Journal of System Simulation

Volume 31 | Issue 10 Article 9

12-12-2019

A Hierarchical Integrated Soft Sensing Modeling Method for Gauss Process Regression

Zhao Shuai
1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China; ;

Xudong Shi
1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China; ;

Weili Xiong
1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China; ;2. Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China;

A Hierarchical Integrated Soft Sensing Modeling Method for Gauss Process Regression

Abstract

Abstract: Chemical processes are often characterized by nonlinearity and multi-phase, a soft sensor model based on the hierarchical ensemble of Gaussian process regression is proposed. *First, the Gaussian mixture model is used to divide the process data into different operation phases. Then, the principal component analysis of each stage is carried out, and the model data are divided into several subspaces, according to the contribution of each auxiliary variable in the principal component space, and the corresponding Gaussian process regression model is built. The subspace model output is fused by means to obtain the first level ensemble output. Finally, the posterior probability is used to fuse the model local prediction to obtain the second level ensemble output*. The validity of the proposed method is verified by the experimental simulation of industrial data.

Keywords

Gaussian mixture model, subspace PCA, Gaussian process regression, hierarchical ensemble, soft sensing

Recommended Citation

Zhao Shuai, Shi Xudong, Xiong Weili. A Hierarchical Integrated Soft Sensing Modeling Method for Gauss Process Regression[J]. Journal of System Simulation, 2019, 31(10): 2042-2051.

第 31 卷第 10 期 系统仿真学报© Vol. 31 No. 10
2019 年 10 月 Journal of System Simulation Oct., 2019

一种分层集成的高斯过程回归软测量建模方法

赵帅[1]，史旭东[1]，熊伟丽[1,2]
(1. 江南大学物联网工程学院自动化研究所，江苏 无锡 214122；2. 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)

摘要：针对一些化工过程呈现显著的非线性和多阶段特点，*提出一种分层集成的高斯过程回归软测量建模方法。采用高斯混合模型将过程数据划分为不同的操作阶段；对各阶段的数据进行主元分析，依据各辅助变量在主元空间上的贡献度，将各阶段数据划分成若干子空间并建立相应的高斯过程回归模型；再对子空间模型输出进行均值融合，得到第一层集成输出；采用后验概率对各阶段局部预测进行融合，得到第二层集成输出。*通过对工业数据的实验仿真，验证了所提方法的有效性。

关键词：高斯混合模型；子空间 PCA；高斯过程回归；分层集成；软测量
中图分类号：TP274 文献标识码：A 文章编号：1004-731X (2019) 10-2042-10
DOI: 10.16182/j.issn1004731x.joss.17-0374

A Hierarchical Integrated Soft Sensing Modeling Method for Gauss Process Regression

Zhao Shuai[1], *Shi Xudong*[1], *Xiong Weili*[1,2]
(1. School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China;
2. Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China)

Abstract: Chemical processes are often characterized by nonlinearity and multi-phase, a soft sensor model based on the hierarchical ensemble of Gaussian process regression is proposed. *First, the Gaussian mixture model is used to divide the process data into different operation phases. Then, the principal component analysis of each stage is carried out, and the model data are divided into several subspaces, according to the contribution of each auxiliary variable in the principal component space, and the corresponding Gaussian process regression model is built. The subspace model output is fused by means to obtain the first level ensemble output. Finally, the posterior probability is used to fuse the model local prediction to obtain the second level ensemble output*. The validity of the proposed method is verified by the experimental simulation of industrial data.

Keywords: Gaussian mixture model; subspace PCA; Gaussian process regression; hierarchical ensemble; soft sensing

引言

化工、冶金和发酵等工业过程中的一些重要质量变量，往往无法或难以通过在线仪表测量[1-2]。基于数据驱动的软测量建模方法，无需深入了解过程的机理知识，在工业过程建模中得到了广泛应用。常用的线性建模方法如主成分回归[3-4] (Principal component regression，PCR)、偏最小二乘[5-6] (Partial least squares，PLS)等可以很好的处理过程数据间的线性关系并进行建模。然而，化工过程常呈现显著的非线性特征，因此非线性建模方法如人工神经网络[7-8] (Artificial neural networks,

收稿日期：2017-08-04 修回日期：2017-08-28；
基金项目：国家自然科学基金(61773182)，江苏高校优势学科建设工程资助项目(PAPD)；
作者简介：赵帅(1992-)，男，安徽淮北，硕士生，研究方向为工业过程建模；史旭东(1993-)，男，江苏常州，硕士生，研究方向为工业过程建模。

ANN)、支持向量机[9-10] (Support vector machine, SVM)和高斯过程回归[11-13] (Gaussian process regression, GPR)等得到广泛应用。针对化工过程中的多阶段特征，各种聚类算法如 K 均值[14] (K-means)、模糊 C 均值[15] (Fuzzy C means，FCM)和高斯混合模型[16](Gaussian mixture model，GMM)等被广泛用于过程数据的阶段划分，从而提取多阶段特征，进而提高建模精度。

集成学习是将一系列有差异的模型进行组合，以提高模型的泛化能力与预测精度。集成学习的首要任务是划分子数据集，常用的划分方法有 Bagging、Boosting、聚类和子空间方法等。Chen 等[17]采用 Bagging 算法提升高斯过程回归模型的泛化能力与预测精度，通过聚丙烯聚合过程验证了所提方法的有效性；Cao 等[18]采用 Boosting 算法产生若干有差异的样本子集，并将所提方法应用于心脏病数据的分类和近红外光谱数据的回归预测，可以有效的增强模型的分类性能和预测精度；Li 等[19]采用聚类算法将原始数据集进行划分，通过对子模型的融合得到最终的预测输出；Ge 等[20]采用子空间 PLS 方法将原始数据集划分为不同的子样本集，并建立相应的 PLS 回归模型，采用均值融合方式实现了对近红外光谱蛋白质含量的预测。Bagging、Boosting 和聚类 3 种方法在样本维度上划分子数据集，子空间方法在变量维度构建子数据集。常规的集成学习仅从单一的样本或变量维度建立软测量模型，没有充分利用样本所包含的两个维度信息，因此 Wang 等[21]提出了一种分层集成的软测量模型结构，采用 GMM 进行样本划分，利用随机重采样策略和偏互信息准则进行变量的划分与选择，增强模型的多样性。

