

研究论文

# 多尺度小波核支持向量回归及其对 丙烯浓度的估计与应用

余艳芳, 杜文莉, 钱 锋

(华东理工大学化工过程先进控制和优化技术教育部重点实验室, 上海 200237)

**摘要:** 多尺度小波核支持向量回归方法具有较强的鲁棒性和较好的泛化能力, 而模型选择是其获得良好泛化性能的关键, 其中采用多尺度核方法参数选择的复杂度比单个核方法的参数选择大得多。这里提出了一种构造多尺度 Morlet 小波核的支持向量回归机的方法, 它采用量子聚类方法划分样本类别以确定多尺度核的尺度个数, 依赖支持向量数据描述方法对相应样本计算其核宽度, 然后用文化算法优化剩下的少量模型参数。结果表明所得到的多尺度小波核模型的泛化能力明显优于单小波核或高斯核情形。分别用 3 个标准回归数据集 Boston housing、Bodyfat 和 Santa 作仿真, 结果表明, 相对于高斯核方法, 多尺度小波核支持向量回归方法的测试集均方差分别减小了 6.8%、62.0% 和 91.3%。同时, 该方法对丙烯精馏塔的塔釜丙烯浓度预估表现出较好的泛化能力。它不仅使丙烯浓度训练集模型输出与实际输出基本吻合, 而且使丙烯浓度测试集相对误差为 0.211, 与其他方法相比, 其预测误差是最小的。

**关键词:** 多尺度小波核; 量子聚类; 支持向量数据描述; 文化算法; 模型选择

**中图分类号:** TP 18

**文献标识码:** A

**文章编号:** 0438—1157 (2010) 06—1486—09

## Support vector regression based on multi-scale wavelet kernel for propylene concentration estimation and application

YU Yanfang, DU Wenli, QIAN Feng

(Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education,  
East China University of Science and Technology, Shanghai 200237, China)

**Abstract:** Support vector regression based on multi-scale wavelet kernel has strong robustness and good generalization ability, but it is critical for it to choose appropriate model parameters. Obviously, the multi-scale kernel method has more difficulty in model selection than the single kernel approach. This paper proposed an approach about how to develop support vector regression based on multi-scale wavelet kernel. It applied quantum clustering to data partition in order to determine the scale parameter of the multi-scale kernel, resorted to the support vector data description algorithm to calculate the kernel width of the corresponding data points, and then used cultural algorithms to optimize the kernel width and the remaining parameters. The results showed that the multi-scale kernel method outperformed the single wavelet

2009—10—26 收到初稿, 2010—03—24 收到修改稿。

**联系人:** 杜文莉, 钱锋。第一作者: 余艳芳 (1979—), 女, 博士研究生。

**基金项目:** 国家杰出青年科学基金项目 (60625302); 国家重点基础研究发展计划项目 (2009CB320603); 国家自然科学基金项目 (60804029); 长江学者和创新团队发展计划项目 (IRT0721)。

**Received date:** 2009—10—26.

**Corresponding author:** DU Wenli, wldu @ ecust.edu.cn;  
QIAN Feng, fqian@ecust.edu.cn

**Foundation item:** supported by the National Science Fund for Distinguished Young Scholars (60625302), the National Basic Research Program of China (2009CB320603), the National Natural Science Foundation of China (60804029) and the Program for Changjiang Scholars and Innovative Research Team in University (IRT0721).

approach and the Gaussian method. The experiments about three regression data sets—Boston housing, Bodyfat and Santa demonstrated that in contrast with the Gaussian approach, the present multi-scale wavelet support vector regression made the mean squared error of test sets decrease by 6.8%, 62.0% and 91.3%, respectively. Meanwhile, the proposed approach exhibited good generalization ability for propylene concentration estimation in the bottom byproduct of propylene fractionation tower. It not only enabled the model output of training set for propylene concentration to show little difference with the actual output, but also made the relative error of test set down to 0.211. Compared with other methods, it had the minimal prediction error.

**Key words:** multi-scale wavelet kernel; quantum clustering; support vector data description; cultural algorithms; model selection

## 引 言

在广受关注的支持向量机算法中，高斯核是个普适的核函数，但是对于复杂多峰的函数拟合或实际的工业过程建模，其模型往往难以呈现出良好的泛化性能。而构造多尺度核可以有效地建立精确的数学模型。如 Zheng 等<sup>[1]</sup> 和 Pozdnoukhov 等<sup>[2]</sup> 均采用多个不同核宽度的高斯径向基函数的线性组合构成多尺度核，前者用之解决复杂多峰的函数估计问题，后者用于突发的放射性尘埃的检测与建模。段崇雯<sup>[3]</sup> 专门研究了多尺度核函数支持向量机算法，并分别用 Harr 函数、二次 B 样条函数和三次 B 样条函数构造多尺度核。王鹏等<sup>[4]</sup> 提出了多尺度 Morlet 小波核，并与最小二乘支持向量机结合对红外弱小目标图像进行灰度曲面拟合。相对于高斯核的支持向量回归 (support vector regression, SVR)<sup>[5]</sup> 方法，小波核 SVR 方法表现出较好的泛化性能。如 Zhang 等<sup>[6]</sup> 不仅提出了小波核 SVR，而且通过两个复杂数学函数仿真来验证 Morlet 小波核 SVR 比高斯核 SVR 具有较小的泛化误差。Wen 等<sup>[7]</sup> 构造了 Mexico 帽小波核最小二乘支持向量机，并与高斯核性能比较得知 Mexico 帽小波核泛化精度略优于高斯核，同时也指出小波核 SVR 具有较强的抗噪声能力。综上所述，多尺度小波核 SVR 可以进一步提高算法的泛化能力和增强其鲁棒性。到目前为止，对多尺度核的研究多限于其核形式的变化，甚少涉及多尺度核的尺度个数确定。尺度个数越多，模型越复杂，多尺度核方法未必能成为解决回归估计问题的有效途径；反之，尺度个数越少，多尺度核方法也未必能有效改善算法的推广能力。另外，与单尺度核相比，多尺度核函数无

