

引用格式:林琳,王建林,随恩光,等. 基于 DCC-RVM 的多模态间歇过程质量变量在线预测方法[J]. 北京化工大学学报(自然科学版),2026,53(4):116-124.

LIN Lin, WANG JianLin, SUI EnGuang, et al. An online prediction of quality variables based on DCC-RVM for multi-mode batch processes[J]. Journal of Beijing University of Chemical Technology (Natural Science),2026,53(4):116-124.

# 基于 DCC-RVM 的多模态间歇过程质量变量在线预测方法

林琳 王建林\* 随恩光 辛维 张泽斌

(北京化工大学 信息科学与技术学院, 北京 100029)

**摘要:** 间歇过程具有多模态特性,现有的间歇过程模态划分方法忽略了模态中心与过程特性的因果关系,制约了模态划分结果准确性和合理性的提升,直接影响了质量变量预测模型的精度和泛化性。提出一种基于深度因果聚类相关向量机(deep causal clustering-relevance vector machine, DCC-RVM)的多模态间歇过程质量变量在线预测方法。首先结合密度峰值聚类,考虑间歇过程数据的强非线性特征,利用深度自编码器提取过程数据的深层特征,构建基于长短时记忆网络(long short-term memory, LSTM)间歇过程的过程模型,以模态中心数据样本与 LSTM 过程模型输出的因果关系为约束,构建间歇过程模态中心选取策略,获得合理的模态中心;然后考虑过程数据的时序特征,依据样本间的相对距离将非模态中心样本分配至对应模态,获得各个模态的切换点及数据集;最后,通过各模态 RVM 预测模型在线预测间歇过程质量变量。青霉素发酵过程的实验结果表明, DCC 模态划分结果较合理,所建立的多模态 RVM 模型预测均方根误差  $RMSE$  为 0.010 4,  $R^2$  为 0.999 5,有效提高了质量变量的预测精度。

**关键词:** 间歇过程; 模态划分; 质量变量预测; 密度峰值聚类; 因果关系

**中图分类号:** TP274; TP181 **DOI:** 10.13543/j.bhxbzr.2026.04.013

## 引言

间歇过程具有灵活性高、适应性强的优势,已成为现代制造业重要的生产方式之一<sup>[1-2]</sup>。在线测量间歇生产过程质量变量是实现质量控制的重要前提<sup>[3]</sup>。软测量技术利用过程数据建立易测辅助变量和不易测主导变量的数学模型,从而在线测量主导变量<sup>[4]</sup>,已被用于间歇过程质量变量在线预测。现有的间歇过程质量变量在线预测方法大多是建立全局预测模型,忽略了间歇过程的多模态特性,限制了质量变量预测模型精度的提升<sup>[5]</sup>。因此,准

确获得间歇过程模态切换点及各模态的数据集,并建立多模态质量变量预测模型,有助于提高间歇过程质量变量的预测精度,对实现间歇过程质量控制具有重要意义。

现有的基于时序特征分析的模态划分方法通过提取及分析过程变量的时序特征,将变量特征发生显著变化的点作为模态切换点。但该方法对异常值敏感,易产生不合理的模态划分结果。基于模型识别的模态划分方法通过建立统计模型捕捉过程变量的特征信息,从而确定模态切换点<sup>[6-7]</sup>。然而,该方法需要原始数据满足特定分布,难以直接用于具有非高斯数据特征的间歇过程。

基于聚类分析的模态划分方法通过无监督学习将具有相似特征的数据样本划分为同一模态<sup>[8-9]</sup>,无需原始数据满足特定分布。Lu 等<sup>[10]</sup>将 k 均值聚类(k-means)算法应用于模态划分,通过分析聚类负

收稿日期:2025-03-12

基金项目:国家自然科学基金(62373036)

第一作者:女,1999年生,硕士生

\*通信联系人

E-mail:wangjl@mail.buct.edu.cn

载矩阵并添加阈值实现了模态划分;Ng等<sup>[11]</sup>利用模糊c均值(fuzzy c-means, FCM)聚类通过隶属度函数量化数据点对各模态的隶属程度,通过最小化目标函数获得了合理的模态划分结果。但以上方法均忽略了过程数据的时序特征,使得模态划分结果存在时序错乱。Luo等<sup>[12]</sup>提出一种基于时序约束模糊C均值的模态划分方法,然而该方法易受初始模态中心及输入参数的影响,模态划分结果的一致性较差。

密度峰值聚类(density peaks clustering, DPC)算法通过非迭代的方式获得聚类结果<sup>[13-14]</sup>。刘聪等<sup>[15]</sup>将DPC引入间歇过程模态划分,通过样本点间的K近邻关系定义样本点分配策略,将非模态中心样本分配至所属模态,然后利用各模态的数据集建立预测模型。然而间歇过程数据的高维特征限制了质量变量预测模型的精度。针对该问题,周新杰等<sup>[16]</sup>提出一种基于改进密度峰值聚类的模态划分方法,提高了间歇过程模态划分的合理性,建立的多模态质量变量预测模型具有较高的预测精度。受操作切换的影响,间歇过程的过程特性存在阶段性变化,使得整个过程具有多个模态中心,过程特性与模态中心具有稳定的因果关系,引入因果关系约束模态中心的选择并分配非模态中心样本点,获得准确的模态划分结果,能够进一步提高质量变量预测模型的预测精度。

为提高间歇过程质量变量预测的准确性,本文提出了一种基于深度因果聚类相关向量机(DCC-RVM)的多模态间歇过程质量变量在线预测方法。在密度峰值聚类的基础上,利用深度自编码器获得原始数据的深层特征,建立长短时记忆网络(long short-term memory, LSTM)过程模型,以模态中心数据样本与过程模型输出的因果关系为约束,构建模态中心选取策略,并将非模态中心样本分配至对应模态,获得各个模态的切换点和数据集。然后,利用各模态数据集建立多模态RVM质量变量预测模型。最后通过实验验证了本文所提方法的有效性。

