基于径基函数-加权偏最小二乘回归的干点软测量

颜学峰[1]

摘 要 针对影响石油馏分产品干点因素众多且呈高度非线性的特征，提出了一种径基函数 (Radial basis function，RBF) 和加权偏最小二乘回归 (Weighted partial least squares regression，WPLSR) 相结合的建模方法建立干点软测量模型. 该组合方法首先应用 RBF 实现样本数据的非线性变换；然后根据非线性变换后样本在结构参数空间中的分布，分析它们对预测对象的预报能力，自适应地为各个样本分配权值，并进而从中提取和选用 PLS 成分，实施加权 PLSR，以获得预报性能良好的模型. 在实际应用于初顶石脑油干点软测量建模中，RBF-WPLSR 获得比 PLSR、WPLSR 及 RBF-PLSR 更高精度的模型.

关键词 径基函数，加权，偏最小二乘回归，干点
中图分类号 O212.4 TQ021.8

Radial Basis Function-weighted Partial Least Square Regression and Its Application to Develop Dry Point Soft Sensor

YAN Xue-Feng[1]

Abstract A novel approach of integrating the radical basis function (RBF) with weighted partial least squares regression (WPLSR) has proposed to develop the dry point sensor in petroleum distillation products. Many operation factors have effect on the dry point products and correlation among them. Firstly, this approach uses RBF to carry out the nonlinear transformation for the sample data. Secondly, the space distribution a of a nonlinear transformation sample data set is analyzed, and each nonlinear transformation sample is self-adaptively weighted according to its different ratios of predicting contribution for the predicting sample. Thirdly, PLSR is applied to weighted nonlinear transformation sample data set to remove the correlation and develop a model with high predicting precision. Finally, PLSR, WPLSR, RBF-PLSR and RBF-WPLSR are utilized to develop the naphtha dry point soft sensor. The comparison results show that the prediction by RBF-WPLSR is the most precise.

Key words Radial Basis Function, Weighted, Partial Least Square Regression, Dry Point

1 引言

炼油厂中馏分产品的干点是重要的质量指标，但该指标无合适的在线分析仪进行测定，只能离线分析且时间长. 因此建立预报性能良好的干点软测量模型，可以实现馏分产品干点质量指标的实时预测，并为生产操作优化提供指导[1]. 软测量模型的建立，已提出了很多方法. 有各种全局多元线性回归方法[2, 3]，具有简洁明确的解析表达形式，但难以描述高度非线性系统. 有多种神经网络方法[4, 5]，可以处理高度非线性体系，但易出现过拟合现象. 为此，提出了一种径基函数 (Radial basis function，RBF) 和加权偏最小二乘回归 (Weighted partial least squares regression，WPLSR) 相结合的建模方法. 该组合方法首先应用 RBF 实现样本数据的非线性变换；然后对于已知操作条件的工况 (称为预测对象)，将根据非线性变换后样本在结构参数空间中的分布，分析它们对预测对象的预报能力，自适应地为各个样本分配权值. 又考虑到变量间存在交互作用，将从加权样本数据中提取并选定参与回归的 PLS 成分，从而实施加权的偏最小二乘回归，以建立针对该预测对象的，预报性能良好的软测量模型. 将 RBF-WPLSR 应用于初馏塔石脑油干点软测量，获得满意的结果.