疑会增加参数选择的个数，进一步扩大参数搜索的范围，降低了学习效率。

针对上述问题，本文提出了一种构造多尺度 Morlet 小波核的线性规划支持向量回归方法，它依据量子聚类<sup>[8]</sup> 算法选择多尺度小波核的尺度个数，用支持向量数据描述<sup>[9]</sup> (support vector data description, SVDD) 方法确定小波核参数，并运用文化算法<sup>[10-11]</sup> (cultural algorithm, CA) 搜索剩余的模型参数。相对于标准支持向量回归机，线性规划 SVR 具有支持向量稀疏和计算过程简单的特点，本文采用线性规划  $v$ -支持向量回归 ( $v$ -LPSVR)<sup>[12-13]</sup> 来估计回归问题。该方法有效确定了多尺度核的尺度个数及相关参数，通过对 Boston housing、Bodyfat 和 Santa 3 个数据集的仿真表明，与高斯核方法相比，该方法的测试集均方差 (MSE) 分别减小了 6.8%、62.0% 和 91.3%。与此同时，本文方法对实际工业丙烯精馏塔的丙烯浓度作出了较为准确的估计。尽管由于受操作条件等因素影响，采集到很多不合理的丙烯数据，然而本文方法仍使丙烯浓度训练集和测试集相对误差 (RE) 分别为 0.209 和 0.211，比常用于工业过程建模的 RBF 神经网络方法的测试集相对误差减少了 23.6%，表现出较强的鲁棒性。

## 1 多尺度小波核

可生成小波框架的母小波都可用来构造小波核函数<sup>[14]</sup>，部分小波核函数的形式有：一维 Harr 小波、Mexico 帽小波、Morlet 小波、双正交 B 样条小波等，这里及下面提到的小波核函数都采用文献中使用较多的 Morlet 小波，即

$$K(x, x') = \prod_{i=1}^d \left[ \cos\left(\frac{1.75(x_i - x'_i)}{a}\right) \exp\left(-\frac{\|x_i - x'_i\|^2}{2a^2}\right) \right] \quad (1)$$

式中  $d$  为样本数据  $x \in (R^d)$  的维数,  $a \in R$  为伸缩因子。根据核函数的封闭性可知, 多个核函数的叠加仍然是核函数。因此, 多尺度小波核形式如下

$$K(x, x') = \sum_{j=1}^L \prod_{i=1}^d \left[ \cos\left(\frac{1.75(x_i - x'_i)}{a_j}\right) \exp\left(-\frac{\|x_i - x'_i\|^2}{2a_j^2}\right) \right] \quad (2)$$

式中  $L$  为多尺度核的尺度个数。

### 1.1 尺度个数的确定

多尺度核是从多个尺度上来描述核函数。为此作者从样本分布特点出发, 按距离远近分类, 将每类样本用某一核函数表示, 从而构成多尺度核。这里采用量子聚类方法划分原始样本, 所得类别数为尺度个数。在量子聚类算法中, 唯一的一个自由参数  $\sigma$  的取值是聚类的关键。到目前为止, 还没有合适的理论确定  $\sigma$  的大小。为了减少算法对  $\sigma$  的依赖, 本文引入惰性阈值  $\mu$  作聚类划分, 下面简单介绍量子聚类的思想。

薛定谔于 1926 年提出在势场  $V$  中运动的粒子, 其波函数满足薛定谔方程 (Schrödinger equation)

$$H\psi = \left(-\frac{\sigma^2}{2}\nabla^2 + V(x)\right)\psi(x) = E\psi(x) \quad (3)$$

其中,  $H$  为 Hamiltonian 算子,  $\psi(x)$  为波函数,  $V(x)$  为粒子的势能函数,  $E$  为  $H$  算子的能量特征值,  $\sigma$  是方程中唯一的自由参数,  $\nabla$  是偏微分算子。波函数可由尺度为  $\sigma$  的高斯函数构成的 Parzen 窗函数来估计, 即

$$\psi(x) = \sum_{i=1}^{\ell} e^{-(x-x_i)^2/2\sigma^2} \quad (4)$$

则按照式 (4) 即可求解势能函数  $V(x)$

$$V(x) = E + \frac{\sigma^2 \nabla^2 \psi}{2\psi} \quad (5)$$

势能  $V(x)$  是非负函数。假设最小的势能  $V=0$ ,  $E$  可以根据式 (5) 求解得

$$E = -\min \frac{\sigma^2 \nabla^2 \psi}{2\psi} \quad (6)$$

为寻求最小势能作为聚类中心, 采用式 (7) 梯度下降公式进行迭代计算

$$y_i(k+1) = y_i(k) - \eta(k) \nabla V(y_i(k)) \quad (7)$$

式中  $\eta(k)$  为学习速率,  $\nabla V$  表示势能  $V(x)$  的梯度,  $k$  为梯度下降的迭代次数, 并设初始点变量  $y_i(0)=x$ , 使  $y_i(k)$  最终到达一个固定值逐渐逼近某个类中心点。

定义一个惰性阈值  $\mu(t)$  来划分上述量子聚类算法计算出的数据矩阵以获得样本数据的类别划分。

$$\mu(t) = \max_{1 \leq j \leq \ell} D(x, x_j) - (\max_{1 \leq j \leq \ell} D(x, x_j) - \min_{1 \leq j \leq \ell} D(x, x_j))t/T \quad (8)$$

