DOI: 10.19650/j.cnki.cjsi.J2209166

基于时空相似性的即时学习在线建模*

施锦涛[1,2],陈 磊[1,2],秦 凯[1,2],李振兴[1,2],郝矿荣[1,2]
(1.东华大学数字化纺织服装技术教育部工程研究中心 上海 201620;2.东华大学信息科学与技术学院 上海 201620)

摘 要:流程工业数据具有较大的时变性以及非线性,传统的离线模型难以应对实际生产过程中的工况变化,而即时学习是在线建模的有效方法。已有研究对即时学习的相似度度量方法大多只侧重于样本的空间距离,忽略了工业数据时序性的特点。为此,提出基于时空相似性的即时学习建模方法。首先,将样本点延拓成样本序列,结合动态时间规整计算样本间的时序距离。其次,提出时空相似性度量准则,通过对时序距离和空间距离进行非线性加权,构建时空相似性度量指标。最后,提出基于时空相似性的即时学习在线建模方法。将所提算法应用于公共数据集及聚酯纤维聚合过程,拟合优度分别达到了 91.6% 和 98.6%,实验结果验证了算法的有效性和优越性。

关键词:时空相似性;即时学习;在线建模;流程工业;数据驱动
中图分类号:TH86 TP274 **文献标识码**:A **国家标准学科分类代码**:510.80

Online modeling of just-in-time learning based on spatial-temporal similarity

Shi Jintao[1,2], Chen Lei[1,2], Qin Kai[1,2], Li Zhenxing[1,2], Hao Kuangrong[1,2]
(*1. Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, Donghua University, Shanghai 201620, China; 2. College of Information Science and Technology, Donghua University, Shanghai 201620, China*)

Abstract: Data in the process industry are highly time-varying and nonlinear. Traditional offline models can hardly cope with the changing working conditions in the actual production process, while the just-in-time learning (JITL) is an effective online modeling method. Most of the studied similarity measurements of JITL only focus on samples' spatial distance, which ignore the time-series characteristics of industrial data. To address this issue, a JITL method based on spatial-temporal similarity is proposed. First, the sample point is extended into a sample sequence, and the temporal-sequence distance among samples is calculated by combining dynamic time warping. Then, the spatial-temporal similarity metric (SSM) is proposed, and the SSM is constructed by nonlinearly weighting the temporal and spatial distances. Finally, the online modeling method for just-in-time learning based on spatial-temporal similarity (SS-JITL) is proposed. The algorithm is applied to a public dataset and an actual polyester fiber polymerization process. Experiment results show that the goodness of fit reaches 91.6% and 98.6%, which demonstrates the effectiveness and superiority of the proposed algorithm.
Keywords: spatial-temporal similarity; just-in-time learning; online modeling; process industry; data-driven

0 引 言

随着计算机、传感器、数据存储等技术的发展,流程工业行业产生了大量的工业数据,这些数据对于流程工业的建模具有较高价值[1]。流程工业建模方法分为机理建模和数据驱动建模,机理建模方法需要全面准确的理论支撑,且基于较多简化和假设之上,因而机理模型在工业过程中的应用受到很大阻力。而数据驱动建模方法只涉及过程的输入输出变量,不必详细分析其内部机理,因此更适用于具有非线性和不确定性的复杂工业过程建模[2-3]。工业过程本身具有多变量、强干扰、大滞后和强耦合等特点,使得数据检测和采样存在着不可避免的误差。传统的离线模型一旦在线应用,通常很难保持预期

收稿日期:2022-01-10 Received Date: 2022-01-10
*基金项目:上海市自然科学基金面上项目(19ZR1402300)、中央高校基本科研业务费专项资助

的性能,因此需要实时更新模型参数以保持模型良好的性能[4]。即时学习(just-in-time learning, JITL)作为一种局部建模技术,在离线训练结束之后不会丢弃所有历史数据,而是会反复利用数据库中的历史数据,局部模型根据生产环境的变化做出相应的调整,从而克服了全局建模方法参数一成不变的缺点,被广泛应用于流程工业在线建模中。

针对 JITL 的研究主要是局部模型和相似性度量方法的选取。局部模型选取的关键在于要在满足过程动态性能的同时兼顾算法运行效率[5],相似性度量方法则需全面地描述查询样本与历史样本的相似性,不同的相似性度量考虑了历史样本与查询样本不同方面的相似性,从而会导致不同的预测性能。常用的相似性度量方法大多基于空间距离,比如欧氏距离[6]、马氏距离[7-8]等,这些度量准则在 JITL 中取得了较好的应用效果,然而由于没有考虑特征间的角度信息,在空间上距离较近的样本会具有较大的角度差异。Xia 等[9]将余弦相似度用于样本的相似性度量中,充分利用了样本间的角度信息。杨潇谊等[10]将样本的空间距离和角度距离相结合,采用余弦欧氏距离衡量样本间的相似性。Song 等[11]和 Zhao 等[12]考虑了输入与输出间的联系,将两者间的互信息作为权重因子对输入变量加权,有效利用了输出变量的信息。

当今的流程工业具有较大的时滞性,不同输入的作用时间有所差别,对输出产生的影响也有一定延迟,而传统 JITL 的相似性度量方法忽略了工业数据的时序性。动态时间规整(dynamic time warping, DTW)算法是一种处理时间序列数据的常用算法,对时间序列进行局部调整,通过拉伸、弯曲等操作将目标时间序列之间的相似度调整至最优匹配,并将其作为相似度距离,近年来被广泛应用于工业过程的时间序列数据处理。Shen 等[13]使用 DTW 同步不均匀的批次样本,处理了运行轨迹预测时的批处理长度不均匀问题。Li 等[14]基于 DTW 的距离度量方法进行工业时间序列数据的聚类分析。于蕾等[15]利用 DTW 对各个变量组分别进行同步化,提出一种基于变量分组 DTW-MCVA 的不等长间歇过程故障检测方法。

数据的时序特性对 JITL 建模结果具有重要影响,如何结合时间序列数据特性进行 JITL 建模优化是一个重要问题,然而,目前没有明确的理论指导时间和空间距离结合的有效方法,Yuan 等[16]把时间间隔作为相似性的一种衡量依据,却没有进一步分析不同的时间间隔对相似性产生的影响的变化情况,而对于时序数据来说,相邻样本间也会相互产生影响。针对以上问题,本文提出时空相似性度量准则,并发展了结合时空相似性度量准则的 JITL 建模方法,本文主要贡献如下:1)将样本在时间轴上延拓成样本序列,结合 DTW 序列距离度量方法,引入时间间隔系数,整合了样本间的时序距离计算方法。2)提出了时空相似性度量准则(spatial-temporal similarity metric, SSM),通过对样本间的时序距离和空间距离进行非线性加权,构建时空相似性度量指标。3)建立了基于时空相似性的即时学习建模方法(spatial-temporal similarity just-in-time learning, SS-JITL),根据延拓后样本间的时空相似性,建立在线模型。所提方法的有效性和优越性在公共数据集及聚酯纤维聚合过程中得到了验证。