## 1 基于深度因果聚类的间歇过程模态划分

### 1.1 深度因果聚类

密度峰值聚类算法<sup>[17]</sup>的步骤如下。

1) 计算所有样本点的局部密度 $\rho$ , 计算式为

$$\rho_i = \sum_{j \neq i} \chi(d_{ij} - d_c), \chi(d) = \begin{cases} 1, & d < 0 \\ 0, & d \geq 0 \end{cases} \quad (1)$$

式中, $d_{ij}$ 为样本 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 间的欧氏距离; $d_c$ 表示截断距离; $d$ 表示 $d_{ij}$ 和 $d_c$ 的距离之差。

2) 计算每个样本点到更高密度样本点的相对距离 $\delta$

$$\delta_i = \begin{cases} \min_{\rho_j > \rho_i} (d_{ij}), & \rho_i \text{非最大} \\ \max (d_{ij}), & \rho_i \text{最大} \end{cases} \quad (2)$$

构建样本 $\mathbf{x}_i$ 的决策值 $\gamma_i$ 的计算式,将 $\gamma_i$ 值较大的样本作为聚类中心,并依据相对距离将非聚类中心样本分配至对应类别。决策值 $\gamma_i$ 的计算式为

$$\gamma_i = \rho_i \times \delta_i \quad (3)$$

#### 1.1.1 深度特征提取

间歇过程原始数据中存在的冗余特征会降低聚类算法的性能,因此本文利用深度自编码器(deep autoencoder, DAE)提取过程数据的深层特征,之后再进行聚类,以提高聚类结果的准确性。

DAE包含编码器和解码器两个部分,编码器 $f_\varphi(\mathbf{x})$ 将可测过程变量 $\mathbf{x}$ 映射到 $\mathbf{z}$ ,可表示为

$$\mathbf{z} = f_\varphi(\mathbf{x}) \quad (4)$$

$$f_\varphi(\mathbf{x}) = \tanh(\mathbf{W}_3 \times \text{ReLU}(\mathbf{W}_2 \times \text{ReLU}(\mathbf{W}_1 \mathbf{x} + b_1) + b_2) + b_3) \quad (5)$$

式中, $\mathbf{W}_1$ 、 $\mathbf{W}_2$ 和 $\mathbf{W}_3$ 为编码器权重; $b_1$ 、 $b_2$ 和 $b_3$ 为偏置项。

解码器 $g_\theta(\mathbf{z})$ 对 $\mathbf{z}$ 进行重构,其计算式为

$$\hat{\mathbf{x}} = g_\theta(\mathbf{z}) \quad (6)$$

$$g_\theta(\mathbf{z}) = \tanh(\mathbf{W}_6 \times \text{ReLU}(\mathbf{W}_5 \times \text{ReLU}(\mathbf{W}_4 \mathbf{z} + b_4) + b_5) + b_6) \quad (7)$$

式中, $\mathbf{W}_4$ 、 $\mathbf{W}_5$ 和 $\mathbf{W}_6$ 为解码器权重; $b_4$ 、 $b_5$ 和 $b_6$ 为偏置项。

DAE损失函数 $L$ 的计算式为

$$L = \sqrt{\frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \hat{\mathbf{x}}_i)^2} \quad (8)$$

式中, $N$ 表示样本量; $\mathbf{x}_i$ 为第 $i$ 个样本; $\hat{\mathbf{x}}_i$ 为重构后的第 $i$ 个样本。

DAE网络结构如图1所示,使用ReLU激活函数加速模型训练并避免梯度消失。在网络最后一层使用tanh激活函数以压缩特征至 $[-1, 1]$ 的范围内,增强模型的表达能力。

#### 1.1.2 因果关系约束的样本相似性度量和模态中心选择

间歇过程具有多个运行模态,不同模态具有不

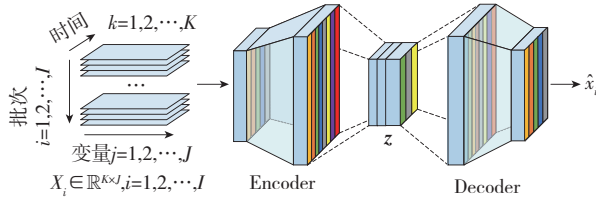


图1 深度自编码器网络结构

Fig. 1 Network structure of DAE

同的过程特性,模态中心是过程特性在数据上的集中体现,因此过程特性的变化会导致模态中心的变化,模态中心和过程特性间存在稳定的因果关系。间歇过程的过程特性可以通过过程模型进行表征。由于LSTM捕捉时间序列中的长期依赖关系能力较强<sup>[18]</sup>,因此本文利用LSTM建立时间序列模型以表征间歇过程的过程特性。给定间歇过程变量  $X_i \in \mathbb{R}^{K \times J}, i = 1, 2, \dots, I$  ( $I$ 为批次数,  $K$ 为样本数,  $J$ 为可测过程变量数),基于LSTM的过程模型可表示为

$$\hat{\mathbf{x}}(t) = f(p, \mathbf{x}(t-1), \mathbf{x}(t-2), \dots, \mathbf{x}(t-w)) \quad (9)$$

式中,  $\hat{\mathbf{x}}(t)$ 是过程模型第  $t$  个时刻的输出;  $p$  为模型参数;  $t$  为时间;  $w$  为时间窗。

建立的过程模型包含多个LSTM单元,LSTM单元的结构如图2所示。每个LSTM单元根据式(10)中的遗忘门  $f_t$ 、输入门  $i_t$ 、输出门  $o_t$  和记忆细胞  $C_t$  来处理每个时间步的数据,生成的隐藏特征向量  $\mathbf{h}_t$  将作为下一个时间步的输入。通过LSTM单元捕捉时间序列的长期依赖关系和动态变化,从而建立不同时刻样本间的联系。LSTM单元可表示为

$$\begin{aligned} f_t &= \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_f) \\ i_t &= \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_i) \\ \tilde{C}_t &= \tanh(\mathbf{W}_C \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_C) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ o_t &= \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + b_o) \\ \mathbf{h}_t &= o_t \odot \tanh(C_t) \end{aligned} \quad (10)$$

