha}}) + \mathbf{y} - \hat{b}\mathbf{e} - \hat{\gamma} + \hat{\lambda} = 0, \quad (8b)$$

$$\hat{\gamma}^{(*)} \geq 0, \quad (8c)$$

$$\hat{\lambda}^{(*)} \geq 0, \quad (8d)$$

$$\hat{\gamma}^{(*)T} \hat{\boldsymbol{\alpha}}^{(*)} = 0, \quad (8e)$$

$$\hat{\lambda}^{(*)T} (\hat{\boldsymbol{\alpha}}^{(*)} - C\mathbf{e}) = 0. \quad (8f)$$

由式(8a)(8b)和(8c)可得:

$$\begin{aligned} \mathbf{K}'(\hat{\boldsymbol{\alpha}}^* - \hat{\boldsymbol{\alpha}}) - \mathbf{y} + \hat{b}\mathbf{e} &\geq -\hat{\lambda}^*, \\ -\mathbf{K}'(\hat{\boldsymbol{\alpha}}^* - \hat{\boldsymbol{\alpha}}) + \mathbf{y} - \hat{b}\mathbf{e} &\geq -\hat{\lambda}. \end{aligned}$$

即 $y_i - (\mathbf{K}' \hat{\boldsymbol{\alpha}})_i - \hat{b} \leq \hat{\lambda}_i^*$ ,  $(\mathbf{K}' \hat{\boldsymbol{\alpha}})_i + \hat{b} - y_i \leq \hat{\lambda}_i$ . 所以,

取 $[\alpha; b; \xi^{(*)}] = [\hat{\alpha}; \hat{b}; \lambda^{(*)}]$ 即为式(7)的可行解.

从上面的定理可以看出, 当 $K'$ 非半正定时, 用求解(4)去代替求解(2)是可行的. 这说明, 在某些特定的场合可以考虑用非半正定的核函数去替代正定的核函数来进行回归. 而本文的目的正是要设计出针对非半正定核Huber-SVR的SMO算法.

#### 4 非半正定核Huber-SVR的求解(Solving Huber-SVR with non-PSD kernels)

##### 4.1 Huber-SVR的SMO模型(Model of SMO for Huber-SVR)

就公式(4)的优化模型而言, 每一步可以把 $\bar{\alpha}$ 的分量分成两部分: 一部分属于工作集; 另一部分属于非工作集. 记它们对应的下标集分别为 $B$ 和 $N$ . 需要调整的是 $\bar{\alpha}_B$ , 而 $\bar{\alpha}_N$ 是固定不变的. 如果用上述分解算法解决Huber-SVR的问题, 则Huber-SVR的二次规划问题就转变为:

$$\min_{\bar{\alpha}} F(\bar{\alpha}_B \bar{\alpha}_N) = \frac{1}{2} \begin{bmatrix} \bar{\alpha}_B^T & \bar{\alpha}_N^T \end{bmatrix} \begin{bmatrix} K'_{BB} & K'_{BN} \\ K'_{NB} & K'_{NN} \end{bmatrix} \begin{bmatrix} \bar{\alpha}_B \\ \bar{\alpha}_N \end{bmatrix} - \begin{bmatrix} y_B^T & y_N^T \end{bmatrix} \begin{bmatrix} \bar{\alpha}_B \\ \bar{\alpha}_N \end{bmatrix}, \quad (9)$$

s.t.  $-C \leq \bar{\alpha}_i \leq C, i = 1, \dots, l,$   
 $e_B \bar{\alpha}_B = -e_N \bar{\alpha}_N.$

当移去常数项后, 即为:

$$\min_{\bar{\alpha}_B} F(\bar{\alpha}_B \bar{\alpha}_N) = \frac{1}{2} \bar{\alpha}_B^T K'_{BB} \bar{\alpha}_B + \bar{\alpha}_B^T K'_{BN} \bar{\alpha}_N - y_B^T \bar{\alpha}_B, \quad (10)$$

s.t.  $-C \leq \bar{\alpha}_i \leq C, i \in B, e_B \bar{\alpha}_B = -e_N \bar{\alpha}_N.$

