文章编号: 1001-0920(2005)05-0549-04

基于混合SVR-PLS方法的丙烯腈收率软测量建模

王华忠, 俞金寿
(华东理工大学 自动化研究所, 上海 200237)

摘 要: 为了更有效地处理过程非线性、多输入和数据共线性等复杂特性, 提高模型的推广能力和精度, 提出了混合支持向量回归机-偏最小二乘法(SVR-PLS)方法. 该方法兼具SVR和PLS的优点, 用PLS进行特征提取, 用SVR建立PLS的内部模型. 对工业丙烯腈生产过程丙烯腈收率软测量建模的应用表明, 采用该方法建立的软测量模型, 在模型精度、推广能力等方面明显优于一些传统软测量建模方法, 满足工业应用要求.

关键词: 支持向量回归机; 偏最小二乘法; 丙烯腈; 软测量
中图分类号: TP183; TQ063 **文献标识码**: A

Soft sensor modeling of acrylonitrile yield based on hybrid SVR-PLS approach

WANG Hua-zhong, YU Jin-shou
(Research Institute of Automation, East China University of Science and Technology, Shanghai 200237, China. Correspondent: WANG Hua-zhong, E-mail: hzwang@ecust.edu.cn)

Abstract: A hybrid SVR-PLS method is proposed to deal with complicated process with nonlinearity and a large number of correlated inputs. The SVR-PLS method, which has merits of both SVRs and PLS, is an integration of support vector regression machine and partial least squares. The PLS outer projection is used as a dimension reduction tool to remove collinearity and the SVRs are trained to capture the nonlinearity in the projected latent space. Soft sensor modeling of acrylonitrile yield is established using SVR-PLS method. The generalization ability and accuracy of the soft sensor using the method proposed is superior to traditional methods.

Key words: support vector regressor(SVR); partial least squares(PLS); acrylonitrile; soft sensing

1 引 言

现代市场竞争日趋激烈, 产品质量在某种程度上决定了企业的生死存亡. 但在过程工业中, 许多重要的质量指标却很难在线获得, 软测量技术试图以过程可测辅助变量来估计直接质量指标. 软测量技术包括软测量建模方法、软测量工程化实施和软测量模型校正, 其中软测量建模是软测量技术的核心. 工业过程通常具有非线性、时变、多变量、过程机理复杂等十分复杂的特性, 采用机理方法建立过程软测量模型十分困难. 基于数据驱动的软测量建模方法在过程工业中得到了广泛研究和应用[1], 但由于与直接质量指标相关的许多工艺参数之间具有较强的相关性, 单纯用这些数据进行软测量建模, 通常会导致模型的推广能力较差. 因此, 对历史数据进行特征提取、消除冗余信息对于数据驱动的软测量建模方法就显得十分重要.

一些化学计量学方法(如PCA和PLS)在消除过程数据共线性以及特征提取中得到了广泛的应用. 与PCA等特征提取方法不同, PLS方法同时对输入和输出数据进行分解, 使得PLS模型能从较少的载荷向量(Load Vector)得到尽可能多的信息. 标准PLS方法在回归中采用的是线性回归, 因此对于非线性建模, 其性能较差.

本文提出了将SVR与PLS相结合的混合SVR-PLS方法, 在保持PLS的结构和优点的基础上, 提高模型的非线性处理能力和精度. 介绍了SVR的基

收稿日期: 2004-05-13; **修回日期**: 2004-08-23.
作者简介: 王华忠(1969—), 男, 江苏南京人, 副教授, 博士生, 从事过程模型化控制的研究; 俞金寿(1939—), 男, 浙江海宁人, 教授, 博士生导师, 从事工业过程模型化与控制等研究.