综上所述，本文提出基于 GMM 与子空间主成分分析(Subspace principle component analysis, SPCA)的分层集成高斯过程回归(hierarchical ensemble Gaussian process regression, HEGPR)软测量建模方法，从样本和变量两个角度出发建立了分层集成软测量模型，既考虑到样本含有的信息，又考虑了输入变量的多样性信息。采用 SPCA 方法在变量维度进行集成，确保在主元方向上贡献度大的变量被用于建立模型，所建立的模型具有良好的稳定性与泛化能力。HEGPR 模型结构上有两层集成，第一层集成针对采用 GMM 方法得到的子样本集，利用 SPCA 划分子空间数据集，并建立集成高斯过程回归(SPCA-GPR)模型，采用均值融合方式得到第一层的预测输出。第二层集成针对 SPCA-GPR 模型得到的各操作阶段的局部输出进行再次融合，采用后验概率融合方式得到最终的模型输出。通过对实际工业过程脱丁烷塔塔底丁烷浓度和青霉素发酵过程青霉素浓度的仿真实验验证了本文方法具有良好的预测精度。

1 预备知识

1.1 高斯混合模型(GMM)

GMM 是一种有效的概率聚类方法[21]，GMM 假设数据服从具有未知参数的混合高斯分布。给定训练样本集 $\boldsymbol{X}\in \mathrm{R}^{n\times m}$ 和 $\boldsymbol{y}\in \mathrm{R}^{n\times 1}$，其中 n 是样本点个数，m 是样本维数。训练样本 $\boldsymbol{X}$ 的概率密度可以表示为

$$p(\boldsymbol{X}\mid\boldsymbol{\Theta})=\sum_{i=1}^{K}\pi_i N(\boldsymbol{X}\mid\boldsymbol{\Theta}_i) \tag{1}$$

式中：K 是高斯成分个数，$\boldsymbol{\Theta}=\{\pi_1,\ \ldots,\ \pi_K,\ \mu_1,\ \ldots,\mu_K,\ \mathrm{C}_1,\ \ldots,\ \mathrm{C}_K\}$ 是高斯混合模型的超参数，π_i 代表第 i 个高斯成分的权值，μ_i 和 C_i 是第 i 个高斯成分的均值和协方差，参数 π_i 满足 $\sum_{i=1}^{K}\pi_i=1$ 和 $0\leqslant\pi_i\leqslant 1$。$N(\boldsymbol{X}\mid\boldsymbol{\Theta}_i)$ 表示多元高斯分布，对应的概率密度函数如式(2)所示

$$\mathrm{P}(\boldsymbol{X}\mid\boldsymbol{\Theta}_i)=\frac{1}{\sqrt{(2\pi)^m\det(C_i)}}\times \exp\left[-\tfrac{1}{2}\left(\boldsymbol{X}-\mu_i\right)^T C_i^{-1}\left(\boldsymbol{X}-\mu_i\right)\right] \tag{2}$$

GMM 的模型参数可以通过期望最大化算法(expectation maximization，EM)[22-23]估计得到，对于新来样本 $\boldsymbol{x}_q$ 对于每个高斯成分的后验概率可通过式(3)计算得到。

DOI: 10.16182/j.issn1004731x.joss.17-0374

第 31 卷第 10 期 系统仿真学报 Vol. 31 No. 10
2019 年 10 月 Journal of System Simulation Oct., 2019

$$\mathrm{P}(\boldsymbol{\Theta}_i \mid \boldsymbol{x}_q)=\frac{\pi_i N(\boldsymbol{x}_q \mid \mu_i, \mathrm{C}_i)}{\sum_{i=1}^{K}\pi_i N(\boldsymbol{x}_q \mid \mu_i, C_i)} \tag{3}$$

1.2 子空间主成分分析(SPCA)

假设有训练样本集 $\mathrm{D}=\{\boldsymbol{X},\boldsymbol{y}\}$，其中 $\boldsymbol{X}=\{\boldsymbol{x}_i\in R^m\}_{i=1,2,\cdots,n}$ 是 n 个训练输入样本，m 是辅助变量个数，$\boldsymbol{y}=\{y_i\in R\}_{i=1,2,\cdots,n}$ 是 n 个输出样本。PCA 模型如式(4)和(5)所示。

$$\boldsymbol{X}=\boldsymbol{T}\boldsymbol{P}^T+\boldsymbol{E} \tag{4}$$

$$\boldsymbol{T}=\boldsymbol{X}\boldsymbol{P} \tag{5}$$

式中：$\boldsymbol{P}\in \mathrm{R}^{m\times t}$ 和 $\boldsymbol{T}\in \mathrm{R}^{n\times t}$ 分别表示第 k 个操作阶段下的载荷矩阵和得分矩阵，$\boldsymbol{E}\in \mathrm{R}^{n\times m}$ 表示残差矩阵，t 表示选择的主成分数目。不同主元之间互不相关，根据主元方向构建不同的子模型，增加模型间的差异性，在每个主元方向上，由每个变量的贡献指数(contribution index, CI)来衡量其重要性，贡献指数的定义如式(6)所示。

$$\boldsymbol{CI}(i,j)=\frac{p_{ij}}{\left|p_{1j}\right|+\left|p_{2j}\right|+\ldots+\left|p_{ij}\right|+\ldots+\left|p_{mj}\right|} \tag{6}$$

式中：$i=1,2,\cdots,m$，$j=1,2,\cdots,t$，p_{ij} 是载荷矩阵对应的第 j 个主元上的第 i 个元素。CI 的值越大，表示变量包含的信息越多，与主元方向的相关性越强。根据 CI 的值进行降序排列，设置不同的 CI 阈值确定各子空间模型选取的辅助变量个数。

基于主元方向的不相关性，根据 d 个不同主元方向将训练样本集 $\boldsymbol{X}$ 划分为 d 个不同的子空间。

$$\boldsymbol{X}\xrightarrow{division}\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_d\} \tag{7}$$

1.3 高斯过程回归(GPR)

