文章编号: 1001-0920(2006)12-1326-06

基于SVM 的精确数-区间数回归模型建模方法

任世锦, 吴铁军
(浙江大学 智能系统与决策研究所, 杭州 310027)

摘　要: 分析了现有的精确数输入和区间数输出回归算法存在的问题,提出了基于支持向量机的区间数回归建模算法. 该算法把支持向量机从精确数回归分析方法推广到区间数回归分析建模方法,在小样本训练集下回归模型具有良好的泛化性能,有效地避免了现有算法中回归模型的下界可能大于上界的问题. 以连续退火生产过程中冷却段出口带钢温度预测为例,通过仿真说明了该算法的有效性.

关键词: 区间数; 支持向量机; 回归分析; 数据挖掘
中图分类号: TP18　　**文献标识码**: A

SVM Based Algorithm for Regressive Modeling with Accurate Data Input-interval Number Output

REN Shi-jin, WU Tie-jun
(Institute of Intelligent Systems and Decision Making, Zhejiang University, Hangzhou 310027, China. Correspondent: REN Shi-jin, E-mail: sjren@iipc.zju.edu.cn)

Abstract: A SVM based algorithm is proposed to cope with regressive modeling problems with accurate data input-interval number output. A regressive modeling approach based on the support vector machine (SVM) theory is generalized from real number domain to interval number domain. The proposed algorithm can promote the generalization properties of regressive models in the case of small sampling sets and effectively overcomes the shortage of the existing algorithms that the lower limits of model outputs may exceed their upper limits. An example for the prediction of temperature distribution of export strap steel in continuous annealing process demonstrates the effectiveness and efficiency of the proposed method.

Key words: Interval number; Support vector machine(SVM); Regression analysis; Data mining

1 引　　言

工程中常常遇到以下情况: 1) 由于噪声及仪器精度等原因,测量数据只能给出一个近似值及其准确性的范围; 2) 过去和现在的数据可以准确得到,但由于受到一些不确定因素的影响,未来的数据不可能准确预测,只能给出其估计值及其误差限. 对于此类不精确的数据,既可用包含准确值的区间表示(即区间数),也可使用模糊数表示[1, 2]. 但是数据模糊化受主观因素的影响很大,而区间数是不确定数的客观表示,不受主观因素的影响.

实际生产中常常存在自变量相对准确而因变量难以确定的情况. 例如,加热炉内钢锭的温度难以直接测量,只能采用多个非接触式的红外测温仪进行观测,给出钢锭温度所在的可能区间范围; 而炉膛温度的测量精度则要高得多. 炉内钢锭温度对于炉膛温度的依赖关系,可归结为精确数输入和区间数输出的区间数回归建模问题.

目前已出现了一些研究这类区间数回归分析的结果. 文献[2, 3]通过线性规划(LP)求得区间数回归模型参数,但是LP 方法容易受到噪声干扰,并且难以进行非线性回归建模. 文献[4, 5]使用神经网络分别逼近区间数样本的上下界来实现区间数的回

收稿日期: 2005-09-26; **修回日期**: 2006-03-13.
基金项目: 国家973 计划项目(2002CB312203).
作者简介: 任世锦(1971-),男,江苏徐州人,博士生,从事数据挖掘、智能控制等研究; 吴铁军(1950-),男,南京人,教授,博士生导师,从事复杂系统智能控制与决策等研究.

归. 虽然这类方法具有较好的鲁棒性能, 但却存在以下缺点: 1) 由于分别对区间数样本的上下界逼近建模, 可能出现下界回归模型的输出大于上界回归模型输出的错误情况; 2) 模型的泛化性能不佳.

本文针对上述问题, 根据支持向量机(SVM)的思想, 提出一种精确数输入和区间数输出回归模型, 使用一个模型实现精确数-区间数的回归建模, 克服了模型输出的下界可能大于上界的情况, 并且继承了精确数支持向量机回归建模的优点, 在小样本集的情况下使回归模型具有较好的泛化性能和鲁棒性能.

2　区间数及其运算的定义[6, 7]

本节简要介绍区间数的相关知识. 首先定义区间数如下:

定义1(区间数)　给定$\overline{x}\in R$, $\underline{x}\in R$ 且$\underline{x}\le\overline{x}$, 则定义

$$A=\{u\mid \underline{x}\le u\le\overline{x}\}\triangleq[\underline{x},\overline{x}] \tag{1}$$

为一区间数, 称$\overline{x}$和$\underline{x}$分别为区间数A的上界和下界, 记作$A_L=\underline{x}, A_R=\overline{x}$.