1 动态时间规整

在时间序列处理中,通常需要比较两个序列的相似性,当序列长度不等时就无法使用欧氏距离等方法来计算。DTW 作为一个典型的动态规划问题,它通过把时间序列在时间轴上扭曲对齐计算两个序列间的相似性,用满足一定条件的时间规整函数表示两个序列在时间上的对应关系,进而求解累加距离最小时所对应的规整函数[17]。

假设两个时间序列分别为 $\boldsymbol{S}=\{s_1,s_2,\cdots,s_N\}$ 和 $\boldsymbol{R}=\{r_1,r_2,\cdots,r_M\}$,长度分别为 N 和 M。当 N 和 M 相等时,可以通过将这两个序列元素一一对应,直接计算出它们的距离;当 N 和 M 不相等时,则需要通过将这两个序列对齐,采用 DTW 进行求解。为此,需要构造一个 $N\times M$ 的距离矩阵 $\boldsymbol{D}$,如图 1 所示,矩阵元素 $D(i,j)$ 表示 s_i 与 r_j 两点间的距离,定义如下:

$$D(i,j)=\|s_i-r_j\|_l \tag{1}$$

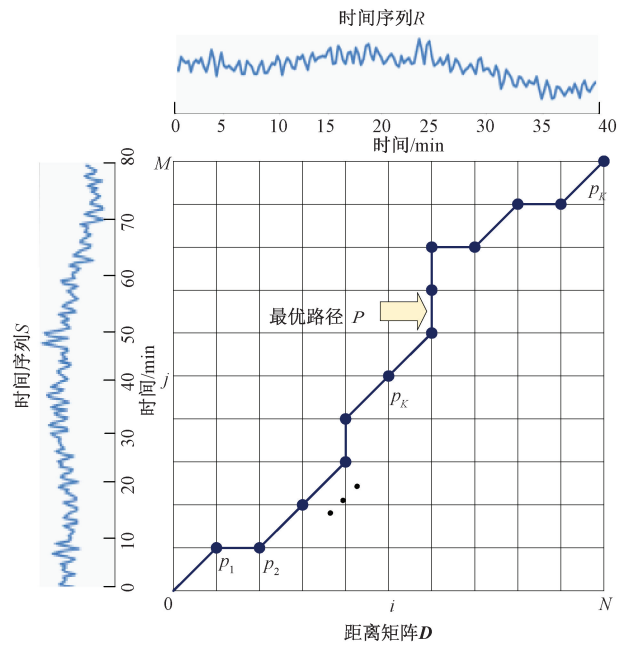


图 1 DTW 动态规划路径

Fig. 1 Dynamic planning path for DTW

当 $l=2$ 时,$D(i,j)$ 便表示两点间的欧氏距离。DTW 的目标是寻找一条通过该距离矩阵的最优路径 $\boldsymbol{P}=\{p_1,p_2\cdots p_K\}$,$p_k$ 表示该路径元素在距离矩阵 $\boldsymbol{D}$ 中的位置,K 表示路径长度,满足如下约束:

$$\max(N,M)\leqslant K\leqslant N+M-1 \tag{2}$$

有效路径可以有多个,但必须满足两个条件:

1) 边界条件:$p_1=(1,1)$ 以及 $p_K=(N,M)$,即路径起止点是确定的;

2) 单调性和连续性:给定 $p_k=(i,j)$ 和该点的下一点 $p_{k+1}=(i',j')$,需满足 $i\leqslant i'\leqslant i+1,j\leqslant j'\leqslant j+1$,即在对齐过程中,只能按照时间顺序逐点进行匹配,以保证时间序列中的每个元素都能得到正确的匹配。

最优路径 $\boldsymbol{P}$ 满足如下代价函数,$C(\boldsymbol{S},\boldsymbol{R})$ 表示序列 $\boldsymbol{S}$ 和 $\boldsymbol{R}$ 的最优路径所对应的代价:

$$C(\boldsymbol{S},\boldsymbol{R})=\min_{\boldsymbol{P}}\Big\{\sum_{k=1}^{K}D(p_k)/K\Big\} \tag{3}$$

其中,P_k 表示路径 $\boldsymbol{P}$ 中的第 k 个元素,K 表示路径 $\boldsymbol{P}$ 的长度,用于消除路径长度对相似性的影响。$D(p_k)$ 表示 P_k 所对应元素间的距离,计算方式如式(1) 所示。累加距离 $\gamma(i,j)$ 定义为当前匹配的距离加上一步累加距离的最小值,表示如下:

$$\gamma(i,j)=D(i,j)+\min\{\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1)\} \tag{4}$$

从起点开始匹配并把距离逐次累加,匹配到终点时的累加距离就是序列 $\boldsymbol{S}$ 和 $\boldsymbol{R}$ 之间的最小距离。

2　时空相似性度量准则

时空相似性度量准则 SSM 是协同样本间的时序相似性和空间相似性的指标。本节结合 DTW 序列相似性度量方法衡量查询样本与历史样本的相似性,并引入时间间隔系数,以区分不同时间间隔对相似性影响的变化情况,进而计算查询样本与历史样本间的时序距离;其次,将样本之间的时序距离与空间距离进行非线性加权,权衡样本间的时空组合相似性,构建时空相似性度量指标。

2.1　时序距离

首先,对于建模对象的查询样本和历史样本,需要对其在时间轴上进行延拓。对于 t 时刻的查询样本 $\boldsymbol{x}_q$ 以及历史样本 $\boldsymbol{x}_h$,延拓后得到的查询序列 $\boldsymbol{S}_q$ 及历史序列 $\boldsymbol{S}_h$ 满足如下公式:

$$\boldsymbol{S}_q=\begin{cases}\{\boldsymbol{X}_Q^i\}_{i=0}^{q}, & q<N\\ \{\boldsymbol{X}_Q^i\}_{i=q-N}^{q}, & q\geqslant N\end{cases} \tag{5}$$

$$\boldsymbol{S}_h=\begin{cases}\{\boldsymbol{X}_H^j\}_{j=0}^{h}, & h<M\\ \{\boldsymbol{X}_H^j\}_{j=h-M}^{h}, & h\geqslant M\end{cases} \tag{6}$$