给定数据集 $\{\boldsymbol{X},\boldsymbol{y}\}$，其中 $\boldsymbol{X}\in \mathrm{R}^{n\times m}$，$\boldsymbol{y}\in \mathrm{R}^{n\times 1}$，$n$ 样本点数，m 是样本维数。输入输出之间满足

$$y=f(x)+\varepsilon \tag{8}$$

式中：ε 是均值为 0，方差为 σ_n^2 的高斯噪声，f 是未知的函数形式。GPR 假设回归函数 $y=f(x)$ 具有零均值的高斯先验分布，描述如式(9)

$$y\sim N(0,\boldsymbol{C}) \tag{9}$$

式中：$\boldsymbol{C}$ 为 $n\times n$ 的协方差矩阵，其中 $\mathrm{C}_{ij}=C(\boldsymbol{x}_i,\boldsymbol{x}_j;\boldsymbol{\theta})$，协方差矩阵通过核函数计算得到，文中选取平方指数协方差核函数，定义如式(10)所示

$$C(\boldsymbol{x}_i,\boldsymbol{x}_j;\boldsymbol{\theta})=\sigma_f^2\exp\left(-\frac{1}{2l^2}(\boldsymbol{x}_i-\boldsymbol{x}_j)^2\right)+\sigma_n^2\delta_{ij} \tag{10}$$

式中：$\delta_{ij}=1$ 仅在 $i=j$ 时成立，否则 $\delta_{ij}=0$，l 为方差尺度，σ_f^2 为信号方差，σ_n^2 为噪声方差，$\boldsymbol{\theta}=\{\sigma_f^2,l^2,\sigma_n^2\}$ 为 GPR 的超参数，超参数的选取对模型性能有很大影响，采用最大似然估计法可以获得最优的超参数[24,25]。

对于新来的样本 $\boldsymbol{x}_q$，其对应的 GPR 模型输出均值和方差如式(11)和(12)所示

$$\hat{y}_q=\boldsymbol{c}^T(\boldsymbol{x}_q)\boldsymbol{C}^{-1}\boldsymbol{y} \tag{11}$$

$$\sigma_q^2=C(\boldsymbol{x}_q,\boldsymbol{x}_q)-\boldsymbol{c}^T(\boldsymbol{x}_q)\boldsymbol{C}^{-1}\boldsymbol{c}(\boldsymbol{x}_q) \tag{12}$$

式中：$\boldsymbol{c}(\boldsymbol{x}_q)=[C(\boldsymbol{x}_q,\boldsymbol{x}_1),\ \cdots,\ C(\boldsymbol{x}_q,\boldsymbol{x}_n)]$ 是新来样本与训练样本间的协方差矩阵，$\boldsymbol{C}$ 是训练样本之间的协方差矩阵，$C(\boldsymbol{x}_q,\boldsymbol{x}_q)$ 是新来样本的自协方差。

2 分层集成(HEGPR)的软测量建模方法

当化工过程呈现显著的非线性和多阶段特征时，全局模型无法准确的描述过程的阶段与时变特征。因此，选用分层集成的软测量模型，采用 GMM 划分过程操作阶段，针对每个操作阶段建立 SPCA-GPR 模型，并进行第一层集成输出，即可得到各操作阶段的局部输出，最后通过新来样本对于各操作阶段的后验概率进行融合输出，得到最终的质量预测。整体的 HEGPR 建模包括 3 个关键步骤，(1) 通过 GMM 划分操作阶段；(2) 第一层集成输出；(3) 第二层集成输出。

对于新来样本 $\boldsymbol{x}_q$，根据 SPCA 方法对测试样本进行变量选择，对应离线建立的各操作模态下的 SPCA-GPR 模型，得到各子空间下的输出，此时采用均值融合进行第一层集成输出得到各操作阶段模式下的局部预测输出；根据测试样本对于各操作阶段的后验概率，再对各操作阶段下的局部输出采用贝叶斯融合方式进行第二层集成输出，得到最终的集成预测。测试样本基于 HEGPR 的方法在线预测流程如图 1 所示。

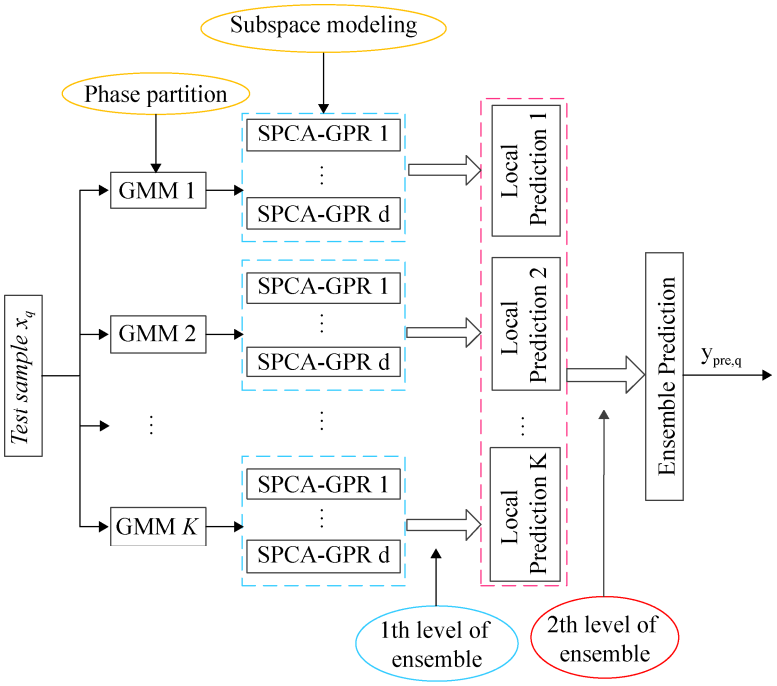


图 1 HEGPR 测试样本在线预测流程图

Fig. 1 Flow diagram of test sample online prediction based on HEGPR method

2.1 第一层集成输出

化工过程具有的多个阶段或操作模态，在建立集成模型之前需要对原始数据集进行阶段划分，采用 GMM 将过程数据划分为 K 个子数据集 $\boldsymbol{S}_k=\{\boldsymbol{X}^k,\boldsymbol{y}^k\};k=1,2,\cdots,K$，采用子空间 PCA 方法对数据集 $\boldsymbol{S}_k$ 在主元空间上进行集成，划分为 d 个子空间数据集 $\boldsymbol{M}_i=\{\boldsymbol{X}_i^k,\boldsymbol{y}_i^k\}$；$k=1,2,\cdots,K$；$i=1,2,\cdots,d$，对子空间数据集建立对应的 GPR 模型，对于新来测试样本 $\boldsymbol{x}_q$ 在第 k 个操作阶段的 i 个子空间下建立的 GPR 模型如式(13)所示。

$$GPR_{k,i}:\begin{cases}y_{q,i}^k=\boldsymbol{c}_{k,i}^T(\boldsymbol{x}_q)\boldsymbol{C}_{k,i}^{-1}\boldsymbol{y}\\ \sigma_{q,k,i}^2=C(\boldsymbol{x}_q,\boldsymbol{x}_q)-\boldsymbol{c}_{k,i}^T(\boldsymbol{x}_q)\boldsymbol{C}_{k,i}^{-1}\boldsymbol{c}_{k,i}(\boldsymbol{x}_q)\end{cases}\tag{13}$$