定义2(区间数相加)　设两个区间数为$X=[X_L, X_R]$, $Y=[Y_L, Y_R]$, 它们之间的相加定义为

$$X+Y=[X_L+Y_L, X_R+Y_R] \tag{2}$$

定义3(区间数距离)　设两个区间数为$X=[X_L, X_R]$, $Y=[Y_L, Y_R]$, 它们之间的距离定义为

$$d(X, Y)=|X_L-Y_L|+|X_R-Y_R| \tag{3}$$

由定义3可知, 当X和Y均为实数时, 即$X_L=X_R$, $Y_L=Y_R$, 则$d(X, Y)=2|X-Y|$. 可见它与实数的距离定义是等价的, 只是相差一个常系数项.

定义4(精确数与区间数的乘积)　对于区间数$X=[X_L, X_R]$, 它与一个常数c相乘可得

$$cX=\begin{cases}[cX_L, cX_R], & c\ge 0;\\ [cX_R, cX_L], & c<0.\end{cases} \tag{4}$$

定义5(区间数的范数)　对于d维区间数$\mathrm{x}=[x_1, x_2, \cdots, x_d]$, 其中$x_i=[x_{iL}, x_{iR}]$, $i=1, 2, \cdots, d$, 其范数定义为

$$\|\mathrm{x}\|=\left(\sum_{i=1}^{d}\left(|x_{iL}|^2+|x_{iR}|^2\right)\right)^{1/2}. \tag{5}$$

容易验证, 区间数的范数满足正定性、齐次性和三角不等式.

定义6(区间数向量与实数向量的乘积)　设$\mathrm{w}=[w_1, w_2, \cdots, w_d]$为$d$维实数行向量, $\mathrm{X}=[\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_d]^{\mathrm{T}}$为$d$维区间数列向量, 其中$\mathrm{X}_i=[X_{iL}, X_{iR}]$, 则它们的乘积定义为

$$\mathrm{wX}=\sum_{i=1}^{d}w_i\mathrm{X}_i \tag{6}$$

如果$w_i\ge 0$, $i=1, 2, \cdots, d$, 则有

$$\mathrm{wX}=\left[\sum_{i=1}^{d}w_iX_{iL}, \sum_{i=1}^{d}w_iX_{iR}\right]=\left[\langle \mathrm{w}^{\mathrm{T}}, \mathrm{X}_L\rangle, \langle \mathrm{w}^{\mathrm{T}}, \mathrm{X}_R\rangle\right]$$

其中: $\mathrm{X}_L=[\mathrm{X}_{1L}, \cdots, \mathrm{X}_{dL}]^{\mathrm{T}}$, $\mathrm{X}_R=[\mathrm{X}_{1R}, \cdots, \mathrm{X}_{dR}]^{\mathrm{T}}$, $\langle\cdot, \cdot\rangle$为精确数的内积.

3　基于 SVM 的区间数线性回归建模

设有区间数样本(x_i, y_i), 其中$\mathrm{x}_i\in R^d$的列向量, 且经过合适变换可使$x_{ij}\ge 0$, $i=1, 2, \cdots, n$, $j=1, 2, \cdots, d$; 输出$y_i=[y_{iL}, y_{iR}]$为区间数.

设精确数输入和区间数输出的线性回归模型为

$$f(\mathrm{x})=[f^-(\mathrm{x}), f^+(\mathrm{x})]\triangleq \mathrm{w}x+[w_{0L}, w_{0R}]=[w_L\mathrm{x}+w_{0L}, w_R\mathrm{x}+w_{0R}] \tag{7}$$

其中

$$\mathrm{x}=[x_1, x_2, \cdots, x_d]^{\mathrm{T}},$$
$$\mathrm{w}=[w_{1L}, w_{1R}], \cdots, [w_{dL}, w_{dR}],$$
$$\mathrm{w}_L=[w_{1L}, \cdots, w_{dL}], \mathrm{w}_R=[w_{1R}, \cdots, w_{dR}];$$

$f^+(\mathrm{x})$和$f^-(\mathrm{x})$分别为区间数上界回归函数和下界回归函数.