其中,q 和 h 分别为 $\boldsymbol{x}_q$ 和 $\boldsymbol{x}_h$ 的索引;N 和 M 分别为 $\boldsymbol{S}_q$ 和 $\boldsymbol{S}_h$ 的长度;当 $q<N$ 时 $\boldsymbol{S}_q$ 的长度是自适应的,$\boldsymbol{S}_h$ 亦是如此;$\boldsymbol{X}_Q$ 表示测试集的输入且样本数量为 Q;$\boldsymbol{X}_H$ 表示历史数据集的输入且样本数量为 H;$\{\boldsymbol{X}_Q^i\}_{i=0}^{q}$ 表示测试集前 q 条数据所组成的序列。考虑到时间间隔对相似性的影响,在式(3) 的基础引入关于时间间隔 Δt 的控制系数 $f(\Delta t)$,表示如下:

$$f(\Delta t)=1/(1+\exp(-\alpha\Delta t)),\ (\alpha>0,\Delta t\geqslant 0) \tag{7}$$

其中,Δt 表示 $\boldsymbol{x}_q$ 和 $\boldsymbol{x}_h$ 的时间间隔,α 为可调参数,用于调整控制系数随着 Δt 的变化率。如果 $\boldsymbol{x}_q$ 和 $\boldsymbol{x}_h$ 在时间上相距越远,那么 $\boldsymbol{x}_h$ 对 $\boldsymbol{x}_q$ 产生的影响就越小,反之亦然。同时,$f(\Delta t)$ 应该对较小的时间间隔更加敏感,当 Δt 大到一定程度时,$\boldsymbol{x}_h$ 对 $\boldsymbol{x}_q$ 产生的影响可以忽略不计。新的代价函数可以表示为:

$$C_{\text{new}}(\boldsymbol{S}_q,\boldsymbol{S}_h)=f(\Delta t)\times\min_{\boldsymbol{P}}\Big\{\sum_{k=1}^{K}D(p_k)/K\Big\} \tag{8}$$

为了降低计算的复杂度,根据定义的代价函数并结合式(4),$\boldsymbol{S}_q$ 和 $\boldsymbol{S}_h$ 的最小累加距离 $\gamma(N,M)$ 表示如下:

$$\gamma(N,M)=D(N,M)+\min\{\gamma(N-1,M-1),\gamma(N-1,M),\gamma(N,M-1)\} \tag{9}$$

其中,N 为查询序列 $\boldsymbol{S}_q$ 的长度,M 为历史序列 $\boldsymbol{S}_h$ 的长度。实际计算时从 $\gamma(1,1)$ 开始,最优路径 $\boldsymbol{P}$ 则可以通过回溯的方法确定。最终 $\boldsymbol{x}_q$ 和 $\boldsymbol{x}_h$ 的时序距离可以表示为:

$$D_{\text{T}}(\boldsymbol{x}_q,\boldsymbol{x}_h)=\gamma(N,M)/(1+\exp(-\alpha\Delta t)) \tag{10}$$

这里的 N 与 M 以及参数 α 的取值对于相似性度量的效果有较大影响,本文采用贝叶斯方法优化超参数[18]。

2.2　构建时空相似性度量指标

将上一节提出的时序距离与传统的空间距离相结合,对两种距离进行非线性加权,综合考虑样本间的时空相似性来提高相似样本集的质量。其中空间相似性度量方法采用马氏距离,马氏距离是建立在总体样本的基础上计算的,不受量纲的影响,表示如下:

$$D_{\text{M}}(\boldsymbol{x}_q,\boldsymbol{x}_h)=\sqrt{(\boldsymbol{x}_q-\boldsymbol{x}_h)^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_q-\boldsymbol{x}_h)} \tag{11}$$

其中,$\boldsymbol{x}_q$ 和 $\boldsymbol{x}_h$ 分别表示查询样本以及历史样本,$\boldsymbol{\Sigma}$ 表示数据的协方差矩阵。

加权方式通过交叉验证集确定,对于每种相似性度量方法选出来的相似样本集,通过同一个局部模型在交叉验证集上得到相应的输出。计算出对应的均方误差(mean squared error, MSE):

$$MSE=\frac{1}{V}\sum_{i=1}^{V}(y_i^v-\hat{y}_i^v)^2 \tag{12}$$

其中,y_i^v 表示验证集上查询样本对应的真实输出,$\hat{y}_i^v$ 表示局部模型的输出,V 表示验证集样本数。

在非线性加权步骤中,关键在于如何确定加权函数从而为时序距离 D_{T} 和空间距离 D_{M} 分配合适的权重 ω_{T} 及 ω_{M}:

$$\omega_{\mathrm{T}}=\frac{1}{MSE_{\mathrm{T}}}\Big/\left(\frac{1}{MSE_{\mathrm{T}}}+\frac{1}{MSE_{\mathrm{M}}}\right) \tag{13}$$

$$\omega_{\mathrm{M}}=\frac{1}{MSE_{\mathrm{M}}}\Big/\left(\frac{1}{MSE_{\mathrm{T}}}+\frac{1}{MSE_{\mathrm{M}}}\right) \tag{14}$$

其中,ω_{T} 和 ω_{M} 分别表示时序距离和空间距离对应的权重,MSE_{T} 和 MSE_{M} 则分别表示相应的均方误差。

文献[19]列出了几种常用的加权函数,由于 MSE 一般都接近于 0,所以要求在原点附近的斜率变化较大,从而放大不同相似性度量方法间的差异。MSE 越小,说明对应的相似性度量方法精度越高,那么该方法占据着主导地位,应分配较大的权重。将 ω_{T} 和 ω_{M} 分别对式(10)中的时序距离 D_{T} 以及式(11)中的空间距离 D_{M} 进行非线性加权,最终查询样本 $\boldsymbol{x}_q$ 与历史样本 $\boldsymbol{x}_h$ 的时空距离可以表示为:

$$D_{\mathrm{TM}}=\omega_{\mathrm{T}}D_{\mathrm{T}}+\omega_{\mathrm{M}}D_{\mathrm{M}}=$$
$$(MSE_{\mathrm{T}}\times D_{\mathrm{M}}+MSE_{\mathrm{M}}\times D_{\mathrm{T}})/(MSE_{\mathrm{T}}+MSE_{\mathrm{M}}) \tag{15}$$