收稿日期 2005-8-26 收修改稿日期 2006-3-20
Received August 26, 2005; in revised form March 20, 2006
国家自然科学基金 (20506003)，教育部科学技术研究重点项目 (106073)，上海启明星项目 (04QMX1433) 资助
Supported by National Natural Science Foundation of P. R. China (20506003), Key Project of Chinese Ministry of Education (106073), Science and Technology Phosphor of Shanghai (04QMX1433)
1. 华东理工大学自动化研究所 上海 200237
1. Automation Institute, East China University of Science and Technology, Shanghai 200237
DOI: 10.1360/aas-007-0193

2 径基函数–加权偏最小二乘回归

2.1 径基函数网络

RBF 网络[6, 7] 本质是多变量插值的径基函数方法. 设给定一个 p 维点集 $\{\boldsymbol{x}_i\}$ 和与之对应的 q 维点集 $\{\boldsymbol{y}_i\}$，$i = 1, 2, \cdots, n$，插值问题就是要求一个函数 $f(\boldsymbol{x})$，使之满足插值条件：$f(\boldsymbol{x}_i) = \boldsymbol{y}_i$. 以一维输出为例，并采用高斯函数作为径基函数，则隐层输

出加权求和为

$$f(\boldsymbol{x}) = \sum_{j=1}^{m} \lambda_j \exp(-\|\boldsymbol{x} - \boldsymbol{c}_j\|^2/\sigma_j^2) \tag{1}$$

其中，λ_j 是插值系数；m 是径基个数；$\boldsymbol{c}_j$ 是径基向量；$\|\cdot\|$ 是欧氏范数；σ_j^2 是第 j 个隐节点的归一化参数.

2.2 加权偏最小二乘回归

2.2.1 偏最小二乘回归

设样本容量为 n，自变量维数为 p，因变量维数为 q，则自变量数据矩阵 X 为 $n \times p\,(n > p)$ 维，因变量数据矩阵 Y 为 $n \times q\,(n > q)$ 维. 针对各自变量间可能存在复共线性，可以采用偏最小二乘回归 (PLSR)[8, 9]，PLS 成分的提取可用 NIPALS 算法[9]. 设 T 是前 k 个 PLS 成分组成的 $n \times k$ 维隐变量矩阵，且有 $T = XU$，则 PLSR 模型

$$Y = TC + E = XUC + E \tag{2}$$

其中 U 是 $p \times k$ 维转换矩阵；C 是 $k \times q$ 维回归系数矩阵；E 是 $n \times q$ 维的残差矩阵. 在 NIPALS 算法中可以同时计算出 C 与 U，则 q 个因变量的预报值可用如下的回归方程计算

$$\hat{\boldsymbol{y}} = \boldsymbol{x}CU \tag{3}$$

对于采集的数据，由于种种原因，各个样本在建模中处于不同的地位，为反映这种情况，可以采用加权回归方法[10]，即对各个样本赋予不同的权值，此后式 (2) 回归模型为

$$WY = WXUC + E \tag{4}$$

其中 W 为对角矩阵 $diag(\sqrt{w_1}, \sqrt{w_2}, \cdots, \sqrt{w_n})$，要求权值 $w_i \geq 0\ (i = 1, 2, \cdots, n)$.

2.2.2 加权偏最小二乘回归算法

WPLSR 以被预测的对象为出发点，并认为在建立适用于该预测对象的模型时，各个样本处于不同的地位，应分配不同的权值. 而权值的大小将根据样本与预测对象之间欧氏距离的大小自适应地进行分配. 即，样本和预测对象之间的欧氏距离越小，则认为它对预测对象的线性预测能力也越大，在建模中应分配较大的权值[8~11].

依据样本和预测对象之间欧氏距离的大小可以有多种权值分配方案，本文拟采用一种较为简单易行的方案：设权值分配参数为 m (m 为正整数)，则与预测对象欧氏距离最近的前 m 个样本权值为 1，其余样本权值为 0. WPLSR 算法的计算步骤如下：

1) 设提取的隐变量 (PLS 成分) 数为 k ($k = 1, 2, \cdots, p$)；权值分配参数为 m ($m = k+1, k+2, \cdots, n-1$).

2) 对于每对 k 和 m 值，采用留一交叉验证法，算出所有校验样本相对误差平方和.

2.1) 从 n 个样本中选出第 j 个样本 $\boldsymbol{x}_j$ ($j = 1, 2, \cdots, n$) 为校验样本，其余的均为建模样本.