本原理, 重点分析了混合SVR-PLS 方法的原理和建模步骤, 并以某大型化工厂 5 万吨丙烯腈装置软测量建模为例, 验证了本文方法的有效性

2 混合SVR-PLS 方法的原理

2.1 SVR 基本原理及特点

与传统机器学习方法相比, 支持向量机具有小样本学习能力强、模型推广性能好以及高维输入数据处理能力等特性, 是一种新的机器学习方法. 支持向量机按其输出可分为两类, 即输出为指示函数集的支持向量分类机(SVC)和输出为实函数集的支持向量回归机(SVR). 由于引入了结构风险最小化和核函数的思想, 支持向量机较传统的非线性数据处理方法具有更好的性能. 用结构风险最小化代替传统算法中的经验风险最小化, 提高了支持向量机的模型推广能力, 避免了过拟合. 核函数的引入提高了支持向量机非线性处理能力, 而且通过选择不同的核函数, 可以隐式地使得分类或回归在不同维数的高维特征空间进行, 甚至可以使得特征空间为无穷大, 大大提高了分类或回归能力. 更重要的是, 核函数技术使得算法的实现只要在原始输入空间进行, 便可大大减小计算量, 避免维数灾难. 限于篇幅, 更详细的内容可参见文献[2, 3].

在支持向量机等核函数方法中, 正则化参数、核函数类型和参数对核函数方法的应用效果有很大的影响. 通常将确定这些参数使得模型测试误差最小的过程称作模型选择. 常用的模型选择方法有交叉检验法、Bayesian 学习法和自举法等[4]. 在模型选择中, 核函数的选择和构造十分重要, 常用的核函数有: 高斯核函数 $K(x, x_i) = \exp\left(\frac{-\|x - x_i\|^2}{2\sigma^2}\right)$; 指数型核函数 $K(x, x_i) = \exp\left(\frac{-\|x - x_i\|}{2\sigma^2}\right)$; 多项式核函数 $K(x, x_i) = (x \cdot x_i + 1)^d$, $d = 1, 2, \cdots, N$; 感知器核函数 $K(x, x_i) = \tanh(\beta x_i + b)$ 等.

2.2 混合SVR-PLS 方法

混合SVR-PLS 方法的基本原理如图 1 所示, 它与PLS 方法的根本不同在于用SVR 而不是线性回归来建立PLS 的内部模型. 混合SVR-PLS 方法通过PLS提取过程的特征信息, 同时消除了数据的共线性问题; 用 SVR 对输入和输出的得分向量(Score Vector)进行非线性回归, 可更有效、更方便地建立载荷向量与得分向量之间的非线性关系, 提高内部模型的非线性处理能力. 这样, 混合SVR-PLS 方法便继承了PLS 和SVR 的优点. 与其他采用非线性函数(如神经网络)建立PLS 内部模型的非线性PLS 方法[5]相比, 混合SVR-PLS 方法中非线性内部模型结构可通过训练算法自动确定, 内部模型推广能力强, 而且可以得到全局最优解而不是局部优化. 本文选用最小二乘支持向量机来建立PLS 的内部模型, 其理由如下:

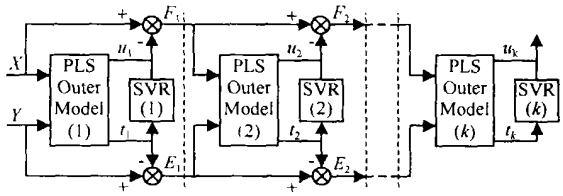

图1 混合SVR-PLS 方法原理图

1) 由于上述SVR 算法计算复杂度与训练样本个数有关, 当样本数目越来越大时, 求解相应的二次规划问题越复杂, 计算速度越慢. 而最小二乘SVM 算法(LS-SVMs)与标准SVM 的主要区别在于损失函数项的不同, 以及将不等式约束修改成等式约束, 从而使得该算法具有较好的实时性. 这一点对于过程建模十分重要.

2) 与标准支持向量机算法相比, 虽然LS-SVMs 的解不具有稀疏性, 但这对于过程建模并非很重要. 因为过程建模中, 有效的样本通常较难获得, 因此其数量较少, 不象数据挖掘和模式识别等应用中样本数量可以达到上万个, 甚至更多.