对于不同的数据集,样本间的相似性关系会有所差异。空间距离可能足以衡量部分样本的相似性,时间距离又可能是其它样本的真实相似性。而一般很难获得流程工业中数据的真实非线性关系,因此单一的相似性度量方法存在一定的局限性。式(15)通过交叉验证集确定权重,将空间距离和时序距离进行非线性加权,更加全面地衡量样本间多方面的相似性。此外,不同的相似性度量方法可以获得不同的预测精度,通过为 MSE 较小的方法分配更大的权重,以此来提高其对输出的影响。ω_{T} 和 ω_{M} 经过归一化处理,如果 $\omega_{\mathrm{M}}=0$,那么时空相似性就退化成了单一的时序相似性;反之,如果 $\omega_{\mathrm{T}}=0$,该方法就成为了空间相似性度量。实际上不同相似性度量方法的 MSE 值并不会相差过大,因此该方法的意义在于在时间维度和空间维度找到平衡点,权衡这两方面的相似性,使得相似性度量能够更加贴近于样本间真实的非线性关系。

进一步将时空距离转化为相似度 Sim 如下:

$$Sim=\exp(-D_{\mathrm{TM}}/\varphi\sigma_{\mathrm{D}}) \tag{16}$$

其中,σ_{D} 表示时空距离 D_{TM} 的标准差,φ 表示相似度随着距离增加的衰减率。相似度越高,意味着该历史样本将会被优先分配给相似样本集[20]。将得到的相似度按降序排列构建相似样本集并进行局部建模即可得到查询样本 $\boldsymbol{x}_q$ 的预测输出 $\hat{y}_q$。

3 基于时空相似性的在线建模

JITL 是非线性过程中常用的软测量建模策略[21],将 SSM 应用到基于 JITL 的在线建模中,构建基于时空相似性的即时学习 SS-JITL 建模方法,局部模型选择高斯贝叶斯网络(Gaussian bayesian network, GBN)。GBN 将高斯分布与贝叶斯网络相结合,假设所有变量服从高斯分布,采用图形化的语言来表示变量间的依赖关系,简洁明了且通俗易懂。此外,GBN 采用贝叶斯方法处理不确定性事件,能够较好地处理工业过程中的非线性以及时变性。

假设历史数据集表示为 $\{(\boldsymbol{X}_H,\boldsymbol{Y}_H)\}$,交叉验证集表示为 $\{(\boldsymbol{X}_V,\boldsymbol{Y}_V)\}$,测试集表示为 $\{(\boldsymbol{X}_Q,\boldsymbol{Y}_Q)\}$,$\boldsymbol{X},\boldsymbol{Y}\in\boldsymbol{R}$ 表示输入输出变量,H 表示历史样本的数量,V 表示交叉验证样本数量,Q 表示测试样本数量。整个 JITL 的建模流程如图 2 所示,伪代码如算法 1 所示。

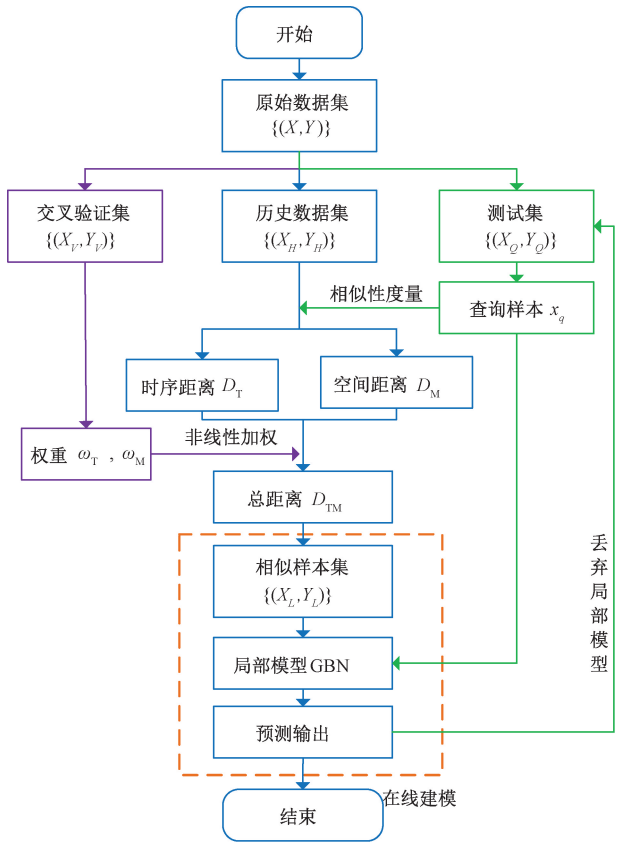


图 2 即时学习建模流程

Fig. 2 Modeling process of JITL

算法主要步骤如下:

1)将数据集分为历史数据集、交叉验证集以及测试集,当查询样本 $\boldsymbol{x}_q$ 到来时,利用式(10)以及式(11)计算出 $\boldsymbol{x}_q$ 与每条历史样本 $\boldsymbol{x}_h$ 的时序距离 D_{T} 和空间距离 D_{M}。

2)在交叉验证集上计算出两种相似性度量方法的权重 ω_{T} 和 ω_{M},根据式(15)对 $D_{\mathrm{T}}(\boldsymbol{x}_q,\boldsymbol{x}_h)$ 和 $D_{\mathrm{M}}(\boldsymbol{x}_q,\boldsymbol{x}_h)$ 进行非线性加权得到 $\boldsymbol{x}_q$ 与 $\boldsymbol{x}_h$ 的时空距离 D_{TM}。

3)利用式(16)将距离转化为相似度 Sim,并将所有相似度降序排列,挑选出前 L 个相似度所对应的历史样本组成相似样本集 $\{(\boldsymbol{X}_L,\boldsymbol{Y}_L)\}$。

4)构建局部模型并训练,利用贝叶斯优化算法确定超参数 N、M 以及 α 的取值,得到 $\boldsymbol{x}_q$ 的预测输出 $\hat{y}_q$。

5)丢弃该局部模型,当下一时刻的查询样本到来时,返回步骤 1)。

算法 1 SS-JITL 建模流程

输入:$\{(\boldsymbol{X}_H,\boldsymbol{Y}_H)\}$:历史数据集;$\{(\boldsymbol{X}_V,\boldsymbol{Y}_V)\}$:交叉验证集;
$\{(\boldsymbol{X}_Q,\boldsymbol{Y}_Q)\}$:测试集;$G$:高斯贝叶斯网络结构;
N:查询序列长度;M:历史序列长度;
$size$:相似样本集容量;α:控制系数的变化率;