式中  $D(x, x_j)$  表示样本  $x$  与  $x_j$  之间的欧氏距离,  $\ell$  为原始样本个数,  $t$  为当前迭代次数,  $T$  为事先定义的最大迭代次数。划分的基本思想是: 选取一个未划分样本  $x$ , 计算其与所有其他样本之间的距离  $D(x, x_j)$ , 将距离小于  $\mu(t)$  的样本与样本  $x$  归为一类。继续取下一个未划分样本重复上述操作。如果同一个样本被归属于多个类别, 则保留距离最小的那个类别标签。每个阈值都对应一种聚类划分, 然后根据 DB Index 准则<sup>[15]</sup>确定最优聚类划分结果。DB Index 准则使用了类内离散度和类间距离, 其具体内容如下。

#### (1) 类内离散度

$$S_i = \frac{1}{N_{C_i}} \sum_{x \in C_i} \|x - z_i\| \quad (9)$$

其中,  $z_i$  是  $C_i$  类的类中心,  $N_{C_i}$  表示  $C_i$  类样本个数。

#### (2) 类间距离

$$d_{ij} = \|z_i - z_j\| \quad (10)$$

即用两个类中心的距离表示类间距离。

#### (3) DB Index

$$DB_k = \frac{1}{k} \sum_{i=1}^k R_i \quad (11)$$

其中,  $R_i = \max_{j=1, \dots, k, j \neq i} \frac{S_i + S_j}{d_{ij}}$ ,  $k$  是分类数目。

DB Index 准则:  $DB_k$  值越小, 分类效果就越好。基于量子聚类算法的样本分类流程图如图 1。

### 1.2 小波核宽度的确定

与单小波核一样, 多尺度小波核的核宽度并不是任意选择就能达到满意的回归精度。这里采用支持向量数据描述方法计算所包含样本的最小直径, 并把这个直径定义为这些样本的核宽度。每个尺度上的核宽度都只与该尺度对应的聚类样本相关。通过下列对偶式求解最小直径

$$\max \sum_i \alpha_i k(x_i, x_i) - \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \quad 0 \leq \alpha_i \leq C \quad (12)$$

式中  $C$  为惩罚参数, 核函数  $k(x_i, x_j)$  选择高斯核函数,  $\alpha$  为拉格朗日乘子。

$$k(x_i, x_j) = \exp(-\|x_i - x_j\|^2/s^2) \quad (13)$$

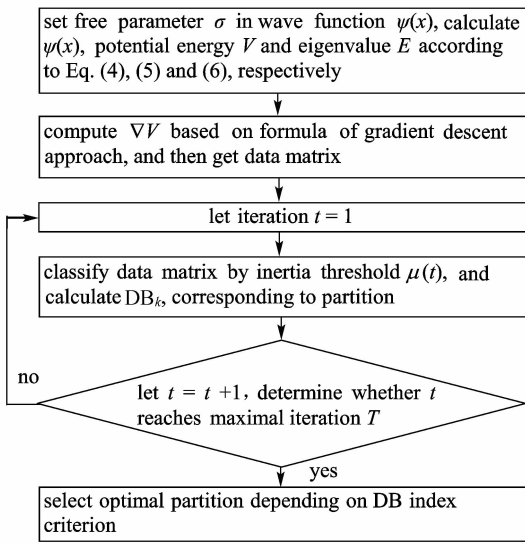


图1 基于量子聚类算法的样本分类流程图

Fig. 1 Flow chart for example classification based on quantum clustering algorithm

计算最小直径  $D$

$$D = \frac{2}{N_{sv}} \sum_i \left\| x_i - \sum_j \alpha_j x_j \right\|^2 \quad (14)$$

其中,  $N_{sv}$  为支持向量个数。

高斯核宽度  $s$  越大, 支持向量个数就越小; 惩罚参数  $C$  越小, 支持向量个数越多。因此选择恰当的参数  $C$  和  $s$  是至关重要的。这里采用 1.3 节介绍的文化算法来搜索最优参数  $C$  和  $s$ , 以直径最小为优化目标, 优化得到的最小直径  $D$  即为该尺度上所包含样本的小波核宽度。

### 1.3 文化算法优化其他参数

线性规划  $v$ -支持向量回归机对应的最优化目标问题的对偶式<sup>[13]</sup>如下

$$\begin{aligned} \min \quad & \frac{1}{\ell} \sum_{i=1}^{\ell} (\alpha_i + \alpha_i^*) + C^* v \varepsilon + \frac{C^*}{\ell} \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \\ \text{subject to} \quad & \sum_{j=1}^{\ell} (\alpha_j - \alpha_j^*) K(x_j, x_i) + b - y_i \leq \varepsilon + \xi_i \\ & y_i - \sum_{j=1}^{\ell} (\alpha_j - \alpha_j^*) K(x_j, x_i) - b \leq \varepsilon + \xi_i^* \\ & \alpha_i, \alpha_i^*, \xi_i, \xi_i^*, \varepsilon \geq 0 \end{aligned} \quad (15)$$

式中 核函数  $K(x_i, x_j)$  为上文定义的多尺度小波核, 尺度个数及小波核宽度的确定如 1.1 节和 1.2 节所述, 惩罚参数  $C^*$  和控制参数  $v$  需作优化选择。这里采用文化算法搜索最优参数。文化算法通过种群空间和信念空间的双重进化与通信以获取问题的最优解, 即首先在种群空间个体按规则进化, 然后根据接受函数 accept() 提取最优秀的一些个体信息, 更新信念空间, 再通过影响函数 influence() 将信念空间的知识指导种群空间个体

