

基于非线性偏鲁棒 M-回归的萃余液 pH 值软测量

贾润达¹ 毛志忠^{1,2} 常玉清^{1,2}

摘要 提出了一种径向基函数网络 (Radial basis function networks, RBFNs) 与偏鲁棒 M-回归 (Partial robust M-regression, PRM) 相结合的非线性 PRM (Nonlinear PRM, NLPRM) 建模方法, 用以解决鲁棒非线性系统建模问题. 该方法首先通过 RBF 变换获得扩展的输入数据矩阵; 接下来 PRM 算法通过反复迭代计算, 自适应地为变换后的数据分配不同的连续权值, 用以克服离群点对模型的影响. 本文通过仿真实验, 验证了方法的有效性; 并将其应用于湿法冶金萃取过程萃余液 pH 值软测量建模问题, 获得了相比于偏最小二乘法 (Partial least squares, PLS)、PRM 以及 RBF-PLS 方法更高的预测精度.

关键词 湿法冶金, 萃取, 径向基函数网络, 偏鲁棒 M-回归, 软测量
中图分类号 O212.4 TQ021.8

Soft Sensing for pH Value of Raffinate Solution Based on Nonlinear Partial Robust M-regression

JIA Run-Da¹ MAO Zhi-Zhong^{1,2} CHANG Yu-Qing^{1,2}

Abstract A nonlinear partial robust M-regression (NLPRM) modeling method combining radial basis function networks (RBFNs) and partial robust M-regression (PRM) is presented to solve the robust nonlinear system modeling problem. First, an extended input data matrix is formed by RBF transforming. Then, PRM algorithm is used, so that through iterative computation, consecutive weights are adaptively distributed to transformed data to diminish the effects of outliers on modeling. Simulation experiment is performed to show the effectiveness of this method. And it is utilized to develop a soft sensor model for pH value of raffinate solution in hydrometallurgy extraction process, and highly precise prediction results are obtained compared with partial least squares (PLS), PRM, and RBF-PLS methods.

Key words Hydrometallurgy, extraction, radial basis function networks (RBFNs), partial robust M-regression (PRM), soft sensor

萃取是湿法冶金过程重要的除杂手段, 根据有价金属与杂质金属在性质上的差别, 利用有机溶剂达到分离除杂的目的. 萃余液的 pH 值是反映萃取过程的重要指标, 较低的 pH 值会使杂质去除的不够彻底, 而较高的 pH 值则会造成有价金属的流失. 由于料液对 pH 计存在严重的腐蚀性, 以及萃取过程所具有的大滞后、非线性特性^[1], 建立一个预报性能良好的软测量模型, 实现萃余液 pH 值的实时预测, 对生产过程的优化控制具有十分重要的意义.

为了解决非线性系统的建模问题, 提出了一种径向基函数网络 (Radial basis function networks, RBFNs) 与偏最小二乘法 (Partial least squares, PLS) 相结合的非线性 PLS 建模方法^[2], 该方法利

用 PLS 算法克服了由于 RBF 变换所引发的多重共线性, 并能有效抑制样本数据中正态分布的噪声. 为了提高预测精度, 相继提出了一些改进方法^[3-5], 并实际应用于非线性系统的软测量建模问题, 取得了较好的预测效果. 然而来自工业现场的数据往往受到各种形式的干扰 (如传感器故障、过程扰动等), 这直接导致样本数据中存在离群点 (Outliers). 由于普通 PLS 算法易受离群点影响^[6], 因此上述基于普通 PLS 算法的非线性方法也会由于离群点的出现而失去其应有的泛化能力.