输出:$\{\hat{\boldsymbol{Y}}_q\}_{q=1}^{Q}$:特性粘度的估计值;

设置 $N,M,size,\alpha$;
计算两种相似性度量方法的权重:ω_T,ω_M;
for $q = 0;q < Q;q++$ do
　if $q < N$ then
　　$S_q \leftarrow \{\boldsymbol{X}_Q^i\}_{i=0}^{q}$;
　else
　　$S_q \leftarrow \{\boldsymbol{X}_Q^i\}_{i=q-N}^{q}$;
　end if
　for $h = 0;h < H;h++$ do
　　if $h < M$ then
　　　$S_h \leftarrow \{\boldsymbol{X}_H^i\}_{i=0}^{h}$;
　　else
　　　$S_h \leftarrow \{\boldsymbol{X}_H^i\}_{i=h-M}^{h}$;
　　end if
　　$\Delta t \leftarrow H - h + q$;
　　$D_M \leftarrow \sqrt{(\boldsymbol{x}_q - \boldsymbol{x}_h)^T \Sigma^{-1} (\boldsymbol{x}_q - \boldsymbol{x}_h)}$;
　　$D_T \leftarrow \gamma(N,M)/(1 + \exp(-\alpha\Delta t))$;
　　$D_{TM} \leftarrow \omega_T D_T + \omega_M D_M$;
　　$\{Sim\} \leftarrow \exp(-D_{TM}/\varphi\sigma_D)$;
　end if
　将 $\{Sim\}$ 按降序排列得到 $\boldsymbol{x}_q$ 的相似样本集;
　构建模型 GBN,得到 $\boldsymbol{x}_q$ 的预测输出 $\hat{y}_q$;
　$\{\hat{\boldsymbol{Y}}_q\} \leftarrow \hat{y}_q$
end for

4　实验结果及讨论

4.1　评价标准

采用均方误差(mean squared error, MSE)、平均绝对误差(mean absolute error, MAE)以及决定系数 R^2 作为评价标准衡量实验结果。三种评价标准的定义如下所示:

$$MSE = \frac{1}{Q}\sum_{i=1}^{Q}(y_i - \hat{y}_i)^2 \tag{17}$$

$$MAE = \frac{1}{Q}\sum_{i=1}^{Q}|y_i - \hat{y}_i| \tag{18}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{Q}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{Q}(y_i - \bar{y})^2} \tag{19}$$

其中,y_i 表示查询样本对应的真实输出,$\hat{y}_i$ 表示局部模型的估计值,$\bar{y}$ 表示真实输出的平均值,Q 表示测试集的样本数。这里的 MSE 和 MAE 越小,表明模型的精度越高,R^2 越接近于 1,表明模型的拟合效果越好。

4.2　仿真数据建模

为了验证所提方法的有效性,将 SS-JITL 建模方法在公共数据集上进行实验。公共数据集来自于 UCI 机器学习库的空气质量数据集[22],该时序数据集包含了位于意大利城市内一个严重污染地区的多个金属氧化物传感器响应值及非金属化合物浓度值,数据记录时间为 2004 年 3 月—2005 年 2 月。选取一氧化碳浓度、氧化锡传感器响应值、苯浓度、二氧化钛传感器响应值、氮氧化合物浓度、氧化钨传感器响应值、氧化铟传感器响应值共 7 个特征作为输入变量,每小时平均二氧化氮浓度作为输出变量。经缺失值处理后选取前 1 700 条实例作为实验数据集,其中 1 600 条用作训练集,其余 100 条用作测试集。

相似样本集的大小设置为固定值 150,式(8)中的参数 α 以及式(10)中的查询序列长度 N 与历史序列长度 M 通过贝叶斯优化获得,由于贝叶斯优化只能对连续值进行优化,对 N 和 M 进行了取整操作,实际最优参数组合为:$\alpha = 0.809$,$N = 50$,$M = 49$。离线模型为 GBN 模型,GBN 的网络结构采用多输入单输出的方式,即每个输入变量都是输出变量的直接父节点,而这些输入变量之间遵循条件独立的假设,相互之间不存在依赖关系。对比方法为全局建模方式,即使用 GBN 模型进行全局建模并输出。基于 SS-JITL 的局部建模及全局建模的预测结果如图 3 所示,从图中可以看出,相比于全局建模方式,局部建模在工况发生变化的情况下通过构建相似样本集,能够更好地拟合真实曲线。两种方法的预测误差如表 1 所示,局部建模方法的 3 个指标均优于全局建模方法,验证了 SS-JITL 建模方法应用于时序数据在线建模的有效性。

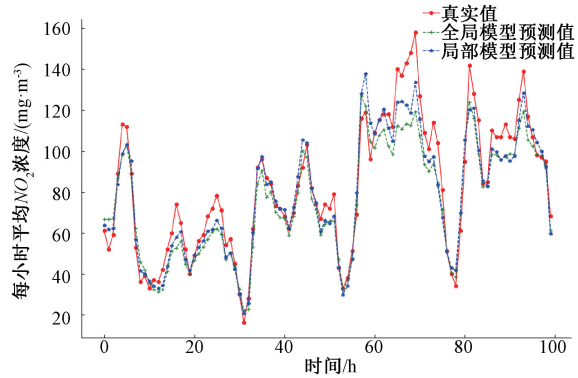


图 3　不同建模方式在空气质量数据集上的预测结果

Fig. 3　Prediction results of different modeling methods on the air quality data set

表 1 不同建模方法的预测误差

Table 1 Comparison of prediction errors between two modeling methods

编号	建模方式	*MAE*	*MSE*	R^2
1	全局建模	9.02	134.5	0.870 2
2	局部建模	7.16	86.8	0.916 3

4.3 聚酯纤维聚合过程建模

聚合过程是聚酯纤维生产的第一步,该过程中杜邦公司的三釜工艺流程如图 4 所示,包括酯化单元、预缩聚单元和终缩聚单元[23]。具体包括 5 个部分:1)对苯二甲酸(terephthalic acid, TPA)进料及浆料配比:精对苯二甲酸(pure terephthalic acid, PTA)氧化装置生产的粗对苯二甲酸通过相关输送装置进入浆料罐,注水溶解后通过增压泵运送到预热器中。2)酯化:对苯二甲酸和乙二醇的混合物进入换热器与管内的低聚物混合,然后发生酯化反应。3)齐聚物输送和添加剂注入:将酯化反应生成的齐聚物通过管道输送到预缩聚反应釜,并在管道中加入添加剂。4)预缩聚:在真空条件下,酯化部分的产物在预缩聚釜中完成酯化反应并进行首次缩聚。5)终缩聚:将输入的预缩聚物再一次缩聚,得到高质量的成品聚合物。

