

一种动态校正的 AGMM-GPR 多模型软测量建模方法

熊伟丽^{*1,2}, 李妍君², 姚乐², 徐保国²

(1. 江南大学轻工过程先进控制教育部重点实验室, 江苏无锡 214122;

2. 江南大学物联网工程学院自动化研究所, 江苏无锡 214122)

摘要: 工业过程常常是强非线性的, 并有多个工况, 传统的软测量方法存在预测能力差, 不能有效利用误差信息等缺点. 为了有效解决这些问题, 提出一种基于自适应高斯混合模型-高斯过程回归 (AGMM-GPR) 的多模型动态校正软测量建模方法. 首先, 通过贝叶斯信息准则构建自适应高斯混合模型 (AGMM), 得到优化的子模型个数; 然后, 利用 GPR 方法建立各局部模型, 当新的数据到来时, 将其隶属于各局部模型的后验概率和预测值融合得到多模型输出; 最后, 为了进一步提高模型的精度, 构建自回归积分滑动平均 (ARIMA) 模型对多模型输出进行动态反馈校正. 通过数值仿真和硫回收装置 (SRU) 中 H_2S 浓度的估计, 验证了所提方法具有良好的预测精度和泛化性能.

关键词: 自适应; 多模型; 动态校正; 高斯过程回归; ARIMA 模型

中图分类号: TP391.9 **文献标识码:** A **doi:** 10.7511/dllgxb201601012

0 引言

随着软测量技术的发展, 在机理模型难以建立的时候, 基于数据驱动的建模方法已成功地应用于实际的工业过程中并发挥着重大作用^[1-3]. 常见的方法有最小二乘支持向量机^[4]、神经网络^[5]、偏最小二乘法^[6]等. 近年来, 高斯过程回归方法由于其输出具有概率意义、精度高和灵活等优点, 被广泛用于软测量建模研究领域^[7-8]. 对于实际工业过程的高维度和大样本数据, 采取传统的单模型方法建模往往训练时间较长, 其泛化性能和精度较低, 对具有多工况的对象特性及扰动特性拟合不佳. 因此众多研究者将数据驱动和多模型建模的思想应用于软测量建模中^[9-10].

一般情况下, 多模型软测量建模方法是采用某种规则对需要分析的数据进行聚类, 然后建立局部回归模型, 最后再进行融合得到全局模型. 聚类分析时, 聚类结果的好坏对软测量的输出有着很大的影响. 在聚类方法的研究中, 采用

K-means、核模糊等方法进行聚类分析时, 需要明确知道聚类个数, 有时聚类的效果不理想. 如何确定一个“准确”的聚类个数去改善聚类效果, 一直以来是研究的热点^[11-12].

通过将各个子模型的输出进行加权得到全局模型的输出, 综合考虑了样本的局部特性, 提高了模型的鲁棒性. 但在多模型软测量建模方法中, 传统方法不能有效利用误差信息, 造成模型精度下降, 泛化性能不高. 因此, 在模型动态性能的改善上, 杜文莉等^[13]将 ARMA 模型用于乙烯精馏过程乙烷浓度的时序补偿, 提高了预测精度. 王振雷等^[14]提出了一种基于 D-S 理论和 ARIMA 建模的动态最小二乘支持向量机的建模方法, 用于酯化过程酯化率软测量中对静态多模型输出进行动态校正, 得到更加精确的估计结果.

本文结合多模型和模型校正的思想, 提出一种基于自适应高斯混合模型 (adaptive Gaussian mixture model, AGMM) 和高斯过程回归 (Gaussian

收稿日期: 2015-10-05; 修回日期: 2015-11-25.

基金项目: 国家自然科学基金资助项目(21206053, 21276111); 江苏省“六大人才高峰”计划资助项目(2013-DZXX-043); 江苏省产学研资助项目(BY2014023-27); 江苏高校优势学科建设工程资助项目(PAPD).

作者简介: 熊伟丽^{*} (1978-), 女, 博士, 教授, E-mail: greenpre@163.com.

process regression, GPR)的多模型动态校正软测量方法. 先利用贝叶斯信息准则确定样本特征空间的 n 个最佳高斯成分个数, 构建 AGMM, 自适应地确定聚类数目, 然后利用 GPR 方法建立局部模型. 当新的数据到来时, 用每个局部模型进行预测输出, 同时计算新数据隶属于每个子模型的后验概率, 最后对子模型的输出进行融合, 得到全局输出. 为进一步提高模型的精度, 在此基础上, 采用自回归积分滑动平均 (autoregressive integrated moving average, ARIMA) 模型对多模型输出进行动态反馈补偿. 通过硫回收装置 (SRU) 中 H_2S 浓度的软测量仿真, 验证所提方法的精度和外推能力.