式中,  $\sigma$  表示 sigmoid 激活函数;  $\mathbf{x}_t$  为  $t$  时刻的样本;  $\mathbf{h}_{t-1}$  表示  $t-1$  时刻的隐藏特征;  $\tilde{C}_t$  表示候选记忆细胞,  $b_f, b_i, b_C, b_o$  为偏置项。

$t-1$  时刻的隐藏特征由LSTM单元传递至输出层,得到  $t$  时刻的预测结果,计算式为

$$\hat{\mathbf{x}}(t) = \mathbf{W}_L \times \mathbf{h}_{t-1} + b_L \quad (11)$$

式中,  $\mathbf{W}_L$  和  $b_L$  分别为输出层的权重和偏置项。

将  $t-w$  至  $t-1$  时刻的样本输入至LSTM过程模型,对  $t$  时刻样本进行预测,将  $K$  个时刻的模型预

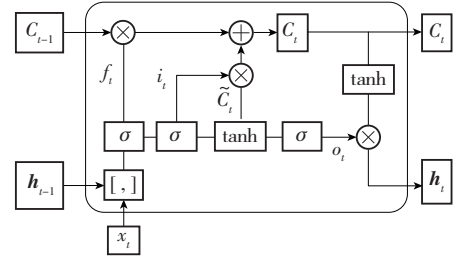


图2 LSTM单元结构

Fig. 2 Structure of the LSTM unit

测值和真实值之间的误差作为衡量不同样本对应时刻模型预测行为差异程度的依据。 $t$ 时刻过程模型的预测误差  $m_t$  计算式为

$$m_t = \sqrt{\sum_{i=1}^w (\hat{\mathbf{x}}(t) - \mathbf{x}(t))^2} \quad (12)$$

由于过程模型的输出从第  $w$  时刻开始,因此前  $w$  时刻过程模型的输出用真值代替,修正后的预测误差  $m_t$  的计算式为

$$m_t = \begin{cases} 0, & t \in [1, w] \\ \sqrt{\sum_{i=1}^w (\hat{\mathbf{x}}(t) - \mathbf{x}(t))^2}, & t \in [w+1, K] \end{cases} \quad (13)$$

最后将过程预测误差矩阵  $\mathbf{m} = [m_1, m_2, \dots, m_K]$  进行 z-score 标准化得到  $\hat{\mathbf{m}}$ 。

DPC 算法利用欧式距离度量样本间相似性,然而随着样本维数的增加,基于欧氏距离的相似性度量易出现“距离集中”的现象,导致距离度量失去区分性。针对该问题,本文引入 Close 函数<sup>[19]</sup>度量高维空间中两样本的相似度,Close 函数的计算式为

$$\text{Close}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum e^{-|\mathbf{x}_i - \mathbf{x}_j|}}{d} \quad (14)$$

式中,  $\mathbf{x}_i$  和  $\mathbf{x}_j$  分别表示第  $i$  和  $j$  个样本,  $d$  为样本维度,  $\text{Close}(\mathbf{x}_i, \mathbf{x}_j)$  越大,则  $\mathbf{x}_i$  和  $\mathbf{x}_j$  越相似。

相似样本的变化趋势应保持一致,其对应时刻的过程模型应具有相近的预测行为,即过程模型在两相似样本对应时刻的预测误差应接近。因此,本文改进的高维样本距离计算式为

$$d_{ij} = \frac{1}{\text{Close}(\mathbf{x}_i, \mathbf{x}_j)} + \left| \frac{1}{\hat{m}_i} - \frac{1}{\hat{m}_j} \right| \quad (15)$$

考虑间歇过程数据的时序特征,定义过程数据样本间的相对距离  $\delta'_i$  为

$$\delta'_i = \begin{cases} \min(d_{ij} + |i - j|), & \rho_i \text{ 非最大} \\ \max(d_{ij}), & \rho_i \text{ 最大} \end{cases} \quad (16)$$

以样本采样时间的差值作为时序约束,确保聚

类结果在时间上的连续性,从而有效避免了时间上不连续但相关性较高的非相似样本被分在同一个类别中的问题,使得聚类结果更加准确。

DPC直接将 $\gamma_i$ 较大的样本作为模态中心,当高密度区域非模态中心样本的决策值高于低密度区域模态中心样本的决策值时,难以准确选择模态中心。而模态中心数据样本与LSTM过程模型输出间存在因果约束,与过程模型输出偏离较大的数据样本不应被选为模态中心数据样本。因此,本文构建了一种因果关系约束的模态中心选择策略,将每个样本的局部密度、相对距离和预测误差组成新的样本 $p_i = (\rho_i, \delta_i, \hat{m}_i^{-1})$ ,改进后的决策值 $\theta_i$ 的计算式为

$$\theta_i = (p_i - \bar{p})\Sigma^{-1}(p_i - \bar{p})^T \quad (17)$$

式中, $\Sigma^{-1}$ 为协方差矩阵的逆; $\bar{p}$ 为样本均值。

最后将非模态中心样本分配给与其相对距离最小的样本所属类别,完成所有样本的聚类。

## 1.2 基于DCC的间歇过程模态划分方法

基于DCC的间歇过程模态划分步骤如下:

(1)对历史数据集 $X_i \in \mathbb{R}^{K \times J \times I}$ ( $I$ 为批次个数, $K$ 为采样点个数, $J$ 为过程变量个数)进行标准化和累计批次标准化,得到 $\tilde{X}_i \in \mathbb{R}^{K \times J \times I}$ 和 $\bar{X} \in \mathbb{R}^{K \times J}$ ;

(2)将 $\tilde{X}_i$ 输入式(10)、(11)训练得到LSTM过程模型,输入式(4)~(8)训练得到DAE模型;

(3)将 $\bar{X}$ 输入LSTM过程模型,依据式(12)得到 $\hat{m}$ ,将 $\bar{X}$ 输入DAE模型得到 $\hat{X}$ ;