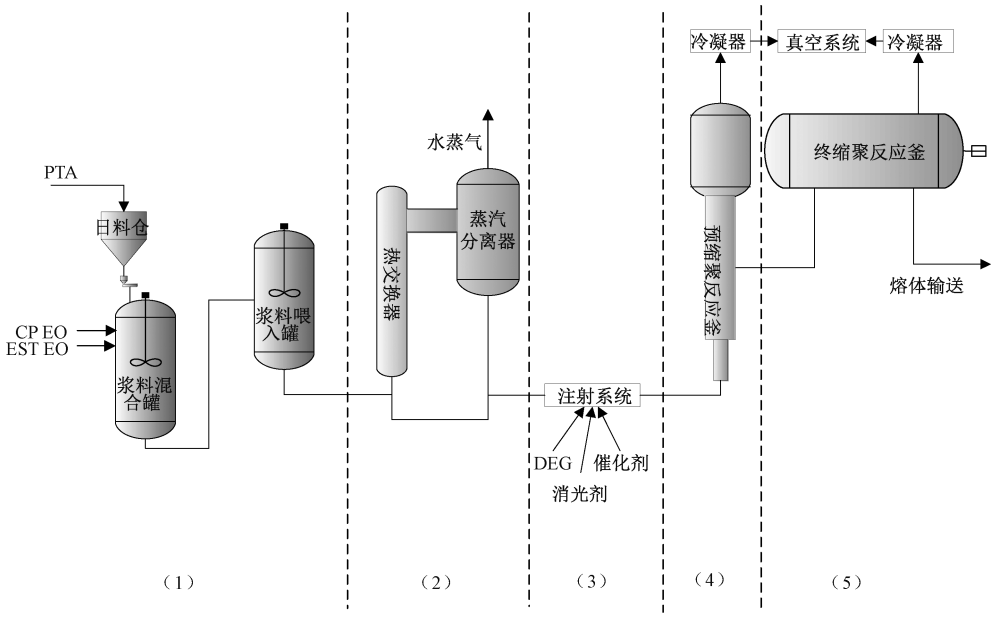


图 4 杜邦三釜工艺流程图

Fig. 4 Flowchart of DuPont triple-kettle

使用的数据采集于聚酯纤维聚合过程,采样间隔为 1 min,数据集总共包含 294 个特征,包括温度、压力、流量、液位等特征。根据皮尔逊相关系数,选择与特性粘度具有较高相关系数的特征作为输入变量,最终选择 TPA 旋转阀转速、冷却器压差、聚合物搅拌器力矩、乙二醇(ethylene glycol, EG)喷淋温度、终缩聚釜搅拌器力矩、终缩聚釜压力共 6 个特征作为输入变量,聚合物的特性粘度作为输出变量。选取 1 000 条样本,其中前 800 条样本用作历史数据集,后 200 条样本用作测试数据集,交叉验证集和测试数据集相同,输入变量的变化趋势如图 5 所示。

聚酯纤维生产过程具有较大的不确定性,而高斯贝叶斯网络能够较好地处理这种不确定性,因此将 SS-JITL 与 GBN 结合建立在线软测量模型。GBN 的网络结构采用多输入单输出的方式,即每个输入变量都是输出变量的直接父节点,而这些输入变量之间遵循条件独立的假设,相互之间不存在依赖关系。相似样本集的大小设置为固定值 150,式(8)中的参数 α 以及式(10)中的查询序列长度 N 与历史序列长度 M 通过贝叶斯优化获得,实际最优参数组合为:$\alpha = 0.656\ 8$,$N = 33$,$M = 49$,通过交叉验证集将式(15)中的权重设置为:$\omega_T = 0.635$, $\omega_M = 0.365$。

根据第 3 节提到的 SS-JITL 建模流程可以得到特性粘度的估计值,与真实输出进行比较,对比结果如图 6 所示。从图 6 中可以明显看出,整体的拟合效果较好,说明所提出的 SS-JITL 方法在聚酯纤维聚合过程的应用是有效的。第 60~70 组数据的拟合效果略差,具体表现为部分数据点的预测值与真实值的偏差增大,从图 5 中可以看到,由于工况出现异常,作为输入变量之一的终缩聚釜压力在该数据段出现了大幅度的突变,说明 SS-JITL 通过模型重构的方式能够对生产环境的变化做出相应的调整,但是面对大幅度突变的处理能力有待进一步优化。

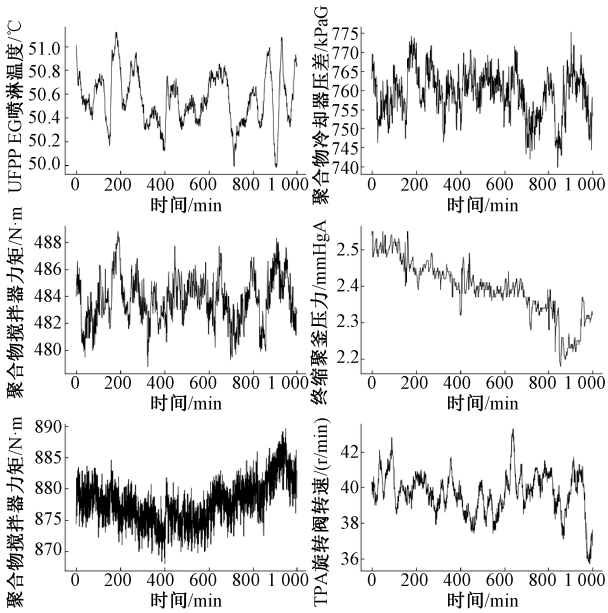


图 5　输入变量的变化趋势

Fig. 5　Trends of input variables

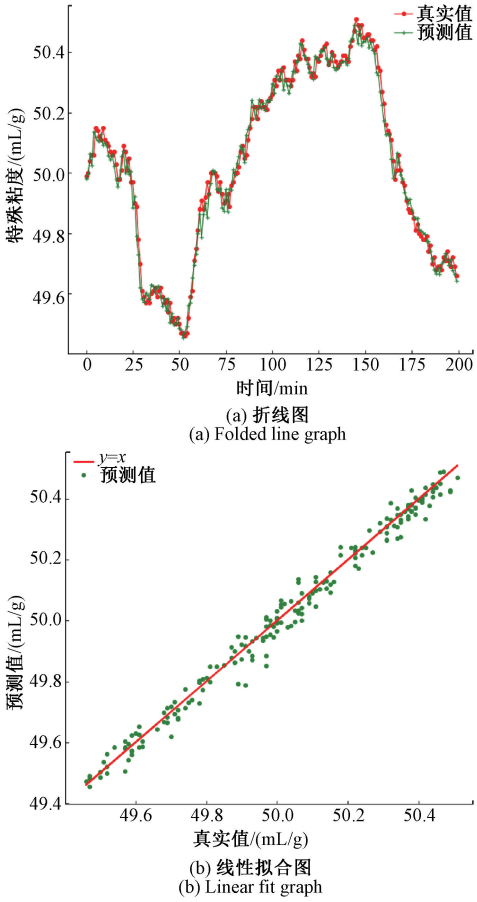


图 6　基于时空 JITL 方法的预测结果对比

Fig. 6　Comparison of prediction results based on the spatial-temporal JITL method