1 基于 AGMM-GPR 的多模型建模

1.1 GMM 原理及参数估计

高斯混合模型^[15]在用于数据聚类分析时, 主要思想是样本空间的数据分布可以利用多个独立同分布的高斯成分进行混合. 对多个高斯成分赋予不同的权重, GMM 可以表示为式(1) K 个高斯成分的加权和, 其中 $p(\mathbf{x}, \Theta_{\text{GMM}})$ 为混合模型的概率密度, a_j 为第 j 个高斯成分所占的权重.

$$p(\mathbf{x}, \Theta_{\text{GMM}}) = \sum_{j=1}^K a_j N(\mathbf{x} | \theta_j);$$

$$\sum_{j=1}^K a_j = 1, a_j > 0, j = 1, 2, \dots, K \quad (1)$$

高斯成分密度函数 $N(\mathbf{x} | \theta_j)$ 表达式如式(2)所示, 其中 $\theta_j = (\boldsymbol{\mu}_j \quad \boldsymbol{\Sigma}_j)$ 为第 j 个高斯成分的参数. 样本 $\mathbf{x} \in \mathbf{R}^D$, $\boldsymbol{\mu}$ 为 D 维均值向量, $\boldsymbol{\Sigma}$ 为 $D \times D$ 维正定协方差矩阵.

$$N(\mathbf{x} | \theta_j) = \frac{\exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right]}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_j|^{1/2}} \quad (2)$$

混合参数 $\Theta_{\text{GMM}} = (a_1 \quad \dots \quad a_K \quad \theta_1 \quad \dots \quad \theta_K)$, 若求解出来, 就能对样本进行聚类分析. 本文采用 EM 算法^[16]迭代求解 GMM 的参数集. 在某些观测数据缺失的情况下, EM 算法能够在贝叶斯框架的后验分布模式下反复迭代计算出密度函数的极大似然估计, 似然函数记为式(3), 算法的迭代主要包括两个步骤.

E-step:

关于未知的参数变量, 通过参数的当前估计和关于观测样本的条件参数估计建立一个 Q 函数(式(4))来求取样本极大似然函数的期望.

$$L(\Theta_{\text{GMM}} | \mathbf{X}) = \prod_{i=1}^n p(\mathbf{x}_i, \Theta_{\text{GMM}}) \quad (3)$$

$$Q(\Theta_{\text{GMM}}, \Theta_{\text{GMM}}^i) = E(\ln L(\Theta_{\text{GMM}} | \mathbf{X})) =$$

$$\sum_{j=1}^K \sum_{i=1}^n \lambda_j(\mathbf{x}_i) \left[\ln \frac{a_j}{\sqrt{2\pi} \boldsymbol{\Sigma}_j} \cdot \frac{(\mathbf{x}_i - \boldsymbol{\mu}_j)^\top (\mathbf{x}_i - \boldsymbol{\mu}_j)}{2\boldsymbol{\Sigma}_j} \right] \quad (4)$$

M-step:

对建立的 Q 函数, 通过式(5)最大化步骤反复迭代新的参数估计值, 直到参数收敛. 迭代后的 GMM 参数如式(6)所示.

$$\Theta^{i+1\text{GMM}} = \arg \max Q(\Theta_{\text{GMM}}, \Theta_{\text{GMM}}^i) \quad (5)$$

$$a_j = \frac{1}{n} \sum_{i=1}^n \lambda_j(\mathbf{x}_i)$$

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^n \lambda_j(\mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^n \lambda_j(\mathbf{x}_i)} \quad (6)$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^n \lambda_j(\mathbf{x}_i) (\mathbf{x}_i - \boldsymbol{\mu}_j)^\top (\mathbf{x}_i - \boldsymbol{\mu}_j)}{\sum_{i=1}^n \lambda_j(\mathbf{x}_i)}$$

第 i 个样本对应于第 j 个高斯成分的后验概率 $\lambda_j(\mathbf{x}_i)$ 可由下式计算:

$$\lambda_j(\mathbf{x}_i) = \frac{a_j N(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{j=1}^K a_j N(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \quad (7)$$

1.2 GMM 高斯成分个数的 BIC 优化

传统的 GMM 进行聚类时, 需要通过经验事先确定高斯成分个数, 当工况阶段特征不明显时, 无法实现样本空间的自适应拟合. 在对样本特征了解不充分的情况下, K 值太大, 会造成模型的过拟合问题; 反之, K 值太小, 则不能够充分解释样本信息. 两种结果皆造成模型精度的下降.

对于 GMM 模型聚类优化问题, 以选择合适高斯成分个数来拟合样本空间特征为出发点, 本文采用贝叶斯信息准则^[17] (Bayesian information criterion, BIC) 来优化 GMM 的有限混合高斯成分个数.

L 是式(3)所示 GMM 的似然函数, M 为混合模型参数的个数, n 为训练样本个数, K 指选择的混合的高斯成分个数, 由于 K 过大或者过小会分别造成模型的过拟合与欠拟合, 当混合模型分布进入稳态时, BIC 值越小, 模型拟合程度越好.

BIC 指标计算公式为

$$BIC(K) = -2\log L + M\ln n \quad (8)$$

$$\log L = \sum_{i=1}^n \log \sum_{j=1}^K a_j N(\mathbf{x}_i | \boldsymbol{\theta}_j)$$

1.3 GPR 子模型的建立

在采用 1.2 节 AGMM 方法实现对样本特征的自适应聚类后, 本文采用 GPR 方法建立子模型. GPR 是一种基于统计学习理论的非参数概率模型, 适合处理复杂的高维数、小样本及非线性问题, 近年来, 在机器学习领域有着深入的发展. 通过给定的训练样本的输入输出数据得到映射关系, 便可由新的输入数据得到相应的预测值 and 不确定程度^[18].