的变化趋势或范围。

#### (1) 种群空间的进化

这里运用经典的遗传算法作为种群空间的进化算法, 定义个体并编码, 确定种群规模和迭代次数, 给出初始解, 计算适应度函数和评价各个体的适应值。

#### (2) 信念空间的定义

本文将信念空间结构定义为  $\langle N[j], C[k] \rangle$ 。  $N[j]$  用来存储信念空间最能产生优秀个体的区域 (称为优胜区域) 的信息, 包括  $\langle I_j, L_j, U_j \rangle$  3 个参数,  $j=1, 2, \dots, n$ ,  $n$  表示自变量的个数, 其中参数  $I_j = [l_j, u_j] = \{x_{i,j} \mid l_j \leq x_{i,j} \leq u_j, x_{i,j} \in R\}$ , 表示第  $j$  个自变量的优胜区域, 其中  $x_{i,j}$  表示第  $i$  个个体的第  $j$  个自变量,  $L_j$  和  $U_j$  分别表示该自变量的优胜区域的下限和上限对应的适应值。  $C[k]$  表示第  $k$  个信念细胞, 包含  $R_{k,j} = [l_{k,j}, u_{k,j}] = \{x_{i,j} \mid l_{k,j} \leq x_{i,j} \leq u_{k,j}, x_i \in C[k]\}$ ,  $k=1, 2, \dots, m$ ,  $m$  表示信念细胞的个数,  $j=1, 2, \dots, n$ ,  $l_{k,j}$  和  $u_{k,j}$  表示第  $k$  个信念细胞的第  $j$  个自变量的优胜区域的下限和上限。

#### (3) 信念空间的更新

假设信念空间中第  $i$  个优秀个体作用于第  $j$  个自变量的下限, 第  $i$  个优秀个体作用于第  $j$  个自变量的上限, 则更新规则如下

$$\begin{aligned} l_j^{t+1} &= \begin{cases} x_{i,j}^t, & x_{i,j}^t \leq l_j^t \quad \text{或} \quad \text{objective}(x_i^t) < L_j^t \\ l_j^t, & \text{其他} \end{cases} \\ L_j^{t+1} &= \begin{cases} \text{objective}(x_i^t), & x_{i,j}^t \leq l_j^t \quad \text{或} \quad \text{objective}(x_i^t) < L_j^t \\ L_j^t, & \text{其他} \end{cases} \\ u_j^{t+1} &= \begin{cases} x_{i,j}^t, & x_{i,j}^t \geq u_j^t \quad \text{或} \quad \text{objective}(x_i^t) < U_j^t \\ u_j^t, & \text{其他} \end{cases} \\ U_j^{t+1} &= \begin{cases} \text{objective}(x_i^t), & x_{i,j}^t \geq u_j^t \quad \text{或} \quad \text{objective}(x_i^t) < U_j^t \\ U_j^t, & \text{其他} \end{cases} \end{aligned} \quad (16)$$

其中,  $l_j^t$  和  $u_j^t$  表示信念空间中第  $t$  次迭代第  $j$  个自变量的下限和上限,  $L_j^t$  和  $U_j^t$  分别表示与其对应目标函数的适应值,  $x_{i,j}^t$  表示第  $t$  次迭代第  $i$  个个体的第  $j$  个自变量。

#### (4) 接受函数 accept()

接受函数用于选择一部分最优秀个体, 以修改信念空间的信息。这里根据适应值的大小选择最优秀个体的 20% 用于更新信念空间。

#### (5) 影响函数 influence()

影响函数通过信念空间的信息指导种群空间个

体的变化趋势。例如，在信念空间优胜区域外，个体通过变异转化到信念空间优胜区域内；在信念空间优胜区域内，则个体在所属信念细胞的优胜区域内随机变化。

$$x_{i,j}^{t+1} = \begin{cases} x_{i,j}^t + |(u_j^t - l_j^t)N_{i,j}(0,1)| & x_{i,j}^t < l_j^t \\ x_{i,j}^t - |(u_j^t - l_j^t)N_{i,j}(0,1)| & x_{i,j}^t > u_j^t \\ x_{i,j}^t + (u_{p,j}^t - l_{p,j}^t)N_{i,j}(0,1) & l_{p,j}^t < x_{i,j}^t < u_{p,j}^t \end{cases}$$

(17)

其中， $l_j^t$ 和 $u_j^t$ 表示信念空间中第 $t$ 次迭代第 $j$ 个自变量的下限和上限， $l_{p,j}^t$ 和 $u_{p,j}^t$ 表示第 $t$ 次迭代第 $p$ 个信念细胞的第 $j$ 个自变量的优胜区域的下限和上限，细胞 $p$ 代表个体 $x_i$ 所属的信念细胞， $x_{i,j}^t$ 表示第 $t$ 次迭代第 $i$ 个个体的第 $j$ 个自变量， $N_{i,j}(0,1)$ 表示满足一维正态分布的随机值。

文化算法优化 $v$ -LPSVR的参数 $C^*$ 和 $v$ 的具体学习步骤如下。

- ① 设置种群空间的大小和最大迭代次数，初始化各个体，构造适应度函数，评价各个体的适应值，同时初始化信念空间。
- ② 根据接受函数获取部分最优秀个体。
- ③ 根据式(16)，利用②得到的优秀个体更新信念空间的优胜区域等信息。
- ④ 根据式(17)，对信念空间优胜区域外的个体变异，对信念空间优胜区域内的个体在其所属信念细胞的优胜区域内作随机变化，选择并产生新一代个体，计算新个体的适应值。
- ⑤ 如果迭代次数超过预定值，或者适应值达到预定的限值，则停止搜索；否则转到②继续执行。

## 2 实验仿真及分析

为了验证本文方法的有效性，首先用标准回归数据集作实验，然后将其用于丙烯浓度估计及应用。

### 2.1 标准数据集仿真及分析

分别用3个标准回归数据集——Boston housing、Bodyfat及Santa进行仿真。Boston housing包含了506个样本。整个样本集被随机划分100次，每次都被划分成481个训练样本和25个测试样本。Bodyfat数据集被随机划分30次，每次被分成202个训练样本和50个测试样本。Santa数据集<sup>[17]</sup>将当前样本作为目标输出，当前样本的前12个样本作为输入变量，新的样本集就成为包含具有12个输入属性的988样本模式的数据集合。其中前面100个样本模式作为测试集，剩余样本模式作为训