为此, 本文提出了一种 RBFNs 与偏鲁棒 M-回归 (Partial robust M-regression, PRM) 相结合的非线性 PRM (Nonlinear PRM, NLPRM) 建模方法. 该方法首先通过 RBF 变换获得扩展的输入数据矩阵, 该矩阵由输入数据矩阵以及激活矩阵^[2]组成, 在实现样本数据从非线性关系到线性关系转变的同时, 也便于识别数据中的离群点; 接下来 PRM 算法通过反复迭代计算, 自适应地为变换后的数据分配不同的权值 (为离群点分配接近于 0 的权值, 为正常数据点分配接近于 1 的权值), 用以克服离群点对模型的影响, 同时保留离群点所提供的部分有用信息. 最后本文还通过仿真实验, 验证了方法的有效

收稿日期 2008-7-17 收修改稿日期 2008-10-28
Received July 17, 2008; in revised form October 28, 2008
国家高技术研究发展计划 (863 计划) (2006AA060201) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2006AA060201)
1. 东北大学信息科学与工程学院 沈阳 110004 2. 东北大学流程工业综合自动化教育部重点实验室 沈阳 110004
1. School of Information Science and Engineering, Northeastern University, Shenyang 110004 2. Key Laboratory of Integrated Automation of Process Industry, Northeastern University, Ministry of Education, Shenyang 110004
DOI: 10.3724/SP.J.1004.2009.00583

性; 并将其实际应用于湿法冶金萃取过程萃余液 pH 值软测量建模问题, 获得了相比于 PLS^[7]、PRM^[8] 以及 RBF-PLS 方法^[2] 更高的预测精度。

1 偏鲁棒 M-回归算法

1.1 偏最小二乘法

偏最小二乘法是一种多元线性回归方法, 用以解决输入数据间存在多重相关性的建模问题. 设有 $n \times m$ 维输入数据 X (假设已进行标准化处理) 和 $n \times 1$ 维输出数据 \mathbf{y} , 其中 n 代表样本数据的个数, m 代表输入数据的维数. 若 T 是由前 k 个得分向量组成的 $n \times k$ 维得分矩阵, 则 PLS1 (单因变量 PLS) 模型可以利用下式进行描述

$$X = TP^T + E \quad (1)$$

$$\mathbf{y} = T\mathbf{q} + \mathbf{r} = X\mathbf{b} + \mathbf{r} \quad (2)$$

其中 P 是相应的载荷矩阵, \mathbf{q} 是 k 个得分向量的回归系数向量, E 和 \mathbf{r} 分别是相应的残差矩阵和向量, \mathbf{b} 是 PLS1 模型的回归系数向量. 常用的 PLS 算法有 NIPALS (Nonlinear iterative partial least squares) 算法^[7] 和 SIMPLS (Simple partial least squares) 算法^[9], 由于后者的计算速度较快, 因此在接下来的 PRM 方法中, 选用 SIMPLS 算法进行回归计算.

1.2 偏鲁棒 M-回归算法

偏鲁棒 M-回归是一种鲁棒形式的 PLS1 算法, 源于迭代再加权 PLS (Iteratively reweighted PLS, IRPLS)^[10]. 该方法通过反复迭代计算, 自适应地为样本数据分配不同的权值, 用以消除离群点对回归模型的影响. 该方法将离群点分成两类, 一类是高杠杆点, 另一类是高残差点. 高杠杆点是远离输入数据中心的样本点; 而高残差点是输出预测值与实际值相差较大的样本点. PRM 算法利用不同的加权方法对上述两类离群点进行加权处理. 设第 i 个样本数据的杠杆权值 w_i^x , 可由下式进行定义

$$w_i^x = f\left(\frac{\|\mathbf{t}_i - \text{med}_{L1}(T)\|}{\text{med}_i\|\mathbf{t}_i - \text{med}_{L1}(T)\|}, c\right), \quad i=1, 2, \dots, n \quad (3)$$