相应于SMO算法的Huber-SVR模型为:

$$\min_{\bar{\alpha}_i, \bar{\alpha}_j} F(\bar{\alpha}_B \bar{\alpha}_N) = \frac{1}{2} [\bar{\alpha}_i \ \bar{\alpha}_j] \begin{bmatrix} K'_{ii} & K'_{ij} \\ K'_{ji} & K'_{jj} \end{bmatrix} \begin{bmatrix} \bar{\alpha}_i \\ \bar{\alpha}_j \end{bmatrix} - \left( \begin{bmatrix} y_i \\ y_j \end{bmatrix} + K'_{BN} \bar{\alpha}_N \right)^T \begin{bmatrix} \bar{\alpha}_i \\ \bar{\alpha}_j \end{bmatrix}, \quad (11)$$

s.t.  $-C \leq \bar{\alpha}_k \leq C, k \in \{i, j\},$   
 $\bar{\alpha}_i + \bar{\alpha}_j = -\sum_{k \neq i, j} \bar{\alpha}_k, \bar{\alpha}_k = \text{其他}.$

##### 4.2 工作集的选取及停机准则的设定(Selection of working set and setting of stopping criteria)

工作集的选取是SMO算法中最为重要的一个环节, 工作集选取的好坏直接影响算法的收敛速度. 在文献[6]及文献[7]中, SVC和SVR的SMO算法寻找工作集的简要过程是: 在外循环中找出违反KKT(Karush-Kuhn-Tucker)条件最严重的训练点作为第1个点, 然后在内循环中找到能使目标函数值有

一定下降的点作为第2个训练点. 而文献[8]则是通过 $b_{low}, b_{up}$ 两个指标来选择工作集, 它们所基于的思想也是找出违反KKT条件最为严重的点, 但收敛速度明显优于文献[7]中的方法. 本文将文献[8]求解 $\epsilon$ -SVR中工作集的选取思想用于非半正定核Huber-SVR工作集选取中.

停机准则从不同的角度一般可为分3种, 具体可参见文献[9]. 在本文中, 是通过观察KKT条件的违背程度来判断是否可以停机. 以下给出工作集的选取及停机准则的推导过程.

对式(4)引入Lagrange函数

$$\bar{L} = F(\bar{\alpha}) + \beta \sum_{i=1}^l \bar{\alpha}_i - \sum_{i=1}^l \phi_i (\bar{\alpha}_i + C) - \sum_{i=1}^l \delta_i (C - \bar{\alpha}_i), \quad (12)$$

则相应的KKT条件为:

$$\begin{cases} \frac{\partial \bar{L}}{\partial \bar{\alpha}_i} = \nabla F(\bar{\alpha})_i + \beta - \phi_i + \delta_i = 0, \\ \phi_i (\bar{\alpha}_i + C) = 0, \phi_i \geq 0, \\ \delta_i (\bar{\alpha}_i - C) = 0, \delta_i \geq 0, \\ i = 1, \dots, l. \end{cases} \quad (13)$$

对于上式分3种互不相交的情形加以考虑:

i)  $\bar{\alpha}_i = -C$ , 则  $\phi_i \geq 0, \delta_i = 0,$   
 $\nabla F(\bar{\alpha})_i + \beta \geq 0, \quad (14a)$

ii)  $-C < \bar{\alpha}_i < C$  则  $\phi_i = 0, \delta_i = 0,$   
 $\nabla F(\bar{\alpha})_i + \beta = 0, \quad (14b)$

iii)  $-C < \bar{\alpha}_i = C$  则  $\phi_i = 0, \delta_i \geq 0,$   
 $\nabla F(\bar{\alpha})_i + \beta \leq 0. \quad (14c)$

综上所述可得:

当 $-C \leq \bar{\alpha}_i < C$ 时,  $\beta \geq -\nabla F(\bar{\alpha})_i;$

当 $-C < \bar{\alpha}_i \leq C$ 时,  $\beta \leq -\nabla F(\bar{\alpha})_i.$

若定义:

$$\begin{cases} I_{up}(\bar{\alpha}) = \{i \mid -C < \bar{\alpha}_i \leq C\}, \\ I_{low}(\bar{\alpha}) = \{i \mid -C \leq \bar{\alpha}_i < C\}. \end{cases} \quad (15)$$

有如下定理:

**定理 2** 如果 $\hat{\alpha}$ 是式(4)的稳定点, 当且仅当不存在有序对 $\{i, j\}$ , 其中,  $i \in I_{up}(\hat{\alpha}), j \in I_{low}(\hat{\alpha}).$  使得

$$-\nabla F(\bar{\alpha})_i < -\nabla F(\bar{\alpha})_j, \quad (16)$$

定理的结论可以直接从上面的推导过程得到.

从上面的定理可以看出, 如果式(4)没有达到稳定点, 本文可以选取违犯KKT条件最为严重的点作

为工作集. 因此, 可以得到工作集的选取方法及停机准则.

