Journal of System Simulation

Volume 30 | Issue 1 Article 2

1-2-2019

Online Modeling with Semi-Supervised Locally Weighted Partial Least Squares Based on Expectation Maximization Algorithm

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

Mingchen Xue
1.School of Internet of Things Engineering, Institute of Automation, Jiangnan University, Wuxi 214122, China; ;

Yanjun Li
1.School of Internet of Things Engineering, Institute of Automation, Jiangnan University, Wuxi 214122, China; ;

Follow this and additional works at: https://dc-china-simulation.researchcommons.org/journal

 Part of the Artificial Intelligence and Robotics Commons, Computer Engineering Commons, Numerical Analysis and Scientific Computing Commons, Operations Research, Systems Engineering and Industrial Engineering Commons, and the Systems Science Commons

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation.

Online Modeling with Semi-Supervised Locally Weighted Partial Least Squares Based on Expectation Maximization Algorithm

Abstract

Abstract: *As only small proportion of labeled data* can be obtained from chemical processes, *an online soft sensing modeling method based on semi-supervised locally weighted partial least squares is proposed* . The labeled and unlabeled historical data are accumulated to construct training database. *The similarity between the newly measured data and the sample points in the database are calculated and used as the weight of each data point. The semi-supervised locally weighted partial least squares model is constructed, and the Expectation Maximization (EM) algorithm is employed to estimate the parameters of the model.* Online prediction achieves simulation results of debutanizer distillation processes, which suggests that the proposed method has good prediction accuracy and stable generalization performance.

Keywords

semi-supervised, locally weighted partial least squares, EM algorithm, online modeling

Recommended Citation

Xiong Weili, Xue Mingchen, Li Yanjun. Online Modeling with Semi-Supervised Locally Weighted Partial Least Squares Based on Expectation Maximization Algorithm[J]. Journal of System Simulation, 2018, 30(1): 8-10.

This paper is available in Journal of System Simulation: https://dc-china-simulation.researchcommons.org/journal/vol30/iss1/2

第 30 卷第 1 期 系统仿真学报© Vol. 30 No. 1
2018 年 1 月 Journal of System Simulation Jan., 2018

基于 EM 算法的半监督局部加权 PLS 在线建模方法

熊伟丽[1,2]，薛明晨[1]，李妍君[1]
（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：针对化工过程采样分析获得的*有标签样本数量较少*的问题，提出一种*基于半监督学习的局部加权偏最小二乘在线软测量建模方法*。将过程收集到的有标签及无标签训练样本放入同一数据库中；*对于在线测得的新数据点，计算其与数据库中各样本点之间的相似度，将其作为各数据点的权重；建立半监督局部加权偏最小二乘在线软测量模型，并采用 EM(Expectation Maximization)算法估计模型的参数，得到模型的在线预测输出*。通过对脱丁烷塔过程的仿真研究，验证了所提方法具有良好的预测精度和泛化性能。

关键词：半监督；局部加权偏最小二乘；EM 算法；在线建模
中图分类号：TP391 文献标识码：A 文章编号：1004-731X (2018) 01-0008-10
DOI: 10.16182/j.issn1004731x.joss.201801002

Online Modeling with Semi-Supervised Locally Weighted Partial Least Squares Based on Expectation Maximization Algorithm

Xiong Weili[1,2], *Xue Mingchen*[1], *Li Yanjun*[1]
(1. School of Internet of Things Engineering, Institute of Automation, Jiangnan University, Wuxi 214122, China;
2. Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China)

Abstract: *As only small proportion of labeled data* can be obtained from chemical processes, *an online soft sensing modeling method based on semi-supervised locally weighted partial least squares is proposed* . The labeled and unlabeled historical data are accumulated to construct training database. *The similarity between the newly measured data and the sample points in the database are calculated and used as the weight of each data point. The semi-supervised locally weighted partial least squares model is constructed, and the Expectation Maximization (EM) algorithm is employed to estimate the parameters of the model.* Online prediction achieves simulation results of debutanizer distillation processes, which suggests that the proposed method has good prediction accuracy and stable generalization performance.
Keywords: semi-supervised; locally weighted partial least squares; EM algorithm; online modeling

引言

在实际化工生产过程中，由于操作条件复杂、工况变化快或测量技术的局限，一些与产品质量密切相关的过程变量的在线测量往往很困难，软测量技术可以用来解决某些复杂工业生产过程中质量变量的在线预测，所以建立高精度的软测量模型显得十分重要。目前常采用的软测量建模方法主要有基于机理模型的方法、基于数据驱动以及两者相结合的混合建模方法。基于机理模型的建模方法需要了解过程的详细信息，运用物料守恒、能量守恒、化学反应动力学等原理，对对象的操作过程和工艺

收稿日期：2015-11-17 修回日期：2016-01-30；
基金项目：国家自然科学基金(61773182)，江苏省“六大人才高峰”计划(2013-DZXX-043)；
作者简介：熊伟丽(1978-)，女，河南洛阳，博士，教授，硕导，研究方向为复杂工业过程建模及优化，智能优化算法及应用。

Published by Journal of System Simulation, 2018

原理进行分析，建立机理模型，对于复杂的化工过程而言，这些是很难获得的，建立相应的机理模型也变得非常困难，由于通过 DCS 系统获得的过程数据越来越多，基于数据驱动的建模方法得到越来越多的关注。

近几十年来，基于数据驱动的软测量模型无论在工业领域或学术领域均受到了广泛的关注。常用的建模方法有多元线性回归(Multiple Linear Regression，MLR)[1]、主元回归法(Principal Component Regression，PCR)[2]、部分最小二乘(Partial Least Squares，PLS)[3-4]等，这些均为线性建模方法，对于非线性严重的过程，常利用非线性建模方法如支持向量机(Support Vector Machines，SVM)[5]、非线性 PLS[6]、局部加权回归(Locally Weighted Regression，LWR)[7]等。局部加权偏最小二乘(Locally Weighted Partial Least Squares，LWPLS)是基于即时(Just-In-Time，JIT)建模思想在传统 PLS 上的改进[8]，根据数据库中不同样本点与新数据点之间的相似度，建立局部模型时考虑它们的权重，从而减小了相似度较低的样本对建模的影响，提高了建模精度。另外，由于局部加权偏最小二乘建模方法是对每一在线测得的数据均建立局部模型，用局部线性模型来逼近非线性过程，所以能够较好地描述非线性过程的动态特征。

但是，传统的基于数据驱动的软测量建模方法往往需要既有输入又有输出的样本数据来进行建模，即有标签样本，将只有输入的样本数据，即无标签样本丢弃，过程数据的信息得不到充分利用，建立的模型性能往往也得不到保证。而对应品质变量的输出样本一般很难测量，这样会导致采集到的输入样本与输出样本数量不对应，在获得大量输入样本数据的同时获得的输出样本数据仅占小部分。针对此类问题，章军等[9]利用无标签样本中的支持向量，训练出来的 SVM 回归模型具有更好的估计精度和泛化性能；葛志强等[10]提出基于半监督学习思想的 SBPCR 建模方法，充分利用有标签样本及无标签样本，提高了模型的预测精度及鲁棒性；吴继明等[11]提出一种基于 Boosting 思想的半监督学习算法 IMSB，解决了多分类问题，并在训练样本集中有标签样本比较少的情况下取得较好的效果。