练集。用均方差来计算测试集误差。

2.1.1 标准数据集仿真结果 在量子聚类过程中，令唯一的自由参数 $\sigma$ 在0.1~5之间变化，使惰性阈值 $\mu(t)$ 中最大迭代次数 $T=30$ ，选择具有最小 $DB_k$ 值的聚类划分。对于housing数据集，随机选择某一次划分的训练集进行聚类，结果分成2类，则其多尺度核函数由2个Morlet小波核构成。每个小波核宽度都与该尺度所对应的那类训练样本相关。每类训练样本都用1.2节描述的SVDD方法求直径，并用文化算法对其进行优化，求得最小直径。在参数搜索过程中，以 $(\lg C, s)$ 作为优化个体，采用实数编码，个体各自变量的取值范围分别为 $[-2, 6]$ 和 $[0.1, 0.9]$ ，种群规模为20；迭代次数为20；以直径 $D$ 构造适应度函数，即 $\text{Fitness} = 1/(1 + D)$ ；迭代停止条件为直径不再减小。分别对housing、Bodyfat、Santa及丙烯精馏塔的丙烯浓度数据进行优化，结果如表1。

表 1 SVDD 核宽度优化结果  
Table 1 Kernel width selection by SVDD

| Dataset   | $C$    | $s$     | Diameter |
|-----------|--------|---------|----------|
| housing   | 2.16   | 15.1    | 4.17     |
|           | 3.9    | 7120    | 2.21     |
| Bodyfat   | 3.28   | 797.3   | 2.51     |
|           | 147.46 | 4577.6  | 1.93     |
| Santa     | 0.94   | 60430   | 0.87     |
|           | 59.96  | 81079.5 | 0.92     |
|           | 0.58   | 10601   | 0.73     |
| propylene | 0.025  | 1008.6  | 1.89     |
|           | 0.021  | 3954.4  | 0.6      |

Note:  $C$  and  $s$  are model parameters of SVDD. Diameter represents kernel width of present method.

多尺度核 $v$ -LPSVR方法中，剩余的模型参数 $C^*$ 和 $v$ 也通过CA方法进行优化。在参数搜索过程中，以 $(\lg C^*, v)$ 作为优化个体，采用实数编码，个体各自变量的取值范围、种群规模、迭代次数同上；用5-交叉验证<sup>[16]</sup>均方差构造适应度函数，即 $\text{Fitness} = 1/(1 + \text{MSE})$ ；迭代停止条件为训练集误差和测试集误差都不再减小。

表2中 $v$ -LPSVR方法均采用高斯核，single wavelet-LPSVR和multi-scale wavelet-LPSVR算法分别采用单小波核和本文方法构造的多尺度小波核。使单小波核方法的参数 $C^*$ 和 $v$ 与多尺度核方法的参数取值相同，以与多尺度核情况作对比。对比表2中各方法的100次随机划分的测试集误差可知，多尺度小波核方法的预测均方差平均值最小为

8.943，比  $v$ -LPSVR、单小波核方法的误差均值分别减少了 6.8%和 19.2%。从图 2 可知，在绝大多数划分中，housing 数据的多尺度核方法的测试集误差均比高斯核的要小。至于 Bodyfat 和 Santa 数据，它们分别被分成 2 类和 3 类，核宽度分别如表 3 和表 4 所示，单小波核方法的泛化性能要优于高斯核方法，且本文方法仍具有最小平均均方差。与  $v$ -LPSVR 方法相比，它们的多尺度核方法的测试集平均误差分别减少了 62.0%和 91.3%。图 3 展示了各种核方法在 Bodyfat 数据的 30 次划分中的测试集误差，几乎在每次划分中多尺度核方法的测试集误差都是最小的。

表 2 Boston housing 数据集的参数选择及泛化性能对比  
Table 2 Parameter selection and generalization ability comparison for Boston housing dataset

| Method                         | $C^*, v, \sigma^{*2}(a_i^2)$ | nSVs  | Mean±Std     |
|--------------------------------|------------------------------|-------|--------------|
| Gaussian $v$ -LPSVR            | 100, 0.2, 1                  | 170.2 | 9.593±6.627  |
| single wavelet $v$ -LPSVR      | 305.3, 0.13, (18.58)         | 77.5  | 11.071±7.487 |
| multi-scale wavelet $v$ -LPSVR | 305.3, 0.13, (17.39, 4.88)   | 166.7 | 8.943±4.489  |

Note: nSVs denotes average number of support vectors for all splits. Mean and Std are averages, standard deviation of mean squared error of test sets for all splits.  $\sigma^*$  is width of Gaussian kernel.  $a_i$  is kernel width of single or multi-scale wavelet.  $C^*$  and  $v$  are model parameters of  $v$ -LPSVR.

表 3 Bodyfat 数据集的参数选择及泛化性能对比  
Table 3 Parameter selection and generalization ability comparison for Bodyfat dataset

| Method                         | $C^*, v, \sigma^{*2}(a_i^2)$ | nSVs  | Mean±Std                                      |
|--------------------------------|------------------------------|-------|-----------------------------------------------|
| $v$ -LPSVR                     | 1000, 0.001, 0.29            | 87.3  | $6.39 \times 10^{-5} \pm 2.98 \times 10^{-5}$ |
| single wavelet $v$ -LPSVR      | 57.84, 0.72, (8.61)          | 90.0  | $3.5 \times 10^{-5} \pm 2.11 \times 10^{-5}$  |
| multi-scale wavelet $v$ -LPSVR | 57.84, 0.72, (6.3, 3.74)     | 129.5 | $2.43 \times 10^{-5} \pm 1.75 \times 10^{-5}$ |

表 4 Santa 数据集的参数选择及泛化性能对比  
Table 4 Parameter selection and generalization ability comparison for Santa dataset

| Method                          | $C^*, v(\epsilon), \sigma^{*2}(a_i^2)$ | nSVs | MSE of test set | NMSE of test set |
|---------------------------------|----------------------------------------|------|-----------------|------------------|
| $\epsilon$ -SVR <sup>[17]</sup> | 10, 1000, 0.01                         | —    | —               | 0.0197           |
| $v$ -LPSVR                      | 100, 0.2, 2.5                          | 37   | 0.000447        | 0.0114           |
| single wavelet $v$ -LPSVR       | 5.23, 0.24, (1.35)                     | 43   | 0.000134        | 0.0034           |
| multi-scale wavelet $v$ -LPSVR  | 5.23, 0.24, (0.753, 0.85, 0.539)       | 86   | 0.000039        | 0.0010           |

Note:  $\epsilon$  is parameter of  $\epsilon$ -SVR.