2.2) 在 $n-1$ 个建模样本 $\boldsymbol{x}_i$ ($i = 1, 2, \cdots, j-1, j+1, \cdots, n$) 中，与校验样本 $\boldsymbol{x}_j$ 距离最近的 m 个样本权值为 1，其余样本权值为 0，组成权值矩阵 W.

2.3) 对建模样本的自变量数据矩阵 X 和因变量数据矩阵 Y (不含校验样本) 进行加权处理，求得 $m \times p$ 维的自变量矩阵 $X_m = WX$ 和 $m \times q$ 维的因变量矩阵 $Y_m = WY$.

2.4) 采用 NIPALS 算法从 X_m 中提取前 k 个 PLS 成分，构成 $m \times k$ 维隐变量数据矩阵 $T_{m,k}$，并求出对应的 $p \times k$ 维转换矩阵 $U_{m,k}$ 和 $k \times q$ 维回归系数矩阵 $C_{m,k}$.

2.5) 计算校验样本 $\boldsymbol{x}_j$ 的预报值矢量 $\hat{\boldsymbol{y}}_j = C_{m,k}^T U_{m,k}^T \boldsymbol{x}_j$，其第 l 个分量为 $\hat{y}_{jl}$. 当 j 依次取 1 到 n 的所有值时，将对每个样本进行一次预报，对于每对 m 和 k 值，按下式计算相应的预报相对误差平方和，并记为 $E_{m,k}$

$$E_{m,k} = \sum_{j=1}^{n} \sum_{l=1}^{q} \left(\frac{y_{jl} - \hat{y}_{jl}}{y_{jl}}\right)^2 \tag{5}$$

其中 y_{jl} 是因变量数据矩阵 $\boldsymbol{Y}$ 中的元素，它们是样本的实际观测值.

3) 将每个 $E_{m,k}$ 值存入相对误差矩阵 R 的第 m 行第 k 列，在矩阵其他空白位置上可添入 $+\infty$. 选取误差矩阵 R 中最小值的元素，它所在的行号和列号分别记为 $m_{optimal}$ 和 $k_{optimal}$.

4) 对于自变量取值为 $\boldsymbol{x}^*$ 的预测对象，将权值分配参数取为 $m_{optimal}$，隐变量数取为 $k_{optimal}$，按 2) 步中 2.2) - 2.5) 步建立专门用于预报预测对象 $\boldsymbol{x}^*$ 的回归模型.

2.3 径基函数–加权偏最小二乘回归算法

径基函数–加权偏最小二乘回归相结合的组合建模方法克服 RBF 网络隐层节点个数难以确定的缺点，并通过自适应加权偏最小二乘回归提高模型预报精度. 算法执行步骤如下：

1) 样本数据的 RBF 非线性变换. 将每个训练样本作为 RBF 函数的基向量 ($\boldsymbol{c}_j = \boldsymbol{x}_j$，$j = 1, 2, \cdots, n$)，对输入矢量 $\boldsymbol{x}$ 执行非线性变换，形成变换矢量 $\boldsymbol{x}_A$，即

$$x_j^A = \exp(-\|\boldsymbol{x} - \boldsymbol{c}_j\|^2/\sigma_j^2) \tag{6}$$

其中，x_j^A 是 $\boldsymbol{x}_A$ 的第 j 个元素. $\sigma_j = \frac{e}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$，$j = 1, 2, \cdots, n$，其中，$e$ 是大于 0 的常数. 对所有样本 $\boldsymbol{x}_i, i = 1, 2, \cdots, n$ 执行 (6) 式的非线性变换，形成非线性变换后样本数据 $\boldsymbol{x}_{A,i}, i = 1, 2, \cdots, n$.