结合实际化工过程，为了解决有标签样本数量较少，充分利用无标签样本包含的过程信息，本文提出一种基于半监督学习的局部加权偏最小二乘在线软测量建模方法，利用有标签样本的同时，充分利用无标签样本所携带的信息，建立半监督局部加权偏最小二乘模型，并采用 EM 算法求取实时更新的半监督局部加权偏最小二乘模型的参数，得到模型的在线预测输出。通过对脱丁烷塔过程的仿真研究，结果表明本文所提方法可以实现对工业过程中关键变量的有效预测。

1 局部加权偏最小二乘

LWPLS 是一种基于 JIT 建模思想的建模方法，在建立局部模型时，考虑了新来数据与数据库中各样本点之间的相似度，计算出它们之间的相似度值并将其作为各数据点的权重，继而建立局部模型。因此，相比 PLS 建模而言，LWPLS 能够更好描述化工过程中的非线性特征。假设存入数据库中的输入输出变量矩阵为 $\boldsymbol{X} \in R^{N\times M}$, $\boldsymbol{Y} \in R^{N\times R}$，其中第 i 个输入输出样本分别为 $\boldsymbol{x_i}=[x_{i1},x_{i2},...,x_{iM}]^{\mathrm{T}}$ 及 $\boldsymbol{y_i}=[y_{i1},y_{i2},...,y_{iR}]^{\mathrm{T}}$，$i=1,2,\ldots,N$，$N$ 为样本个数；M 和 R 分别为输入、输出样本的维度。PLS 模型可以表示成式(1)形式：

$$\begin{cases}\boldsymbol{X}=\boldsymbol{T}\boldsymbol{P}^{\mathrm{T}}+\boldsymbol{E}_{\boldsymbol{X}}\\ \boldsymbol{Y}=\boldsymbol{T}\boldsymbol{C}^{\mathrm{T}}+\boldsymbol{F}_{\boldsymbol{Y}}\end{cases} \tag{1}$$

这里，$\boldsymbol{P} \in R^{M\times K}$ 为负荷矩阵，$\boldsymbol{T} \in R^{N\times K}$ 为主成分矩阵，K 为选择的主成分的个数，$\boldsymbol{C} \in R^{R\times K}$ 为回归矩阵，$\boldsymbol{E}_{\boldsymbol{X}} \in R^{N\times M}$ 及 $\boldsymbol{E}_{\boldsymbol{Y}} \in R^{N\times R}$ 分别对应输入输出变量矩阵的残差矩阵。

对于新的数据点 $\boldsymbol{x_q}=[x_{q1},x_{q2},...,x_{qM}]^{\mathrm{T}}$，首先根据式(2)计算出它与数据库中各样本点之间的相似度：

https://dc-china-simulation.researchcommons.org/journal/vol30/iss1/2
DOI: 10.16182/j.issn1004731x.joss.201801002

Xiong et al.: Online Modeling with Semi-Supervised Locally Weighted Partial Lea

$$\begin{cases}\omega_i = \mathrm{e}^{-\frac{d_i}{\sigma_d \varphi}} \\ d_i = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_q)^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{x}_q)}\end{cases} \tag{2}$$

式中，σ_d 是 d_i 的标准差，φ 是位置参数，一般取 0.1~1.5 之间[12]。将由(2)式计算得到的 ω_i 表示成相似矩阵 $\boldsymbol{\Omega} \in R^{N\times N}$ 的形式为：

$$\boldsymbol{\Omega} \in \mathrm{diag}(\omega_1, \omega_2, ..., \omega_N) \tag{3}$$

LWPLS 的算法步骤如下：

Step1：确定主成分 K 的值，并设置其在算法中的初值 k=1。K 的值通常可通过主元分析法(PCA)来确定。

Step2：根据公式(2)、(3)计算出相似矩阵 $\boldsymbol{\Omega}$。

Step3：对输入输出矩阵及新来数据点进行预处理，计算出 $\boldsymbol{X}_k, \boldsymbol{Y}_k$ 及 $\boldsymbol{x}_{qk}$

$$\begin{cases}\boldsymbol{X}_k = \boldsymbol{X} - \boldsymbol{I}_N\left[\bar{x}_1, \bar{x}_2, ..., \bar{x}_M\right] \\ \boldsymbol{Y}_k = \boldsymbol{Y} - \boldsymbol{I}_N\left[\bar{y}_1, \bar{y}_2, ..., \bar{y}_R\right] \\ \boldsymbol{x}_{qk} = \boldsymbol{x}_q - \left[\bar{x}_1, \bar{x}_2, ..., \bar{x}_M\right]^{\mathrm{T}}\end{cases} \tag{4}$$

$$\bar{x}_m = \frac{\sum_{i=1}^{N}\omega_i x_{im}}{\sum_{i=1}^{N}\omega_i},\ \bar{y}_r = \frac{\sum_{i=1}^{N}\omega_i y_{ir}}{\sum_{i=1}^{N}\omega_i} \tag{5}$$

其中 $\boldsymbol{I}_N \in R^N$ 为全 1 的列向量，$m = 1, 2, ..., M$，$r = 1, 2, ..., R$。

Step4：推断出局部线性模型

$\boldsymbol{X}$ 的主成分向量：$\boldsymbol{t}_k = \boldsymbol{X}_k \boldsymbol{w}_k$ (6)

$\boldsymbol{X}$ 的负荷向量：$\boldsymbol{p}_k = \dfrac{\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{t}_k}{\boldsymbol{t}_k\boldsymbol{\Omega}\boldsymbol{t}_k}$ (7)

模型回归系数向量：$\boldsymbol{c}_k = \dfrac{\boldsymbol{Y}_k^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{t}_k}{\boldsymbol{t}_k\boldsymbol{\Omega}\boldsymbol{t}_k}$ (8)

新来数据点的得分：$t_{qk} = \boldsymbol{x}_{qk}^{\mathrm{T}}\boldsymbol{w}_k$ (9)

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \boldsymbol{t}_k \boldsymbol{p}_k^{\mathrm{T}} \tag{10}$$

$$\boldsymbol{Y}_{k+1} = \boldsymbol{Y}_k - \boldsymbol{t}_k \boldsymbol{c}_k^{\mathrm{T}} \tag{11}$$

$$\boldsymbol{x}_{q(k+1)} = \boldsymbol{x}_{qk} - t_{qk}\boldsymbol{p}_k \tag{12}$$

如果 k=K，则转入下一步；否则，使 k=k+1 并转到 Step4。这里，$\boldsymbol{w}_k$ 是 $\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{Y}_k\boldsymbol{Y}_k^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{X}_k$ 的最大特征值所对应的特征向量。

Step5：通过上面步骤得到输入样本 $\boldsymbol{X}$ 的得分矩阵 $\boldsymbol{T}$=$[\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_k]$，负荷矩阵 $\boldsymbol{P}$ =$[\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_k]$，模型回归系数矩阵 $\boldsymbol{C}$ = $[\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_k]$ 以及新来数据点的得分向量 $\boldsymbol{t}$ = $[t_{q1}, t_{q2}, \cdots, t_{qk}]$，可以计算新来数据点的输出估计值 $\hat{\boldsymbol{y}}_q$：

$$\hat{\boldsymbol{y}}_q = \boldsymbol{t}\boldsymbol{C}^{\mathrm{T}} = [t_{q1}, t_{q2}, ..., t_{qk}]\left[\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_k\right]^{\mathrm{T}} \tag{13}$$