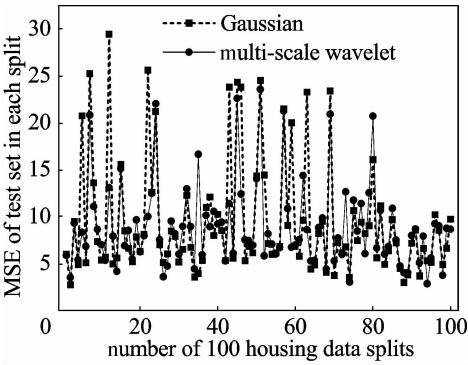


图 2 housing 数据 100 次划分的测试集误差  
Fig. 2 Test set error for 100 housing data splits

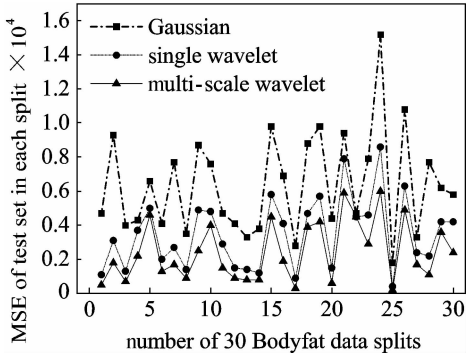


图 3 Bodyfat 数据 30 次划分的测试集误差  
Fig. 3 Test set error for 30 Bodyfat data splits

2.1.2 标准数据集分析与讨论 SVR 的推广性能取决于模型参数的选择。如果任意选取多尺度核的尺度个数，本文方法未必就有良好的学习性能和泛化性能。有时甚至比不上单个核函数的性能。如果通过不断调整核函数个数，比较它们的性能以选取最优的多尺度核，尤其对复杂问题，势必会增加学习复杂度。因此，本文依据样本空间分布特征，运用聚类算法划分训练样本，使每类样本对应一个尺度上的核函数。如果能事先确定多尺度核宽度，那么对  $v$ -LPSVR 方法，就只剩 2 个未知参数，即本文构造的多尺度核方法的参数  $C^*$  和  $v$ 。实验过程中，运用 CA 方法通常迭代 2 或 3 次就能得到满意参数值。而对于高斯核  $v$ -LPSVR 方法，存在 3 个优化参数，运用 CA 方法通常迭代 5 或 6 次才会有较好的解，有时候甚至要迭代 10 多次才行。若多尺度核宽度事先未知，则优化参数至少为 4 个，搜索范围扩大，其模型选择的学习效率将会大大降低。本文采用 SVDD 方法，选择最小直径为核宽度，正好包含了所有样本的分布特征。若将核宽度减小，未必能使所有数据一定得到最优性能。例如，对于 housing 数据，将最小半径作为核宽度，

无论怎么优化参数  $C^*$  和  $v$ ，其 100 次随机划分的测试集误差平均值都比本文方法的大。表 5 比较了相同  $C^*$  和  $v$ ，不同核宽度情况下的泛化性能。若核宽度增大，不能准确反映样本分布，就不可能选择出最佳模型。

表 5 不同核宽度下 Boston housing 数据集泛化性能对比

Table 5 Generalization ability comparison of Boston housing dataset with different kernel width

| Method                           | $C^*, v, \sigma^{*2}(a_i^2)$  | nSVs  | Mean±Std     |
|----------------------------------|-------------------------------|-------|--------------|
| multi-scale<br>(radius kernel)   | 305.3, 0.13,<br>(4.35, 1.22)  | 348.7 | 12.673±7.677 |
| multi-scale<br>(diameter kernel) | 305.3, 0.13,<br>(17.39, 4.88) | 166.7 | 8.943±4.489  |

2.2 丙烯精馏塔的丙烯浓度估计及应用

丙烯是一种有机化合物，主要用来生产用途广泛的聚丙烯，通常可经石油炼制或乙烯生产装置分馏得到。本文研究的丙烯均是通过乙烯生产装置中丙烯精馏塔分离出来的。图 4 就是丙烯精馏塔的工艺流程图。来自 C3 加氢反应器 DC-2431A/S 的物料进入丙烯汽提塔 E-DA-2430，经汽提后 E-DA-2430 塔顶气相产品进入丙烯精馏塔 E-DA-2440，丙烷作为产品从 E-DA-2430 侧线抽出。丙烯汽提塔有一个中间再沸器 E-EA-2441A/B 和底部再沸器 E-EA-2440A/B，二者均为急冷水加热。丙烯产品从丙烯精馏塔 E-DA-2440 侧线抽出，经冷凝器 E-EA-2456 冷却后送至丙烯产品贮罐。塔釜物料送往丙烯汽提塔 E-DA-2430。丙烯精馏塔塔釜出料丙烯浓度要求一般不超过 5.0%。塔釜丙烯组分分析仪经常会出现故障跳变，很难长时间稳定工作，操作人员为保证塔顶聚合级丙烯的纯度，经常设置较大的回流比，塔顶出现质量过剩，造成塔釜丙烯损失增大。因此，从丙烯精馏的机理出发，建立塔釜丙烯浓度软测量模型，在线实时估计塔釜丙烯浓度是非常必要的，也为丙烯精馏塔先进控制系统 (APC) 的实施打下基础。根据工业装置现场操作变量分析，以及丙烯精馏系统的工艺机理模型模拟分析，塔釜中的丙烯浓度和塔釜压力及灵敏板温度有直接的关系，塔釜温度变化较小，因此选择塔釜丙烯（损失）浓度的人工分析值作为软测量模型训练样本的输出，选择对应时刻的 DA2430 的塔顶压力 PI24018、灵敏板温度 TU24123 作为影响 DA2430 塔釜丙烯浓度的主要操作变量，构成