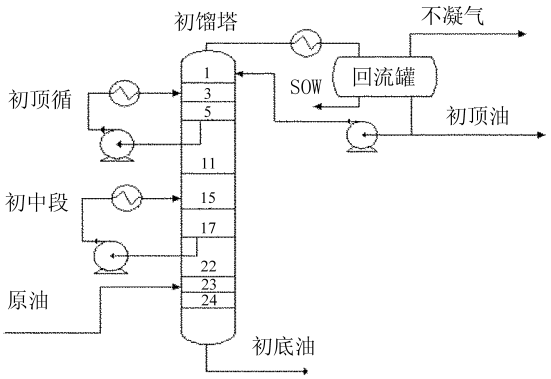


图 1　初馏塔的工艺流程图
Fig. 1　Flow chart of preflash tower

2) 按 WPLSR 算法的 1) - 3) 步，由 $\boldsymbol{y}_i$ 和 $\boldsymbol{x}_{A,i}, i = 1, 2, \cdots, n$，求得 $m_{optimal}$ 和 $k_{optimal}$.

3) 对于自变量取值为 $\boldsymbol{x}^*$ 的预测对象，通过 (6) 式形成非线性变换矢量 $\boldsymbol{x}_A^*$，将权值分配参数取为 $m_{optimal}$，隐变量数取为 $k_{optimal}$，按 WPLSR 算法的 2) 步中 2.2) - 2.5) 步建立专门用于预报预测对象 $\boldsymbol{x}^*$ 的回归模型.

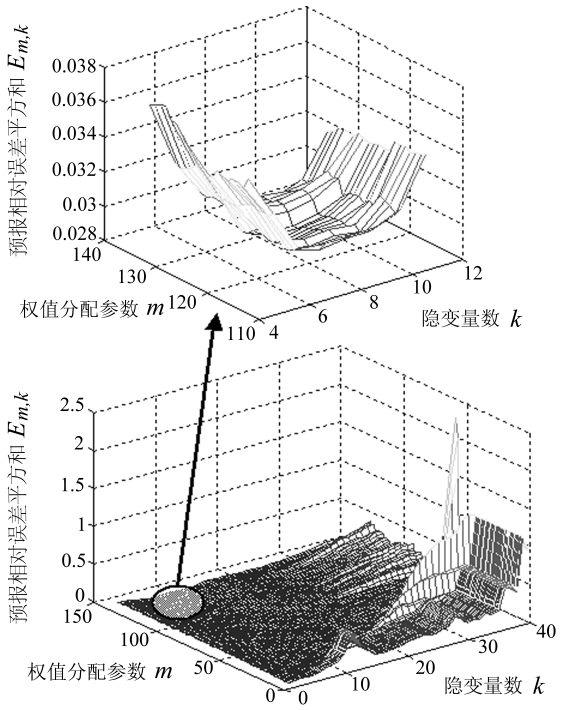


图 2　$E_{m,k}$ 随权值分配参数 m 与隐变量数 k 变化曲面
Fig. 2　Surface graph of $E_{m,k}$ with m and k

3　初顶石脑油干点软测量

3.1　影响因素分析和自变量的选取

某常减压装置初馏塔的工艺流程如图 1 所示. 其中影响初顶石脑油干点 y 的因素主要有：初馏塔处理量 x_1、塔顶温度 x_2、塔顶压力 x_3、顶回流带出能量 x_4、回流比 x_5、初顶石脑油流量 x_6、初顶循带出能量 x_7、初中段带出能量 x_8、及进料温度 x_9，这些影响因素值都可以实时通过 DCS 系统获得. 同时炼制原油性质也直接影响初顶石脑油干点，但其性质经常无法及时获得. 为此，提出引入前一时刻初顶石脑油干点人工分析值作为软测量模型的第十个自变量 x_{10}，以间接表征炼制原油性质.