2 基于局部加权偏最小二乘回归的半监督学习

2.1 半监督局部加权偏最小二乘模型

对实际工业过程中获得的无标签样本，传统的 LWPLS 方法无法利用这些样本来进行建模，常见的处理方法是将这些无标签样本丢弃，这样可能会影响模型的预测精度和泛化性能。半监督学习可以在利用有标签样本的同时，利用无标签样本去增强机器学习能力，以提高模型的性能[13-14]。对于每一个新来样本数据，基于半监督学习的 LWPLS 模型可以表示为[15]：

$$\begin{cases}\boldsymbol{x}_i = \boldsymbol{P}\boldsymbol{t}_i + \boldsymbol{e}_{xi} \\ \boldsymbol{y}_j = \boldsymbol{C}\boldsymbol{t}_j + \boldsymbol{f}_{yj}\end{cases} \tag{14}$$

其中 $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n_1$；n_1 为数据库中有标签的样本的个数，$n_2 = n - n_1$ 为数据库中无标签样本的个数。$\boldsymbol{P} \in R^{m\times K}, \boldsymbol{C} \in R^{r\times K}$ 分别表示对应的负荷矩阵及回归矩阵，其中 m 为输入变量的个数，r 为输出变量的个数，$\boldsymbol{t} \in R^{K\times 1}$ 为主成分向量，$\boldsymbol{e}_x \in R^{m\times 1}, \boldsymbol{f}_y \in R^{r\times 1}$ 分别为过程噪声的输入输出变量。在半监督 LWPLS 模型中，一般假设主成分及过程噪声的概率密度函数均服从高斯分布，即 $p(\boldsymbol{t}) = N(0, \boldsymbol{I})$，$p(\boldsymbol{e}) = N(0, \sigma_x^2/\omega_i\ \boldsymbol{I})$，$p(\boldsymbol{f}) = N(0, \sigma_y^2/\omega_j\ \boldsymbol{I})$，其中 $\boldsymbol{I}$ 为单位矩阵，σ_x^2/ω_i，σ_y^2/ω_j 分别为输入变量和输出变量的噪声方差，ω_i, ω_j 越小，则表明包含的噪声信息越多，对模型的贡献越小。

2.2 基于 EM 算法估计半监督 LWPLS 模型参数

2.2.1 期望最大化算法

期望最大化(Expectation-Maximization，EM)

Published by Journal of System Simulation, 2018

Journal of System Simulation, Vol. 30 [2018], Iss. 1, Art. 2

算法是由 Dempster 提出的一种由已知数据估计未知参数的迭代方法，旨在求解数据有缺失或隐含情况下参数的极大似然估计[16-18]，适用于解决不完全数据建模问题。该算法由 E 步和 M 步组成，在 E 步中根据隐含变量的后验分布计算条件期望，即 Q 函数；在 M 步中根据 E 步得到的条件期望求解参数的极大似然估计。E 步和 M 步间交替进行计算，直到参数的估计值收敛。

2.2.2 EM 算法在半监督 LWPLS 模型参数估计中的应用

E 步：选取主成分向量 $\{\boldsymbol{t}_i\}_{i=1}^{n}$ 作为隐含变量，即 $\boldsymbol{C}_{miss}=\{\boldsymbol{T}\}=\{\boldsymbol{t}_i,i=1,2,\cdots,n\}$；可观测数据集为 $\boldsymbol{C}_{obs}=\{\boldsymbol{X},\boldsymbol{Y}\}$；则完整数据集可定义为 $\boldsymbol{C}=\{\boldsymbol{C}_{obs},\boldsymbol{C}_{miss}\}=\{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{T}\}$，模型参数集为 $\boldsymbol{\Theta}=\{\boldsymbol{P},\boldsymbol{C},\sigma_x^2,\sigma_y^2\}$，根据 EM 算法，首先计算 Q 函数为：

$$
\begin{aligned}
&\mathrm{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^S)=E_{\boldsymbol{C}_{miss}|\boldsymbol{C}_{obs},\boldsymbol{\Theta}^S}\ln \mathrm{P}(\boldsymbol{C}_{obs},\boldsymbol{C}_{miss}|\boldsymbol{\Theta})=\\
&E_{\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S}\ln P(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{T}|\boldsymbol{\Theta})=\\
&E_{\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S}\sum_{j=1}^{n_1}\ln P(\boldsymbol{x}_j|\boldsymbol{t}_j,\boldsymbol{\Theta})+E_{\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S}\sum_{j'=1}^{n_2}\ln P(\boldsymbol{x}_{j'}|\boldsymbol{t}_{j'},\boldsymbol{\Theta})+\\
&E_{\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S}\sum_{j=1}^{n_1}\ln P(\boldsymbol{y}_j|\boldsymbol{t}_j,\boldsymbol{\Theta})+E_{\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S}\sum_{j=1}^{n_1}\ln P(\boldsymbol{t}_j|\boldsymbol{\Theta})+\\
&E_{\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S}\sum_{j'=1}^{n_2}\ln P(\boldsymbol{t}_{j'}|\boldsymbol{\Theta})=\\
&\int_{\boldsymbol{T}}\sum_{j=1}^{n_1}\ln P(\boldsymbol{x}_j|\boldsymbol{t}_j,\boldsymbol{\Theta})P(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S)d\boldsymbol{T}+\\
&\int_{\boldsymbol{T}}\sum_{j'=1}^{n_2}\ln P(\boldsymbol{x}_{j'}|\boldsymbol{t}_{j'},\boldsymbol{\Theta})P(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S)d\boldsymbol{T}+\\
&\int_{\boldsymbol{T}}\sum_{j=1}^{n_1}\ln P(\boldsymbol{y}_j|\boldsymbol{t}_j,\boldsymbol{\Theta})P(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S)d\boldsymbol{T}+\\
&\int_{\boldsymbol{T}}\sum_{j=1}^{n_1}\ln P(\boldsymbol{t}_j|\boldsymbol{\Theta})P(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S)d\boldsymbol{T}+\\
&\int_{\boldsymbol{T}}\sum_{j'=1}^{n_2}\ln P(\boldsymbol{t}_{j'}|\boldsymbol{\Theta})P(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S)d\boldsymbol{T}
\end{aligned}
\tag{15}
$$