若令

$$i = \arg \min_{i \in I_{\text{up}}(\bar{\alpha})} -\nabla F(\bar{\alpha})_i, j = \arg \max_{j \in I_{\text{low}}(\bar{\alpha})} -\nabla F(\bar{\alpha})_j,$$

则工作集取为  $B = \{i, j\}$ .

若再令

$$\begin{cases} b_{\text{up}} = \min\{-\nabla F(\bar{\alpha})_i : i \in I_{\text{up}}(\bar{\alpha})\}, \\ b_{\text{low}} = \max\{-\nabla F(\bar{\alpha})_j : j \in I_{\text{low}}(\bar{\alpha})\}. \end{cases} \quad (17)$$

则停机准则取为:  $b_{\text{low}} \leq b_{\text{up}}$ , 由于在数值解中, 不可能完全达到最优, 因此本文可以用下式代替  $b_{\text{low}} \leq b_{\text{up}} + 2\text{tol}$ , 其中,  $\text{tol}$  是表示误差所能容忍的范围.

### 4.3 求解非半正定核Huber-SVR的SMO算法设计(Design of the algorithm of SMO for solving Huber-SVR with non-PSD kernels)

为便于问题的求解及收敛性的判定, 作如下变换: 在式(9)中, 令

$$\bar{\alpha}_i = \bar{\alpha}_i^k + d_i, \bar{\alpha}_j = \bar{\alpha}_j^k + d_j, \quad (18)$$

则含两变量的子最优化问题可化为:

$$\begin{aligned} \min_{\bar{\alpha}_i, \bar{\alpha}_j} F(\bar{\alpha}_B, \bar{\alpha}_N) = & \frac{1}{2} [d_i \ d_j] \begin{bmatrix} K'_{ii} & K'_{ij} \\ K'_{ji} & K'_{jj} \end{bmatrix} \begin{bmatrix} d_i \\ d_j \end{bmatrix} + \\ & [\nabla F(\bar{\alpha}^k)_i \ \nabla F(\bar{\alpha}^k)_j] \begin{bmatrix} d_i \\ d_j \end{bmatrix}, \end{aligned} \quad (19)$$

$$\text{s.t. } -C \leq \bar{\alpha}_i^k + d_i, \bar{\alpha}_j^k + d_j \leq C, d_i + d_j = 0.$$

其中:

$$\begin{bmatrix} \nabla F(\bar{\alpha}^k)_i \\ \nabla F(\bar{\alpha}^k)_j \end{bmatrix} = \begin{bmatrix} K'_{ii} & K'_{ij} \\ K'_{ji} & K'_{jj} \end{bmatrix} \begin{bmatrix} \bar{\alpha}_i^k \\ \bar{\alpha}_j^k \end{bmatrix} + \begin{bmatrix} y_i \\ y_j \end{bmatrix} + \mathbf{Q}_{BN} \bar{\alpha}_N,$$

上式的  $k$  代表第  $k$  次迭代. 由  $(\bar{\alpha}_i^k + d_i) + (\bar{\alpha}_j^k + d_j) = \bar{\alpha}_i^k + \bar{\alpha}_j^k$  得  $d_i + d_j = 0$ , 即,  $d_i = -d_j$ . 再由  $-C \leq \bar{\alpha}_i^k + d_i, \bar{\alpha}_j^k + d_j \leq C$ , 令

$$\begin{cases} L = \max(\bar{\alpha}_i^k - C, -C - \bar{\alpha}_j^k), \\ H = \min(\bar{\alpha}_i^k + C, C - \bar{\alpha}_j^k), \end{cases} \quad (20)$$

则式(19)可化为:

$$\begin{aligned} \min_{d_j} & \frac{1}{2} (K'_{ii} - 2K'_{ij} + K'_{jj}) d_j^2 + \\ & (-\nabla F(\bar{\alpha}^k)_i + \nabla F(\bar{\alpha}^k)_j) d_j, \end{aligned} \quad (21)$$

$$\text{s.t. } L \leq d_j \leq H.$$

这样就把优化问题转化为在给定区间求抛物线的极值点了, 从而可以解析求解式(4).

求解非半正定核Huber-SVR的SMO算法:

初始化: 设定精度  $\text{tol}$ ,  $\bar{\alpha}^0 = \mathbf{0}$ , 令  $k = 0$ , 并初始化  $b_{\text{low}}, b_{\text{up}}$ ;

**While** (停机准则不满足)

- 1) 选取工作集;
- 2) 解子优化问题(21);
- 3) 计算  $\bar{\alpha}^{k+1}$ ;
- 4) 令  $k = k + 1$ ;

**end while**

返回: 近似最优解  $\hat{\alpha}$ .