假设有训练集 $\{\mathbf{x}, \mathbf{y}\}$, 其中 $\mathbf{X} = \{\mathbf{x}_i \in \mathbf{R}^D\}$, $\mathbf{y} = \{y_i \in \mathbf{R}\}$, $i = 1, 2, \dots, n$, 通常样本观测值 y_i 和噪声 ε 满足式(9)关系:

$$y_i = f(\mathbf{x}_i) + \varepsilon \quad (9)$$

$$\varepsilon \sim N(0, \sigma_n^2)$$

若确定均值函数 $m(\mathbf{x})$ 和协方差函数 $k(\mathbf{x}, \mathbf{x}')$, 高斯过程就能唯一确定. 为了方便, 通常将均值函数预处理为 0. 协方差函数能够把输出间的相关关系转化为输入数据之间的函数关系, 由于相近的输入产生相近的输出, 协方差函数的选择可以根据样本分布的特征选取, 要符合距离相近的样本间相关性大, 反之相关性小的特征. 本文选择的协方差函数形式为

$$k(\mathbf{x}_p, \mathbf{x}_q) = \nu \exp\left[-\frac{1}{2} \sum_{d=1}^D \pi_d (x_p^d - x_q^d)^2\right] \quad (10)$$

式中: $\mathbf{x}_p, \mathbf{x}_q \in \mathbf{R}^D$, ν 控制协方差函数的量度, π_d 刻画了每个 x^d 的相对重要性. 高斯过程的超参数 $\boldsymbol{\Theta}_{\text{gp}} = (\nu \quad \pi_1 \quad \dots \quad \pi_D \quad \sigma_n^2)$ 的确定一般通过 MLE 方法对式(11)进行估计, 超参数的优化可以通过共轭梯度法实现. 基于测试样本和训练数据, 可以计算出测试数据 \mathbf{x}_* 预测值的后验分布服从式(12)的联合高斯分布, $K(\mathbf{X}, \mathbf{X})$ 为训练样本

间的 n 维协方差矩阵, $k(\mathbf{X}, \mathbf{x}_*)$ 为训练样本与测试样本的协方差向量, $k(\mathbf{x}_*, \mathbf{x}_*)$ 为测试样本的自协方差值, f_{gp} 为 GPR 预测值.

$$L(\boldsymbol{\Theta}_{\text{gp}}) = -\frac{1}{2} \mathbf{y}^T [K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}_n]^{-1} \mathbf{y} - \frac{1}{2} \log \det[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}_n] - \frac{n}{2} \log 2\pi \quad (11)$$

$$f_{\text{gp}} | \mathbf{X}, \mathbf{y}, \mathbf{x}_* \sim N(\overline{f_{\text{gp}}}, \text{cov}(f_{\text{gp}}))$$

$$\text{s. t. } \overline{f_{\text{gp}}} = k(\mathbf{X}, \mathbf{x}_*) [K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}_n]^{-1} \mathbf{y}$$

$$\text{cov}(f_{\text{gp}}) = k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{X}, \mathbf{x}_*)^T \cdot [k(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}_n]^{-1} k(\mathbf{X}, \mathbf{x}_*) \quad (12)$$

2 自回归积分滑动平均模型

对样本聚类分析及回归建模后, 在扰动及误差影响下, 建模精度会受到很大的影响. 为提高模型的估计性能, 模型校正的引入有其必要性, 本文采用 ARIMA 模型对输出时序误差进行动态校正.

2.1 ARIMA 模型简述

ARIMA 模型^[19]是一类用于时间序列分析的参数模型, 通过搜集某一待预测变量的历史观测数据来产生一个描述其潜在关系的模型. 此模型中, 变量的将来值被认为是历史观测数据和随机误差的线性函数, 也就是说, 产生时间序列的潜在过程有如下形式:

$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \quad (13)$$

式中: y_t 与 ε_t 分别是在 t 时刻的真实值和随机误差; $\phi_i (i = 1, 2, \dots, p)$, $\theta_j (j = 0, 1, \dots, q)$ 为 ARIMA 的模型参数, p, q 指模型的阶次. 随机误差 ε_t 独立同分布, 其均值为 0, 方差为 σ^2 .

ARIMA 模型可以表示一些不同的时间序列类型, 式(13)中如果 $q = 0$, 则变为 p 阶的 AR 模型; 当 $p = 0$, 模型转化为 q 阶的 MA 模型, 否则, 表示为 ARMA 模型.

2.2 时间序列模型定阶

建立时间序列的 ARIMA 模型, 其核心任务是确定合适的模型阶数. ARIMA 模型仅用于平稳时间序列, 是否达到平稳条件可用 ADF 根检

验的方法判断.若时间序列不平稳,对样本序列进行差分变换是有效的方法之一.引入滞后算子 L ,满足 $y_t \cdot L = y_{t-1}$, $y_t \cdot L^2 = y_{t-2}$,对待分析的时间序列进行 d 阶差分使其平稳化得到如式(14)所示的结果,式中的 $d \in \mathbf{N}$,在这种情况下,得到的模型记为 $ARIMA(p, d, q)$.