DA2430 塔釜丙烯浓度支持向量回归软测量模型的训练样本输入。作者从上海石化采集了 2007 年的丙烯数据，从中挑选了 176 个样本，将其随机划分为 120 个训练样本和 56 个预测样本，部分训练样本模式如表 6 所示，其中丙烯用体积分数表示。用相对误差来评价性能。

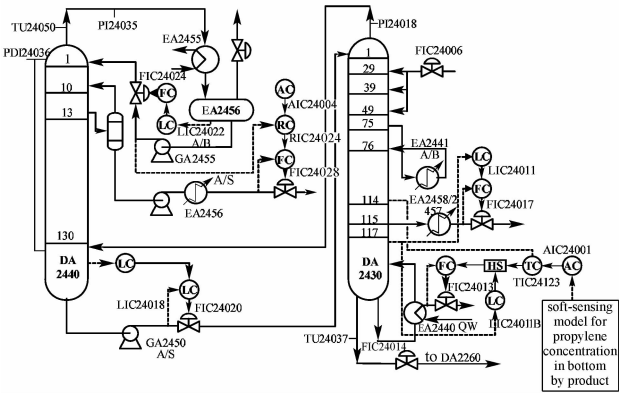


图 4 丙烯精馏塔工艺流程图

Fig. 4 Flow chart for propylene rectification tower

表 6 丙烯数据的训练模式

Table 6 Training patterns for propylene data

| Output variable           | Input attributes              |                             |
|---------------------------|-------------------------------|-----------------------------|
| Propylene concentration/% | Sensitive plate temperature/℃ | Column reactor pressure/MPa |
| 1.88                      | 60.87663                      | 1.98383                     |
| 4.01                      | 60.79488                      | 1.99318                     |
| 10.24                     | 61.23131                      | 2.05157                     |
| 23.78                     | 55.47893                      | 1.93808                     |

Note: Propylene data is derived from industrial equipment data.

2.2.1 塔釜丙烯浓度估计结果 对于丙烯数据，在运用量子聚类和文化算法时均选择与标准数据集相同的参数设置，所得模型参数如表 7 所示。多尺度小波核  $v$ -LPSVR 用于估计塔釜丙烯浓度时，依据量子聚类方法确定其尺度个数为 2，使用 SVDD 方法并结合文化算法得到的最小直径为 1.89 和 0.6，根据表 7 中的模型参数计算得到丙烯训练集的相对误差为 0.209，丙烯测试集的相对误差为 0.211。从表 7 中可以看出，本文方法的丙烯浓度测试集相对误差是最小的，与高斯核  $v$ -LPSVR 相比，减少了 18.8%；与 RBF 神经网络方法相比，减少了 23.6%。图 5 展示了本文方法的丙烯浓度训练集的模型输出与目标输出基本吻合，且本文方法的丙烯浓度测试集的模型输出十分逼近于目标输出，如图 6 所示。

表 7 丙烯数据的性能比较

Table 7 Performance comparison for propylene data

| Method                         | $C^*, v, \sigma^{*2}(a_i^2)$ | nSVs | MSE of test set | RE of test set |
|--------------------------------|------------------------------|------|-----------------|----------------|
| $v$ -SVR                       | 90, 0.35, 0.476              | 49   | 2.701           | 0.237          |
| $v$ -LPSVR                     | 300.023, 0.179, 0.512        | 12   | 2.749           | 0.260          |
| single wavelet $v$ -LPSVR      | 306.3, 0.71, (4.41)          | 9    | 2.839           | 0.272          |
| multi-scale wavelet $v$ -LPSVR | 306.3, 0.71, (3.572, 0.36)   | 27   | 2.052           | 0.211          |
| RBF-NN                         | —                            | —    | 2.885           | 0.276          |

Note: Parameter setting of RBF-NN in matlab toolbox is: goal= $1 \times 10^{-5}$ , spread=0.5, MN=7, DF=25.

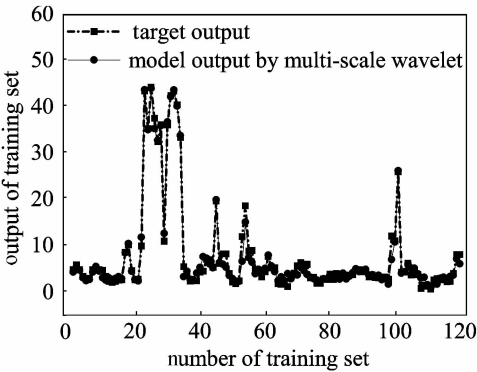


图 5 丙烯训练集预测

Fig. 5 Training set prediction for propylene

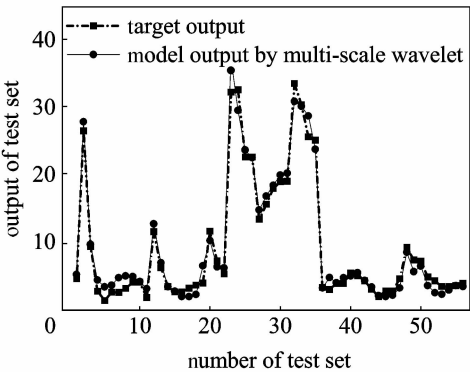


图 6 丙烯测试集预测

Fig. 6 Test set prediction for propylene

2.2.2 塔釜丙烯浓度估计分析 通过分析实际操作工况，凡是小于 10% 的丙烯浓度样本都是合理的。而在实际采样数据中，大约 1/5 的样本的丙烯浓度都大于 10%，如表 6 中第三和第四行训练模式所示。这主要是由于高压塔中大于 4 及以上的碳重组分没有脱干净，导致过多的重组分作为进料进入丙烯精馏塔，使丙烯精馏塔塔釜再沸器的温度升不上去，很多丙烯就不能蒸发到塔顶，从而造成塔