将以上分析的影响初顶石脑油干点 y 的 9 个主要因素 $(x_1, x_2, \cdots, x_9)$ 和前一时刻初顶石脑油干点人工分析值 x_{10} 作为软测量模型的自变量，初顶石脑油干点 y 作为因变量，采集到 138 个样本数据，形成 138×10 维自变量样本数据矩阵 X 和 138×1 维因变量数据矩阵 Y.

3.2　最佳模型参数的确定

采用 RBF-WPLSR 算法对 X 和 Y 进行回归建模计算，求得误差矩阵 R，获得图 2 所示的预报相对误差平方和 $E_{m,k}$ 随权值分配参数与隐变量数变化曲面. 图中网格曲面的每个交叉点表示一对权值分配参数 m 和隐变量数 k 对应的 $E_{m,k}$. 图 2 的上半部分是下半部分的子图，是图 2 下半部分中圆圈部分的放大. 从图 2 中可以看出，当隐变量数取定值时，$E_{m,k}$ 值随着权值分配参数 m 值的增大，先呈下降趋势，后又开始上升；当权值分配参数取定值时，$E_{m,k}$ 值随着隐变量数的增加，先呈下降趋势，后又开始上升. 在 $m = 127, k = 8$ 时，$E_{m,k}$ 值达到最小值，即 $E_{127,8} = 0.02843$. 则，最佳权值分配参数为 $m_{optimal} = 127$，最佳隐变量数为 $k_{optimal} = 8$.

3.3　软测量模型性能分析

初顶石脑油干点软测量模型为：对预测对象 $\boldsymbol{x}^*$，首先，通过 (6) 式形成非线性变换矢量 $\boldsymbol{x}_A^*$；然后，基于采集的经过 (6) 式非线性变换的 138 个样本数据，从中挑选与 $\boldsymbol{x}_A^*$ 空间最邻近的 $m_{optimal} = 127$ 个样本分配权值 1，其余的样本分配权值 0，形成加权建模样本. 最后，基于加权建模样本，采用PLSR选取 $k_{optimal} = 8$ 个隐变量，建立专门用于预报预测对象 $\boldsymbol{x}^*$ 的回归模型. 因此，不同的预测对象，将根据样本数据对预测对象的预报能力，对样本进行自适应加权形成建模样本数据，建立不同的模型.

3.2 节中模型最佳参数确定是采用留一交叉验证法进行的，为此，为了直观说明建立的干点软测量模型 ($m_{optimal} = 127, k_{optimal} = 8$) 的性能，以初顶石脑油干点人工分析值为横坐标，以干点软测量模型对 138 个样本的预测值 (即，留一交叉验证的校验值) 为纵坐标作图，结果如图 3 所示. 从图 3 可以

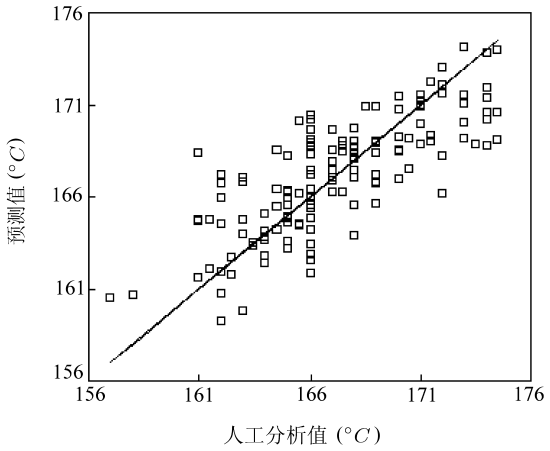


图 3 干点软测量模型的预测值与人工分析值对比
Fig. 3 Analysis value of dry point versus predicting value

看出，图中方形点分布在对角线的两侧 (若方形点落在对角线上，则预测值与分析值相同)，模型具有良好预测性能.