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) (1-L)^d y_t = \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t \quad (14)$$

对平稳化时间序列,通过检验 ACF 和 PACF 初步得到 ARIMA 模型的阶次范围,然后利用 AIC 进行模型选择^[20],采用系统辨识方法(如最小二乘法)辨识出模型的参数,选择信息量损失最少的模型.衡量每个模型信息损失量的 AIC 准则判断公式如下:

$$AIC(M) = -2 \log L(\boldsymbol{\theta} | \mathbf{y}) + 2M \quad (15)$$

$$\log L(\boldsymbol{\theta} | \mathbf{y}) = \sum_{i=1}^n \log L(\boldsymbol{\theta} | y_i)$$

式中: $\log L(\boldsymbol{\theta} | \mathbf{y})$ 为每个模型所对应的对数似然函数, M 是每个模型中独立参数的个数, n 为时间序列长度.

3 动态校正的 AGMM-GPR 建模步骤

本文软测量建模方法建模流程如图 1 所示.具体建模步骤描述如下:

Step 1 采集软测量建模的输入输出数据,分为训练样本和测试样本,进行适当的预处理.

Step 2 训练样本采用 AGMM 聚类,通过 BIC 选取合适的高斯成分个数 K .

Step 3 分别建立上述 K 个高斯成分相对应的局部 GPR 子模型,每个 GPR 子模型的输出分别记为 $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_K$.

Step 4 采用式(7)对多个子模型进行融合.对于新的测试数据 \mathbf{x}_* , $\lambda_i(\mathbf{x}_*)$ 与 $\hat{y}_i(\mathbf{x}_*)$ 分别指其在第 i 个阶段输出的后验概率权值和估计值,式(16)为每个测试数据的融合方式:

$$y_{\text{pre}}(\mathbf{x}_*) = \sum_{i=1}^K \lambda_i(\mathbf{x}_*) \hat{y}_i(\mathbf{x}_*) \quad (16)$$

Step 5 对多模型的输出进行校正,得到模型的时序误差序列 $\Delta y = y_{\text{pre}} - y_{\text{true}}$, y_{true} 为主导变

量的人工分析值, y_{pre} 为多模型融合估计值.判断误差序列是否满足平稳序列的条件,初始差分次数 $d_0 = 0$,时间序列平稳时,选择 AIC 指标最小的 $ARIMA(p, d, q)$,得时序误差预测值 Δy_{pre} .否则,进行 $d+1$ 次差分,反复进行单位根检验直到时间序列平稳为止.

Step 6 最终的主导变量输出值 \hat{y} 由式(17)估计:

$$\hat{y} = y_{\text{pre}} - \Delta y_{\text{pre}} \quad (17)$$

选择如式(18)和(19)所示的均方根误差(e_{rms})和平均绝对误差(e_{ma})来评价模型的综合预测能力:

$$e_{\text{rms}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (18)$$

$$e_{\text{ma}} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (19)$$

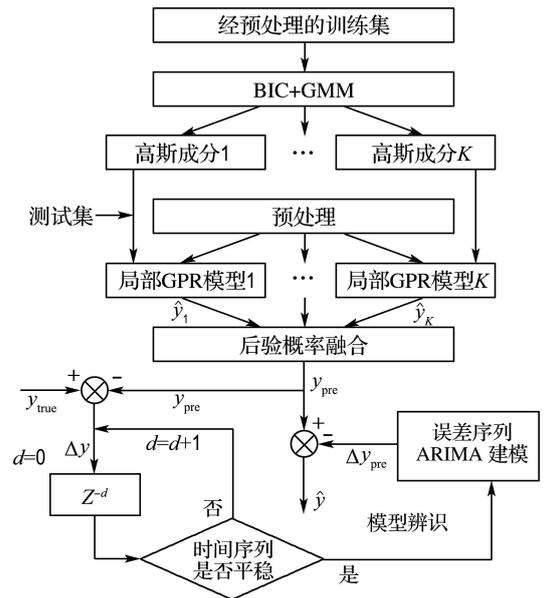


图 1 动态校正的 AGMM-GPR 方法建模图

Fig. 1 Modeling diagram of dynamically corrected AGMM-GPR method

4 数值仿真实验

为验证上述方法的有效性,本文采用文献[21]中的非线性函数进行数值仿真实验:

$$y(t) = \frac{x_1(t)}{1 + 0.5 \sin(x_2(t))} +$$

$$\frac{x_2(t)}{1+0.5\sin(x_1(t))} + \varepsilon(t)$$

$$x_1(t+1) = \left(\frac{x_1(t)}{1+x_1^2(t)} + 1 \right) \sin(x_2(t))$$

$$x_2(t+1) = x_2(t) \cos(x_2(t)) + \frac{\exp\left(-\frac{x_1^2(t)+x_2^2(t)}{8}\right)x_1 + u^3(t)}{1+u^2(t)+0.5\cos(x_1(t)+x_2(t))}$$