釜中部分丙烯损失浓度超过 10%。尽管如此，本文方法在估计丙烯浓度时表现出了较强的鲁棒性和良好的泛化能力，不仅使得训练集的模型输出与观测输出基本一致，而且使其测试集相对误差比表中其他方法的误差要小得多。表 8 是采用异常点去除法后，再进行建模得到的性能比较结果，从表中可以看出本文方法的测试集均方差和相对误差也都是最小的。

表 8 去除异常点后的丙烯数据的性能比较

Table 8 Performance comparison for propylene data after exceptions removal

| Method                         | $C^*, v, \sigma^{*2}(a_i^2)$  | nSVs | MSE of test set | RE of test set |
|--------------------------------|-------------------------------|------|-----------------|----------------|
| $v$ -SVR                       | 15.204, 0.611, 0.067          | 72   | 1.438           | 0.264          |
| $v$ -LPSVR                     | 23.688, 0.256, 1.408          | 5    | 1.487           | 0.274          |
| single wavelet $v$ -LPSVR      | 71.671, 0.676, (0.518)        | 16   | 1.345           | 0.261          |
| multi-scale wavelet $v$ -LPSVR | 71.671, 0.676, (0.457, 0.275) | 16   | 1.309           | 0.256          |
| RBF-NN                         | —                             | —    | 2.221           | 0.349          |

Note: Parameter setting of RBF-NN in matlab toolbox is: goal= $1 \times 10^{-5}$ , spread=0.5, MN=7, DF=25.

3 结 论

恰当的模型选择是多尺度小波核支持向量回归方法具备良好泛化性能的保证，为此，本文根据量子聚类算法划分样本以定义尺度个数，借助 SVDD 方法确定小波核宽度，它避免了通过反复实验确定尺度个数，间接提高了 CA 方法优化模型参数的学习效率，而且使多尺度小波核支持向量回归方法表现出良好的预测精度。本文方法不仅准确预测了标准回归数据集的输出，也有效估计了实际工业丙烯精馏塔的塔釜丙烯浓度，而且与上述列出的其他方法相比，已体现出其本身的优越性。进一步的研究工作是对其他形式的小波核函数及其在多尺度建模中的应用进行研究，以进一步提高建模精度与模型可靠性。

References

[1] Zheng Danian, Wang Jiabin, Zhao Yannan. Non-flat function estimation with a multi-scale support vector regression. *Neurocomputing*, 2006, **70** (1/2/3): 420-429

[2] Pozdnoukhov A, Kanevski M. Multi-scale support vector algorithms for hot spot detection and modeling. *Stochastic Environmental Research and Risk Assessment*, 2008, **22** (5): 647-660

- [3] Duan Chongwen (段崇雯). The support vector machine based on multiscale kernel and its application research. *National University of Defense Technology*, 2006
- [4] Wang Peng (王鹏), Wang Zhicheng (王志成), Zhang Jun (张钧), Tian Jinwen (田金文). Infrared weak small target detection based on multi-scale wavelet kernel LS-SVM. *Infrared and Laser Engineering*, 2006, **35** (24): 251-257
- [5] Schölkopf B, Bartlett P, Smola A, *et al.* Shrinking the tube: a new support vector regression algorithm// *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*. 1999: 330-336
- [6] Zhang Li, Zhou Weida, Jiao Licheng. Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics*, 2004, **34** (1): 34-39
- [7] Wen Xiangjun, Cai Yunze, Xu Xiaoming. Least squares support vector machine based on continuous wavelet kernel. *Lecture Notes in Computer Science*, 2005, **3496**: 843-850
- [8] Horn D, Gottlieb A. The method of quantum clustering// *Proceedings of the Neural Information Processing Systems*. 2001: 769-776
- [9] Tax D, Duin R. Support vector data description. *Machine Learning*, 2004, **54**: 45-66
- [10] Robert R. An introduction to cultural algorithms// *Proceedings of the 3th Annual Conference on Evolution Programming*. 1994: 131-136
- [11] Reynolds R, Zhu Shinin. Knowledge-based function optimization using fuzzy cultural algorithms with evolutionary programming. *IEEE Transactions on Systems, Man and Cybernetics*, 2001, **31** (1): 1-18
- [12] Pedroso J, Murata N. Support vector machines for linear programming: motivation and formulations: BSIS Technical Report 99-2 [R]. Wako-shi, Saitatma, Japan: Riken Brain Science Institute, 1999
- [13] Smola A, Schölkopf B, Röttsch G. Linear programs for automatic accuracy control in regression// *Ninth International Conference on Artificial Neural Networks*. London: Springer, 1999 (2): 575-580
- [14] Li Yuancheng (李元诚), Fang Tingjian (方廷健). Wavelet support vector machines. *Pattern Recognition and Artificial Intelligence*, 2004, **17** (2): 167-172
- [15] Zhang Li (张莉), Sun Gang (孙刚), Guo Jun (郭军). Unsupervised feature selection method based on K-means clustering. *Application Research of Computers*, 2005, **22** (3): 23-24, 42
- [16] Joachims T. Estimating the generalization performance of a SVM efficiently// *Proceedings of 17th International Conference on Machine Learning*. San Fransisco, CA, 2000
- [17] Cao L J, Chua K S, Chong W K, *et al.* A comparison of PCA, KPCA, and ICA for dimensionality reduction in support vector machine. *Neurocomputing*, 2003, **55**: 321-336