同时，为了进一步分析 RBF-WPLSR 模型的性能，在仿真试验中分别采用 PLSR、WPLSR 及 RBF-PLSR 等 (组合) 回归算法对采集到 138 个样本数据进行回归建模，由留一交叉验证法求得各自的最优预报相对误差平方和 E_{opt} 如表 1 所示. 从表 1 中可以看出，RBF-WPLSR 回归算法建立模型的预报性能明显优于其他三种 (组合) 回归算法.

表 1 不同回归算法的最优预报相对误差平方和
Table 1 Comparison result of optimal predicting performances of different models

回归算法	全局PLSR	WPLSR	RBF-全局PLSR	RBF-WPLSR
E_{opt}	0.04706	0.03849	0.02856	0.02843

4 结论

径基函数 (Radial basis function，RBF) 和加权偏最小二乘回归 (Weighted partial least squares regression，WPLSR) 相结合的建模方法具有很强非线性表达能力，建立的模型预测精度高且稳定. RBF-WPLSR 首先应用 RBF 实现样本数据的非线性变换；然后采用具有自适应加权 PLSR 对非线性变换后的样本数据进行回归建模. 在实际应用于初顶石脑油干点软测量建模中，RBF-WPLSR 算法预报精度明显优于 PLSR、WPLSR 及 RBF-PLSR 等 (组合) 回归算法.

References

1 Chatterjee T, Saraf D N. On-line estimation of product properties for crude distillation units. *Journal of Process Control*, 2004, **14**(1): 61∼77

2 Bastion P, Esposito V V, Michel T. PLS generalised linear regression. *Computational Statistics and Data Analysis*, 2005, **48**(1): 17∼46

3 Eliana Z, Massimiliano B, Seborg Dale E. Estimating product composition profiles in batch distillation *via* partial least squares regression. *Control Engineering Practice*, 2004, **12**(7): 917∼929

4 Kishore V R S, Kishore R, Doble M. Neural network modeling of hydantoinase production. In: Proceedings of International Conference on Intelligent Sensing and Information Processing, IEEE, Chennai, India, 2004, 456∼459

5 Xiilia M G, Barbalace N. Sulphur recovery unit modelling via stacked neural network. In: Proceedings of the IASTED (International Association of Science and Technology for Development) International Conference on Applied Simulation and Modelling, Rhodes, Greece, 2004, 65∼70

6 Hunt K J, Haas R, Murray-Smith R. Extending the functional equivalence of radial basis function networks and fuzzy inference systems. *IEEE Transactions on Neural Networks*, 1996, **7**(3): 776∼781

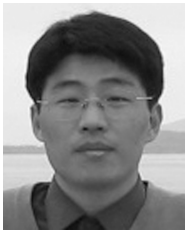
7 Chng E S, Chen S, Mulgrew B. Gradient radial basis function networks for nonlinear and nonstationary time series prediction. *IEEE Transactions on Neural Networks*, 1996, **7**(1): 190∼194

8 Chen De-Zhao. *Multivariate Processing*. Beijing: Chemical Industry Press, 1998, 191∼200
(陈德钊. 多元数据处理. 北京: 化学工业出版社, 1998, 191∼200)

9 Geladi P, Kowalski B R. Partial Least-squares regression: A tutorial. *Anlytica Chemica Acta*, 1986, **185**: 1∼17

10 Centner V, Massart D L. Optimization in locally weighted regression. *Analytical Chemistry*, 1998, **711**(19): 4206∼4211

11 Frank I E. A nonlinear PLS model. *Chemometrics and Intelligent Laboratory Systems*, 1990, **8**: 109∼119



颜学峰 华东理工大学自动化研究所副研究员，2002 年博士毕业于浙江大学，主要研究方向为复杂系统建模与优化、智能信息处理. E-mail: xfyan@ecust.edu.cn

(**Yan Xue-Feng** Received his Ph. D. degree from Zhejiang University. He is now an associate professor of Automation Institute, East China University of Science and Technology. His research interests include complex chemical process modeling and optimizing, and intelligent information processing.)