若损失函数取ϵ-不敏感函数

$$|x|_\epsilon=\begin{cases}0, & |x|\le\epsilon\\ |x|-\epsilon, & |x|>\epsilon\end{cases} \tag{8}$$

根据支持向量回归理论[8, 9], 通过求解如下优化问题:

$$\min_{w_L, w_{0L}, \xi_i^*, \xi_i} L=\frac{1}{2}\|\mathrm{w}_L\|^2+C\sum_{i=1}^{n}(\tau\xi_i^*+\xi_i);$$
$$\mathrm{s.t.}\begin{cases}y_{iL}-\mathrm{w}_L\mathrm{x}_i-w_{0L}\le\epsilon+\xi_i,\\ \mathrm{w}_L\mathrm{x}_i+w_{0L}-y_{iL}\le\epsilon+\xi_i^*,\\ \xi_i^*, \xi_i\ge 0, i=1, 2, \cdots, n.\end{cases} \tag{9}$$

可以得到$f^-(\mathrm{x})$. 其中: ξ_i^*和ξ_i是松弛变量, C是罚系数, τ是正数. 优化问题第1项与回归函数集的VC维有关, 涉及到回归模型对于整个数据空间的泛化能力; 第2项表示在训练样本空间的拟合误差.

如果把上面优化问题中的$\mathrm{w}_L, w_{0L}, y_{iL}$换成$\mathrm{w}_R$, w_{0R}, y_{iR}, 便可得到$f^+(\mathrm{x})$. 但是对于给定的样本, 如果按上述方法单独求解上界和下界模型, 则不能保证在整个可能的输入范围内满足$f^-(\mathrm{x})\le f^+(\mathrm{x})$. 为此, 本文将上述优化问题加以综合, 构成一个新的区间数回归损失函数

$$L(y, f(\mathrm{x}))\triangleq d(y, f(\mathrm{x}))+e(f^-(\mathrm{x}), f^+(\mathrm{x})). \tag{10}$$

其中: $d(y, f+(\mathrm{x}))$用于度量区间数回归模型输出与期望值之间的逼近误差, 由定义3的区间数距离计算; $e(f^-(\mathrm{x}), f^+(\mathrm{x}))$用于度量模型输出下界大于

上界的误差, 定义如下:

$$e(f^{-}(x), f^{+}(x)) = \begin{cases} f^{-}(x) - f^{+}(x), f^{-}(x) > f^{+}(x); \\ 0, f^{-}(x) \leq f^{+}(x). \end{cases} \tag{11}$$

为了便于支持向量机建模, 损失函数取 ϵ- 不敏感函数, 使得样本落在 ϵ- 不敏感区间内, 并使回归模型的下界小于上界. 根据样本 (x_i, y_i), 当对应的估计函数为 $\hat{y}_i = f(x_i)$ 时, 则损失函数变为

$$L_\epsilon(y_i, \hat{y}_i) = |y_{iL} - \hat{y}_{iL}|_\epsilon + |y_{iR} - \hat{y}_{iR}|_\epsilon + e_\epsilon(f^{-}(x_i), f^{+}(x_i)). \tag{12}$$

其中

$$e_\epsilon(f^{-}(x_i), f^{+}(x_i)) = \begin{cases} f^{-}(x_i) - f^{+}(x_i) - \epsilon, \\ f^{-}(x_i) - f^{+}(x_i) > \epsilon, \\ 0, f^{-}(x_i) - f^{+}(x_i) \leq \epsilon, \end{cases} \tag{13}$$

$$\hat{y}_{iL} = w_L x_i + w_{0L}, \tag{14}$$

$$\hat{y}_{iR} = w_R x_i + w_{0R}. \tag{15}$$

给定区间数样本 (x_i, y_i), $i = 1, 2, \cdots, n$, 使得区间数回归模型对样本拟合误差最小, 并考虑模型的泛化性能 即要求样本外的数据的误差最小, 并保证模型输出下界小于上界. 于是区间数回归模型可通过下面的优化问题解得:

$$\min_{w_L, w_R, w_0, \xi_{3i}, \xi_{ki}, \xi^{*}_{ki}} \frac{1}{2}\|w_L\|^2 + \frac{1}{2}w_R\|^2 + C_2\sum_{i=1}^{n}(\xi_{3i}) + C_1\sum_{i=1}^{n}(\tau\xi^{*}_{2i} + \xi_{2i}) + C_1\sum_{i=1}^{n}(\xi^{*}_{1i} + \tau\xi_{1i}); \tag{16}$$