式中：$k=1,2,L,K;i=1,2,L,d$。

对于新来样本 $\boldsymbol{x}_q$ 得到 $K\times d$ 个预测输出值 $\boldsymbol{y}_q^k=\{y_{q,1}^k,y_{q,2}^k,\cdots,y_{q,d}^k\}$，其中 $y_{q,i}^k$ 是第 k 个操作阶段第 i 个子空间模型的预测输出。对子空间模型预测输出采用均值方式融合，得到第一层集成输出，即第 k 个操作阶段的局部输出 y_q^k，由式(14)计算得到。

$$y_q^k=\frac{1}{d}\sum_{i=1}^{d}y_{q,i}^k\tag{14}$$

第一层集成输出 SPCA-GPR 建模流程如图 2 蓝色虚线框中所示。

2.2 第二层集成输出

第一层集成输出得到各操作阶段下的局部预测值，此时根据新来样本 $\boldsymbol{x}_q$ 对于各操作阶段的后验概率，采用贝叶斯融合方式进行第二层集成输出，得到最终的预测输出，如式(15)所示。

$$y_{pre}=\sum_{k=1}^{K}y_q^k\,\mathrm{P}(S_k\mid\boldsymbol{x}_q)\tag{15}$$

新来测试样本 $\boldsymbol{x}_q$ 对于各操作阶段的后验概率可由式(3)计算得到。

HEGPR 整体建模过程如图 2 所示。

第 31 卷第 10 期 系统仿真学报 Vol. 31 No. 10
2019 年 10 月 Journal of System Simulation Oct., 2019

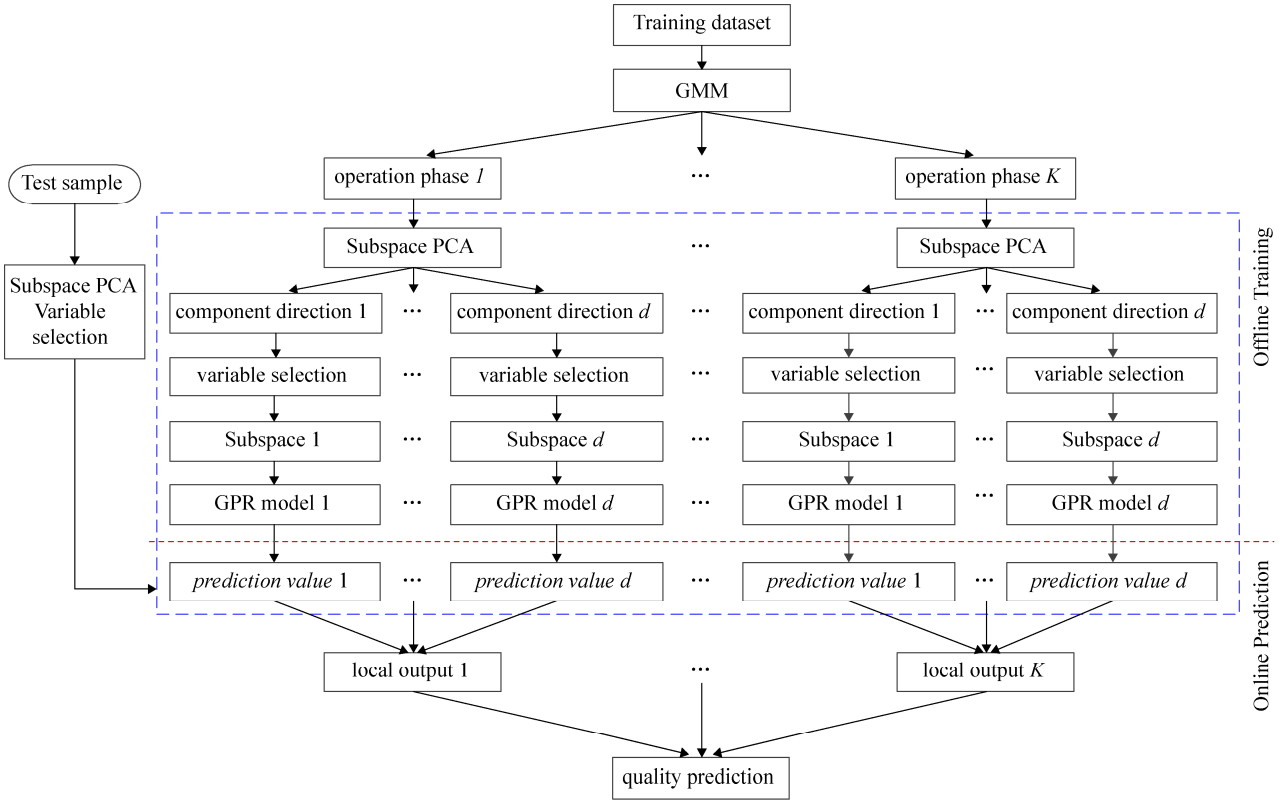


图 2 HEGPR 建模流程图

Fig. 2 Flow diagram of HEGPR model

为更好说明整体的建模过程，建模步骤描述如下：

(1) 收集过程输入输出数据，组建历史数据库；

(2) 基于历史训练数据，采用 GMM 对数据进行阶段划分，划分为 K 个子数据集；

(3) 对 K 个操作阶段的数据集采用子空间 PCA 方法进行分解；

(4) 在 d 个不同主元方向将第 k 个操作阶段下的样本集 $\boldsymbol{S}_k$ 划分为 d 个不同的子空间；

(5) 对不同的子空间建立离线的局部 SPCA-GPR 模型；

(6) 对于新来样本点，根据式(13)所示建立好的离线 SPCA-GPR 模型，得到子空间下的预测值，由式(14)得到第一层集成输出；

(7) 计算新来样本点对于各操作阶段的后验概率，根据式(15)进行融合输出，得到最终的质量预测，完成第二层集成输出。

本文选取均方根误差(root mean square error, RMSE)、相关性系数(correlation coefficient, COR)和跟踪性能指标(tracking precision, TP) 评估模型的预测能力与跟踪性能，计算公式如式(16)~(18)所示。