将式(15)最后一个等式等号右端第一项进一步展开可得：

$$
\begin{aligned}
&\int_{\boldsymbol{T}}\sum_{j=1}^{n_1}\ln P(\boldsymbol{x}_j|\boldsymbol{t}_j,\boldsymbol{\Theta})P(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S)d\boldsymbol{T}=\\
&\sum_{j=1}^{n_1}\int P(\boldsymbol{t}_j|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)\ln P(\boldsymbol{x}_j|\boldsymbol{t}_j,\boldsymbol{\Theta})d\boldsymbol{t}_j=\\
&\sum_{j=1}^{n_1}\int P(\boldsymbol{t}_j|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)\left\{\ln(2\pi)^{-\frac{m}{2}}\left|\frac{\sigma_x^2}{\omega_j}\boldsymbol{I}\right|^{-\frac{1}{2}}-\right.\\
&\left.\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{P}\boldsymbol{t}_j)^{\mathrm{T}}\left(\frac{\sigma_x^2}{\omega_j}\boldsymbol{I}\right)^{-1}(\boldsymbol{x}_j-\boldsymbol{P}\boldsymbol{t}_j)\right\}d\boldsymbol{t}_j=\\
&\sum_{j=1}^{n_1}\left\{-\frac{m}{2}\ln 2\pi-\frac{1}{2}\ln\left|\frac{\sigma_x^2}{\omega_j}\boldsymbol{I}\right|\right\}-\\
&\sum_{j=1}^{n_1}\frac{\omega_j}{2\sigma_x^2}\int(\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j-\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{P}\boldsymbol{t}_j-\boldsymbol{t}_j^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{x}_j+\boldsymbol{t}_j^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{t}_j)\cdot\\
&P(\boldsymbol{t}_j|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)d\boldsymbol{t}_j=\\
&\sum_{j=1}^{n_1}\left\{-\frac{m}{2}\ln 2\pi-\frac{1}{2}\ln\left|\frac{\sigma_x^2}{\omega_j}\boldsymbol{I}\right|\right\}-\\
&\sum_{j=1}^{n_1}\frac{\omega_j}{2\sigma_x^2}\left\{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j-\int\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{P}\boldsymbol{t}_jP(\boldsymbol{t}_j|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)d\boldsymbol{t}_j-\right.\\
&\int\boldsymbol{t}_j^{\mathrm{T}}(\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{P})^{\mathrm{T}}P(\boldsymbol{t}_j|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)d\boldsymbol{t}_j+\\
&\left.\int\boldsymbol{t}_j^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{t}_jP(\boldsymbol{t}_j|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)d\boldsymbol{t}_j\right\}=\\
&\sum_{j=1}^{n_1}\left\{-\frac{m}{2}\ln 2\pi-\frac{1}{2}\ln\left|\frac{\sigma_x^2}{\omega_j}\boldsymbol{I}\right|\right\}-\sum_{j=1}^{n_1}\frac{\omega_j}{2\sigma_x^2}\left\{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j-\right.\\
&2\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{P}E(\boldsymbol{t}_j|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)+\\
&\left.\mathrm{trace}\left[\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}E(\boldsymbol{t}_j\boldsymbol{t}_j^{\mathrm{T}}|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)\right]\right\}=\\
&\sum_{j=1}^{n_1}\left\{-\frac{m}{2}\ln 2\pi-\frac{1}{2}\ln\left|\frac{\sigma_x^2}{\omega_j}\boldsymbol{I}\right|\right\}-\sum_{j=1}^{n_1}\frac{\omega_j}{2\sigma_x^2}\left\{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j-\right.\\
&2\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{P}E(\boldsymbol{t}_j|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)+\\
&\left.\mathrm{trace}[\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}E(\boldsymbol{t}_j\boldsymbol{t}_j^{\mathrm{T}}|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S)]\right\}
\end{aligned}
\tag{16}
$$

同理，将公式(15)最后一个等式等号右端第二、三、四、五展开分别可得：

https://dc-china-simulation.researchcommons.org/journal/vol30/iss1/2
DOI: 10.16182/j.issn1004731x.joss.201801002

第 30 卷第 1 期 系统仿真学报 Vol. 30 No. 1
2018 年 1 月 Journal of System Simulation Jan., 2018

$$\int_{\boldsymbol{T}}\sum_{j'=1}^{n_2}\ln P(\boldsymbol{x}_{j'}\left|\boldsymbol{t}_{j'},\boldsymbol{\Theta}\right.)P(\boldsymbol{T}\left|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S\right.)d\boldsymbol{T}=$$
$$\sum_{j'=1}^{n_2}\left\{-\frac{m}{2}\ln 2\pi-\frac{1}{2}\ln\left|\frac{\sigma_x^2}{\omega_{j'}}\boldsymbol{I}\right|\right\}-$$
$$\sum_{j'=1}^{n_2}\frac{\omega_{j'}}{2\sigma_x^2}\left\{\boldsymbol{x}_{j'}^{\mathrm{T}}\boldsymbol{x}_{j'}-2\boldsymbol{x}_{j'}^{\mathrm{T}}\boldsymbol{P}E(\boldsymbol{t}_{j'}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}^S\right.)+\right.$$
$$\left.\mathrm{trace}[\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}E(\boldsymbol{t}_{j'}\boldsymbol{t}_{j'}^{\mathrm{T}}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}^S\right.)]\right\} \quad (17)$$

$$\int_{\boldsymbol{T}}\sum_{j=1}^{n_1}\ln P(\boldsymbol{y}_j\left|\boldsymbol{t}_j,\boldsymbol{\Theta}\right.)P(\boldsymbol{T}\left|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S\right.)d\boldsymbol{T}=$$
$$\sum_{j=1}^{n_1}\left\{-\frac{r}{2}\ln 2\pi-\frac{1}{2}\ln\left|\frac{\sigma_y^2}{\omega_j}\boldsymbol{I}\right|\right\}-$$
$$\sum_{j=1}^{n_1}\frac{\omega_j}{2\sigma_y^2}\left\{\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{y}_j-2\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{C}E(\boldsymbol{t}_j\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)+\right.$$
$$\left.\mathrm{trace}\left[\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}E(\boldsymbol{t}_j\boldsymbol{t}_j^{\mathrm{T}}\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)\right]\right\} \quad (18)$$

$$\int_{\boldsymbol{T}}\sum_{j=1}^{n_1}\ln P(\boldsymbol{t}_j\left|\boldsymbol{\Theta}\right.)P(\boldsymbol{T}\left|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S\right.)d\boldsymbol{T}=$$
$$\sum_{j=1}^{n_1}\left\{-\frac{K}{2}\ln 2\pi-\frac{1}{2}\mathrm{trace}\left[E(\boldsymbol{t}_j\boldsymbol{t}_j^{\mathrm{T}}\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)\right]\right\} \quad (19)$$

$$\int_{\boldsymbol{T}}\sum_{j'=1}^{n_2}\ln P(\boldsymbol{t}_{j'}\left|\boldsymbol{\Theta}\right.)P(\boldsymbol{T}\left|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\Theta}^S\right.)d\boldsymbol{T}=$$
$$\sum_{j'=1}^{n_2}\left\{-\frac{K}{2}\ln 2\pi\text{-}\frac{1}{2}\mathrm{trace}[E(\boldsymbol{t}_{j'}\boldsymbol{t}_{j'}^{\mathrm{T}}\left|\boldsymbol{x}_{j'},\boldsymbol{y}_{j'},\boldsymbol{\Theta}^S\right.)]\right\} \quad (20)$$