式中： $x_1(t)$ 、 $x_2(t)$ 为系统状态，系统的输入、输出和白噪声分别为 $u(t)$ 、 $y(t)$ 、 $\varepsilon(t)$ 。假定系统状态不可测，通过已知的输入输出信息来预测系统的输出 $y(t)$ 。数值仿真中选择用于软测量建模的输入为 $\varphi(t-1) = (y(t-1) \quad y(t-2) \quad y(t-3) \quad u(t-1) \quad u(t-2) \quad u(t-2))^T$ 。以 $u(t) \in [-2.5, 2.5]$ 的随机信号与 $\varepsilon(t) \in N(0, 0.1)$ 的白噪声作用于系统，得到 3 000 组时间序列训练样本；以测试信号 $u(t) = \sin(0.2\pi t) + \sin(0.08\pi t)$ 作用于系统，得到 200 组时间序列测试样本。为方便从图中看到各个方法的仿真效果，在上述 3 000 组训练样本的基础上，对原始训练样本的输出人为加入均值 0.2、方差 0.4 的高斯噪声，构成新的训练样本。基于此，分别对单模型 GPR 方法、AGMM-GPR 的多模型方法以及本文所提的加入动态校正的 AGMM-GPR+ARIMA 方法进行了仿真，拟合指标如表 1 所示。

表 1 不同模型的拟合性能

Tab. 1 Fitting performances of different models

方法	均方根误差	平均绝对误差
GPR	0.203 3	0.203 2
AGMM-GPR	0.160 1	0.155 3
本文方法	0.032 2	0.025 1

由图 2 可知，高斯成分个数为 10 时，BIC 指标值基本收敛，即使随着高斯成分个数的增加，未来有更加低的指标值，但是为了防止模型的过拟合造成精度下降，认为 $K=10$ 是一个合适的聚类数目。

由图 3~5 所示的仿真结果可知，单模型建模方法的均方根误差和平均绝对误差依次低于多模型建模方法和经过误差校正的软测量方法，充分反映了自适应地选择聚类个数有助于更加准确地

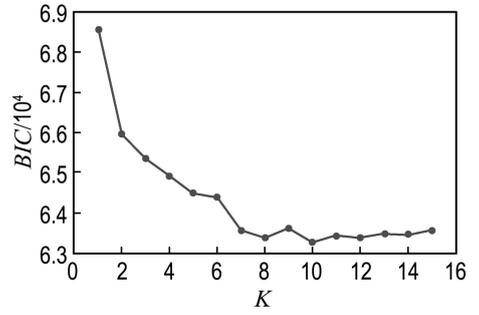


图 2 高斯成分个数的 BIC 优化

Fig. 2 BIC optimization of Gaussian component number

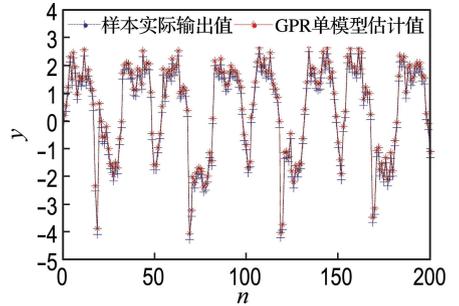


图 3 GPR 单模型仿真结果

Fig. 3 Single GPR model simulation result

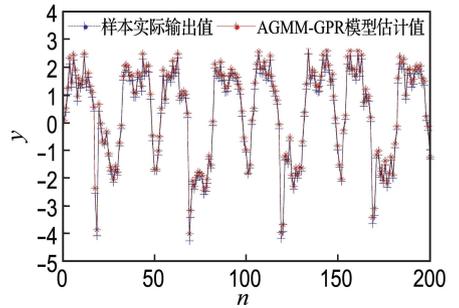


图 4 AGMM-GPR 模型仿真结果

Fig. 4 AGMM-GPR model simulation result

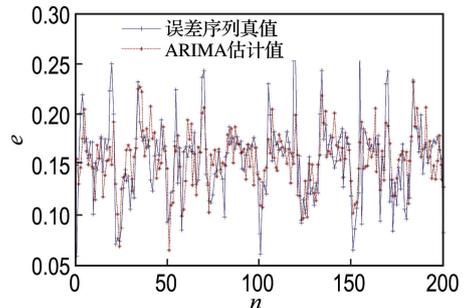


图 5 时序误差的 ARIMA 建模

Fig. 5 ARIMA modeling of time series error

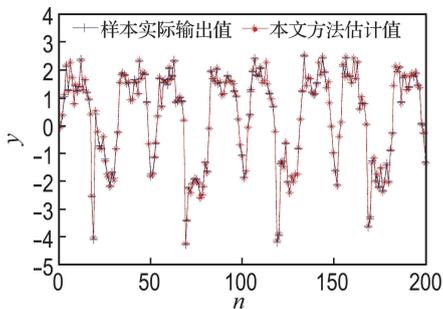


图 6 本文方法仿真结果

Fig. 6 Simulation result of proposed method in this paper

对样本局部特性进行拟合. 图 5 结果表明, 基于 ARIMA 的校正方法能够追踪过程的动态时序误差信息. 图 6 所示的经过模型校正的软测量真值与最终估计值几乎重合, 更加肯定了进行多模型输出误差补偿的有效性.

