文章编号:1671-4598(2021)07-0136-04 DOI:10.16526/j.cnki.11—4762/tp.2021.07.027 中图分类号:TP274.2 文献标识码:A

基于栈式降噪自编码器的发酵过程回归建模

岳向阳[1,2],赵忠盖[1,2],刘 飞[1,2]
(1.江南大学 轻工过程先进控制教育部重点实验室,江苏 无锡 214122;
2.江南大学自动化研究所,江苏 无锡 214122)

摘要:精确有效的发酵过程模型不仅能够定量揭示过程信息间的关联,实现对难以实时监测变量的预测,而且是进一步控制和优化的前提;基于数据驱动的发酵过程建模方法得到了广泛研究与应用,然而其仅考虑发酵过程的非线性特征和数据具有多采样率的特点,忽略了过程数据中测量噪声对模型的影响;为此,提出基于栈式降噪自编码器的发酵过程回归建模方法,该方法不仅具有较强的非线性拟合能力,半监督的学习策略也能够充分挖掘发酵过程中的所有数据信息,同时可以从含噪声的过程数据中提取出鲁棒性的特征,使模型具有噪声适应性;通过青霉素仿真对比实验结果表明,该模型的预测性能更好。

关键词:发酵过程;建模;栈式降噪自编码器;半监督学习;噪声

Modeling of Fermentation Process Based on Stacked Denoising Autoencoder

YUE Xiangyang[1,2], ZHAO Zhonggai[1,2], LIU Fei[1,2]
(1. Ministerial Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China; 2. Institute of Automation, Jiangnan University, Wuxi 214122, China)

Abstract: An accurate and effective fermentation process model can not only quantitatively reveal the correlation between process information and realize the prediction of variables that are difficult to monitor in real time, but also is a prerequisite for further control and optimization. The data-driven modeling method has been widely researched and applied. However, it only considers the nonlinear characteristics of the fermentation process and the characteristics of the data with multiple sampling rates, and ignores the influence of measurement noise in the process data on the model. For this reason, a regression modeling method for fermentation process based on stacked denoising autoencoder is proposed. This method not only has strong nonlinear fitting ability, but also semi-supervised learning strategy can fully mine all data information in the fermentation process. At the same time, robust features can be extracted from the noisy process data, so that the model has certain noise adaptability. The results of penicillin simulation and comparison experiments show that the prediction performance of this model is better.

Keywords: fermentation process; stacked denoising autoencoder; semi-supervised; modeling; noise

0 引言

发酵过程中的生化反应十分复杂,过程中的非线性、不确定性严重[1],缺乏对重要生物学参数的在线监测设备,导致发酵过程的自动化水平远不如其它工业生产过程[2]。对发酵过程进行建模不仅能够揭示过程信息间的关联,实现对难以实时监测变量的预测,而且精确有效的数学模型是进一步实施发酵过程自动控制和优化的前提。

目前,常用的发酵过程建模方法可分为机理建模和数据建模。Trelea 等人对啤酒发酵过程的机理模型进行研究[3],张玲玲则构建了诺西肽发酵过程的机理模型[4]。机理建模涉及对微生物复杂生长代谢活动的分析,常需进行简化,从而导致模型泛化能力不足。伴随智能控制技术的发展,数据建模方法在发酵过程建模中得到广泛应用,如支持向量机(support vector machine,SVM)和人工神经网络(artificial neural network,ANN)。Dach 等人使用传统 ANN 成功预测浆液发酵中甲烷排放的水平[5],Zhong 等人使用 SVM 构建了抗生素发酵模型[6]。尽管 ANN 和 SVM 在发酵过程建模中得到应用,但传统 ANN 中随机初始化权值的策略容易陷入局部最优或梯度消失,SVM 则属于浅层的神经网络[7],非线性表示性能较弱,而且计算复杂度随训练样本数量呈指数增长。另外,SVM 和 ANN 在发酵过程建模中的不足之处是二者均为有监督学习策略,只能利用发酵过程中的标签数据进行数据建模,未挖掘无标签数据中可能含有的丰富过程信息。