M 步：根据 E 步化简得到的 Q 函数，分别对模型参数 $\boldsymbol{\Theta}$ 求偏导并置 0，得到：

$$\frac{\partial[Q(\boldsymbol{\Theta}\left|\boldsymbol{\Theta}^S\right.)]}{\partial\boldsymbol{P}}=0\Rightarrow$$
$$\boldsymbol{P}^{\mathrm{new}}=\left[\sum_{j=1}^{n_1}\omega_j\boldsymbol{x}_jE^{\mathrm{T}}(\boldsymbol{t}_j\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)+\right.$$
$$\left.\sum_{j'=1}^{n_2}\omega_{j'}\boldsymbol{x}_{j'}E^{\mathrm{T}}(\boldsymbol{t}_{j'}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}^S\right.)\right]$$
$$\cdot\left[\sum_{j=1}^{n_1}\omega_jE(\boldsymbol{t}_j\boldsymbol{t}_j^{\mathrm{T}}\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)+\sum_{j'=1}^{n_2}\omega_{j'}E(\boldsymbol{t}_{j'}\boldsymbol{t}_{j'}^{\mathrm{T}}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}^S\right.)\right]^{-1} \quad (21)$$

$$\frac{\partial\left[Q(\boldsymbol{\Theta}\left|\boldsymbol{\Theta}^S\right.)\right]}{\partial\boldsymbol{C}}=0\Rightarrow$$
$$\boldsymbol{C}^{new}=\left[\sum_{j=1}^{n_1}\omega_j\boldsymbol{y}_jE^{\mathrm{T}}(\boldsymbol{t}_j\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)\right]$$
$$\left[\sum_{j=1}^{n_1}\omega_jE(\boldsymbol{t}_j\boldsymbol{t}_j^{\mathrm{T}}\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)\right]^{-1} \quad (22)$$

$$\frac{\partial\left[Q(\boldsymbol{\Theta}\left|\boldsymbol{\Theta}^S\right.)\right]}{\partial\sigma_x^2}=0\Rightarrow(\sigma_x^2)^{new}=$$
$$\frac{1}{m\left(\sum_{j=1}^{n_1}\omega_j+\sum_{j'=1}^{n_2}\omega_{j'}\right)}\left\{\sum_{j=1}^{n_1}\omega_j\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j+\sum_{j'=1}^{n_2}\omega_{j'}\boldsymbol{x}_{j'}^{\mathrm{T}}\boldsymbol{x}_{j'}+\right.$$
$$\mathrm{trace}\left\{\boldsymbol{P}^{new\mathrm{T}}\boldsymbol{P}^{new}\left[\sum_{j=1}^{n_1}\omega_jE(\boldsymbol{t}_j\boldsymbol{t}_j^{\mathrm{T}}\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)+\right.\right.$$
$$\left.\left.\sum_{j'=1}^{n_2}\omega_{j'}E(\boldsymbol{t}_{j'}\boldsymbol{t}_{j'}^{\mathrm{T}}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}^S\right.)\right]\right\}-$$
$$2\mathrm{trace}\left\{\boldsymbol{P}^{new\mathrm{T}}\left[\sum_{j=1}^{n_1}\omega_j\boldsymbol{x}_jE^{\mathrm{T}}(\boldsymbol{t}_j\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)+\right.\right.$$
$$\left.\left.\left.\sum_{j'=1}^{n_2}\omega_{j'}\boldsymbol{x}_{j'}E^{\mathrm{T}}(\boldsymbol{t}_{j'}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}^S\right.)\right]\right\}\right\} \quad (23)$$

$$\frac{\partial\left[Q(\boldsymbol{\Theta}\left|\boldsymbol{\Theta}^S\right.)\right]}{\partial\sigma_y^2}=0\Rightarrow$$
$$(\sigma_y^2)^{new}=\frac{1}{r\sum_{j=1}^{n_1}\omega_j}\left\{\sum_{j=1}^{n_1}\omega_j\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{y}_j+\right.$$
$$\mathrm{trace}\left\{\boldsymbol{C}^{\mathrm{newT}}\boldsymbol{C}^{new}\left[\sum_{j=1}^{n_1}\omega_jE(\boldsymbol{t}_j\boldsymbol{t}_j^{\mathrm{T}}\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)\right]\right\}-$$
$$\left.2\mathrm{trace}\left\{\boldsymbol{C}^{\mathrm{newT}}\left[\sum_{j=1}^{n_1}\omega_j\boldsymbol{y}_jE^{\mathrm{T}}(\boldsymbol{t}_j\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}^S\right.)\right]\right\}\right\} \quad (24)$$

其中，对于有标签样本以及无标签样本，分别有[19]：

$$E\left(\boldsymbol{t}_j\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}\right.\right)=\left(\frac{\omega_j}{\sigma_x^2}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}+\frac{\omega_j}{\sigma_y^2}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}+\boldsymbol{I}\right)^{-1}$$
$$\cdot\left(\frac{\omega_j}{\sigma_x^2}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{x}_j+\frac{\omega_j}{\sigma_y^2}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{y}_j\right)$$
$$E(\boldsymbol{t}_j\boldsymbol{t}_j^{\mathrm{T}}\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}\right.)=\left(\frac{\omega_j}{\sigma_x^2}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}+\frac{\omega_j}{\sigma_y^2}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}+\boldsymbol{I}\right)^{-1}+$$
$$E(\boldsymbol{t}_j\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}\right.)E^{\mathrm{T}}(\boldsymbol{t}_j\left|\boldsymbol{x}_j,\boldsymbol{y}_j,\boldsymbol{\Theta}\right.) \quad (25)$$

$$E\left(\boldsymbol{t}_{j'}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}\right.\right)=\left(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}+\frac{\sigma_x^2}{\omega_{j'}}\boldsymbol{I}\right)^{-1}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{x}_{j'}$$
$$E\left(\boldsymbol{t}_{j'}\boldsymbol{t}_{j'}^{\mathrm{T}}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}\right.\right)=\frac{\sigma_x^2}{\omega_{j'}}\left(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}+\frac{\sigma_x^2}{\omega_{j'}}\boldsymbol{I}\right)^{-1}+$$
$$E\left(\boldsymbol{t}_{j'}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}\right.\right)E^{\mathrm{T}}\left(\boldsymbol{t}_{j'}\left|\boldsymbol{x}_{j'},\boldsymbol{\Theta}\right.\right) \quad (26)$$

Published by Journal of System Simulation, 2018

Journal of System Simulation, Vol. 30 [2018], Iss. 1, Art. 2

将 $\boldsymbol{\Theta}$ 代入 E 步重新计算期望值，如此 E 步、M 步反复循环进行，直至 $\left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^S\right\|$ 收敛为止。

对于新来数据点 $\boldsymbol{x}_q$，模型的输出估计为：

$$\hat{\boldsymbol{y}}_q = \boldsymbol{C}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P} + \sigma_x^2\boldsymbol{I})^{-1}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{x}_q \tag{27}$$

3 基于 EM 算法的半监督局部加权偏最小二乘建模算法步骤

本文基于半监督 LWPLS 算法进行软测量建模，该方法能充分融合无标签样本所携带的信息，从而避免丢弃无标签样本导致训练所得模型预测精度不高等缺陷。对一个在线测得的新数据点 $\boldsymbol{x}_q$ 的建模流程如图 1 所示，算法步骤描述如下：

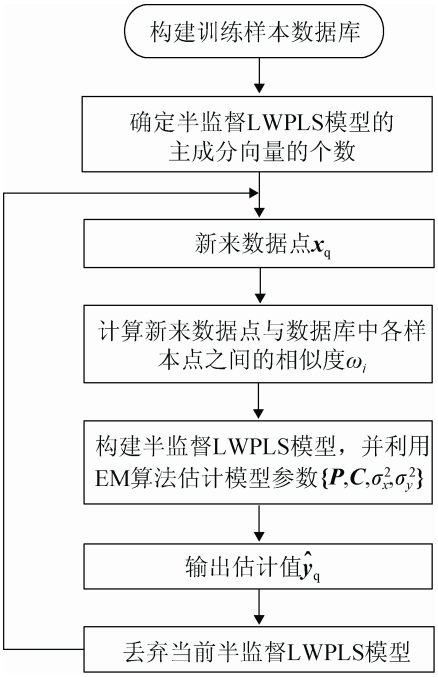

图 1 基于半监督 LWPLS 的在线软测量建模流程
Fig. 1 Online soft sensor modeling process based on semi-supervised LWPLS