为了进一步评估本文所提方法,将 SSM 与在 JITL 中较为常见的相似性度量方法以及近几年新兴的方法进行对比,分别为欧氏距离、马氏距离、余弦距离以及文献[12]中的基于加权互信息的方法、文献[24]中的基于高斯混合模型及马氏距离的方法、文献[16]中的基于时间距离的方法以及文献[8]中的基于标度马氏距离的方法。对比方法的数据集、相似样本容量、局部模型等实验条件均与 SS-JITL 的实验条件相同。比较结果的误差箱线图如图 7 所示。

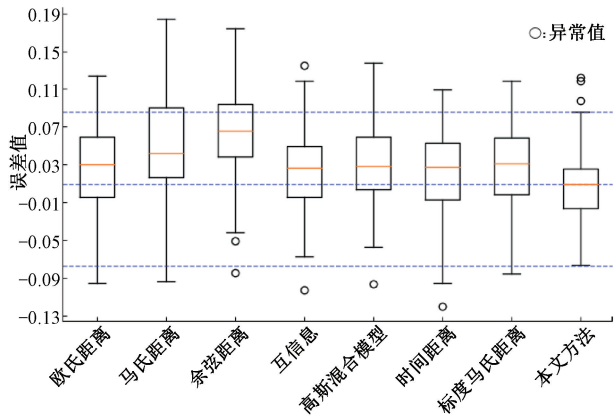


图 7　误差对比箱线图

Fig. 7　Box line diagram of the error comparison

在图 7 中,浅色实线代表每组数据的中位数,箱体的上下底分别表示第三四分位数以及第一四分位数,箱体两侧的短划线代表每组数据的上下限,空心小圆点代表异常值,3 条虚线为本文方法的误差基准线。本文方法的误差中位数约为 0.01,误差上限约为 0.85,均优于对比方法。图 7 中有 5 种方法的误差存在异常值,这也印证了图 6 中部分数据段因生产工况异常导致的拟合效果较差的情况。另外,本文方法的误差分布集中且均匀,而其它方法的整体误差分布偏向于纵坐标的正半部分,说明非线性加权能够权衡样本间不同方面相似性。由此可见,相较于 7 种对比方法,本文方法的预测效果更优越。

上述方法的预测误差指标如表 2 所示,其中基于余弦距离的方法预测效果较差,本文所提方法均比其它方法有着较大的提升,MAE 和 MSE 均为最低,相较于对比方法中效果最好的互信息方法分别降低了约 27% 以及 42%,R^2 最高为 0.986 2。基于余弦距离的方法只考虑了样本间的角度信息,单独把它作为相似性度量方法会丢失样本间的部分信息;欧氏距离和马氏距离考虑了样本间的空间距离,效果虽然比基于余弦距离的方法有所提升,但总体预测效果不好;基于时间距离的方法考虑到了时间间隔对样本间相似性的影响,却没有考虑不同的时间间隔对相似性产生的影响的变化情况;基于互信息的

方法、基于高斯混合模型的方法以及基于标度马氏距离的方法分别考虑到了输入与输出间的互信息、工业数据中的非高斯成分以及输入与输出间的相关系数,相较于常规方法效果有所提升,但是仍然存在较大优化空间。本文方法不仅考虑了样本间的时间距离,还考虑到了时间因素对相似性影响的变化趋势,更加符合时序数据的特征,在各项评价指标中均为最优,可见,本文提出的 SS-JITL 建模方法在进行流程工业建模时的有效性与优越性,相较于对比方法更适应复杂的工业环境。

表 2 对比方法的预测误差

Table 2 Prediction errors of comparison methods

编号	相似性度量方法	$MAE(\times10^{-2})$	$MSE(\times10^{-3})$	R^2
1	欧氏距离	4.22	2.72	0.968 7
2	马氏距离	5.89	5.26	0.939 4
3	余弦距离	6.77	6.04	0.930 3
4	互信息[12]	3.68	2.07	0.976 1
5	高斯混合模型[24]	3.95	2.38	0.972 5
6	时间距离[16]	4.07	2.37	0.972 7
7	标度马氏距离[8]	4.03	2.36	0.972 7
8	本文方法	2.70	1.20	0.986 2

5 结 论

本文提出一种时空相似性度量准则 SSM,并结合 SSM 提出基于时空相似性的即时学习在线建模方法 SS-JITL。通过将样本点延拓为样本序列,结合 DTW 序列距离度量方法,引入时间间隔系数,计算样本间的时序距离,从而更好的描述样本数据间的非线性关系;进一步地,对时序距离和空间距离进行非线性加权,计算时空相似性,以权衡样本间不同方面的相似性;最后,将时空相似性作为相似性度量方法引入 JITL 在线建模。所提方法的有效性和优越性在公开数据集以及聚酯纤维的特性粘度实际工业数据集中得到了验证,为流程工业在线建模提供了有效的参考。以 DTW 为代表的动态规划算法在求解时较为耗时,未来的工作可以进一步考虑结合此类算法进行建模时的效率优化问题。

参考文献

[1] 苗强, 蒋京, 张恒, 等. 工业大数据背景下的航空智能发动机:机遇与挑战[J]. 仪器仪表学报, 2019, 40(7): 1-12.
MIAO Q, JIANG J, ZHANG H, et al. Development of aviation intelligent engine under industrial big data: Chances and challenges[J]. Chinese Journal of Scientific Instrument, 2019, 40(7): 1-12.

[2] YUAN X F, GE Z Q, HUANG B, et al. A probabilistic just-in-time learning framework for soft sensor development with missing data[J]. IEEE Transactions on Control Systems Technology, 2016, 25(3): 1124-1132.