$$\text{s.t.} \begin{cases} y_{iL} - w_L x_i - w_{0L} \leq \epsilon + \xi_{1i}, \\ w_L x_i + w_{0L} - y_{iL} \leq \epsilon + \xi^{*}_{1i}, \\ y_{iR} - w_R x_i - w_{0R} \leq \epsilon + \xi_{2i}, \\ w_R x_i + w_{0R} - y_{iR} \leq \epsilon + \xi^{*}_{2i}, \\ w_L x_i + w_{0L} - w_R x_i - w_{0R} \leq \epsilon + \xi_{3i}, \\ \xi_{3i}, \xi_{ki}, \xi^{*}_{ki} \geq 0, i = 1, 2, \cdots, n, k = 1, 2 \end{cases} \tag{17}$$

上述问题中, $\xi_{3i}, \xi_{ki}, \xi^{*}_{ki} \geq 0 (k = 1, 2)$ 为松弛变量; $C_1, C_2 > 0$ 为惩罚系数, 且 $C_1 < C_2$ 表示对回归模型的下界大于上界产生的误差惩罚力度更大; τ 为小于 1 的正数, 表示对上界回归逼近模型超出期望上界的误差, 以及下界回归逼近模型小于期望下界的惩罚力度较小, 使得回归模型尽量包含期望区间数输出的上下界, 并具有一定的抗噪声干扰能力. 这样便从两个方面防止了区间数回归模型出现下界大于上界的不合理情况

式(16) 和(17) 所示的优化问题, 其目标函数由两部分之和组成: 第一部分 $\frac{1}{2}\|w_L\|^2 + \frac{1}{2}\|w_R\|^2$ 与区间数回归函数集的 VC 维有关, 涉及到回归模型对整个数据空间的泛化能力; 第二部分表示回归模型对于给定样本的拟合误差以及回归模型输出下界大于上界的误差 该算法将 SVM 从精确数回归建模推广到区间数回归建模, 并且继承了 SVM 的优点

为求解上述优化问题, 根据目标函数和约束条件构造 Lagrange 函数, 可将其转化为对偶问题

$$\max_{a_{1i}, a^{*}_{1i}, a_{2i}, a^{*}_{2i}, a_{3i}} L = \frac{1}{2}\sum_{i,j=1}^{n}(a^{*}_{1i} - a_{1i} + a_{3i})(a^{*}_{1j} - a_{1j} + a_{3j})\langle x_i, x_j\rangle - \frac{1}{2}\sum_{i,j=1}^{n}(a^{*}_{2i} - a_{2i} - a_{3i}) \times (a^{*}_{2j} - a_{2j} - a_{3j})\langle x_i, x_j\rangle - \epsilon\sum_{i=1}^{n}a_{3i} - \epsilon\sum_{k=1}^{2}\sum_{i=1}^{n}(a_{ki} + a^{*}_{ki}) - \sum_{i=1}^{n}(a_{1i} - a^{*}_{1i})y_{iL} - \sum_{i=1}^{n}(a_{2i} - a^{*}_{2i})y_{iR}; \tag{18}$$

$$\text{s.t.} \begin{cases} \sum_{i=1}^{n}(a^{*}_{1i} - a_{1i} + a^{*}_{3i}) = \\ \sum_{i=1}^{n}(a^{*}_{2i} - a_{2i} - a_{3i}) = 0, \\ a^{*}_{1i}, a_{2i} \in [0, C_1], a^{*}_{2i}, a_{1i} \in [0, \tau C_1], \\ a_{3i} \in [0, C_2], i = 1, 2, \cdots, n. \end{cases} \tag{19}$$

其中: $a_{1i}, a_{2i}, a^{*}_{1i}, a^{*}_{2i}, a_{3i}$ 是 Lagrange 乘子, $i = 1, 2, \cdots, n$.