对于算法循环部分的第2步中用(21)求解得到  $\hat{d}_j$ , 并进而得到  $\bar{\alpha}^{k+1}$  后, 接下来有3个问题需要回答:

- 1) 在迭代的过程中  $F(\bar{\alpha})$  是否有效下降的?
- 2) 在迭代的过程中  $\bar{\alpha}$  是否收敛的?
- 3) 如果  $\bar{\alpha}$  是收敛的, 那么是否收敛于  $F(\bar{\alpha})$  的稳定点?

这三个问题是支撑整个算法的关键问题所在, 为了使上述3个问题能得到肯定的回答, 这里采取了如下处理策略.

式(21)的求解策略. 以往的收敛分析通常要求  $F(\bar{\alpha})$  满足如下重要的性质:

$$F(\bar{\alpha}^{k+1}) \leq F(\bar{\alpha}^k) - \tau \|\bar{\alpha}^{k+1} - \bar{\alpha}^k\|^2, \quad (22)$$

其中  $k$  是指第  $k$  次迭代. 对于当核矩阵为半正定的情形时文献[10]证明了在SVC中式(22)成立, 然而当核矩阵的正定性未知时就很难保证上式成立. 为了得到式(22), 文献[12]针对SVC的情形对分解算法中的子问题的形式进行了改写. 改写后的形式能非常方便地验证(22)成立, 但不足之处在于, 改进后的形式仍不能保证目标函数是凸函数, 因此在实现的时候还需要对凸函数和非凸函数两种情形分开加以考虑. 本文将文献[11]的思想引入到非半正定Huber-SVR的SMO算法求解当中来, 并通过适当的系数选择使子问题成为凸规划问题, 从而简化了算法的实现. 为此, 可以将式(11)写成如下形式:

$$\min_{\bar{\alpha}_i, \bar{\alpha}_j} F(\bar{\alpha}_B, \bar{\alpha}_N) + \tau \|\bar{\alpha}_B - \bar{\alpha}_B^k\|^2, \quad (23)$$

$$\text{s.t. } -C \leq \bar{\alpha}_i \leq C, i = 1, \dots, 2l,$$

$$\mathbf{e}_B \bar{\alpha}_B = -\mathbf{e}_N \bar{\alpha}_N, B \in \{i, j\}.$$

因为  $\|\bar{\alpha}_B - \bar{\alpha}_B^k\|^2 = 2d_j^2$ , 将其代入式(23)并结合式(21)可得:

$$\begin{aligned} \min_{d_j} & \frac{1}{2} (K'_{ii} - 2K'_{ij} + K'_{jj} + 4\tau) d_j^2 + \\ & (-\nabla F(\bar{\alpha}^k)_i + \nabla F(\bar{\alpha}^k)_j) d_j, \end{aligned} \quad (24)$$

$$\text{s.t. } L \leq d_j \leq H.$$

为了将式(24)转换为凸规划问题, 可对附加项作如下

处理:

令

$$\tau = \begin{cases} \frac{1}{4}(\sigma - (K'_{ii} - 2K'_{ij} + K'_{jj})), \\ \text{当 } K'_{ii} - 2K'_{ij} + K'_{jj} < 0, \\ \sigma, \text{ 当 } K'_{ii} - 2K'_{ij} + K'_{jj} \geq 0. \end{cases}$$

其中 $\sigma$ 取任意小的正数,这样就保证了式(24)的二次项的系数大于0,也就保证了(23)的凸性.这样算法第2步就可以通过解(24)去代替解(21),从而能有效地保证算法的收敛性.

算法收敛性分析.下面逐个来回答上一节提到的3个问题.对于第1个问题有如下定理:

**定理 3** 如果上述SMO算法没有终止,并让 $\{\bar{\alpha}^k\}$ 序贯产生,则有:

$$F(\bar{\alpha}^{k+1}) \leq F(\bar{\alpha}^k) - \tau \|\bar{\alpha}^{k+1} - \bar{\alpha}^k\|^2, \quad (25)$$

证 从(23)可知:

$$\begin{aligned} & F(\bar{\alpha}^{k+1}) + \tau \|\bar{\alpha}^{k+1} - \bar{\alpha}^k\|^2 = \\ & F(\bar{\alpha}_B^{k+1}, \bar{\alpha}_N^k) + \tau \|\bar{\alpha}_B^{k+1} - \bar{\alpha}_B^k\|^2 \leq \\ & F(\bar{\alpha}_B^k, \bar{\alpha}_N^k) = F(\bar{\alpha}^k). \end{aligned}$$

定理得证. 证毕.