[3] SHAO W M, GE Z Q, SONG Z H. Bayesian just-in-time learning and its application to industrial soft sensing[J]. IEEE Transactions on Industrial Informatics, 2020, 16(4): 2787-2798.

[4] 车笑卿, 熊伟丽. 带时延估计的加权高斯模型软测量建模[J]. 仪器仪表学报, 2018, 39(9): 195-202.
CHE X Q, XIONG W L. Weighted gaussian model soft sensor modeling and its application with time delay estimation[J]. Chinese Journal of Scientific Instrument, 2018, 39(9): 195-202.

[5] CHEN Z W, LIU C, DING S X, et al. A just-in-time-learning-aided canonical correlation analysis method for multimode process monitoring and fault detection[J]. IEEE Transactions on Industrial Electronics, 2021, 68(6): 5259-5270.

[6] XIONG W L, ZHANG W, XU B G, et al. JITL based MWGPR soft sensor for multi-mode process with dual-updating strategy[J]. Computers & Chemical Engineering, 2016, 90: 260-267.

[7] KE T, LYU H, SUN M J, et al. A biased least squares support vector machine based on Mahalanobis distance for PU learning[J]. Physica A: Statistical Mechanics and its Applications, 2018, 509: 422-438.

[8] YUAN X F, GE Z Q, HUANG B, et al. Semisupervised JITL framework for nonlinear industrial soft sensing based on locally semisupervised weighted PCR[J]. IEEE Transactions on Industrial Informatics, 2016, 13(2): 532-541.

[9] XIA P P, ZHANG L, LI F Z. Learning similarity with cosine similarity ensemble[J]. Information Sciences, 2015, 307: 39-52.

[10] 杨潇谊, 吴建德, 马军. 基于散布熵和余弦欧氏距离的滚动轴承性能退化评估方法[J]. 电子测量与仪器学报, 2020, 34(7): 15-24.
YANG X Y, WU J D, MA J. Rolling bearing performance degradation assessment method based on dispersion entropy and cosine Euclidean distance[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(7): 15-24.

[11] SONG Y L, REN M L. A novel just-in-time learning strategy for soft sensing with improved similarity measure based on mutual information and pls[J]. Sensors, 2020, 20(13): 3804.

[12] ZHAO D, PAN T H, SHENG B Q. Just-in-time learning algorithm using the improved similarity index[C]. 2016 35th Chinese Control Conference, IEEE, 2016: 9065-9068.

[13] SHEN F F, YE L J, FAN S T, et al. Run-to-run trajectory prediction of uneven-length batch processes using DTW-LSTM[C]. 2019 IEEE 8th Data Driven Control and Learning Systems Conference, IEEE, 2019: 1183-1188.

[14] LI G, QIN S J, YUAN T. Data-driven root cause diagnosis of faults in process industries[J]. Chemometrics and Intelligent Laboratory Systems, 2016, 159: 1-11.

[15] 于蕾, 邓晓刚, 曹玉苹, 等. 基于变量分组 DTW-MCVA 的不等长间歇过程故障检测方法[J]. 化工学报, 2019, 70(9): 3441-3448.
YU L, DENG X G, CAO Y P, et al. Fault detection method of unequal-length batch process based on VGDTW-MCVA[J]. CIESC Journal, 2019, 70(9): 3441-3448.

[16] YUAN X F, ZHOU J, WANG Y L. A spatial-temporal LWPLS for adaptive soft sensor modeling and its application for an industrial hydrocracking process[J]. Chemometrics and Intelligent Laboratory Systems, 2020, 197: 103921.

[17] 李海林, 梁叶, 王少春. 时间序列数据挖掘中的动态时间弯曲研究综述[J]. 控制与决策, 2018, 33(8): 1345-1353.
LI H L, LIANG Y, WANG SH CH. Review on dynamic time warping in time series data mining[J]. Control and Decision, 2018, 33(8): 1345-1353.

[18] LINDAUER M, EGGENSPERGER K, FEURER M, et al. SMAC3: A versatile Bayesian optimization package for hyperparameter optimization[J]. Journal of Machine Learning Research, 2022, 23(54): 1-9.

[19] ATKESON C G, MOORE A W, SCHAAL S. Locally weighted learning[J]. Artificial Intelligence Review, 1997, 11(1/5): 11-73.

[20] ZHENG J Q, SHEN F F, YE L J. Improved mahalanobis distance based JITL-LSTM soft sensor for multiphase batch processes[J]. IEEE Access, 2021, 9. DOI: 10.1109/ACCESS.2021.3079184.

[21] YANG Z Y, GE Z Q. Rethinking the value of just-in-time learning in the era of industrial big data[J]. IEEE Transactions on Industrial Informatics, 2022, 18(2): 976-985.

[22] DE VITO S, MASSERA E, PIGA M, et al. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario[J]. Sensors and Actuators B: Chemical, 2008, 129(2): 750-757.

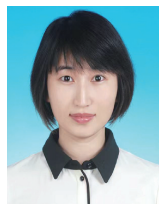
[23] 王建华, 徐松, 王云华, 等. 杜邦三釜预聚至终聚流程优化[J]. 聚酯工业, 2012, 25(2): 30-31.
WANG J H, XU S, WANG Y H, et al. Prepolymerization to final polycondensation process optimization of Du Pont three reactors plant[J]. Polyester Industry, 2012, 25(2): 30-31.

[24] 祁成, 熊伟丽. 基于 BGMM 的即时学习软测量建模方法[J]. 系统仿真学报, 2019, 31(8): 1555-1561.
QI CH, XIONG W L. A just-in-time learning soft sensing modeling method based on bayesian gaussian mixture model[J]. Journal of System Simulation, 2019, 31(8): 1555-1561.

作者简介

施锦涛,2020 年于东华大学获得学士学位,现为东华大学硕士研究生,主要研究方向为工业过程建模和过程控制等。
E-mail: shijintao1997@foxmail.com

Shi Jintao received his B. Sc. degree from Donghua University in 2020. He is currently a master student at Donghua University. His main research interests include industrial process modeling and process control.



陈磊(通信作者),分别在 2006 年,2014 年于江南大学获得学士学位和博士学位,现为东华大学副教授,主要研究方向为工业过程建模、工业大数据分析、过程控制和系统辨识等。
E-mail: leichen@dhu.edu.cn

Chen Lei (Corresponding author) received her B. Sc. degree and Ph. D. degree both from Jiangnan University in 2006 and 2014, respectively. She is currently an associate professor at Donghua University. Her main research interests include industrial process modeling, industrial big data analytics, process control and system identification.