设 $\alpha = [a_1^{\mathrm{T}}, a_1^{*\mathrm{T}}, a_2^{\mathrm{T}}, a_2^{*\mathrm{T}}, a_3^{\mathrm{T}}]^{\mathrm{T}}$, e 为 $n \times 1$ 的全 1 列向量, $H = [\langle x_i, x_j\rangle]_{n\times n}$, $y_R = [y_{iR}]_{n\times 1}$, $y_L = [y_{iL}]_{n\times 1}$. 其中: $a_1 = [a_{1i}]_{n\times 1}$, $a_1^{*} = [a^{*}_{1i}]_{n\times 1}$, $a_2 = [a_{2i}]_{n\times 1}$, $a_2^{*} = [a^{*}_{2i}]_{n\times 1}$, $a_3 = [a_{3i}]_{n\times 1}$, $i = 1, 2, \cdots, n$. 于是, 优化问题(18) 和(19) 可以写成

$$\min_{\alpha} q(\alpha) + c\alpha; \quad \text{s.t.} \begin{cases} A\alpha = 0, \\ 0 \leq \alpha \leq C. \end{cases} \tag{20}$$

其中

$$c = [y_L^{\mathrm{T}} + \epsilon, -y_L^{\mathrm{T}} + \epsilon, y_R^{\mathrm{T}} + \epsilon, -y_R^{\mathrm{T}} + \epsilon, \epsilon],$$

$$A = \begin{bmatrix} -e & e & 0 & 0 & e \\ 0 & 0 & -e & e & -e \end{bmatrix},$$

$$B = \begin{bmatrix} H & -H & 0 & 0 & -H \\ -H & H & 0 & 0 & H \\ 0 & 0 & H & -H & H \\ 0 & 0 & -H & H & -H \\ -H & H & H & -H & 2H \end{bmatrix},$$

$C = [e^T\tau C_1, e^T C_1, e^T C_1, e^T\tau C_1, e^T C_2]^T$,

$q(\alpha) = 1/2\alpha^T B\alpha$

可以看出, 式(20) 所示的优化问题是一个凸二次优化问题 对于 n 个区间数的训练样本, 需要通过上述优化问题求出 $5n$ 个解, 并且矩阵 B 需要占用的内存空间较大 对于大规模样本的回归建模问题, 必须寻找一种高效的求解算法

块方法[10] 是通过逐步扩大样本数量进行建模的一种大规模样本回归建模方法 该方法首先使用较小样本子集对回归模型进行建模; 然后把剩余子集中不满足KKT 条件的样本加入到支持向量集中, 重新建模并不断迭代, 直到所有子集的样本满足KKT 条件: 这种方法能大大降低对内存的需求, 但是每次重新训练模型的速度较慢

内点法[10] 是求解式(20) 的凸二次规划问题的一种高效算法 该算法首先对式(20) 的优化问题增加松弛因子, 把约束条件中的不等式变为等式, 得到其对偶问题; 然后寻找对原问题和对偶问题均可行且满足KKT 条件的一个变量集合, 迭代收敛到可行解中, 直至原函数与对偶目标函数之差小于设定的阈值

将块方法与内点法结合起来, 可以适应大规模样本的回归建模 其依据是: 模型中的支持向量个数远远小于训练样本集的数量, 每个训练子集所含潜在的支持向量也很少, 所以利用少量潜在的支持向量并用内点法来重新训练模型, 不仅能提高运算速度, 而且可降低内存的使用 算法的要点是在块算法的每次迭代中, 使用内点法求取局部区间数回归函数, 这样可充分发挥块方法和内点法各自的优点

在求解问题(20) 的基础上, 可求出回归函数中的偏置量 w_{0L} 和 w_{0R}. 对于 w_{0L} 而言, 根据KKT 条件可得

$y_{iL} - w_L x_i + \epsilon, a_{1i} \in (0, \tau C_1), a_{1i}^* = 0, a_{3i} = 0;$

$y_{iL} - w_L x_i - \epsilon, a_{1i}^* \in (0, C_1), a_{1i} = 0, a_{3i} = 0;$

$w_R x_i + w_{0R} - w_L x_i - \epsilon, a_{3i} \in (0, C_2), a_{1i} = 0, a_{1i}^* = 0;$

$y_{iL} - w_L x_i, a_{1i}^* \in (0, C_1), a_{1i} \in (0, \tau C_1), a_{3i} = 0;$

$(y_{iL} - 2w_L x_i - 2\epsilon + w_R x_i + w_{0R})/2, a_{1i}^* \in (0, C_1), a_{3i} \in (0, C_2), a_{1i} = 0;$

$(y_{iL} - 2w_L x_i + w_R x_i + w_{0R})/2, a_{1i} \in (0, \tau C_1), a_{3i} \in (0, C_2), a_{1i}^* = 0;$

$(2y_{iL} - 3w_L x_i + w_R x_i + w_{0R} - \epsilon)/3, a_{1i}^* \in (0, C_1), a_{1i} \in (0, \tau C_1), a_{3i} \in (0, C_2).$

其中

$$w_L^T = \sum_{i=1}^{n} (a_{1i}^* - a_{1i} + a_{3i}) x_i,$$

$$w_R^T = \sum_{i=1}^{n} (a_{2i}^* - a_{2i} - a_{3i}) x_i.$$

使用类似的方法可求出 w_{0R}, 在此不再列出.