且

$$f(z, c) = \frac{1}{\left(1 + \left|\frac{z}{c}\right|\right)^2} \quad (4)$$

其中 $\|\cdot\|$ 代表欧氏距离, med 代表中位值, med_{L1} 代表 L1 中位值^[11], \mathbf{t}_i 是第 i 个样本数据的 PLS 得分 (T 的第 i 行), c 是常数 (通常取 4). 设第 i 个样本数据的残差权值 w_i^r , 可由下式进行定义

$$w_i^r = f\left(\frac{r_i}{\bar{r}}, c\right) \quad (5)$$

其中 r_i 代表第 i 个样本数据预测值与实际值之间的残差, \bar{r} 代表残差的鲁棒尺度估计^[8], 可由下式进行计算

$$\bar{r} = \text{med}_i|r_i - \text{med}_j(r_j)|, \quad i, j = 1, 2, \dots, n \quad (6)$$

综合考虑上述两种权值, 则第 i 个样本数据的权值 w_i , 可由下式进行确定

$$w_i = \sqrt{w_i^x w_i^r} \quad (7)$$

上述鲁棒 PLS1 模型反复建立, 样本数据的权值不断更新, 直到算法满足收敛条件. PRM 算法的步骤简述如下:

步骤 1. 利用式 (3)、(5) 和 (7) 初始化权值 w_i ;

步骤 2. 对输入数据 X 和输出数据 \mathbf{y} 进行加权处理, 得到加权后的输入数据 X_W 和输出数据 \mathbf{y}_W , 并对加权后的样本数据建立 PLS1 回归模型, 同时修正得分向量;

步骤 3. 计算每个数据样本的残差 r_i , 利用式 (3)、(5) 和 (7) 更新权值 w_i ;

步骤 4. 如果连续两次计算出 \mathbf{q} 的相对差小于某一阈值 (如 10^{-2}), 则算法终止, 否则返回步骤 2.

2 非线性偏鲁棒 M-回归算法

由于 PRM 算法本质上一种线性方法, 因此它并不能描述系统的非线性特性. 然而许多实际生产过程具有很强的非线性特性 (如萃取过程), 因此仅仅利用线性鲁棒回归方法难以建立高精度的软测量模型. 为此, 本文提出了一种 RBFNs 与 PRM 相结合的 NLPRM 建模方法.

2.1 NLPRM 模型结构

NLPRM 模型由 RBFNs 与 PRM 结合而成, RBFNs 已被证明可以以任意精度逼近任意连续函数^[12], 因此可以利用 RBFNs 来描述系统的非线性特性. 仿造 RBF-PLS 方法^[2], 选用高斯径向基函数, 将自变量数据矩阵 X 转化为激活矩阵 X_A . X_A 的元素可以用下式进行定义

$$a_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{c}_j\|^2}{\sigma_j^2}\right), \quad i, j = 1, 2, \dots, n \quad (8)$$

其中 \mathbf{x}_i 是第 i 个数据样本的输入向量, a_{ij} 是 X_A 第 i 行、第 j 列的元素, \mathbf{c}_j 和 σ_j 分别是高斯函数的中心和宽度参数. 中心参数 \mathbf{c}_j 选为每个数据样本的输入向量, 即

$$\mathbf{c}_j = \mathbf{x}_j \quad (9)$$

而宽度参数 σ_j 可由下式进行计算^[4]

$$\sigma_j = \frac{\mu}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|, \quad i, j = 1, 2, \dots, n \quad (10)$$

这里 μ 为大于 0 的常数 (通常取 1). 因此矩阵 X_A 是一个对角元素为 1 的 $n \times n$ 维方阵.

为了提高模型的预测精度以及对离群点的敏感度, 将输入数据矩阵以及激活矩阵组成扩展的输入数据矩阵 $X_E = [X \ X_A]$, 而输出数据仍为 y , 采用扩展输入数据矩阵的优势将在接下来的仿真研究中得以验证.

在进行上述变换之后, 利用 PRM 算法建立扩展的输入数据矩阵与输出数据之间的线性鲁棒回归模型. 最终的 NLPRM 模型可以利用下式进行描述

$$y = X_E b^{\text{PRM}} \quad (11)$$

其中 b^{PRM} 是利用 PRM 算法求得的回归系数.