(4)依据式(15)得到 $d_{ij}$ ,代入式(1)和式(16)得到 $\rho$ 和 $\delta$ ,再依据式(17)得到决策值 $\theta$ ;

(5)对 $\theta$ 降序,将前 $c$ 个样本点作为模态中心,依据式(16)实现非模态中心点分配,完成模态划分。

间歇过程模态划分流程如图3所示,算法伪代码如下。

算法:基于DCC的间歇过程模态划分

输入:间歇过程的可测过程变量 $X \in \mathbb{R}^{K \times J \times I}$ ,  $epoch$ ,学习率 $lr$ ,模态中心数量 $u$ 。

输出:模态中心 $c$ ,所有数据点的模态标签 $label \in \mathbb{R}^{K \times I}$ 。

标准化 $X$ 获得 $\tilde{X} \in \mathbb{R}^{K \times J \times I}$ ;

For  $i=1$  to  $epoch$  do

$\hat{X} \leftarrow \text{LSTM}(\tilde{X}, W_1), \tilde{X} \leftarrow \text{DAE}(\tilde{X}, W_2)$ ;

$W_1 \leftarrow W_1 - lr \times \nabla \text{loss}(\hat{X} - \tilde{X}), W_2 \leftarrow W_2 - lr \times \nabla \text{loss}(\tilde{X} - \hat{X})$ ;

End for

沿批处理方向取平均并标准化获得 $\bar{X} \in \mathbb{R}^{K \times J}$ ;

$\hat{X} \leftarrow \text{LSTM}(\bar{X}, W_1), \tilde{X} \leftarrow \text{DAE}(\bar{X}, W_2)$ ;

For  $i=1$  to  $K$  do

计算各样本的 $m_i, \rho_i, \delta_i$ ;

$p_i \leftarrow [\rho_i, \delta_i, m_i^{-1}]$ ;

$\theta_i = (p_i - \bar{p})\Sigma^{-1}(p_i - \bar{p})^T$ ;

End for

$\theta$ 降序排序得到 $\tilde{\theta} \in \mathbb{R}^{1 \times K}$ ;

$c_i \leftarrow \{\tilde{\theta}\}_{i=1}^u$ ;

For  $i=1$  to  $K - u$  do

$\delta_i = \min(m_i + |i - t_c|)$ ;

$c_i \leftarrow x_i, \text{label}_i \leftarrow u$  If  $\delta_i$  is  $\min(\delta)$ ;

End for

Return  $c, \text{label}$

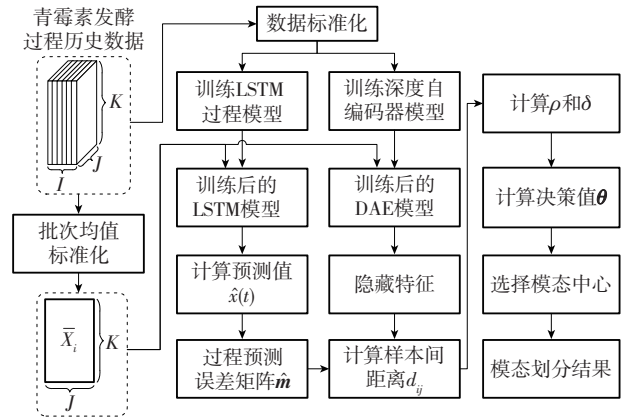


图3 基于DCC的间歇过程模态划分流程图

Fig. 3 Flowchart of mode partition based on DCC for batch processes

## 2 基于DCC-RVM的多模态间歇过程质量变量在线预测

### 2.1 相关向量机质量变量预测建模

相关向量机通过最大化边缘似然选择相关向量,并估计相关向量的权重建立预测模型,在处理高维数据和小样本数据时表现出色<sup>[20]</sup>。

给定输入变量 $\{x_n\}_{n=1}^N$ 和质量变量 $\{t_n\}_{n=1}^N$ ,利用RVM建立的质量变量预测模型可表示为

$$t = y(x, w) + \varepsilon = \left( \sum_{i=1}^N w_i K(x, x_i) + w_1 \right) + \varepsilon \quad (18)$$

式中, $\varepsilon \sim N(0, \sigma^2)$ 为随机噪声; $w_i$ 是模型的权值;

$K(\mathbf{x}, \mathbf{x}_i)$  为高斯核函数。

$\mathbf{t}$  的后验条件概率计算式为

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{-\frac{\|\mathbf{t} - \Phi\mathbf{w}\|^2}{2\sigma^2}\right\} \quad (19)$$

式中,  $\mathbf{w} = (w_1, w_2, \dots, w_{N+1})^T$ ;  $\mathbf{t} = (t_1, t_2, \dots, t_N)^T$ ;  $\Phi$  是  $N \times (N+1)$  维的矩阵。

为避免模型过拟合, 模型参数的约束条件为

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^N N(w_i|0, \alpha_i^{-1}) \quad (20)$$

式中,  $\boldsymbol{\alpha}$  为  $N+1$  维超参数。

RVM 对每个权值参数单独引入超参数, 保证模型的稀疏性。根据贝叶斯公式, 未知参数的后验概率为

$$p(\mathbf{w}, \boldsymbol{\alpha}, \sigma^2|\mathbf{t}) = p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2)p(\boldsymbol{\alpha}, \sigma^2|\mathbf{t}) \quad (21)$$

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \sigma^2) = (2\pi)^{-\frac{N+1}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\right\} \quad (22)$$

式中,  $\boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\Sigma} \Phi^T \mathbf{t}$ ;  $\boldsymbol{\Sigma} = (\sigma^{-2} \Phi^T \Phi + \mathbf{A})^{-1}$ ;  $\mathbf{A} = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_N)$ 。

超参数  $\boldsymbol{\alpha}$  和方差  $\sigma^2$  的计算式为

$$\alpha_i^{\text{new}} = \frac{1 - \alpha_i \Sigma_{ii}}{\mu_i^2} \quad (23)$$

$$(\sigma^2)^{\text{new}} = \frac{\|\mathbf{t} - \Phi\boldsymbol{\mu}\|^2}{N - \sum_{i=1}^N (1 - \alpha_i \Sigma_{ii})} \quad (24)$$