对于第2个问题,有如下定理:

**定理 4** 如果上述SMO算法没有终止,并让 $\{\bar{\alpha}^k\}$ 序贯产生,则有  $\lim_{k \rightarrow \infty} \|\bar{\alpha}^{k+1} - \bar{\alpha}^k\| = 0$ .

证 由定理3可知:

$$F(\bar{\alpha}^{k+1}) + \tau \|\bar{\alpha}^{k+1} - \bar{\alpha}^k\|^2 \leq F(\bar{\alpha}^k). \quad (26)$$

因此 $F(\bar{\alpha}^k)$ 是单调递减的,又由于 $\{\bar{\alpha}^k\}$ 所属的可行集是闭集,所以 $F(\bar{\alpha}^k)$ 有界.单调有界则有极限,即  $\lim_{k \rightarrow \infty} F(\bar{\alpha}^{k+1}) = \lim_{k \rightarrow \infty} F(\bar{\alpha}^k) = \varsigma$ ( $\varsigma$ 为一定值).所以,对式(26)两边取极限得:  $\lim_{k \rightarrow \infty} \|\bar{\alpha}^{k+1} - \bar{\alpha}^k\| = 0$ .

证毕.

在回答第3个问题前,先给出以下定义和引理.

**定义 1** 如果 $d \in \mathbb{R}^l$ ,  $\sum_i d_i = 0$ , 且 $d_i \geq 0, \forall i \in \{i | \alpha_i = -C\}$ 或 $d_i \leq 0, \forall i \in \{i | \alpha_i = C\}$ , 则称 $d$ 为 $\alpha$ 处的可行方向, 记为 $D(\alpha)$ .

对于定义1需要说明的是,因为在SMO算法中,工作集的元素只有两个,即 $B = \{i, j\}$ ,所以在所有的Lagrange乘子中,只有 $i$ 和 $j$ 所对应的Lagrange乘子会改变,而且是等量反向变化的,原因是 $d_i + d_j = 0$ .

**引理 1** 设 $\bar{\alpha}$ 为式(4)的可行点,且 $i \in I_{up}(\bar{\alpha})$ 和 $j \in I_{low}(\bar{\alpha})$ ,如果方向 $d$ 使得下式成立:

$$d_i = -1, d_j = 1, d_k = 0, k \in \{k | k \neq i, j\}, \quad (27)$$

则 $d$ 为 $\bar{\alpha}$ 处的一个可行方向,即 $d \in D(\bar{\alpha})$ .

证 因为 $i \in I_{up}(\bar{\alpha})$ ,则 $\bar{\alpha}_i \neq -C$ ,根据可行方向的定义, $d_i$ 可以取负值,同理,由 $j \in I_{low}(\bar{\alpha})$ ,则 $\bar{\alpha}_j \neq C$ , $d_j$ 可以取正值,并且 $\sum_i d_i = 0$ .所以 $d_i = -1, d_j = 1, d_k = 0, k \in \{k | k \neq i, j\}$ 满足可行方向的定义,即 $d \in D(\bar{\alpha})$ .引理得证. 证毕.

**引理 2** 如果上述SMO算法没有终止,并让 $\{\bar{\alpha}^k\}$ 序贯产生,并令

$$\begin{aligned} i &= \arg \min_{i \in I_{up}(\bar{\alpha}^{k+1})} -\nabla F(\bar{\alpha}^{k+1})_i, \\ j &= \arg \max_{j \in I_{low}(\bar{\alpha}^{k+1})} -\nabla F(\bar{\alpha}^{k+1})_j, \end{aligned}$$

则 $\nabla F(\bar{\alpha}^{k+1})^T d + 2\tau(\bar{\alpha}^{k+1} - \bar{\alpha}^k)^T d \geq 0$ .其中 $d$ 满足式(27).

证 由引理1可以知道 $d$ 是可行点 $\bar{\alpha}^{k+1}$ 处的一个可行方向.记 $\hat{\alpha}_B$ 为(23)中第 $k$ 次迭代的最优解,所以 $\bar{\alpha}_B^{k+1} = \hat{\alpha}_B, \bar{\alpha}_N^{k+1} = \bar{\alpha}_N^k$ .由于式(23)是凸规划问题,所以有

$$\nabla_B F(\hat{\alpha}_B, \bar{\alpha}_N^k)^T d_B + 2\tau(\hat{\alpha}_B - \bar{\alpha}_B^k)^T d_B \geq 0. \quad (28)$$

结合 $\bar{\alpha}_B^{k+1} = \hat{\alpha}_B, \bar{\alpha}_N^{k+1} = \bar{\alpha}_N^k$ 并注意到 $d_k = 0, k \in \{k | k \neq i, j\}$ ,则式(28)可写为:

$$\nabla F(\bar{\alpha}^{k+1})^T d + 2\tau(\bar{\alpha}^{k+1} - \bar{\alpha}^k)^T d \geq 0.$$

引理得证. 证毕.