5 硫回收装置过程建模

硫回收装置(sulfur recovery unit, SRU)是精炼厂处理系统的重要环节, 负责对含硫气体(如 H_2S 和 SO_2)的处理, 避免其对环境造成巨大的危害. SRU 拥有 4 条平行的生产线, 都是以两种酸性气体为输入, 一种是富含 H_2S 的 MEA 气体, 另一种是富含 SO_2 和 NH_3 的 SWS 气体, 具体的反应过程及过程的变量描述说明见文献[22].

为验证所提方法对于 SRU 过程建模的可行性, 采集过程的输入输出数据 10 081 组用于软测量建模研究. 本文选择 H_2S 浓度作为过程需要估计的主导变量进行实验.

采集过程的训练样本 1 680 组, 采用主成分分析(principal component analysis, PCA)算法对软测量的输入辅助变量预处理, 以更好地说明算法的精度. 原始输入矩阵为 $\mathbf{X}^{n \times m}$, n 为输入样本个数, m 为样本特征维数. 通过计算输入样本的协方差矩阵特征值, 保留使累计方差贡献率超过 85% 的样本信息, 将原始的高维空间投影到不相关的低维空间进行特征提取. 如图 7 所示, 采用 PCA 方法提取出 3 个主元进行软测量建模.

特征提取后的得分矩阵不妨记为 \mathbf{T} , 对它进行 AGMM 聚类分析, 贝叶斯信息准则选择结果

见图 8, 当高斯混合模型趋于稳态时, 可以看出, 最合适的高斯成分个数 K 是 11. 在对 AGMM-GPR 输出进行校正时, 对误差时间序列建立 ARIMA 模型, 如图 9 所示, 从仿真结果来看, 这是一种及时且有效的反馈校正方式. 由图 10 可知, 8 401 组测试样本的 H_2S 浓度预测值和人工分析值几乎完全重合, 因此大大提高了模型估计结果的精度.

为验证引入校正模型的必要性, 利用 3 种建模方法对该工业过程进行仿真, 仿真结果如图 11

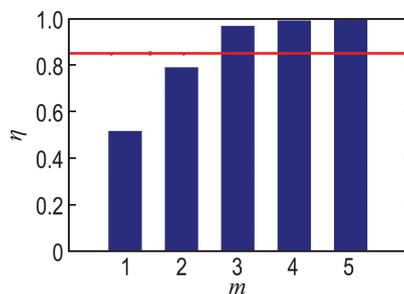


图 7 PCA 方法特征提取结果

Fig. 7 Feature extraction through PCA method

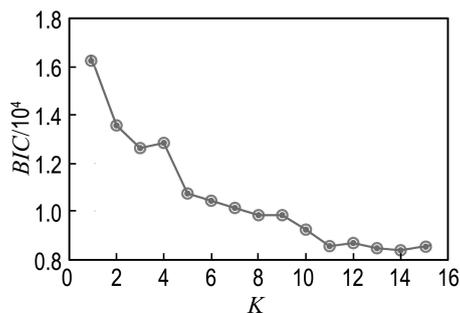


图 8 硫回收装置的高斯成分个数的 BIC 优化
Fig. 8 BIC optimization of Gaussian component number for SRU

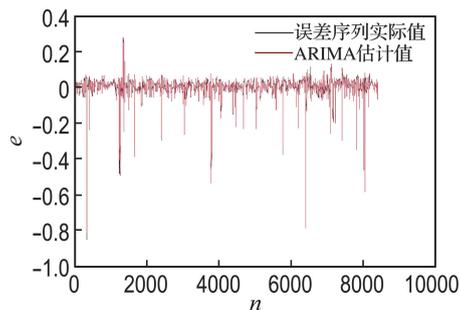
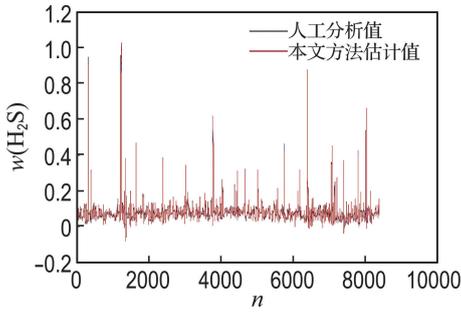
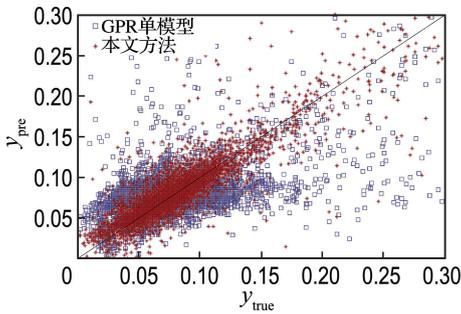
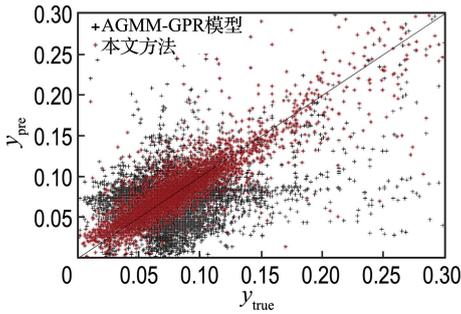