如上所述, 区间数线性回归函数可表示为

$$f^-(x) = \sum_{i=1}^{n} (a_{1i}^* - a_{1i} + a_{3i}) \langle x_i, x \rangle + w_{0L}, \tag{21}$$

$$f^+(x) = \sum_{i=1}^{n} (a_{2i}^* - a_{2i} - a_{3i}) \langle x_i, x \rangle + w_{0R}. \tag{22}$$

4　基于核函数的区间数非线性回归模型

由支持向量机理论可知, 对于实数域的非线性回归问题, 可通过映射函数 $\varphi: R^d \to F$, 把输入样本映射到高维特征空间, 使其可线性回归 映射函数的内积运算可转换为核函数的形式, 即对于实数向量 x 和 y, 有

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle.$$

实际上, 只需获得核函数而无需知道 $\varphi(\cdot)$ 的具体形式, 便可在原样本空间构造非线性回归模型

这一方法原则上可推广到区间数非线性回归问题 即对于输入为精确数、输出为区间数的数据样本 (x_i, y_i), $i = 1, 2, \cdots, n$, 寻找映射函数 $z = \psi(x)$, 使得在转换后的高维精确数特征空间中, 样本数据 (z_i, y_i) 是可线性回归的, 其中 $z_i = \psi(x_i)$. 根据式(21) 和(22), 在特征空间的区间数回归模型应有如下形式:

$$f^-(x) = \sum_{i=1}^{n} (a_{1i}^* - a_{1i} + a_{3i}) K(x_i, x) + w_{0L}, \tag{23}$$

$$f^+(x) = \sum_{i=1}^{n} (a_{2i}^* - a_{2i} - a_{3i}) K(x_i, x) + w_{0R}. \tag{24}$$

其中: w_{0L} 和 w_{0R} 的求法与第 3 节相似, 在此不再具体写出; 核函数 K 可取常用的满足Mercer 条件的高斯核函数、多项式核函数等[8]; $a_{1i}, a_{1i}^*, a_{2i}, a_{2i}^*, a_{3i}$ 是Lagrange 乘子, 可通过下面的对偶优化问题求得:

$$\max_{a_{1i}, a_{1i}^*, a_{2i}, a_{2i}^*, a_{3i}} L =$$

$$\frac{1}{2}\sum_{i,j=1}^{n} (a_{1i}^* - a_{1i} + a_{3i})(a_{1j}^* - a_{1j} + a_{3j}) K(x_i, x_j) - \frac{1}{2}\sum_{i,j=1}^{n} (a_{2i}^* - a_{2i} - a_{3i})(a_{2j}^* - a_{2j} - a_{3j}) K(x_i, x_j) - \epsilon\sum_{i=1}^{n} a_{3i} - \epsilon\sum_{k=1}^{2}\sum_{i=1}^{n} (a_{ki} + a_{ki}^*) - \sum_{i=1}^{n} (a_{1i} - a_{1i}^*) y_{iL} - \sum_{i=1}^{n} (a_{2i} - a_{2i}^*) y_{iR};$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{n}(a_{1i}^{*}-a_{1i}+a_{3i})=\\ \sum_{i=1}^{n}(a_{2i}^{*}-a_{2i}-a_{3i})=0,\\ a_{1i}^{*},a_{2i}\in[0,C_1],a_{2i}^{*},a_{1i}\in[0,\tau C_1],\\ a_{3i}\in[0,C_2],i=1,2,\cdots,n.\end{cases}\tag{25}$$

(26)

上述优化问题的求解可参见第3节关于对偶优化问题的讨论. 令$H=[K(x_i,x_j)]_{n\times n}$], 可获得类似的算法, 在此不再赘述.

5 仿真实例

在冷连轧带钢的连续退火生产线中, 由于冷却段传热方式多, 干扰因素多, 采用非接触高温辐射计测量带钢温度时, 难以得到精确的温度值. 为此把温度值转变成区间数, 以表示可能的温度范围. 令T_{out}表示冷却段带钢的出口温度分布范围, 它是反映带钢质量的最重要指标. T_{out}的区间长度越小, 表示板材温度越均匀, 冷却的效果越好; 反之则表示温度分布不均匀, 板材的质量较差.