Step 1：将实际过程中采集到的有标签样本 $\{(\boldsymbol{x}_j, \boldsymbol{y}_j), \boldsymbol{x}_j \in R^m, \boldsymbol{y}_j \in R^r, j = 1,2,\cdots,n_1\}$ 及无标签样本 $\{\boldsymbol{x}_{j'}, \boldsymbol{x}_{j'} \in R^m, j' = 1,2,\cdots,n_2\}$ 放入数据库中；

Step 2：利用 PCA 确定需要选取的半监督 LWPLS 模型主成分向量的个数 K(即选取的半监督 LWPLS 模型投影的个数)。PCA 可以对高维数据进行降维，通过累计方差贡献率准则得到模型的 K 个投影；

Step 3：对于新数据点 $\boldsymbol{x}_q$，根据式(2)计算其与数据库中各样本点之间的相似度 $\omega_i (i = 1,2,\cdots,n)$；

Step 4：根据式(14)构建半监督 LWPLS 模型，并利用 EM 算法估计半监督 LWPLS 模型参数 $\boldsymbol{\Theta} = \left\{\boldsymbol{P}, \boldsymbol{C}, \sigma_x^2, \sigma_y^2\right\}$，即本文公式(15)~(26)部分；

Step 5：通过公式(27)得到新建立的半监督模型在线预测输出 $\hat{\boldsymbol{y}}_q$；

Step 6：将均方根误差(RMSE)作为性能指标，评价本文建模方法的预测性能，其计算公式如式(28)所示：

$$\mathrm{RMSE} = \sqrt{\frac{1}{n-1}\sum_{q=1}^{n}\left|\boldsymbol{y}_q - \hat{\boldsymbol{y}}_q\right|^2} \tag{28}$$

式中：$\boldsymbol{y}_q$ 为真值；$\hat{\boldsymbol{y}}_q$ 为模型预测值；n 为测得的数据点的个数。

4 仿真实例

脱丁烷塔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[20]，其中脱丁烷塔塔底的丁烷浓度对石油炼制影响较大且难以直接测得，需要对其进行严格的监测与控制。为了提高脱丁烷塔的控制质量，实时监测塔底丁烷浓度，对丁烷浓度的实时估计显示十分重要。该工艺过程如图 2 所示。实际工业过程中选取 7 个测量变量作为辅助变量，主导变量为脱丁烷塔塔底丁烷浓度，选取的辅助变量如表 1 所示。

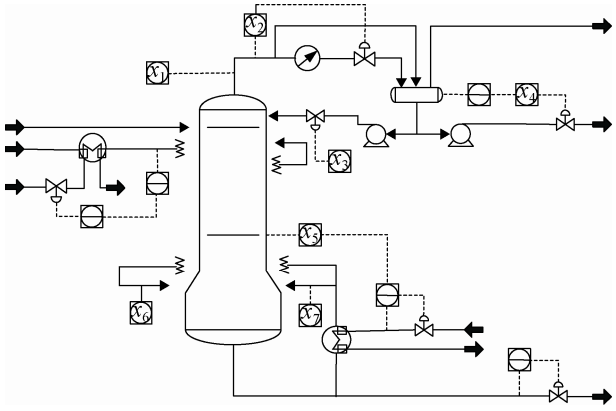

图 2 脱丁烷塔工艺流程图
Fig.2 Diagram of Debutanizer column process

https://dc-china-simulation.researchcommons.org/journal/vol30/iss1/2
DOI: 10.16182/j.issn1004731x.joss.201801002

第 30 卷第 1 期 系统仿真学报 Vol. 30 No. 1
2018 年 1 月 Journal of System Simulation Jan., 2018

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

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

为了预测该过程丁烷浓度，通过脱丁烷塔过程平台，共收集 2 000 组数据，其中 1 000 组作为训练样本集，1 000 组作为测试样本集；同时，训练样本集又被分为有标签样本集及无标签样本集。为了验证本文所提基于半监督学习建模方法的有效性，利用本文方法对有标签训练样本集占总训练样本集的 10%进行建模研究，在 1 000 个测试样本集中每隔 10 个样本选一个作为有标签样本，构成 100 组有标签样本集，即有输入又有输出的样本数据，其余 900 组作为无标签样本集，即只有输入的样本数据，则有标签样本集可表示为：$\{(\boldsymbol{x}_j,\boldsymbol{y}_j),\boldsymbol{x}_j\in R^7,\boldsymbol{y}_j\in R,j=1,2,...,100\}$，无标签样本集可表示为：$\{\boldsymbol{x}_{j'},\boldsymbol{x}_{j'}\in R^7,j'=1,2,...,900\}$。为显示有标签训练样本集占总训练样本集的比例，图 3 给出了总训练样本集中输出变量与样本个数之间的效果图。其中，绿点为有标签样本，红点为无标签样本，这里为突出显示有标签样本与无标签样本的效果，将无标签样本的输出暂时补上，在建模时，红点部分只有相对应的输入变量。

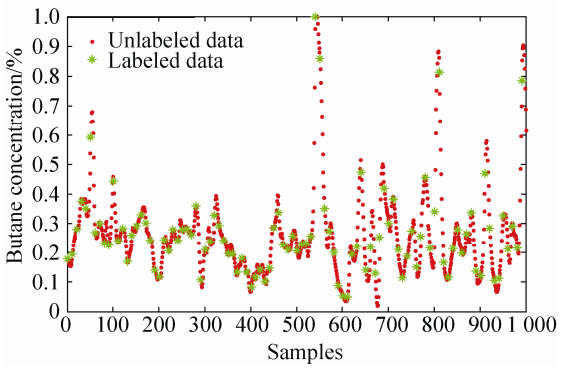

图 3 有标签训练样本与无标签训练样本
Fig. 3 Labeled and unlabeled training samples

利用 PCA 确定需要选取的半监督 LWPLS 模型主成分向量的个数 K，通过累计方差贡献率准则确定主成分个数以保证这些提取的主成分能够解释过程数据 90%以上的信息，根据该准则提取出 5 个主成分，即 K=5，如图 4 所示。