下面的定理给出第3个问题的回答.

**定理 5** 如果上述SMO算法没有终止,并让 $\{\bar{\alpha}^k\}$ 序贯产生,则 $\{\bar{\alpha}^k\}$ 的极限点是(4)的稳定点.

证 采用反证法.令 $\tilde{\alpha} = \lim_{k \rightarrow \infty} \bar{\alpha}^k$ ,如果 $\tilde{\alpha}$ 不是式(4)的稳定点,即其KKT条件不成立,则根据定理2可知,存在 $i \in I_{up}(\tilde{\alpha}), j \in I_{low}(\tilde{\alpha})$ .使得下式成立:

$$-\nabla F(\tilde{\alpha})_i < -\nabla F(\tilde{\alpha})_j, \quad (29)$$

根据工作集的选取准则,在第 $k$ 次迭代中,有

$$\begin{aligned} i &= \arg \min_{i \in I_{up}(\bar{\alpha}^{k+1})} -\nabla F(\bar{\alpha}^{k+1})_i, \\ j &= \arg \max_{j \in I_{low}(\bar{\alpha}^{k+1})} -\nabla F(\bar{\alpha}^{k+1})_j. \end{aligned}$$

由引理2,可得:

$$\nabla F(\bar{\alpha}^{k+1})^T d + 2\tau(\bar{\alpha}^{k+1} - \bar{\alpha}^k)^T d \geq 0.$$

其中 $d$ 满足式(27).

另外,由定理(4),有  $\lim_{k \rightarrow \infty} \|\bar{\alpha}^{k+1} - \bar{\alpha}^k\| = 0$ ,因此,当 $k \rightarrow \infty$ 时,

$$\nabla F(\tilde{\alpha})^T d = -\nabla F(\tilde{\alpha})_i + \nabla F(\tilde{\alpha})_j \geq 0.$$

与式(29)矛盾, 所以原命题成立. 证毕.

上面若干定理表明, 当采用上述策略对(21) 进行变换后再求解能保证算法的收敛性. 并且收敛到(4)的稳定点. 虽然这里提出的求解非半正定核的SMO算法是收敛到稳定点而不是全局最优点, 但是在某些情况下却能达到比正定核或半正定核更好的回归效果, 以下的实验说明了这一点.

### 5 实验及其结果(Experiments and results)

#### 5.1 实验设计(Design for experiments)

##### 实验 1 取函数

$$y = \cos(\exp x), x \in [-4, 2],$$

为基准函数, 令间隔为0.1产生训练样本. 分别用正定核Huber-SVR与非半正定核Huber-SVR对这组数据进行训练, 然后进行回归. 非半正定核Huber-SVR的求解使用本文的算法. 这一实验主要考察本文算法的可行性以及非半正定核Huber-SVR的回归性能.

**实验 2** 在实际的4000个鲍鱼样本数据中, 随机抽取50个作为训练样本, 分别用正定核Huber-SVR与非半正定核Huber-SVR对这组数据进行训练. 再在剩余的样本数据中随机抽取3组每组50个样本数据作为预测测试样本, 利用训练的结果, 分别用正定核Huber-SVR与非半正定核Huber-SVR对这组数据进行预测. 每一个鲍鱼样本数据包括: 性别、长度、直径、厚度、总重量、去外壳的重量、内脏的重量、外壳的重量共8项指标. 这一实验主要考察非半正定核Huber-SVR的泛化性能. 鲍鱼数据取自: <http://archive.ics.uci.edu/ml/datasets/Abalone>.

在以上两个实验中, 非半正定核Huber-SVR的核函数取组合高斯核

$$\exp\left(\frac{-\|s-t\|^2}{\sigma_1}\right) + \exp\left(\frac{-\|s-t\|^2}{\sigma_2}\right) - \exp\left(\frac{-\|s-t\|^2}{\sigma_3}\right)_{[12]},$$

标准差分别取为 $\sigma_1 = 0.3, \sigma_2 = 0.2, \sigma_3 = 1$ . 正定核Huber-SVR的核函数取Gauss径向基核函数

$$\exp\left(\frac{-\|s-t\|^2}{2\sigma^2}\right),$$

其中 $\sigma = 1$ .