在假定钢材的规格和型号不变、均热段板材加热均匀的情况下, 根据理论分析和经验总结[11, 12], 发现T_{out}与以下过程变量有关: 1) 可直接测量的物理量, 包括冷却水温度、机组速度、喷吹温度、板材与冷却辊之间的接触角、后侧喷吹速度、后侧喷吹流量、冷却气体压力、张力和带钢厚度; 2) 生产过程变量的统计量, 包括冷却水平均流量、对侧喷吹平均速度、对侧喷吹平均流量、对侧喷吹流量方差. 这些变量都能准确测量或者准确统计得到. 这样便可将上述变量作为精确数输入, 将板材温度分布范围作为区间数输出, 建立与板材温度有关的回归模型.

从某厂1 550 mm连续退火生产线取关于某型号带钢的退火数据, 以间隔12 min挑选上述变量数据并生成300个样本, 其中前200个样本作为回归建模样本, 后100个样本作为测试样本. 使用本文算法进行回归建模和误差分析, 本例中的核函数取

$$k(x,x')=\exp(-\|x-x'\|^2/2\sigma^2),$$

罚系数C_1和C_2简单地取$C_2=4C_1$, 利用交叉验证的方法确定模型参数σ和C_1[9]. 模型对测试样本的预测结果如图1和图2所示. 通过对预测输出的上下界误差统计发现, 预测误差绝大部分在-5到+5之间, 与实际冷却段出口处带钢温度分布基本吻合, 能为生产人员分析带钢质量提供有价值的参考意见.

为与基于RBF网络的区间数回归方法进行比较, 使用文献[13]提出的平均绝对误差(AAE)、平均相对误差(ARE)和均方误差(MSE)作为预测精度指标, 把测试样本预测输出上下界的预测精度和训练样本输出上下界的训练精度作为预测精度和训练精度, 比较结果如表1所示.

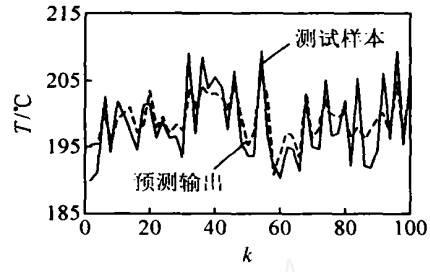

图1 测试样本的下界预测

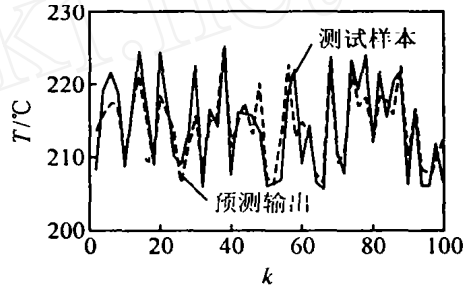

图2 测试样本的上界预测

表1 训练样本训练精度和测试样本预测精度

考察方法	样 本	AAE	ARE/%	MSE
本文方法	训练样本	3.152 7	0.763 8	14.091 7
	测试样本	1.786 5	1.807 1	10.644 6
基于RBF网络方法	训练样本	3.208 4	0.836 2	16.340 9
	测试样本	3.437 0	2.583 5	15.182 7

从表1可以看出, 本文方法与基于RBF网络回归方法在训练精度上比较接近, 但是预测精度却明

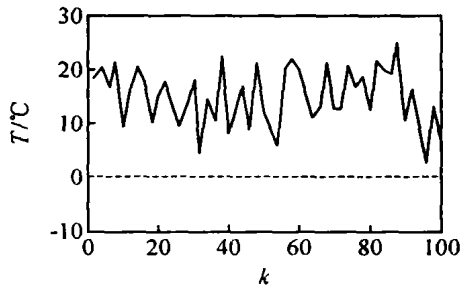

图3 本文方法

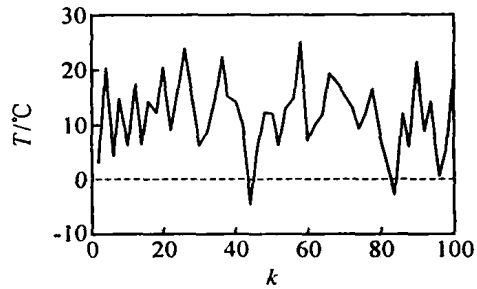

图4 基于RBF网络方法

显提高, 说明本文提出的区间数回归模型具有较好的泛化性能

为了研究是否会出现区间数回归模型输出的下界大于上界的情况, 把本文方法与基于RBF 网络方法的预测输出的上界减去下界, 其差值分别如图3 和图4 所示 从图中可以看出, 本文方法没有出现预测结果上界小于下界的现象, 而基于RBF 网络方法却出现了, 说明本文方法能有效避免回归输出区间数的下界大于上界的现象