式中,  $\Phi = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \dots, \phi(\mathbf{x}_N)]^T$ ;  $\Sigma_{ii}$  为  $\boldsymbol{\Sigma}$  的第  $i$  个对角元素。

给定在线样本  $\mathbf{x}_{\text{test}}$ , RVM 模型的预测值  $\hat{\mathbf{y}}_{\text{test}}$  及其方差  $\hat{\sigma}_{\text{test}}^2$  的计算式为

$$\hat{\mathbf{y}}_{\text{test}} = \boldsymbol{\mu}^T \phi(\mathbf{x}_{\text{test}}) \quad (25)$$

$$\hat{\sigma}_{\text{test}}^2 = \sigma_{\text{MP}}^2 + \phi(\mathbf{x}_{\text{test}})^T \boldsymbol{\Sigma} \phi(\mathbf{x}_{\text{test}}) \quad (26)$$

## 2.2 基于 DCC-RVM 的多模态间歇过程质量变量在线预测方法实现

如图 4 所示, 本文首先利用 DCC 方法划分间歇过程模态; 然后依据各模态的切换点和数据集建立多模态 RVM 质量变量预测模型  $\{\text{RVM}_m\}_{m=1}^M$ ; 最后依据在线样本  $\mathbf{x}_{\text{new}}$  的采样时刻判断  $\mathbf{x}_{\text{new}}$  所属模态  $m$ ,  $\mathbf{x}_{\text{new}}$  时刻模型的预测值为

$$\hat{\mathbf{y}}_{\text{new}} = \text{RVM}_m(\bar{\mathbf{x}}_{\text{new}}) \quad (27)$$

式中,  $\bar{\mathbf{x}}_{\text{new}}$  由  $\mathbf{x}_{\text{new}}$  标准化得到。

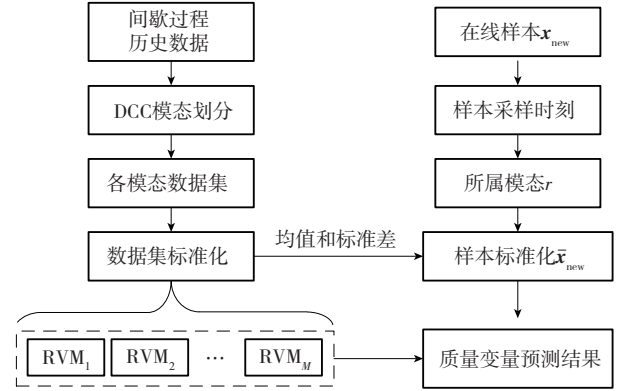


图4 基于 DCC-RVM 的多模态间歇过程质量变量在线预测流程图

Fig. 4 Flow chart of online prediction for quality variable based on DCC-RVM in multimode batch processes

## 3 实验结果与讨论

### 3.1 数值仿真

设计具有 5 个过程变量和 2 个模态的数值仿真过程验证 DCC 方法的有效性。共生成 5 个批次的数据, 每批次的样本数为 100, 前 50 个样本属于第一模态, 后 50 个样本属于第二模态。该过程表示为

$$\text{模态一:} \begin{cases} \mathbf{x}_1(t) = t^2 + \varepsilon \\ \mathbf{x}_2(t) = t^2 + t + \varepsilon \\ \mathbf{x}_3(t) = 0.1 \times t^2 + t + \varepsilon \\ \mathbf{x}_4(t) = 0.2 \times \sqrt{t} + t^2 + \varepsilon \\ \mathbf{x}_5(t) = b + \varepsilon \end{cases} \quad (28)$$

$$\text{模态二:} \begin{cases} \mathbf{x}_1(t) = -t^2 + 0.2 \times t + \varepsilon \\ \mathbf{x}_2(t) = t^3 - 0.2 \times t + \varepsilon \\ \mathbf{x}_3(t) = t^2 + t + \varepsilon \\ \mathbf{x}_4(t) = \sqrt{t} + t^3 + \varepsilon \\ \mathbf{x}_5(t) = b + \varepsilon \end{cases} \quad (29)$$

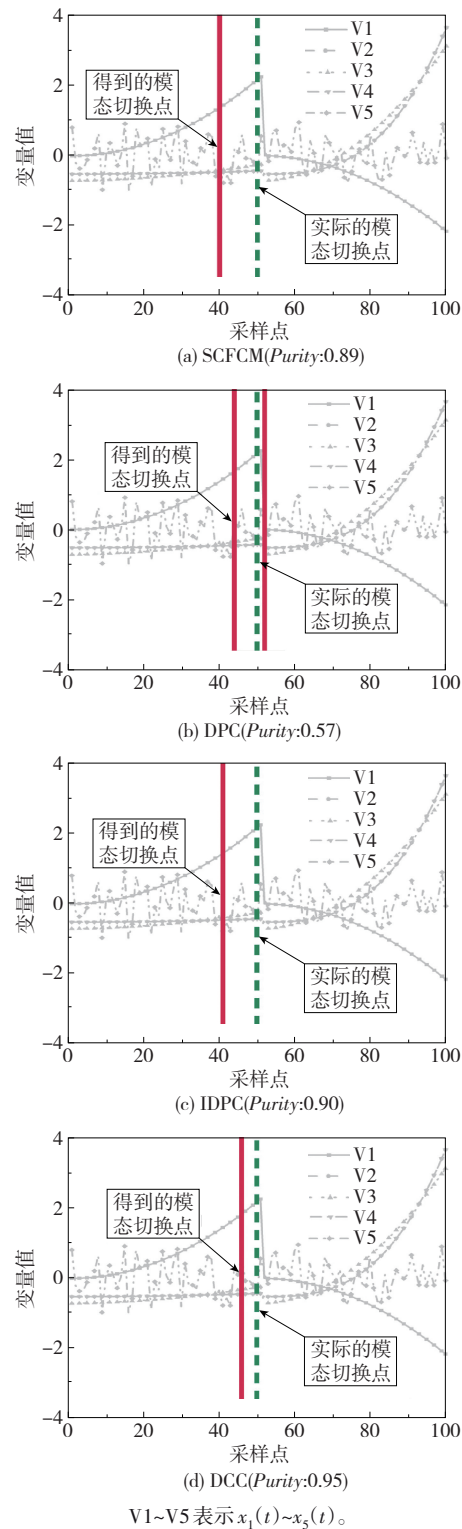
式中,  $t$  表示采样时间;  $\varepsilon \sim N(0, 0.1)$  为高斯噪声。