#### 5.2 实验结果与分析(Results and analysis of experiments)

实验1中, Huber-SVR的参数设置为 $C = 10^5, \mu = 0.05$ . 有关参数的选择方法参见文献 [13, 14]. 组合高斯核的预定精度为 $\text{tol} = 0.1$ . 正定核的预定精度为 $\text{tol} = 0.2$ . 回归的结果见图1和图2. 图中符号“\*”为样本点, 粗实线为回归线.

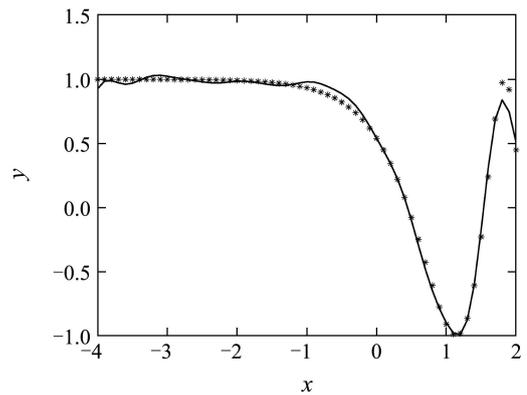


图 1 用非半正定核对 $y = \cos(\exp x)$ 回归的结果  
Fig.1 The regression result for  $y = \cos(\exp x)$  with non-PSD kernel

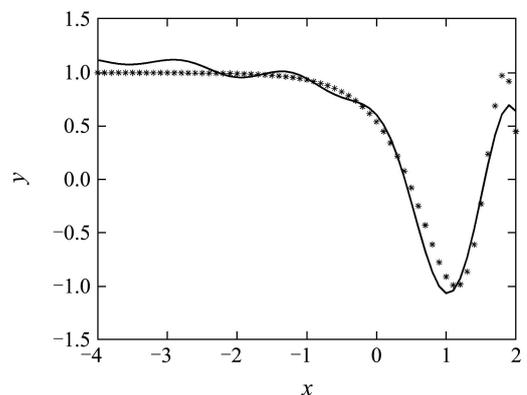


图 2 用正定核对 $y = \cos(\exp x)$ 回归的结果  
Fig.2 The regression result for  $y = \cos(\exp x)$  with PD kernel

同时, 实验2还从以下3个常用的度量准则对实验的回归结果进行定量比较(结果见表1):

1) 最大绝对误差:

$$\text{MAE}(\text{max absolute error}) = \max |y_i - \hat{y}|,$$

其中 $i = 1, \dots, n, n$ 为样本容量.

2) 平均绝对误差:

$$\text{AAE}(\text{average absolute error}) = \sum_{i=1}^{n_{\text{error}}} |y_i - \hat{y}| / n_{\text{error}}.$$

3) 均方根误差:

$$\text{RMSE}(\text{root mean square error}) = \sqrt{\sum_{i=1}^{n_{\text{error}}} (y_i - \hat{y})^2 / n_{\text{error}}}.$$

比较图1和图2可以看出, 对于函数 $\cos(\exp x)$ 用非半正定核回归要优于用正定核回归. 另外, 通过表1的定量比较也可以看出在MAE、AAE及RMSE 3方面, 用非半正定核来进行回归比采用正定核能够

达到更理想的回归精度. 这说明, 对于某些难以回归的函数, 可以考虑选用非半正定核来回归, 同时也说明了本文算法设计的科学性.

表1 用正定核和非半正定核回归的误差比较  
Table 1 The comparison of regression error between using PD kernel and using non-PSD kernel

误差	用正定核回归	用非半正定核回归
MAE	0.3612	0.1733
AAE	0.0877	0.0303
RMSE	0.1153	0.0436

实验2中, Huber-SVR的参数设置为 $C = 10^5$ ,  $\mu = 0.005$ . 组合高斯核的预定精度为 $\text{tol} = 6$ , 正定核的预定精度为 $\text{tol} = 5$ . 同样, 本实验对每一组结果从MAE, AAE, RMSE3个角度来进行分析. 表2给出了实验2的误差比较.

表2 用正定核和非半正定核预测的误差比较  
Table 2 The comparison of prediction error between using PD kernel and using non-PSD kernel

	非半正定核			正定核		
	MAE	AAE	RMSE	MAE	AAE	RMSE
第I组	6.5250	2.0006	2.4554	6.3894	2.1000	2.4867
第II组	6.3662	2.5625	2.9458	6.6567	2.7199	3.2026
第III组	9.2981	2.6548	3.4068	9.5334	2.7992	3.4673

从表2中的非半正定核与正定核的误差分析可以看出: 在第1组实验数据中, 非半正定核的MAE略大于正定核的MAE, 但其他两项数据均小于对应的数据; 在第2组和第3组实验数据中, 前者(非半正核)3项指标(MAE, AAE, RMSE)均小于后者(正定核). 综上可得: 对于实际当中的某些例子, 用非半正定核来预测要优于用正定核, 从而说明研究非半正定核的必要性以及本文算法的有效性.