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(y_i-y_{pre,i})^2}{N}} \tag{16}$$

$$COR=\frac{\sum_{i=1}^{N}(y_i-\bar{y}_i)\cdot(y_{pre,i}-\bar{y}_{pre})}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2\cdot\sum_{i=1}^{N}(y_{pre,i}-\bar{y}_{pre})^2}} \tag{17}$$

$$TP=1-\frac{\sigma_{error}^2}{\sigma_{true}^2} \tag{18}$$

式中：N 是样本点个数，y_i 为过程数据真值，$y_{pre,i}$ 为预测值，$\bar{y}_i$ 为真值的均值，$\bar{y}_{pre}$ 为预测值的均值，σ_{error}^2 为真值与预测值之间的误差方差，σ_{true}^2 为真值的方差。

3 仿真实验

3.1 脱丁烷塔过程

脱丁烷塔过程是石油炼制生产过程中脱硫和石脑油分离装置的重要组成部分。脱丁烷塔塔底丁烷浓度对石油炼制影响较大且难以直接采用仪器测量，工业过程中需要对其进行严格的监测与控制。选取过程中的 7 个变量作为辅助变量，主导变量为脱丁烷塔塔底丁烷浓度。变量选取见表 1。

表 1 脱丁烷塔过程输入变量
Tab. 1 Input variables of debutanizer tower process

Input variables	Description
x_1	Top temperature
x_2	Top pressure
x_3	Reflux flow
x_4	Flow to next process
x_5	6th tray temperature
x_6	Bottom temperature
x_7	Bottom temperature

共收集 2 000 组过程数据，其中 1 000 组作为训练样本集，1 000 组作为测试样本集。图 3 为不同聚类个数对应的均方根误差曲线，考虑模型精度要求，对训练数据采用 GMM 划分为 6 个操作阶段，每个操作阶段下划分 6 个子空间模型(采用遍历的方法，选择建模精度最佳时对应的子空间模型个数)，设定不同的 CI 阈值确定每个子空间模型选取的辅助变量个数，每个子空间模型包含的辅助变量个数如图 4 所示。为验证本文方法的性能，共建立了 4 种软测量模型进行对比：(1) GPR 全局模型；(2) 基于 GMM 的 GPR 多模型(简记为 GMM-GPR)；(3) 基于随机子空间(random subspace method, RSM)的分层集成软测量模型(简记为 GMM-EGPR)；(4) 分层集成高斯过程回归软测量模型(HEGPR)。不同建模方法的预测性能指标如表 2 所示。

从表 2 中可以看出 HEGPR 建模方法的预测精度明显高于其他 3 种建模方法，表明了采用 HEGPR 可以很好的解决工业过程中存在的多阶段和非线性特征，采取 SPCA 集成方法选取在主元方向上的贡献度大的变量进行构建子空间数据集，在保证增加模型泛化能力的同时更加有效的利用变量的信息，提高模型的预测精度。

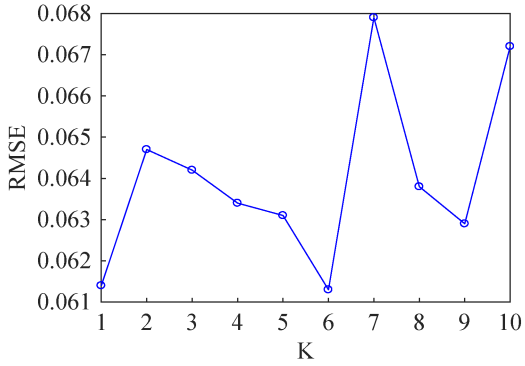


图 3 不同聚类个数的均方根误差
Fig. 3 RMSEs for Different Clusters

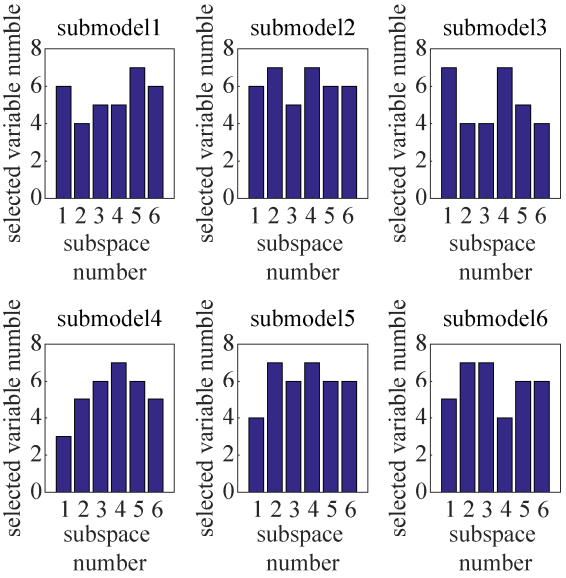


图 4 脱丁烷塔过程各模态子空间选择变量个数
Fig. 4 Selection variable number in each subspace of debutanizer tower process

表 2 不同建模方法的预测性能指标
Tab. 2 Predictive performance indicators for different methods of debutanizer tower process

模型	RMSE	COR/%	TP/%
Single-GPR	0.061 4	93.21	86.65
GMM-GPR	0.061 3	93.45	87.13
GMM-EGPR	0.069 9	91.61	82.69
HEGPR (Proposed)	**0.057 2**	**94.43**	**88.42**

图 5 给出了 3 种不同建模方法的预测结果散

https://dc-china-simulation.researchcommons.org/journal/vol31/iss10/9
DOI: 10.16182/j.issn1004731x.joss.17-0374