2.2.1 用于建立PLS 内部模型的LS-SVMs 算法

设给定 k 个样本数据$\{t_i, u_i\}$, $i = 1, 2, \cdots, k$. 其中: $t_i \in R$ 为样本输入, $u_i \in R$ 为样本输出. 对该样本数据逼近的LS-SVMs 的原理如下[6]:

目标函数

$$\min_{\omega, b, e} J = \frac{1}{2}\omega^{\mathrm{T}}\omega + \frac{1}{2}\gamma\sum_{i=1}^{k} e_i^2, \tag{1}$$

约束条件

$$u_i = \omega^{\mathrm{T}}\varphi(t_i) + b + e_i, \quad i = 1, 2, \cdots, k. \tag{2}$$

相应的拉格郎日函数为

$$L = J - \sum_{i=1}^{k} a_i[\omega^{\mathrm{T}}\varphi(t_i) + b + e_i - u_i], \quad i = 1, 2, \cdots, k. \tag{3}$$

于是所求解的优化问题便转化为求解线性方程

$$\begin{bmatrix} 0 & \vec{1}^{\mathrm{T}} \\ \vec{1} & \Omega + \gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ u \end{bmatrix}. \tag{4}$$

其中: $u = [u_1, \cdots, u_k]^{\mathrm{T}}$, $\vec{1} = [1, \cdots, 1]^{\mathrm{T}}$, $a = [a_1, \cdots, a_k]^{\mathrm{T}}$, $\Omega_{ij} = \varphi(t_i)^{\mathrm{T}}\varphi(t_j) = K(t_i, t_j)$, φ为非线性变换函数, K 为高斯核函数

$$K(t, t_i) = \exp\left(\frac{-\|t - t_i\|^2}{2\sigma^2}\right).$$

通过求解线性方程组(4), 得 SVM 模型

$$u(t) = \sum_{i=1}^{k} a_i K(t, t_i) + b \qquad (5)$$

2.2.2 SVR-PLS 建模步骤

设输入数据构成矩阵 $X \in R^{l\times n}$, 输出数据构成矩阵 $Y \in R^{l}$. 其中: l 为数据长度, n 为输入变量维数. 首先令 $E = X$, $F = Y$; 然后采用如下非线性迭代形式的 PLS 算法[7] 进行建模

1) 用 Y 的第 1 列初始化 u;

2) $w = X^{\mathrm{t}}u/u^{\mathrm{T}}u, w = w/\|w\|$;

3) $t = Xw/(w^{\mathrm{T}}w)$;

4) $c = Y^{\mathrm{t}}t/t^{\mathrm{T}}t, c = c/\|c\|$;

5) $u = Yc/c^{\mathrm{T}}c$;

6) 若第 3) 步的 t 与前次迭代的 t 差值小于某阈值(如 10^{-7}) 则转 7), 否则转 2);

7) 求 X 的负载向量 $p = X^{\mathrm{t}}t/t^{\mathrm{T}}t$;

8) 求 Y 的负载向量 $q = Y^{\mathrm{t}}u/u^{\mathrm{T}}u$;

9) 求 X 与 Y 的内部关系, 即用前文介绍的LS-SVMs 算法建立输入得分向量 t 与输出得分向量 u 的非线性关系 $u = f_{\mathrm{SVR}}(t)$;

10) 计算残差 E, F, 即

$$E = X - tp^{\mathrm{T}},\quad F = F - f_{\mathrm{SVR}}(t)q^{\mathrm{T}};$$

11) 如残差满足收敛条件或主元数达到设定值, 则结束; 否则转 1).

当获得了 SVM-PLS 模型后, 要用其对输入数据进行预测. 设测试数据集合 $X_{\mathrm{new}} \in R^{l_t\times n}$, l_t 为测试数据长度. 为了得到预报值 $\hat{Y}$, 首先求 $\hat{t}_h$.

令 $E_0 = X_{\mathrm{new}}$, 根据训练 SVM-PLS 模型时获得的参数, 依次计算: $\hat{t}_h = E_{h-1}W_h, E_h = E_{h-1} - \hat{t}_h p_h^{\mathrm{T}}$. 最后得 $\hat{Y}$ 的预报值为

$$\hat{Y} = \sum_{h=1}^{k} f_{\mathrm{SVR},h}(\hat{t}_h) q_h^{\mathrm{T}}, \qquad (6)$$

其中 k 为 PLS 主元数目, 其数值通常用交叉检验等方法确定.