2.2 得分向量个数的确定

得分向量的提取个数对于 NLPRM 模型的建立至关重要, 通常的 PLS 采用留一交叉检验^[7]的方式进行确定. 但当数据样本中存在离群点的时候, 离群点会产生较大的残差, 因此鲁棒交叉检验^[13]是一种更好的得分向量个数确定方法. 该方法仅计算正常数据样本的交叉检验均方根误差 (Root mean square error of cross-validation, RMSECV), 称为鲁棒 RMSECV (Robust RMSECV, R-RMSECV), 而忽略离群点对交叉检验的影响.

2.3 NLPRM 算法步骤

为了消除离群点对模型的影响, PRM 算法对变换后的输入输出数据进行加权处理, 为离群点分配接近于 0 的权值, 而为正常数据点分配接近于 1 的权值. 这种加权处理不仅克服了离群点对模型的影响, 同时也保留离群点所提供的部分有用信息. NLPRM 算法步骤如下:

步骤 1. 对输入数据进行标准化处理, 得到输入数据矩阵 X ;

步骤 2. 利用式 (8)~(10) 对标准化处理后的输入数据进行 RBF 变换, 得到激活矩阵 X_A , 并与原输入数据矩阵组成扩展的输入数据矩阵 X_E , 而输出数据仍为 y ;

步骤 3. 对扩展的输入数据矩阵 X_E 和输出数据 y 实施 PRM 回归;

步骤 4. 计算鲁棒交叉检验均方根误差 (R-RMSECV) 值, 判断是否达到最小, 若是, 则停止计算, 记录此时的回归系数; 否则返回步骤 3, 将得分向量提取个数 k 增加 1, 重新计算.

3 仿真研究

考虑利用 NLPRM 算法去逼近正弦目标函数

$$y = \sin(2\pi x) + \varepsilon, \quad 0 \leq x \leq 1 \quad (12)$$

这里 ε 是服从 $N(0, 0.02)$ 分布的噪声; 在 $[0, 1]$ 之间随机产生 30 个数据对用于建模, 从上述数据对中随机选取 4 个, 分别为第 2、9、15、21 个 (在图 1 中已经标明), 对第 2、21 个数据对的输入分别加入 30% 的扰动, 对第 9、15 个数据对的输出分别加入 30% 的扰动, 则这 4 个数据样本即可视为离群点; 另外再产生 $[0, 1]$ 之间均匀分布的 101 个数据对作为测试数据. 在图 1 中, 将 RBF-PLS 方法^[2]与 NLPRM 方法进行比较, 可以看到, RBF-PLS 方法所建立的模型因受到离群点的影响而偏向于离群点; 而 NLPRM 方法对于离群点则具有较好的鲁棒性.

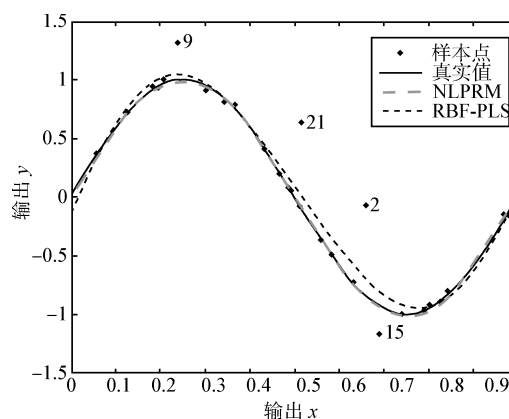


图 1 预测效果比较

Fig. 1 Comparing prediction results

图 2 绘制了鲁棒交叉检验均方根误差 (R-RMSECV) 随得分向量提取个数变化的情况 (为了清晰, 仅绘制了前 10 个得分向量, 之后将继续增大), 从图中可以看出当得分向量提取个数为 6 时获得最小的 R-RMSECV 值, 而此时 RBF-PLS 方法得分向量的提取个数为 3. 两种算法的预测均方根误差 (Root mean square error, RMSE) 分别为 0.0779 和 0.0044, NLPRM 方法明显好于 RBF-PLS 方法.