点图，图中散点越接近黑色对角线说明预测值越接近真值，说明模型的预测精度越高。从图 5 中可以看出 HEGPR 方法得到的散点分布更紧凑的分布在黑色对角线周围，说明本文提出的建模方法对丁烷浓度的预测值更接近真值，模型预测精度更高。图 6 给出了 3 种建模方法的预测误差曲线，比较可以看出本文所提的 HEGPR 方法预测误差在 0 附近波动幅度更小，表明本文方法预测精度更高。

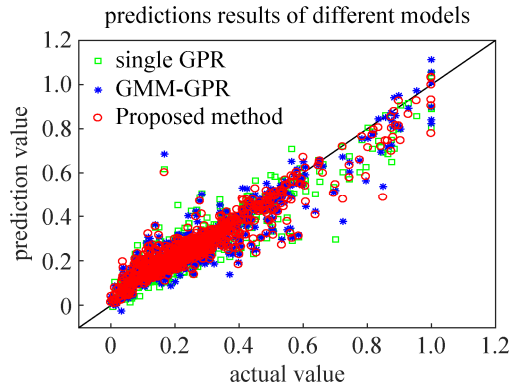


图 5 3 种建模方法对丁烷浓度的预测结果

Fig. 5 Butane concentration prediction result of three modeling methods

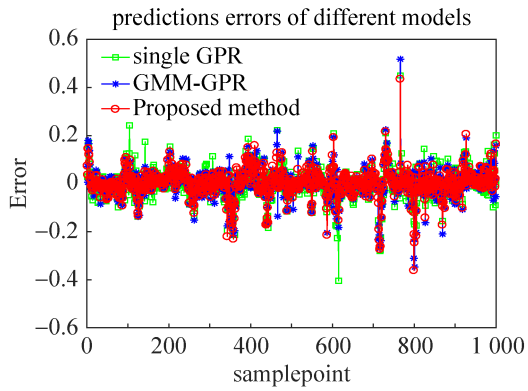
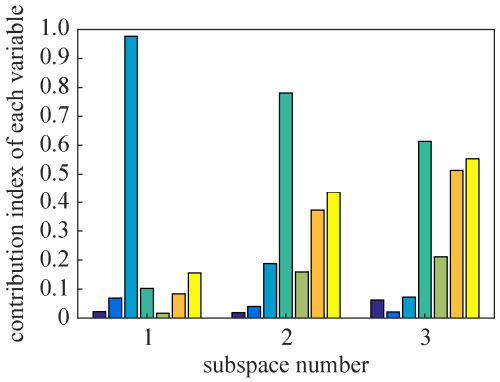


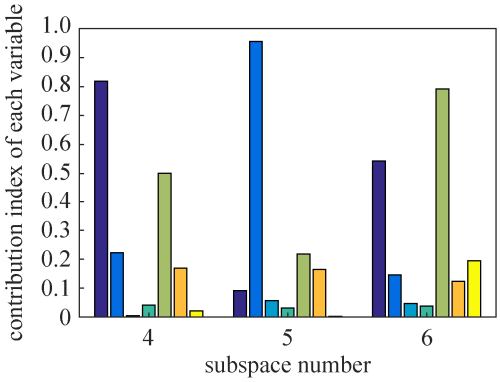
图 6 对丁烷浓度 3 种建模方法的预测误差

Fig. 6 Butane concentration prediction error of three modeling methods

图 7 给出了在 GMM 划分出的第 1 个阶段中各主元子空间中不同变量的贡献程度。不同阶段下分别建立了 6 个子空间模型，图 7(a)对应前 3 个子空间模型的各变量贡献指标，图 7(b)对应后 3 个子空间模型的各变量贡献指标，从图中可以看出，同一种变量在不同的子空间中的信息贡献大小是不同的。因此，相比于 RSM 在子空间中随机选取变量的方式，SPCA 能根据变量贡献指标选择在当前子空间中最重要的变量，从而构建出更精确的子模型。



(a) 子空间 1~3 各变量的贡献率



(b) 子空间 4~6 各变量的贡献率

图 7 阶段 1 中不同子空间下各变量的贡献指标

Fig. 7 Contribution of each variable under different subspaces in phase 1

图 8 给出了 GMM-EGPR 和 HEGPR 两种建模方法的预测结果散点图，从图 8 中可以看出 HEGPR 方法得到的散点分布更紧凑的分布在黑色对角线周围，说明本文提出的建模方法对丁烷浓度的预测值更接近真值，模型预测精度更高。结果表明采用 RSM 随机选取辅助变量，每个变量的权值相同，无法保证重要的变量被用于建模，因此可能会造成模型性能的下降，基于 SPCA 的集成方法，在充分考虑变量的重要程度情况下，对辅助变量进行选取建立集成模型，可以有效的提升模型预测精度和泛化能力。图 9 给出了 GMM-EGPR 和 HEGPR 两种建模方法的预测误差