## 6 结论(Conclusion)

本文提出了用于求解非半正定核Huber-SVR的SMO算法. 由于非半正定核的SVR模型一般认为是无法求解, 但现实中有的情况下却需要用到非半正定核函数, 这就需要能有效地解决非半正定核Huber-SVR的SMO算法. 因此本文提出的求解方法具有一定的理论意义和实用价值. 对于非半正定核, 如sigmoid核函数也包括本文所使用的核函数, 由于只能保证解是局部最优, 并不一定是全局最优, 因此, 如何克服它们的不足并发挥它们的优势仍然是一个有意义的研究课题.

## 参考文献(References):

- [1] VAPNIK V N. *Estimation of Dependences Based on Empirical Data*[M]. Berlin: Springer-Verlag, 1982.
- [2] OSUNA E, FREUND R, GIROSI F. Improved training algorithm for support vector machines[C] // *Neural Networks and Signal Processing*. Amelia Island, FL: IEEE CS Press, 1997: 276 – 285.
- [3] OSUNA E. *Support Vector Machines: Training and Applications*[M]. Cambridge, MA: Massachusetts Institute of Technology, 1998.
- [4] LIN H T, LIN C J. *A study on sigmoid kernels for Svm and the training of non-psd Kernels by Smo-Type methods*[R]. Taipei, Taiwan: Department of Computer Science and Information Engineering, National Taiwan University, 2003.
- [5] HUBER P J. Robust statistics: a review[J]. *Annals of Mathematical Statistics*, 1972, 35: 73 – 101.
- [6] PLATT J C. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization. Advances in Kernel Methods: Support Vector Machines*[M]. Edited by Scholkopf B, Burges C, Smola A. Cambridge MA: MIT Press, 1998: 185 – 208.
- [7] SMOLA J, SCHOLKOPF B. *A tutorial on support vector regression*[R]. Royal Holloway College, Neuro COLT Technology Report TR-1998-030, 1998.
- [8] KEERTHI S, SHEVADE S, BHATTACHARYYA C, et al. Improvements to SMO algorithm for SVM regression[J]. *IEEE Transactions on Neural Networks*, 2000, 11(5): 1188 – 1193.
- [9] 邓乃扬, 田英杰. 数据挖掘的新方法——支持向量机[M]. 北京: 科学出版社, 2004.  
(DENG Naiyang, TIAN Yingjie. *A New Method of Data Mining – Support Vector Machines*[M]. Beijing: Science Press, 2004.)
- [10] LIN C J. Asymptotic convergence of an SMO algorithm without any assumptions[J]. *IEEE Transactions on Neural Networks*, 2002, 13(1): 248 – 250.
- [11] PALAGI L, SCIANDRONE M. On the convergence of a modified version of SVM light algorithm[J]. *Optimization Methods and Software*, 2005, 20(2/3): 315 – 332.
- [12] CHENG SOON ONG, XAVIER MARY, STEPHANE CANU, et al. Learning with non-positive kernels[C] // *Proceedings of International Conference on Machine Learning*. New York: Association for Computing Machinery, 2004: 639 – 646.
- [13] JAMES T KWOK, IVOR W TSANG. Linear dependency between  $\epsilon$  and the input noise in  $\epsilon$ -support vector regression[J]. *IEEE Transaction on Neural Networks*, 2003, 14(3): 544 – 553.
- [14] WANG S T, ZHU J G, CHUNG F L, et al. Theoretically optimal parameter choices for support vector regression machines Huber-SVR and norm.r-SVR with noisy input[J]. *Soft Computing*, 2005, 9(10): 732 – 741.

## 作者简介:

周晓剑 (1979—), 男, 博士研究生, 主要研究方向为人工智能、智能质量控制, 本文通信作者, E-mail: xjzhou2008@yahoo.com.cn;

马义中 (1964—), 男, 博士生导师, 教授, 主要研究方向为质量工程、质量管理, E-mail: yzma-2004@163.com;

朱嘉钢 (1960—), 男, 博士, 副教授, 主要研究方向为人工智能、模式识别, E-mail: zhujg@jiangnan.edu.cn;

刘利平 (1981—), 男, 博士研究生, 主要研究方向为质量控制, E-mail: lpliu1016@163.com;

汪建均 (1979—), 男, 博士研究生, 主要研究方向为质量工程, E-mail: wjj19770818@163.com.