3 基于核函数 PCR 与核函数 PLS 丙烯腈收率软测量模型

3.1 工艺分析

某化工厂的年产 5 万吨丙烯腈装置, 采用美国标准石油公司 Sohio 的生产工艺, 以C-41 为催化剂, 用丙烯、氨、空气为原料, 在沸腾床反应器中一次直接氧化制取丙烯腈(即丙烯氨氧化法). 根据工艺分析, 在催化剂确定后, 对丙烯腈收率影响较大的因素可归结为原料的配比、反应温度、反应压力、接触时间和空塔线速等条件. 根据工艺原理和操作人员经验, 选择反应压力、中段温度、纯丙烯量、空比、氨比、反应线速和触媒量等变量作为软测量建模的辅助变量. 数据的采集通过DCS 系统进行, 每天早 8∶00 采集一次, 并记录丙烯腈收率的人工分析值. 汇总现场采集的数据, 并进行预处理, 共得到建模数据 343 组.

3.2 丙烯腈收率混合SVR-PLS 软测量模型的建立

需要对所采集到的 343 组数据进行分组, 以得到训练集合和测试集合. 在建模过程中发现, 若单纯按样本顺序分组, 得到的模型精度较差. 这在很大程度上是由训练样本和测试样本模式分布不均造成的. 为了尽量减少分组过程的人为干预, 首先设定训练样本容量为 300, 测试样本容量为 43; 然后对 343 组数据采用随机分组, 再对分组集合采用前文所介绍的方法建立混合SVR-PLS 软测量模型, 直到对测试集模型的误差达到满意为止. 在此过程中, 实际上还需要确定模型的参数. 混合SVR-PLS 方法要确定的参数有主元数和每个内部 SVR 模型的参数; 而 SVR 模型的参数有核函数的类型以及核宽度. 本文选择最小二乘支持向量机(LS-SVM) 建立内部模型, 在完成训练和测试数据的分组后, 再对混合软测量模型的参数进一步优化. 最终确定的模型参数为: 主元数 6, 选用高斯核函数

$$K(x, x_i) = \exp\left(\frac{-\|x - x_i\|^2}{2\sigma^2}\right),$$

$\sigma^2 = [25, 25, 25, 20, 20, 20]$,

正则化参数

$\gamma = [400, 400, 400, 350, 300, 250]$.

从以上参数发现, 不是所有的SVR 内部模型参数都相同. SVR-PLS 模型的测试曲线如图 2 所示.

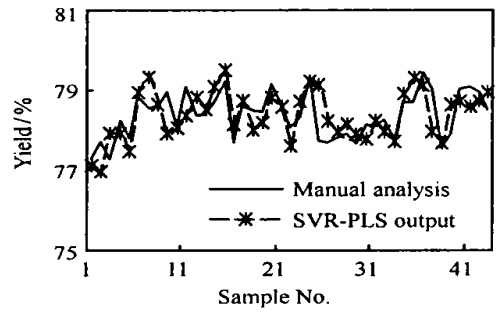

图 2　SVR-PLS 软测量模型预测值与分析值曲线

3.3 不同软测量模型性能分析比较

为了比较不同的非线性 PLS 软测量方法的性能, 以最后得到的训练集合和测试集合为基础, 采用 BPNN-PLS 方法建立软测量模型. BPNN-PLS 模型的测试曲线如图 3 所示.

BPNN-PLS 模型采用反传算法训练前向神经网络(BPNN) 来建立 PLS 内部模型. 该内部模型为 3 层结构, 隐含层神经元数为 8, PLS 模型的主元数为 6. 为进一步比较基于 PLS 的非线性建模方法与

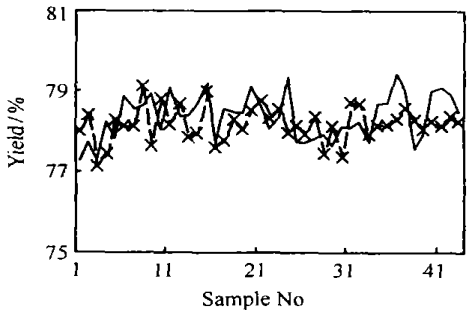

图 3 BPNN-PLS 软测量模型预测值与分析值曲线

常规神经网络方法, 建立了 BPNN 软测量模型, 该模型为 3 层, 隐含层节点数为 16. 表 1 为采用 SVM-PLS, BPNN-PLS 和 BPNN 方法建立的软测量模型的训练与测试结果. 采用预测均方根误差作为比较标准(RMSE). RMSE 的定义为

$$\mathrm{RMSE} = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(\hat{y}_i - y_i)^2}. \tag{7}$$

其中: $\hat{y}_i$ 为软测量模型预测输出值, y_i 为丙烯腈收率人工分析值, l 为数据长度.