曲线，比较可以看出本文所提的 HEGPR 方法预测误差在 0 附近波动幅度更小，进一步表明本文方法预测精度更高。

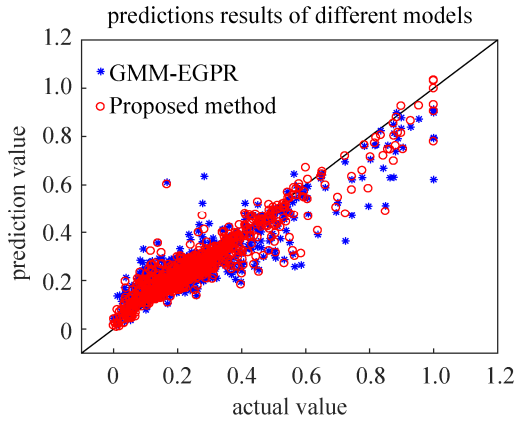


图 8 GMM-EGPR 和 HEGPR 的预测结果散点图
Fig. 8 Prediction result of GMM-EGPR and HEGPR

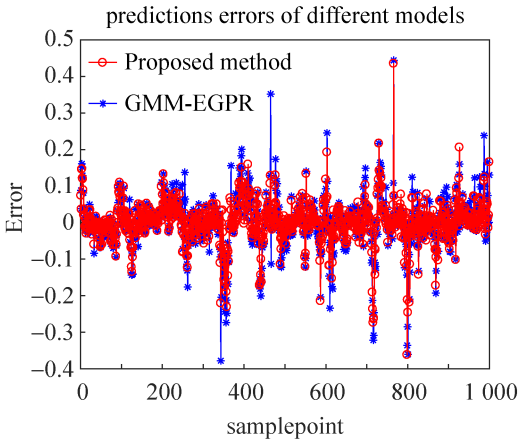


图 9 GMM-EGPR 和 HEGPR 的预测误差
Fig. 9 Prediction error of GMM-EGPR and HEGPR

3.2 青霉素发酵过程

青霉素发酵过程是青霉素产生菌在适宜的温度、PH、溶解氧、基质、搅拌等发酵条件下进行生长和合成抗生素的代谢活动，青霉素发酵过程一般分为 3 个阶段：菌体生长阶段、青霉素合成阶段和菌体自溶阶段[27-28]。通过 Pensim 仿真平台可以模拟青霉素发酵过程，通过使用 Pensim 平台产生 300 组过程数据，其中 200 组作为训练样本集，100 组作为测试样本集。对过程数据采用 GMM 划分为 3 个阶段，每个阶段下面建立 6 个子空间模型，设定不同的 CI 阈值确定每个子空间模型选取的辅助变量个数，每个子空间模型采用的变量个数如图 10 所示。分别对单模型 GPR 方法、GMM-GPR 多模型方法以及本文所提的 HEGPR 方法进行了仿真对比，不同建模方法的性能指标见表 3。

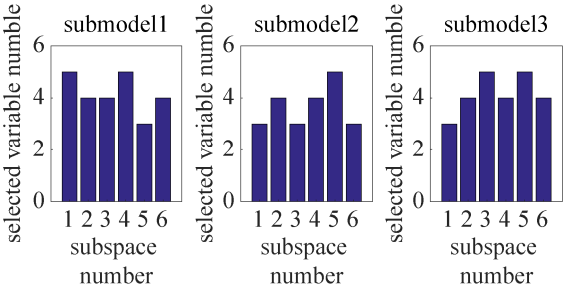


图 10 青霉素发酵过程各模态子空间选择变量个数
Fig. 10 Selection variable number in each subspace of penicillin fermentation process

表 3 青霉素发酵过程不同建模方法的预测性能指标
Tab. 3 Predictive performance indicators for different methods of penicillin fermentation process

	RMSE	COR/%	TP/%
Single-GPR	0.0193	99.91	99.83
GMM-GPR	0.0133	99.96	99.92
HEGPR (Proposed)	**0.0087**	**99.98**	**99.96**

图 11 给出了 3 种不同建模方法的预测结果图，从图中可看出本文提出的 HEGPR 方法预测更为精准，对过程变化跟踪效果更好，采用多模型对过程数据进行阶段划分，可很好的解决多操作阶段的过程建模问题。

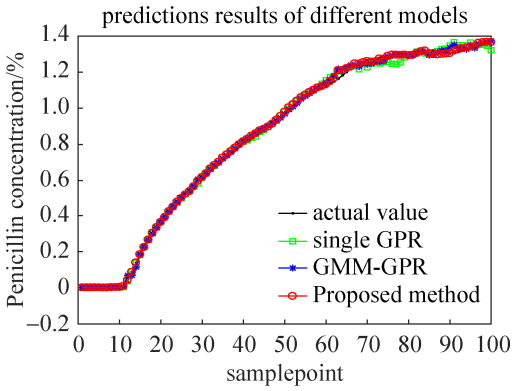


图 11 不同建模方法对青霉素浓度的预测结果
Fig. 11 Penicillin concentration prediction result of different modeling methods

https://dc-china-simulation.researchcommons.org/journal/vol31/iss10/9
DOI: 10.16182/j.issn1004731x.joss.17-0374

图 12 给出了 3 种建模方法的预测误差曲线，比较可以看出本文所提的 HEGPR 方法预测误差在 0 附近波动幅度更小，进一步表明本文方法预测精度更高。

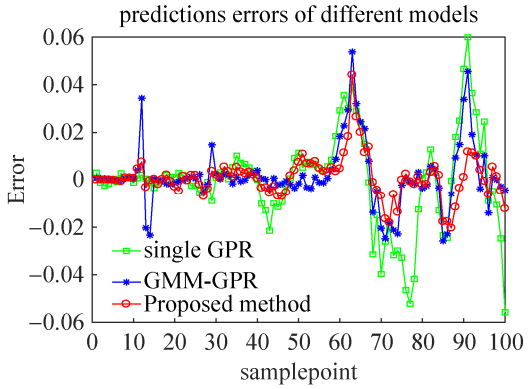


图 12 不同建模方法对青霉素浓度的预测误差

Fig. 12 Penicillin concentration prediction error of different modeling methods

4 结论

针对工业过程中存在的非线性、时变和多阶段特征，提出了分层集成高斯过程回归软测量建模方法，采用 GMM 对过程数据进行阶段划分，对每个操作阶段的数据集采用 SPCA-GPR 建模方法，将变量在主元空间上的贡献度作为变量选取的优先级，在增强模型泛化能力的同时提高了模型的预测精度。通过对脱丁烷塔塔底丁烷浓度和青霉素发酵过程的仿真研究，验证了本文方法的有效性；然而在 SPCA-GPR 建模时变量的选取阈值与子空间模型个数均采用多次实验得到，可以在下一步研究中考虑采用优化算法实现参数的确定。

参考文献:

[1] Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry[J]. Computers & Chemical Engineering (S0098-1354), 2009, 33(4): 795-814.

[2] Shao W, Tian X. Adaptive soft sensor for quality prediction of chemical processes based on selective ensemble of local partial least squares models[J]. Chemical Engineering Research and Design (S0263-8762), 2015, 95: 113-132.

[3] Zhu J, Ge Z, Song Z. Robust supervised probabilistic principal component analysis model for soft sensing of key process variables[J]. Chemical Engineering Science (S0009-2509), 2015, 122: 573-584.

[4] Ge Z, Huang B, Song Z. Mixture semisupervised principal component regression model and soft sensor application[J]. AIChE Journal (S0001-1541), 2014, 60(2): 533-545.