6　结　　论

本文提出一种基于SVM 的精确数输入/区间数输出回归模型建模方法 该方法对区间数上下界回归模型共同建模, 克服了现有方法不能保证模型输出上界不小于下界的缺陷, 使回归模型具有良好的泛化性能, 并且容易从线性回归模型推广到非线性回归模型 文中研究了大规模样本下的模型求解算法, 以适应实际需要, 并通过仿真实例说明了本文方法的有效性

参考文献(References)

[1] Ferson S, Akcakaya H R, Dunham A. Using Fuzzy Intervals to Represent Measurement Error and Scientific Uncertainty in Endangered Species Classification [A]. *Fuzzy Information Processing Society* [C]. Nafips, 1999: 690-694.

[2] Lee Haekwan, Tanaka Hideo. Upper and Lower Approximation Models in Interval Regression Using Regression Quantile Techniques[J]. *European J of Operational Research*, 1999, 116(3): 652-666.

[3] Ishibuchi H, Tanaka H. Fuzzy Regression Analysis Using Neural Networks [J]. *Fuzzy Sets and Systems*, 1992, 50(3): 57-65.

[4] Huang L, Zhang B L, Huang Q, et al. Robust Interval Regression Analysis Using Neural Networks[J]. *Fuzzy Sets and Systems*, 1998, 97(2): 337-347.

[5] Jin-tsong Jeng, Chen-chia Chuang, Shun-feng Su. Support Vector Interval Regression Networks for Interval Regression Analysis[J]. *Fuzzy Sets and Systems*, 2003, 138(2): 283-300.

[6] Moore R E. *Interval Analysis*[M]. New Jersey: Prentice Hall, 1966.

[7] Lai K K, Wang Y, Xu J P. A Class of Linear Interval Programming Problems and Its Application to Portfolio Selection [J]. *IEEE Trans on Fuzzy Systems*, 2002, 10(6): 698-703.

[8] 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, 2001.
(Vapnik V N. *The Nature of Statistical Learning* [M]. Translated by Zhang X G. Beijing: Tsinghua University Press, 2001.)

[9] Vladimir C, Ma Y Q. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression[J]. *Neural Networks*, 2004, 17(1): 113-126.

[10] Smola A J. *Learning with Kernels*[D]. Berlin: Technischen Universitsitat Berlin, 1998.

[11] 潘勋平, 杨杰. 连续退火技术机组辊冷技术及板温控制[J]. ***宝钢技术***, 2001, 18(5): 39-44.
(Pan X P, Yang J. Roll Quench Technology and Strip Temperature Control in CAL [J]. *Baosteel Technology*, 2001, 18(5): 39-44.)

[12] 杨进. ***带钢连续热镀锌退火技术及卧式连续退火炉数学模型研究***[D]. 北京: 北京科技大学, 2002.
(Yang J. *Study of Continuous Annealing Technology of Steel Strip and Mathematical Model of Horizontal Furnace* [D]. Beijing: Beijing University of Science and Technology, 2002.)

[13] Hao X F, Xu D. Time Series Prediction Based on Non-parametric Regression and Wavelet-fractal [A]. 2004 7*th Int Conf on Signal Processing Proc*[C]. Australia, 2004: 386-389.

(上接第1325 页)

[22] Willems J C. Dissipative Dynamical Systems — Part 2: Linear System with Quadratic Supply Rates [J]. *Archive for Rational Mechanics and Analysis*, 1972, 45(4): 352-393.

[23] Trentelman H L, Willems J C. Every Storage Function is a State Function [J]. *System Control Letter*, 1997, 32(3): 249-259.

[24] Doyle J C, Glover K, Khargonekar P, et al. State Space Solution to Standard H_2 and H_∞ Control Problems [J]. *IEEE Trans on Automatic Control*, 1989, 34(8): 831-847.