选择 sequential constrained fuzzy c-means (SCFCM) 方法<sup>[12]</sup>、DPC<sup>[13-14]</sup>方法和 improved density peaks clustering (IDPC) 方法<sup>[16]</sup>作为对比方法, 采用纯度 (Purity)<sup>[21]</sup>作为模态划分的评价指标。

$$\text{Purity} = \sum_{i=1}^k \frac{m_i}{m} p_i \quad (30)$$

式中,  $k$  为模态的数目;  $m$  为批次样本的数目;  $m_i$  为第  $i$  模态样本的数目;  $p_i$  为第  $i$  模态最多数目的样本占总体样本的比例。Purity 反映了正确模态划分的样本数占总样本的比例, 其取值范围为  $[0, 1]$ , 值越大表示模态划分结果越好。

不同方法数值仿真的模态划分结果如图 5 所示,其中绿色虚线为数值仿真的实际模态切换点,红色实线则对应不同方法得到的模态切换点。从



V1~V5 表示  $x_1(t) \sim x_5(t)$ 。

图 5 不同方法数值仿真的模态划分结果

Fig. 5 Mode partition results for numerical simulation using different methods

图 5 可看出,SCFCM 方法的纯度为 0.89;DPC 方法的纯度为 0.57;图 5(c) 的 IDPC 方法将数据划分成两个模态,但纯度与图 5(d) 方法相比降低了 0.05;图 5(d) 为本文 DCC 方法,模态切换点与绿色虚线最为接近,纯度达到 0.95。DPC 没有考虑时序约束,导致模态划分结果出现了多个模态切换点,模态划分准确性较低;SCFCM 和 IDPC 未考虑过程特性与模态中心的因果关系,模态切换点与实际相差较大。本文 DCC 方法的模态划分结果即 *Purity* 值相较于 SCFCM、DPC 和 IDPC 分别提高了 6.74%、66.67% 和 5.56%,表明本文模态划分方法具有较高的准确性。

3.2 青霉素发酵过程

青霉素发酵仿真平台 Pensim 采用 Birol 机理模型<sup>[22]</sup>,该模型不仅包含过程控制变量,如通风率、反应器温度和 pH 值等,也包含生产过程中一些关键的状态量,如产热量、青霉素浓度和二氧化碳浓度等。

本文利用 Pensim 仿真平台生成 23 个批次过程数据,其中 20 个批次作为训练集,3 个批次作为测试集,等间隔采样 400 个数据点。用于预测模型建模的过程变量如表 1 所示,其中青霉素浓度为发酵过程待测质量变量。

| 表 1 青霉素发酵过程变量                                            |              |          |                |
|----------------------------------------------------------|--------------|----------|----------------|
| Table 1 Variables of the penicillin fermentation process |              |          |                |
| 符号                                                       | 变量名称         | 符号       | 变量名称           |
| $x_1$                                                    | 时间(h)        | $x_7$    | 反应器体积(L)       |
| $x_2$                                                    | 通风率(L/h)     | $x_8$    | 二氧化碳浓度(mmol/L) |
| $x_3$                                                    | 搅拌功率(W)      | $x_9$    | pH             |
| $x_4$                                                    | 底物流加速率(L/h)  | $x_{10}$ | 反应器温度(K)       |
| $x_5$                                                    | 底物流温度(K)     | $x_{11}$ | 产热量(kcal/h)    |
| $x_6$                                                    | 溶解氧浓度(mol/L) | $x_{12}$ | 青霉素浓度(g/L)     |

1 cal=4.184 J。

不同方法的模态划分结果如图 6 所示。由图 6 可以看出,由于 DPC 未考虑样本的时序性,模态划分结果出现时序错乱,而 SCFCM、IDPC 和 DCC 方法均加入了时序约束,因此模态划分结果较合理。分别建立各模态的质量变量预测模型,利用网格搜索获得 RVM 模型的最优核参数,搜索范围为[0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1]。质量变量预测模型的评价指标为均方根误差(*RMSE*)<sup>[23]</sup>和判定系数(*R*<sup>2</sup>)<sup>[24]</sup>。*RMSE* 越小、*R*<sup>2</sup> 越接近 1,表明模型的预测精度越高,其计算式分别为

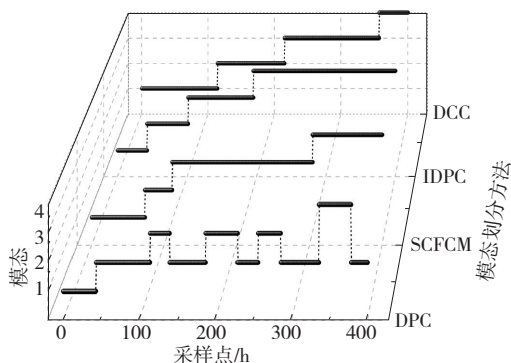
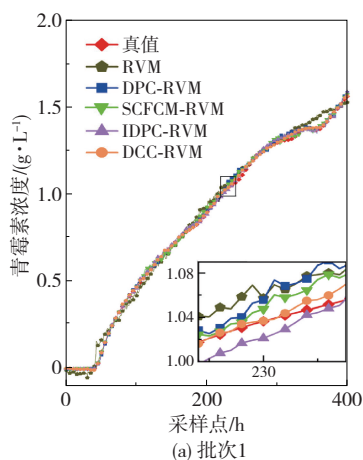