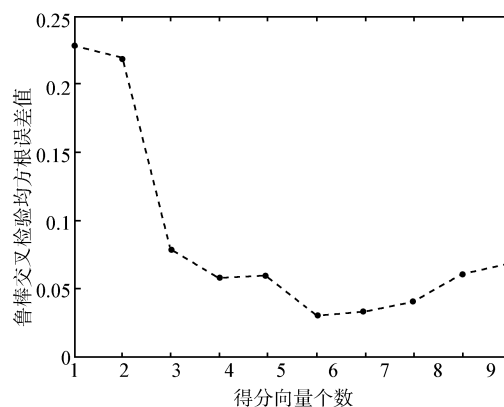


图 2 鲁棒交叉检验均方根误差曲线

Fig. 2 Robust root mean square error of cross-validation curve

为了进一步验证采用扩展输入数据矩阵的优势,在表 1 中同时列出了几种方法的预测均方根误差比较结果.可以看出利用激活矩阵作为输入数据矩阵的非线性鲁棒建模方法 (RBF-PRM),其预测精度虽然好于 RBF-PLS 方法,但却与 NLPRM 方法存在一定的差距.

表 1 预测均方根误差比较

Table 1 Comparing of prediction root mean square error

	RBF-PLS ^[2]	RBF-PRM	NLPRM
RMSE	0.0779	0.0084	0.0044
得分向量个数	3	6	6

在图 3 中绘制了利用 NLPRM 方法获取的各个样本点的权值,可以看到 4 个最小的权值 (分别为 0.0655, 0.1494, 0.1825, 0.0767) 分配给了上述 4 个离群点,因此离群点对模型的影响被很好剔除.利用 RBF-PRM 方法,4 个离群点对应的权值 (分别为 0.0679, 0.1545, 0.1912, 0.0790) 均大于 NLPRM 方法;而对于余下的 26 个非离群点, RBF-PRM 方法获取的权值均值 (其值为 0.6500) 却小于 NLPRM 方法获取的权值均值 (其值为 0.6525).由此可见采用扩展的输入数据矩阵,模型对离群点的敏感度亦有所提高.

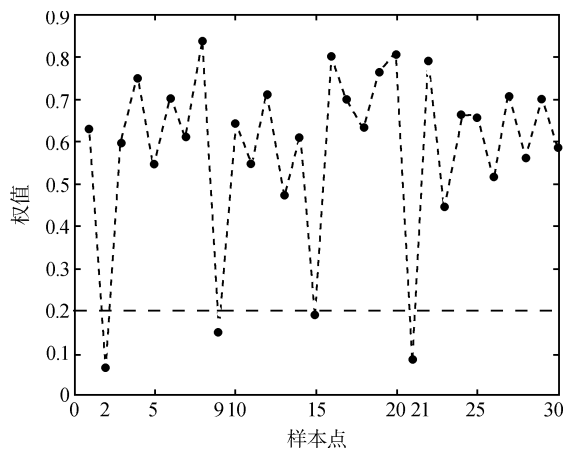


图 3 样本点的权值

Fig. 3 Weights of sample data points

4 基于 NLPRM 的萃余液 pH 值软测量

4.1 过程描述及辅助变量的选择

萃取的目的是除去料液中的有害金属杂质,得到纯净的有价值金属溶液.通常这一过程由多个萃取过程共同完成.在不同的萃取过程,选用不同的有机溶剂,除去相应的金属杂质.本文以北京矿冶研究总院永清钴湿法冶金 P204 萃取除杂过程为例进行研究.

为了除去料液中含量较高的锌、锰、铁杂质,采用分馏萃取的方式,它由以下三段组成:萃取段、洗涤段以及反萃段,如图 4 所示,而每一段又包含几个级.在萃取段,由于钴与其他金属杂质性质上的差别,使得较多的金属杂质和一部分钴同时进入有机相;在洗涤段,控制一定的洗涤条件,可以使洗下的钴远多于其他金属杂质;在反萃段,金属杂质重新返回水相,使有机相得以再生,循环使用.