表 1 不同软测量模型的 RMSE

	SVM-PLS	BPNN-PLS	BPNN
Train	4.16e-3	4.74e-3	6.23e-3
Test	5.12e-3	6.17e-3	7.72e-3

为了对采用 3 种方法所建立的软测量模型的精度进行定量比较, 引入一新的指标(打靶率), 定义为

$$HR = \frac{r}{l} \times 100\%. \tag{8}$$

其中: r 为丙烯腈收率软测量模型预测输出值误差小于某一数值(本文为 0.7%) 的数据个数, l 为数据长度. 表 2 所示为不同软测量模型的打靶率.

表 2 不同软测量模型的打靶率

	SVM-PLS	BPNN-PLS	BPNN
Train	89	80	75
Test	84	74	69

从表 1 和表 2 可以看出, 虽然 3 种模型基本上都能满足工业应用要求, 但 SVM-PLS 模型在推广能力及模型精度上最好, BPNN-PLS 次之, BPNN 模型性能最差. 这可解释为两类非线性 PLS 方法集成了 PLS 方法及 SVR 或 BPNN 的非线性处理强的特点, 而 SVR 能有效地克服神经网络方法中出现的过拟合现象和局部最小点. 因此, 混合 SVM-PLS 方法在复杂工业过程建模中能得到很好的性能.

4 结 论

本文提出了 SVR-PLS 软测量建模方法, 该方法兼具了 SVR 和 PLS 的优点. SVR 在处理复杂非线性数据关系上比神经网络等具有更多的优良性能, 因此 SVR-PLS 模型的性能较 BPNN-PLS 的性能好. 针对工业丙烯腈装置丙烯腈收率软测量建模的对比研究也证明了这一点. 作为新的机器学习方法, SVR 在理论和应用上都体现出优良的性能, 将它与其他方法(如模糊技术、粗糙集技术和各种化学计量法等) 结合进行复杂工业过程的建模研究是值得重视的研究方向.

参考文献(References)

[1] Willis M J, Montague G A, Massimo D C, et al. Artificial neural networks in process estimation and control[J]. *Automatica*, 1992, 28(6): 1181-1187.

[2] Vapnik V N. *The nature of statistical learning theory* [M]. New York: Springer-Verlag, 1999.

[3] David V, Sánchez A. Advanced support vector machines and kernel methods [J]. *Neurocomputing*, 2003, 55(1-2): 5-20.

[4] Müller K-R, Mika S, Rätsch G, et al. An introduction to kernel-based learning algorithms[J]. *IEEE Trans on Neural Networks*, 2001, 12(2): 181-202.

[5] Qin S J, McAvoy T J. Nonlinear PLS modeling using neural networks [J]. *Computers and Chemical Engineering*, 1992, 16(4): 379-391.

[6] Suykens J K A, Gestel T V, Brabanter J D, et al. *Least squares support vector machines* [M]. Singapore: World Scientific Publishing Co Pte Ltd, 2002: 98-99.

[7] Galadi P, Kowalski B R. Partial least-squares regression: A tutorial [J]. *Analytica Chimica Acta*, 1986, 185(1): 1-17.

(上接第 548 页)

[13] Rockafellar R T. Lagrange multiplier and optimality [J]. *SIAM Review*, 1993, 35: 183-238.

[14] Bertsekas D P. *Nonlinear programming* [M]. 2nd ed. Belmont: Athena Scientific, 1999.

[15] Lakshmikantham V, Matrosov V M, Sivasundaram S. *Vector lyapunov functions and stability analysis of nonlinear systems* [M]. Dordrecht: Kluwer Academic Publisher, 1991.

[16] LaSalle J P. *The stability of dynamical systems* [M]. New York: Springer, 1976.

[17] Marquez H J. *Nonlinear control systems-analysis and design* [M]. New Jersey: John Wiley & Sons, 2003.