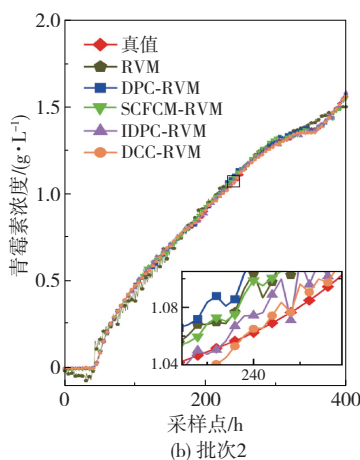
图6 各方法的模态划分结果

Fig. 6 Mode partition results of each method

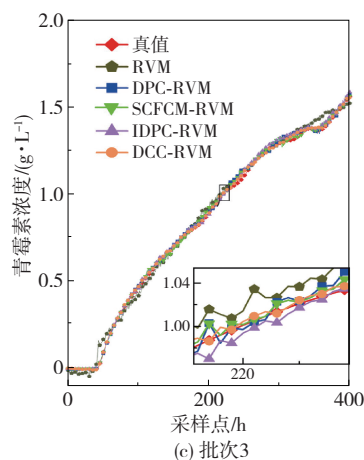
$$RMSE = \sqrt{\sum_{k=1}^K (\hat{y}_k - y_k)^2 / K} \quad (31)$$



(a) 批次1



(b) 批次2



(c) 批次3

图7 不同模型在各批次上的预测结果

Fig. 7 Prediction results using different models in all batches

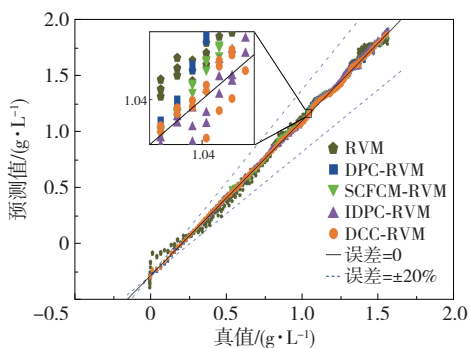


图8 不同模型的质量变量预测误差

Fig. 8 Prediction error of quality variable using different models

如表2所示,由于SCFCM-RVM和DPC-RVM的预测模型在模态划分时均采用欧式距离度量样本间的相似性,影响了模态划分结果的合理性,使得模型预测精度低于IDPC-RVM和DCC-RVM预测模

$$R^2 = 1 - \frac{\sum_{k=1}^K (\hat{y}_k - y_k)^2}{\sum_{k=1}^K (\hat{y}_k - \bar{y})^2} \quad (32)$$

式中, $y_k$ 、 $\hat{y}_k$ 和 $\bar{y}$ 分别表示质量变量的真值、预测值和均值。

不同预测模型在各批次上的预测结果如图7所示。可以看出RVM预测模型未考虑间歇过程的多模态特性,预测精度较差。与RVM、DPC-RVM、SCFCM-RVM和IDPC-RVM相比,DCC-RVM具有较高的预测精度。由图8可知,RVM模型由于忽略了间歇过程的多模态特性,在质量变量预测浓度为0~0.25 g/L时预测误差超过20%,预测精度较低。与DPC-RVM、SCFCM-RVM和IDPC-RVM模型相比,DCC-RVM模型的预测误差较小。

表2 不同模型在所有测试批次上的质量变量预测结果

Table 2 Prediction results for quality variable in all test batches of different models

| 预测模型           | 平均RMSE         | 平均 $R^2$       |
|----------------|----------------|----------------|
| RVM            | 0.034 6        | 0.994 8        |
| DPC-RVM        | 0.015 1        | 0.999 0        |
| SCFCM-RVM      | 0.014 5        | 0.999 1        |
| IDPC-RVM       | 0.013 3        | 0.999 2        |
| <b>DCC-RVM</b> | <b>0.010 4</b> | <b>0.999 5</b> |

加粗表示评价指标为最优。

型。本文方法充分考虑模态中心与过程特性的因果关系,合理地将间歇过程划分为多个模态,建立的多模态RVM质量变量预测模型的平均RMSE和 $R^2$ 优于其他模型,有效提高了质量变量预测模型的预测精度。

## 4 结论

为了提高多模态间歇过程质量变量的预测精度,本文提出了一种基于DCC-RVM的多模态间歇过程质量变量在线预测方法。该方法利用DAE提取间歇过程数据的深层特征,降低数据非线性对模态划分的影响,并建立LSTM过程模型,构建因果关系约束的模态中心选取策略,使得模态中心选择更加合理。引入时序约束分配非模态中心样本,避免模态划分结果出现时序错乱。最后,结合各模态数据集建立多模态RVM质量变量预测模型。青霉素发酵过程实验的结果表明,本文建立的质量变量预测模型在3个测试批次上预测的平均 $RMSE$ 为0.0104,平均 $R^2$ 为0.9995,优于其他比较模型,具有较高的预测精度。