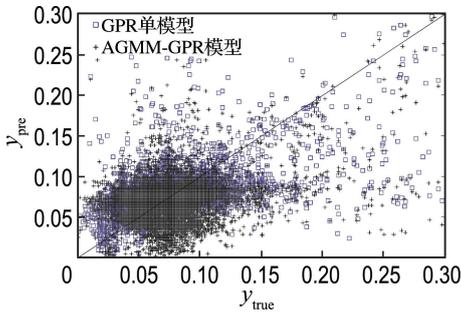
图 9 硫回收装置的时序误差的 ARIMA 建模
Fig. 9 ARIMA modeling of time series error for SRU

图 10 H₂S 浓度估计结果Fig. 10 H₂S concentration estimation result

(a) GPR 单模型与本文方法对比



(b) AGMM-GPR 模型与本文方法对比



(c) GPR 单模型与 AGMM-GPR 模型对比

图 11 不同模型预测结果对比

Fig. 11 Comparisons of different model prediction results

所示。每一种方法的均方根误差和平均绝对误差指标如表 2 所示。由表 2 可知，单模型和多模型建模对于 8 401 组测试数据来说建模精度相差不明

显，经过多模型反馈误差校正后，均方根误差为 0.016 9。在建模效果对比分析图中，由图 11(c)可知，大部分区域中，多模型的估计值相对来说更加集中，逼近真实的 H₂S 浓度。而单模型估计值相对离散，并且局部特性不如多模型拟合得好。如图 11(a)、(b)所示，加入误差校正后估计值比未经过校正的单模型和多模型更贴近 H₂S 真值，这是因为通过对时序误差进行动态跟踪后，对软测量输出构成了负反馈。综上所述，在 3 种方法的比较中，动态校正的 AGMM-GPR 方法在对 SRU 中 H₂S 浓度的估计时，既能够自适应地拟合不同局部工况特性并进行概率意义的融合，又将过程的误差及扰动时序信息补偿至最终输出，具有良好的精度，显著地改善了传统单模型及多模型软测量方法的预测性能。

表 2 SRU 过程不同模型的拟合性能

Tab. 2 Fitting performance of different models in SRU

方法	均方根误差	平均绝对误差
GPR	0.040 6	0.021 7
AGMM-GPR	0.048 2	0.024 8
本文方法	0.016 9	0.011 6

6 结 语

复杂工业过程是强非线性的、时变的且多扰动的。对它进行建模时若采取单模型，要考虑全部样本信息，往往局部特性匹配不佳且估计精度较差，故一般采用多模型的方法对样本信息按照不同的扰动和对象特性进行拟合。为了加强聚类效果，本文对传统的 GMM 混合高斯成分个数进行了 BIC 指标的优化，更加精确地使得样本按照扰动及对象特性“聚集”到一起。为了进一步提高多模型软测量方法的精度，本文利用误差信息的补偿作用，结合 ARIMA 模型对时间序列良好的逼近能力，对静态条件下得到的 AGMM-GPR 多模型输出进行动态时序补偿。仿真实验证明，与传统的单模型建模方法、多模型建模方法相比，本文提出的软测量建模方法把过程误差和扰动考虑在内而进行模型校正，显著提高了估计的准确性。以 SRU 过程的 H₂S 浓度的软测量为例，验证了本

文方法的有效性和外推能力,为解决同类问题提供了新思路.下一步的研究方向就是增强建模的实时性,实现在线校正和优化.

参考文献:

- [1] Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry [J]. **Computers and Chemical Engineering**, 2009, **33**(4):795-814.
- [2] 曹鹏飞, 罗雄麟. 化工过程软测量建模方法研究进展[J]. 化工学报, 2013, **64**(3):788-800.
CAO Peng-fei, LUO Xiong-lin. Modeling of soft sensor for chemical process [J]. **CIESC Journal**, 2013, **64**(3):788-800. (in Chinese)
- [3] YAO Yuan, GAO Fu-rong. Phase and transition based batch process modeling and online monitoring [J]. **Journal of Process Control**, 2009, **19**(5):816-826.
- [4] LIU Yi, WANG Hai-qing. Modelling of the penicillin fermentation process via LS-SVM based on pensim simulator [J]. **Chemical Reaction Engineering and Technology**, 2006, **22**(3):252-258.
- [5] 李大宇, 刘方, 靳其兵. 自增长混合神经网络及其在燃料电池建模中的应用[J]. 化工学报, 2015, **66**(1):333-337.
LI Da-zi, LIU Fang, JIN Qi-bing. Self-growing hybrid neural network and its application for fuel cell modeling [J]. **CIESC Journal**, 2015, **66**(1):333-337. (in Chinese)
- [6] 徐欧官, 陈祥华, 傅永峰, 等. 基于模型性能评估的递推 PLS 建模及应用[J]. 化工学报, 2014, **65**(12):4875-4882.
XU Ou-guan, CHEN Xiang-hua, FU Yong-feng, *et al.* Recursive PLS modeling based on model performance assessment and its application [J]. **CIESC Journal**, 2014, **65**(12):4875-4882. (in Chinese)
- [7] LI Xiu-liang, SU Hong-ye, CHU Jian. Multiple model soft sensor based on affinity propagation, Gaussian process and Bayesian committee machine [J]. **Chinese Journal of Chemical Engineering**, 2009, **17**(1):95-99.
- [8] 何志昆, 刘光斌, 赵曦晶, 等. 高斯过程回归方法综述[J]. 控制与决策, 2013, **28**(8):1121-1129,

1137.