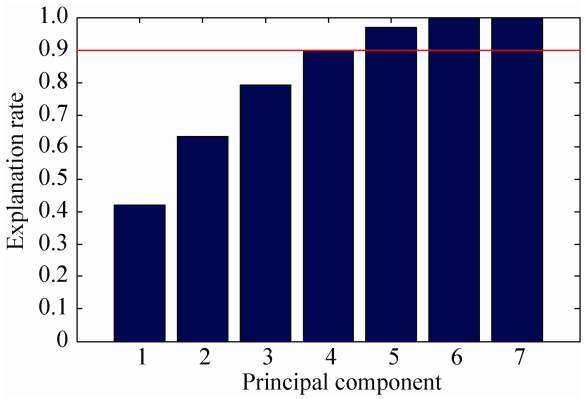

图 4 PCA 特征提取
Fig. 4 Feature extraction of PCA

基于 EM 算法的半监督 LWPLS 建模方法是一种在线建模方法，本仿真共有 1 000 组测试样本，对于其中一组新来数据点 $\boldsymbol{x}_q$，首先根据式(2)计算出其与数据库中有标签训练样本集及无标签训练样本集各样本点之间的相似度 $\omega_j(j=1,2,...,100)$ 及 $\omega_{j'}(j'=1,2,...,900)$；然后，将其带入公式(21)~(24)中不断迭代，公式中迭代算法参数的初值 $\boldsymbol{\Theta}=\{\boldsymbol{P},\boldsymbol{C},\sigma_x^2,\sigma_y^2\}$ 可以任意设定，当 $\left\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^S\right\|$ 收敛时可求出半监督 LWPLS 模型参数 $\boldsymbol{\Theta}=\{\boldsymbol{P},\boldsymbol{C},\sigma_x^2,\sigma_y^2\}$；根据推导出的半监督 LWPLS 模型的参数 $\boldsymbol{\Theta}$，代入公式(27)，即可得到模型的输出估计 $\hat{\boldsymbol{y}}_q$。其中，半监督 LWPLS 模型中的位置参数 φ 通过试验法来确定，φ 一般取 0.1~1.5 之间[12]，通过不断调整位置参数，当 φ 调整为 0.2 时建立的在线模型预测效果较佳。

为验证本文基于半监督学习建模方法的有效性，将本文建模方法与 LWPLS 建模方法进行比较，如图 5 所示。传统的 LWPLS 建模方法只使用有标签样本，将无标签样本丢弃，过程数据的信息得不到充分利用。由图 5 可见，基于 EM 算法的半监督 LWPLS 在线建模方法的预测结果与真值的吻合程度更高，预测效果更好。

Published by Journal of System Simulation, 2018

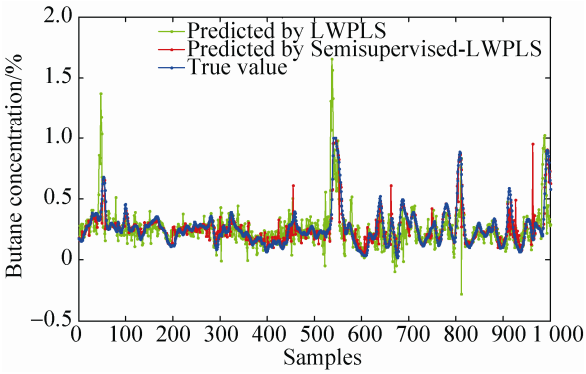

图 5 LWPLS 与半监督 LWPLS 模型对丁烷浓度的预测结果

Fig. 5 Prediction of butane concentration between LWPLS and semi-supervised LWPLS Model

同时，将本文建模方法与 PLS 建模方法(文献[2])及 SBPCR 建模方法(文献[10])进行比较来验证本文建模方法的预测精度。在有标签训练样本集占总训练样本集 10%情况下，3 种不同软测量建模方法的预测结果如图 6 所示。

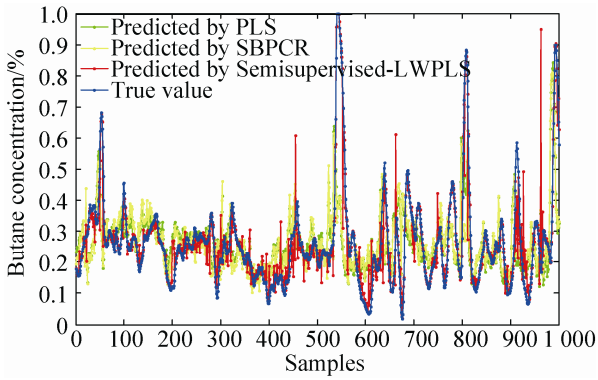

图 6 3 种建模方法对丁烷浓度的预测结果

Fig. 6 Prediction results of Butane concentration by three modeling methods

由图 6 可以看出，基于 EM 算法的半监督局部加权偏最小二乘在线建模方法的预测结果与丁烷浓度的真值基本一致，而采用 PLS 模型以及 SBPCR 模型的预测结果则相对较差。为了进一步说明本文方法的有效性，图 7 给出了有标签训练样本集占总训练样本集 10%情况下丁烷浓度真值与预测值的散点图，图中散点越接近黑色对角线说明预测值更接近真值，模型精度越高。由图 7 可以看出，本文建模方法得到的散点分布更靠近黑色对角线，更加紧凑地分布在对角线两侧，而其它两种建模方法的散点分布则相对分散，说明本文建模方法得到的丁烷浓度预测值更接近真值，模型预测精度更高。

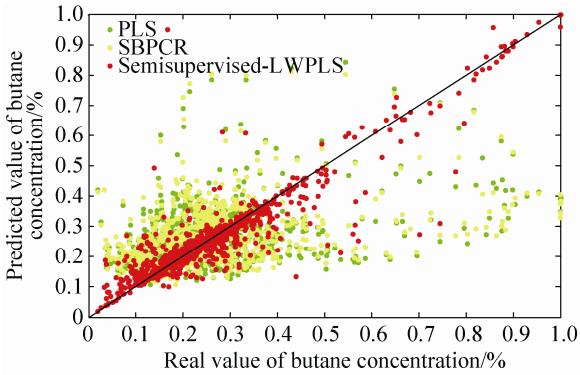

图 7 丁烷浓度真值与预测值散点分布

Fig. 7 Scatters of true and predicted values of Butane concentration

图 8 给出了 3 种建模方法的预测误差曲线。通过比较不难发现，基于 EM 算法的半监督 LWPLS 建模方法所得到的误差曲线在 0 附近波动幅度更小，这表明本文采用的建模方法预测误差较小。同时，仿真过程中还计算了有标签训练样本集占总训练样本集的 10%情况下模型对 1 000 组样本进行预测的均方根误差，如表 2 所示。可以看出，采用基于 EM 算法的半监督局部加权偏最小二乘在线建模方法得到的均方根误差低于其他两种建模方法所得。所以，本文提出的建模方法具有更高的精度和泛化能力。

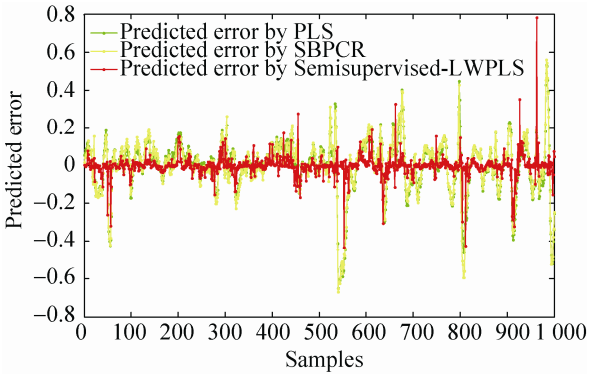

图 8 对丁烷浓度三种建模方法的预测误差

Fig. 8 Prediction errors of three modeling methods for Butane concentration

https://dc-china-simulation.researchcommons.org/journal/vol30/iss1/2
DOI: 10.16182/j.issn1004731x.joss.201801002

表 2 对丁烷浓度预测均方根误差
Tab. 2 Prediction root mean square error for Butane concentration