## 参考文献:

- [1] CHANG P, DING C H. Monitoring multi-domain batch process state based on fuzzy broad learning system [J]. Expert Systems with Applications, 2022, 187: 115851.
- [2] LIU K, ZHAO X Q, MOU M, et al. Quality prediction of multi-stage batch process based on integrated ConvBi-GRU with attention mechanism [J]. Applied Intelligence, 2025, 55: 123.
- [3] SUN Y N, QIN W, XU H W, et al. A multiphase information fusion strategy for data-driven quality prediction of industrial batch processes [J]. Information Sciences, 2022, 608: 81–95.
- [4] REN J C, LIU D, WAN Y. VMD-SEAE-TL-based data-driven soft sensor modeling for a complex industrial batch processes [J]. Measurement, 2022, 198: 111439.
- [5] JI C, MA F Y, WANG J D, et al. Profitability related industrial-scale batch processes monitoring via deep learning based soft sensor development [J]. Computers and Chemical Engineering, 2023, 170: 108125.
- [6] ZHANG S M, BAO X L. Phase partition and online monitoring for batch processes based on Harris hawks optimization [J]. Control Engineering Practice, 2023, 138: 105554.
- [7] CHU F, WANG J C, ZHAO X, et al. Transfer learning for nonlinear batch process operation optimization [J]. Journal of Process Control, 2021, 101: 11–23.
- [8] WANG Z J, CHEN B, ZHENG Y, et al. Sparse representation and modified density peak clustering-based state identification for multimode processes [J]. Measurement Science and Technology, 2023, 34: 025016.
- [9] GUO R X, ZHANG N, WANG J Q, et al. Phase partition and identification based on a two-step method for batch process [J]. Transactions of the Institute of Measurement and Control, 2018, 40(16): 4472–4483.
- [10] LU N Y, GAO F R, WANG F L. Sub-PCA modeling and on-line monitoring strategy for batch processes [J]. AIChE Journal, 2004, 50(1): 255–259.
- [11] NG Y S, SRINIVASAN R. An adjoined multi-model approach for monitoring batch and transient operations [J]. Computers and Chemical Engineering, 2009, 33(4): 887–902.
- [12] LUO L J, BAO S Y, MAO J F, et al. Fuzzy phase partition and hybrid modeling based quality prediction and process monitoring methods for multiphase batch processes [J]. Industrial & Engineering Chemistry Research, 2016, 55(14): 4045–4058.
- [13] YAN H, WANG M Z, XIE J Y. ANN-DPC: density peak clustering by finding the adaptive nearest neighbors [J]. Knowledge-Based Systems, 2024, 294: 111748.
- [14] ZANG W K, LIU X C, MA L L, et al. DPC-MFP: an adaptive density peaks clustering algorithm with multiple feature points [J]. Neurocomputing, 2025, 618: 129060.
- [15] 刘聪, 谢莉, 杨慧中. 基于改进DPC的青霉素发酵过程多模型软测量建模 [J]. 化工学报, 2021, 72(3): 1606–1615.
- LIU C, XIE L, YANG H Z. Multi-model soft sensor development for penicillin fermentation process based on improved density peak clustering [J]. CIESC Journal, 2021, 72(3): 1606–1615. (in Chinese)
- [16] 周新杰, 王建林, 艾兴聪, 等. 基于IDPC-RVM的多模态间歇过程质量变量在线预测 [J]. 化工学报, 2022, 73(7): 3120–3130.
- ZHOU X J, WANG J L, AI X C, et al. IDPC-RVM based online prediction of quality variables for multimode batch processes [J]. CIESC Journal, 2022, 73(7): 3120–3130. (in Chinese)
- [17] XU M L, LI Y H, LI R X, et al. EADP: an extended adaptive density peaks clustering for overlapping community detection in social networks [J]. Neurocomputing, 2019, 337: 287–302.
- [18] FENG Z K, ZHANG J S, NIU W J. A state-of-the-art review of long short-term memory models with applications in hydrology and water resources [J]. Applied Soft Computing, 2024, 167: 112352.
- [19] 邵昌昇, 楼巍, 严利民. 高维数据中的相似性度量算

- 法的改进[J]. 计算机技术与发展, 2011, 21(2): 1-4.
- SHAO C S, LOU W, YAN L M. Optimization of algorithm of similarity measurement in high-dimensional data [J]. Computer Technology and Development, 2011, 21(2): 1-4. (in Chinese)
- [20] QIU K P, WANG J L, ZHOU X J, et al. Soft sensor based on localized semi-supervised relevance vector machine for penicillin fermentation process with asymmetric data[J]. Measurement, 2022, 202: 111823.
- [21] NGUYEN T H T, DINH D T, SRIBOONCHITTA S, et al. A method for  $k$ -means-like clustering of categorical data[J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(11): 15011-15021.
- [22] BIROL G, ÜNDEY C, ÇINAR A. A modular simulation package for fed-batch fermentation: penicillin production [J]. Computers and Chemical Engineering, 2002, 26(11): 1553-1565.
- [23] HODSON T O. Root-mean-square error (RMSE) or mean absolute error(MAE): When to use them or not[J]. Geoscientific Model Development, 2022, 15: 5481-5487.
- [24] 黎宏陶,王振雷,王昕. 基于即时学习的改进条件高斯回归软测量[J]. 化工学报, 2024, 75(6): 2299-2312.
- LI H T, WANG Z L, WANG X. Improved conditional Gaussian regression soft sensor based on just-in-time learning[J]. CIESC Journal, 2024, 75(6): 2299-2312. (in Chinese)

## An online prediction of quality variables based on DCC-RVM for multimode batch processes

LIN Lin WANG JianLin\* SUI EnGuang XIN Wei ZHANG ZeBin

(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

**Abstract:** Batch processes have multimode characteristics, and the existing methods of mode partition for batch processes ignore the causal relationship between the process characteristics and the mode centers, which directly affects the accuracy and reasonableness of the mode partition results and the prediction precision and generalizability of the models. A method for the online prediction of quality variables based on deep causal clustering-relevance vector machine (DCC-RVM) for multimode batch processes is proposed in this work. By combining density peak clustering, with the strong nonlinearity of the batch process data, the deep features of the process data are first extracted using a deep autoencoder. The process model for batch processes based on long short-term memory (LSTM) is then constructed, with the causal relationship between mode centers and the LSTM process model as a constraint, and the mode centers selection strategy is constructed to obtain reasonable mode centers; Subsequently, by considering the temporal features of the process data, the non-mode center samples are assigned to the corresponding modes based on the relative distance between samples, and the partition points and data sets are obtained for each mode; Finally, the quality variable prediction model for each mode is established using RVM, and the batch process quality variable is predicted online. The experimental results show that the mode partition by DCC is reasonable. The root mean square error ( $RMSE$ ) and  $R^2$  for predictions using our DCC-RVM model are 0.010 4 and 0.999 5, respectively.

**Key words:** batch processes; mode partition; quality variable prediction; density peak clustering; causal relationship

(责任编辑:吴万玲)