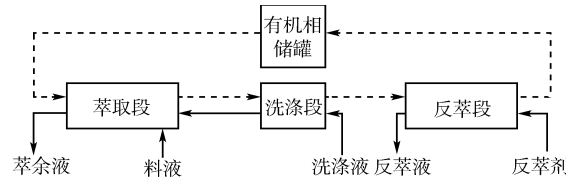


图 4 萃取过程工艺流程图

Fig. 4 Flow chart of extraction process

通过分析,影响 P204 萃取过程萃余液 pH 值的主要因素有:有机相的流量、料液的流量、洗涤液的流量、实测皂化率、料液的 pH 值以及温度.将以上 6 个影响因素作为软测量模型的自变量,将萃余液 pH 值作为因变量,采集到 106 组数据.选取其中 76 组用于建模,组成输入数据矩阵 X 以及输出数据 y ,余下的 30 组数据用于预测.

4.2 软测量模型性能分析

为了直观地说明所建立的 pH 值软测量模型的效果,图 5 以萃余液 pH 值的观测值作为横坐标,以软测量模型的预测值作为纵坐标对余下的 30 组数据进行预测.可以看到,图 5 中的点均匀地分布在对角线的两侧,表明该模型具有良好的预测性能.

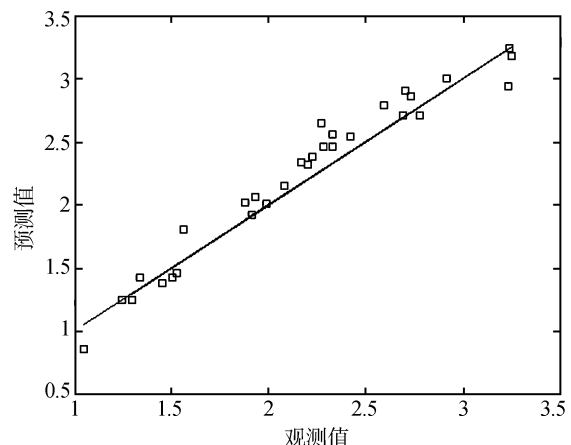


图 5 软测量模型的预测值与测量值对比

Fig. 5 Prediction values of soft sensor model versus observation values

同时,为了进一步检验 NLPRM 方法的性能,分别利用 PLS、PRM 以及 RBF-PLS 方法在相同的条件下进行建模,并将它们的 RMSE 与得分向量

提取个数列于表 2 中进行比较. 从表 2 中可以看到, NLPRM 模型的预报性能明显优于其他三种算法, 而 PRM 模型的效果也优于另外两种方法. 经过分析发现, 在建模数据中由于 pH 计的腐蚀导致其中几个样本点 (离群点) 不能真实地反映过程的本质. 因此 RBF-PLS 方法出现了明显的过拟合现象, 使其在所有方法中效果最差. 而 PLS 方法本质是一种线性方法, 相比于 RBF-PLS 方法, 过拟合现象稍微有所缓解, 但仍会受到上述离群点的影响, 且难以描述过程的非线性特性. 由于 PRM 方法是一种鲁棒线性回归方法, 因此受离群点的影响并不严重. 但由于过程具有较强的非线性, NLPRM 方法表现出了更好的预测效果. 同时也发现, 在鲁棒算法引入之后, 由于克服了离群点对模型的影响, 得分向量的提取个数显著增加, 这正是预测精度提高的原因所在.

表 2 pH 值预测结果比较

Table 2 Comparing prediction results of pH value

	PLS ^[7]	PRM ^[8]	RBF-PLS ^[2]	NLPRM
RMSE	0.2843	0.1296	0.3077	0.0794
得分向量个数	2	5	3	7

5 结论

结合径向基函数网络与偏鲁棒 M-回归, 本文提出了一种非线性 PRM 方法, 用以解决鲁棒非线性系统的建模问题. 该方法首先通过 RBF 变换获得扩展的输入数据矩阵; 接下来利用 PRM 算法, 自适应地为变换后的数据分配不同的连续权值, 建立系统的鲁棒非线性模型. 本文还通过仿真实验, 验证了方法的有效性; 并实际应用于湿法冶金萃取过程萃余液 pH 值的软测量建模问题, 获得了较高的预测精度. 本文所提出的 NLPRM 建模方法, 也同样适用于其他鲁棒非线性系统的软测量建模问题.