预测模型	PLS	SBPCR[10]	半监督 LWPLS
RMSE	0.1948	0.1810	0.0635

结合基于半监督学习的建模方法能够利用无标签样本中所携带信息的特性，利用本文方法、PLS 建模方法及 SBPCR 建模方法分别对有标签训练样本集占总训练样本集的 10%~50%进行建模结果分析，即有标签样本为 100 组至 500 组。对于 5 种不同比例的有标签训练样本集，3 种不同软测量建模方法的预测结果如表 3 所示，进一步验证了本文所提建模方法的预测精度。

表 3 不同有标签训练样本集比例下 3 种建模方法预测结果
Tab. 3 The prediction results of three modeling methods with different label ratios

模型\比例/%	10	20	30	40	50
PLS	0.194 8	0.186 2	0.176 5	0.157 9	0.156 8
SBPCR[10]	0.181 0	0.178 5	0.168 3	0.153 0	0.150 5
半监督 LWPLS	0.063 5	0.061 8	0.060 6	0.059 7	0.059 3

5 结论

对于实际化工过程中通过采样分析获得的有标签的样本数量较少，利用有标签样本建立的软测量模型不能充分反应过程特性的问题，本文提出一种基于半监督学习思想的在线建模方法，在利用有标签样本的同时融合无标签样本中所携带的信息。由于半监督 LWPLS 建模方法是对在线测得的数据不断的构建局部模型，用局部线性模型来逼近非线性过程，所以能够较好地描述非线性过程的动态特征。通过对脱丁烷塔过程的仿真研究，验证了本文所提建模方法的有效性，对于输出变量难以测得的非线性化工过程的动态建模研究具有一定的参考价值。

参考文献：

[1] 王惠文, 孟洁. 多元线性回归的预测建模方法[J]. 北京航空航天大学学报, 2007, 33(4): 500-504.
Wang Huiwen, Meng Jie. Predictive modeling on multivariate linear regression[J]. Journal of Beijing University of Aeronautics and Astronautics, 2007, 33(4): 500-504.

[2] 郑小霞, 钱锋. 基于 PCA 和最小二乘支持向量机的软测量建模[J]. 系统仿真学报, 2006, 18(3): 739-741.
ZHENG Xiaoxia, QIAN Feng. Soft Sensor Modeling Based on PCA and Support Vector Machines[J]. Journal of System Simulation, 2006, 18(3): 739-741.

[3] 王巍, 柴天佑, 赵立杰. 带有稳定学习的递归神经网络动态偏最小二乘建模[J]. 控制理论与应用, 2012, 29(3): 337-341.
WANG Wei, CHAI Tian-you, ZHAO Li-jie. Dynamic partial least squares modeling with recurrent neural networks of stable learning[J]. Control Theory & Applications, 2012, 29(3): 337-341.

[4] Shao Weiming, Tian Xueming. 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(3): 113-132.

[5] 王宏伟, 韩云涛, 彭继慎. 基于 TSPSO 支持向量机红外甲烷传感器动态补偿[J]. 传感技术学报, 2013, 26(9): 1193-1197.
WANG Hongwei, HAN Yuntao, PENG Jishen. Dynamic Compensation of Infrared Methane Sensor Based on TSPSO-ε-SVM[J]. CHINESE JOURNAL OF SENSORS AND ACTUATORS, 2013, 26(9): 1193-1197.

[6] 李春富, 郑松, 葛铭. 基于递推非线性部分最小二乘模型的间歇过程批到批优化[J]. 计算机与应用化学, 2011, 28(7): 939-942.
Li Chunfu, Zheng Song, Ge Ming. Batch-to-batch optimization of batch processes based on recursive nonlinear partial least squares model[J]. Computers and Applied Chemistry, 2011, 28(7): 939-942.

[7] Cleveland W S. Robust Locally Weighted Regression and Smoothing Scatterplots[J]. Journal of American Statistical Association (S0162-1459), 1979, 74(368): 829-836.

[8] Kim S, Kano M, Nakagawa H, et al. Estimation of Active Pharmaceutical Ingredients Content Using Locally Weighted Partial Least Squares and Statistical Wavelength Selection[J]. International Journal of Pharmaceutics (S0378-5173), 2011, 421(2): 269-274.

[9] 章军, 杨慧中. 基于 SVM 的苯酚浓度半监督软测量方法[J]. 计算机与应用化学, 2013, 30(20): 1453-1456.
Zhang Jun, Yang Huizhong. Semi-supervised soft sensor

of Phenol based on SVM[J]. Computers and Applied Chemistry, 2013, 30(20): 1453-1456.

[10] Ge Zhiqiang, Song Zhihuan. Semisupervised Bayesian Method for Soft Sensor Modeling with Unlabeled Data Samples[J]. AIChE Journal (S1547-5905), 2011, 57(8): 2109-2119.

[11] 吴继明. 基于 Boosting 思想的半监督学习算法研究[D]. 吉林: 吉林大学, 2014.
Wu Jiming. The research of semi-supervised learning based on Boosting[D]. Jilin: Jilin University, 2014.

[12] Kim S, Kano M, Hasebe S, et al. Long-term Industrial Applications of Inferential Control Based on Just-in-time Soft Sensors: Economical Impact and Challenges[J]. Industrial and Engineering Chemistry Research (S0888-5885), 2013, 52(35): 12346-12356.

[13] 唐晓亮, 韩敏. 一种基于极端学习机的半监督学习方法[J]. 大连理工大学学报, 2010, 50(5): 771-776.
Tang Xiaoliang, Han Min. A semi-supervised learning method based on extreme learning machine[J]. Journal of Dalian University of Technology, 2010, 50(5): 771-776.

[14] Ge Zhiqiang, Huang Biao, Song Zhihuan. Mixture Semisupervised Principal Component Regression Model and Soft Sensor Application[J]. AIChE Journal (S1547-5905), 2014, 60(2): 533-545.

[15] Ma Ming, Shima K, Huang Biao. A Bayesian Framework for Real-Time Identification of Locally Weighted Partial Least Squares[J]. AIChE Journal (S1547-5905), 2015, 61(2): 518-529.

[16] 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): 4935-4961.
XIONG Weili, YAO Le, XU Baoguo. Multi-stage fusion modeling for penicillin fermentation process based on EM algorithm[J]. CIESC Journal, 2014, 65(12): 4935-4961.

[17] Jin Xing, Huang Biao, Shook D S. Multiple Model LPV Approach to Nonlinear Process Identification with EM Algorithm[J]. Journal of Process Control (S0959-1524), 2011, 21(1): 182-193.

[18] Jin Xing, Huang Biao. Robust Identification of Piecewise/Switching Autoregressive Exogenous Process [J]. AIChE Journal (S1547-5905), 2010, 56(7): 1829-1844.

[19] Yu S P, Yu K, Tresp V, et al. Supervised probabilistic principal component analysis[C]// Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA:ACM, 2006:464–473.

[20] 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.

(上接第 7 页)

[12] Barabasi A L, Albert R, Jeong H. Mean-field theory for scale-free random networks [J]. Physica A (S0378-4371), 1999, 272: 173-187.

[13] 田立新, 贺莹环, 黄益. 一种新型二分网络类局域世界演化模型[J]. 物理学报, 2012, 61(22): 228903.
Tian Lixin, He Yinghuan, Huang Yi. A novel local-world-like evolving bipartite network model[J]. Acta Physica Sinica, 2012, 61(22): 228903.

https://dc-china-simulation.researchcommons.org/journal/vol30/iss1/2
DOI: 10.16182/j.issn1004731x.joss.201801002