HE Zhi-kun, LIU Guang-bin, ZHAO Xi-jing, *et al.* Overview of Gaussian process regression [J]. **Control and Decision**, 2013, **28**(8):1121-1129, 1137. (in Chinese)

- [9] 李卫, 杨煜普, 王娜. 基于核模糊聚类的多模型 LSSVM 回归建模[J]. 控制与决策, 2008, **23**(5):560-562, 566.
LI Wei, YANG Yu-pu, WANG Na. Multi-model LSSVM regression modeling based on kernel fuzzy clustering [J]. **Control and Decision**, 2008, **23**(5):560-562, 566. (in Chinese)
- [10] FENG Rui, ZHANG Yan-zhu, SONG Chun-lin, *et al.* A multiple model approach to modeling based on fuzzy support vector machines [J]. **Journal of Shanghai Jiaotong University**, 2003, **E-8**(2):137-141.
- [11] 王洋课, 费树岷, 翟军勇. 基于聚类的多模型软测量建模及其应用[J]. 化工自动化及仪表, 2010, **37**(1):49-52.
WANG Yang-ke, FEI Shu-min, ZHAI Jun-yong. Clustering-based multi-model soft-sensing modeling and its application [J]. **Control and Instruments in Chemical Industry**, 2010, **37**(1):49-52. (in Chinese)
- [12] Fujita A, Takahashi D Y, Patriota A G. A non-parametric method to estimate the number of clusters [J]. **Computational Statistics and Data Analysis**, 2014, **73**:27-39.
- [13] 杜文莉, 官振强, 钱锋. 一种基于时序误差补偿的动态软测量建模方法[J]. 化工学报, 2010, **61**(2):439-443.
DU Wen-li, GUAN Zhen-qiang, QIAN Feng. Dynamic soft sensor modeling based on time series error compensation [J]. **CIESC Journal**, 2010, **61**(2):439-443. (in Chinese)
- [14] 王振雷, 唐苦, 王昕. 一种基于 D-S 和 ARIMA 的多模型软测量方法[J]. 控制与决策, 2014, **29**(7):1160-1166.
WANG Zhen-lei, TANG Ku, WANG Xin. A multi-model soft sensing method based on D-S and ARIMA model [J]. **Control and Decision**, 2014, **29**(7):1160-1166. (in Chinese)
- [15] Bishop C M. **Pattern Recognition and Machine**

- Learning** [M]. New York:Springer, 2006.
- [16] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm [J]. **Journal of the Royal Statistical Society. Series B (Methodological)**, 1977, **39**(1):1-38.
- [17] Sadanori Konishi, Genshiro Kitagawa. **Information Criteria and Statistical Modeling** [M]. New York: Springer, 2007.
- [18] Rasmussen C E, Williams C K I. **Gaussian Processes for Machine Learning** [M]. Cambridge: The MIT Press, 2006.
- [19] Hassan J. ARIMA and regression models for prediction of daily and monthly clearness index [J]. **Renewable Energy**, 2014, **68**:421-427.
- [20] Akaike Hirotogu. Information theory and an extension of the maximum likelihood principle [C] // **Proceeding of the Second International Symposium on Information Theory**. Budapest: Akademiai Kiado, 1973:267-281.
- [21] 曾 静, 王 军, 郭金玉. 基于向量相似度的多模型局部建模方法研究 [J]. 计算机应用研究, 2012, **29**(5):1631-1634.
ZENG Jing, WANG Jun, GUO Jin-yu. Local multi-model method based on similarity of vector [J]. **Application Research of Computers**. 2012, **29**(5): 1631-1634. (in Chinese)
- [22] Fortuna L, Rizzo A, Sinatra M, *et al.* Soft analyzers for a sulfur recovery unit [J]. **Control Engineering Practice**, 2003, **11**(12):1491-1500.

A dynamically corrected AGMM-GPR multi-model soft sensor modeling method

XIONG Wei-li^{*1,2}, LI Yan-jun², YAO Le², XU Bao-guo²

(1. Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China;

2. Institute of Automation, School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: Industrial processes often encounter strong nonlinearity and multiple operating modes. Traditional soft sensor methods cannot effectively take advantage of error information, which accounts for unsatisfactory predictive results. To effectively address these problems, a dynamically corrected multi-model soft sensor modeling method based on adaptive Gaussian mixture model-Gaussian process regression (AGMM-GPR) is proposed. Firstly, an adaptive Gaussian mixture model is constructed using Bayesian information criterion and optimized sub-model number is obtained. Then, each local model is built through GPR method. For the new data, its posterior probability and prediction value belonging to each local model can be combined to get multi-model output. Finally, to further improve model accuracy, an autoregressive integrated moving average (ARIMA) model is employed to conduct a dynamic feedback correction to multi-model output. Numerical simulation and H₂S concentration estimation in sulfur recovery unit (SRU) indicate that the proposed method has good prediction accuracy and generalization performance.

Key words: adaptive; multi-model; dynamic correction; Gaussian process regression; ARIMA model