References

- 1 Komulainen T, Pekkala P, Rantala A, Jämsä-Jounela S L. Dynamic modelling of an industrial copper solvent extraction process. *Hydrometallurgy*, 2006, **81**(1): 52–61
- 2 Walczak B, Massart D L. The radial basis function-partial least squares approach as a flexible non-linear regression technique. *Analytical Chimica Acta*, 1996, **331**(3): 177–185
- 3 Wang Y, Rong G, Wang S Q. PLS algorithm for radial basis function networks. In: Proceedings of the 37th IEEE Conference on Decision and Control. Tampa, USA: IEEE, 1998. 4748–4753
- 4 Yan X F, Chen D Z, Hu S X. Chaos-genetic algorithms for optimizing the operating conditions based on RBF-PLS model. *Computers and Chemical Engineering*, 2003, **27**(10): 1393–1404
- 5 Yan Xue-Feng. Radial basis function-weighted partial least squares regression and its application to develop dry point soft sensor. *Acta Automatica Sinica*, 2007, **33**(2): 193–196

(颜学峰. 基于径基函数-加权偏最小二乘回归的干点软测量. 自动化学报, 2007, **33**(2): 193–196)

- 6 Lin B, Recke B, Kundsén J K H, Jørgensen S B. A systematic approach for soft sensor development. *Computers and Chemical Engineering*, 2007, **31**(5-6): 419–425
- 7 Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 2001, **58**(2): 109–130
- 8 Serneels S, Croux C, Filzmoser P, Van Espen P J. Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems*, 2005, **79**(1-2): 55–64
- 9 de Jong S. An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 1993, **18**(3): 251–263
- 10 Kruger U, Zhou Y, Wang X, Rooney D, Thompson J. Robust partial least squares regression: Part I, algorithmic developments. *Journal of Chemometrics*, 2008, **22**(1): 1–13
- 11 Daszykowski M, Kaczmarek K, Heyden Y V, Walczak B. Robust statistics in data analysis — a review basic concepts. *Chemometrics and Intelligent Laboratory Systems*, 2007, **85**(2): 203–219
- 12 Walczak B, Massart D L. Local modelling with radial basis function networks. *Chemometrics and Intelligent Laboratory Systems*, 2000, **50**(2): 179–198
- 13 Pell R J. Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems*, 2000, **52**(1): 87–104

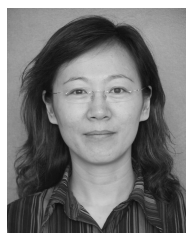


贾润达 东北大学信息科学与工程学院博士研究生. 2004 年获大连理工大学电子与信息工程学院学士学位. 主要研究方向为复杂系统建模与优化. 本文通信作者. E-mail: jiarunda@yahoo.com.cn (JIA Run-Da Ph. D. candidate at the School of Information Science and Engineering, Northeastern University.

He received his bachelor degree at the School of Electronic and Information Engineering, Dalian University of Technology in 2004. His research interest covers complex chemical process modeling and optimizing. Corresponding author of this paper.)



毛志忠 东北大学教授. 主要研究方向为复杂工业系统建模、控制与优化. E-mail: maozhizhong@ise.neu.edu.cn (MAO Zhi-Zhong Professor at Northeastern University. His research interest covers complex chemical process modeling, control and optimizing.)



常玉清 东北大学副教授. 主要研究方向为软测量技术. E-mail: changyuqing@mail.neu.edu.cn (CHANG Yu-Qing Associate professor at Northeastern University. Her main research interest is soft sensor techniques.)