[5] Wang Z X, He Q, Wang J. Comparison of different variable selection methods for partial least squares soft sensor development[C]. American Control Conference (ACC), 2014. Portland, USA: IEEE, 2014: 3116-3121.

[6] Galicia H J, He Q P, Wang J. A reduced order soft sensor approach and its application to a continuous digester[J]. Journal of Process Control (S0959-1524), 2011, 21(4): 489-500.

[7] Gonzaga J C B, Meleiro L A C, Kiang C, et al. ANN-based soft-sensor for real-time process monitoring and control of an industrial polymerization process[J]. Computers & Chemical Engineering (S0098-1354), 2009, 33(1): 43-49.

[8] Forouzantabar A, Talebi H A, Sedigh A K. Adaptive neural network control of bilateral teleoperation with constant time delay[J]. Nonlinear Dynamics (S0924-090X), 2012, 67(2): 1123-1134.

[9] 张倩, 杨耀权. 基于支持向量机回归的火电厂烟气含氧量软测量[J]. 信息与控制, 2013, 42(2): 258-263.
Zhang Qian, Yang Yaoquan. Soft sensor for Oxygen Content in Flue Gas of Coal-fired Power Plant Based on SVR[J]. Information and Control, 2013, 42(2): 258-263.

[10] Yu J. A Bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses[J]. Computers & Chemical Engineering (S0098-1354), 2012, 41: 134-144.

[11] Xiong W, Zhang W, Xu B, et al. JITL based MWGPR soft sensor for multi-mode process with dual-updating strategy[J]. Computers & Chemical Engineering (S0098-1354), 2016, 90: 260-267.

[12] Liu Y, Chen T, Chen J. Auto-switch Gaussian process regression-based probabilistic soft sensors for industrial multigrade processes with transitions[J]. Industrial & Engineering Chemistry Research (S0888-5885), 2015, 54(18): 5037-5047.

[13] Jin H, Chen X, Wang L, et al. Adaptive soft sensor development based on online ensemble Gaussian process regression for nonlinear time-varying batch processes[J]. Industrial & Engineering Chemistry

Research (S0888-5885), 2015, 54(30): 7320-7345.

[14] 张子迎, 周明全, 税午阳, 等. 基于 K 均值聚类和区域匹配的颜色迁移技术[J]. 系统仿真学报, 2015, 27(10): 2359-2364.
Zhang Ziying, Zhou Mingquan, Shui Wuyang, et al. Color Transfer Based on K-means Clustering Algorithm and Region Matching[J]. Journal of System Simulation, 2015, 27(10): 2359-2364.

[15] 嵇小辅, 张翔. 基于FCM与集成高斯过程回归的赖氨酸发酵软测量[J]. 智能系统学报, 2015, 10(1): 156-162.
Ji Xiaofu, Zhang Xiang. Soft measurement of lysine fermentation based on FCM and integrated Gaussian process regression[J]. CAAI Transactions on Intelligent Systems, 2015, 10(1): 156-162.

[16] Yu J. Multiway Gaussian mixture model based adaptive kernel partial least squares regression method for soft sensor estimation and reliable quality prediction of nonlinear multiphase batch processes[J]. Industrial & Engineering Chemistry Research (S0888-5885), 2012, 51(40): 13227-13237.

[17] Chen T, Ren J. Bagging for Gaussian process regression[J]. Neurocomputing (S0925-2312), 2009, 72(7): 1605-1610.

[18] Cao D S, Xu Q S, Liang Y Z, et al. The boosting: A new idea of building models[J]. Chemometrics and Intelligent Laboratory Systems (S0169-7439), 2010, 100(1): 1-11.

[19] Li T, Chen Y. Fuzzy Clustering Ensemble with Selection of Number of Clusters[J]. Journal of Computers (S1796-203X), 2010, 5(7): 1112-1119.

[20] Ge Z, Song Z. Subspace partial least squares model for multivariate spectroscopic calibration[J]. Chemometrics and Intelligent Laboratory Systems (S0169-7439), 2013, 125: 51-57.

[21] Wang L, Jin H, Chen X, et al. Soft sensor development based on the hierarchical ensemble of Gaussian process regression models for nonlinear and non-Gaussian chemical processes[J]. Industrial & Engineering Chemistry Research (S0888-5885), 2016, 55(28): 7704-7719.

[22] 张伟, 熊伟丽, 徐保国. 基于实时学习的高斯过程回归多模型融合建模[J]. 信息与控制, 2015, 44(4): 487-492,498.
Zhang Wei, Xiong Weili, Xu Baoguo. Multi-model Combination Modeling Based on Just-in-time Learning Using Gaussian Process Regression[J]. Information and Control, 2015, 44(4): 487-492,498.

[23] Yu J. Online quality prediction of nonlinear and non-Gaussian chemical processes with shifting dynamics using finite mixture model based Gaussian process regression approach[J]. Chemical Engineering Science (S0009-2509), 2012, 82: 22-30.

[24] 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.
He Zhikun, Liu Guangbing, Zhao Xijing, et al. Overview of Gaussian process regression[J]. Control and Decision, 2013, 28(8): 1121-1129.

[25] Jin H, Chen X, Wang L, et al. Adaptive soft sensor development based on online ensemble Gaussian process regression for nonlinear time-varying batch processes[J]. Industrial & Engineering Chemistry Research (S0888-5885), 2015, 54(30): 7320-7345.

[26] 阮宏镁, 田学民, 王平. 基于联合互信息的动态软测量方法[J]. 化工学报, 2014, 65(11): 4497-4502.
Ruan Hongmei, Tian Xuemin, Wang Ping. Dynamic soft sensor method based on joint mutual information[J]. CIESC Journal, 2014, 65(11): 4497-4502.

[27] Liu Y, Wang H. Pensim simulator and its application in penicillin fermentation process[J]. Journal of System Simulation (S1004-731X), 2006, 12(12): 3524-3527.

[28] 熊伟丽, 姚乐, 徐保国. 基于 EM 算法的青霉素发酵过程多阶段融合建模[J]. 化工学报, 2014, 65(12): 4935-4941.
Xiong Weile, Yao Le, Xu Baoguo. Multi-stage fusion modeling for penicillin fermentation process based on EM algorithm[J]. CIESC Journal, 2014, 65(12): 4935-4941.