随着计算机硬件和神经科学的突破,深度学习技术在语音、图像识别等领域得到成功的应用,其表示性能超过了 SVM、浅层 ANN 等,其标志性突破是 Hinton 提出深度学习概念[8],即先利用无标签数据对深度信念网络进行的逐层预训练,然后利用标签数据对模型进行微调,这种半

收稿日期:2021-01-03; **修回日期**:2021-01-19。

作者简介:岳向阳(1996-),男,河南禹州人,硕士,主要从事基于机器学习的发酵过程建模与优化方向的研究。
赵忠盖(1976-),男,湖北荆州人,博士,教授,主要从事间歇过程建模与估计、工业过程监控与诊断方向的研究。
刘 飞(1965-),男,安徽宣城人,博士,教授,主要从事先进控制、工业系统监控与诊断、智能装备与系统方向的研究。

引用格式:岳向阳,赵忠盖,刘 飞.基于栈式降噪自编码器的发酵过程回归建模[J].计算机测量与控制,2021,29(7):136-139,155.

监督学习策略可以充分挖掘过程中所有数据的信息，Erhan 等人通过大量基准实验说明了无监督预训练的有效性[9]。随后，Shang 等人利用深度信念网络方法对粗蒸馏装置的重柴油 95%切点进行估计[10]，并与偏最小二乘等方法进行比较，证明其有效性。Gopakumar 等人采用半监督学习策略来分别对链激酶和青霉素发酵过程进行建模[11]，并与传统 ANN 和支持向量回归方法进行对比试验，取得了更好的预测性能。虽然上述深度学习方法充分利用了发酵过程中所有的原始数据，但实际发酵生产过程中，由于测量设备性能退化等原因，往往会使测量数据含有噪声[12]，会造成模型预测性能显著下降，这要求发酵过程模型要有一定的噪声适应性。Vincent 等人在自编码器（autoencoder，AE）基础上提出降噪自编码器（denoising autoencoder，DAE)[13]，该方法通过对原始数据加入随机噪声，使得模型提取到的特征具有一定的鲁棒性，而且多个 DAE 逐层堆叠而构成的栈式降噪自编码器（stacked denoising autoencoder，SDAE）能够提取出更深层次的特征，从而提升模型的泛化能力[14]，SDAE 已在故障诊断、图像分类等领域得到了成功的应用[15-16]。

发酵过程具有非线性特征、变量间存在多采样率以及数据含噪声的特点，而 SDAE 方法不仅可以有效拟合发酵过程的非线性，半监督的学习策略也能够充分利用发酵过程的所有数据信息，此外模型还能够提取出具有鲁棒性的深层特征，从而对过程噪声具有一定的适应性。因此本文将 SDAE 方法应用到发酵过程回归建模，通过青霉素仿真对比实验说明基于 SDAE 的发酵过程回归模型能够更好地预测关键生物学参数，可以用于进一步的发酵过程控制和优化。

1　青霉素发酵过程建模问题描述

青霉素发酵过程中由于菌体生长繁殖等都将产生一定的热能，而温度会对酶特性、发酵液物理性质等产生显著影响，因此需要实时改变热水或冷水流量，使发酵罐环境保持在最适发酵温度。另外，菌体的生长代谢会影响培养基的氢离子平衡，从而改变发酵液 pH，而发酵液不同的 pH 会导致菌体细胞膜的通透性等产生明显差异，通过实时调节酸液或碱液流加速率能够使发酵液 pH 稳定在最适的范围。而且，溶氧浓度会影响产物合成以及与菌体呼吸链有关的能量代谢，通过不断调整无菌空气流量和搅拌功率可以满足菌体在不同生长阶段对溶氧浓度的要求。

青霉素发酵生产工艺中需要检测的参数分为三类：物理参数、化学参数和生物学参数。物理参数包括温度、搅拌功率、底物流加速率等。化学参数包括 pH、溶氧溶度和二氧化碳等。这些物理和化学参数都能在线准确测量和控制。为控制菌体的生长、能量代谢等，需要对菌体浓度等生物学参数进行监测，然而生物学参数常要人工取样后离线检测，所得数据无法用于实时控制、优化[17]。为能够实时获取青霉素发酵过程中的生物学参数，需要建立青霉素发酵过程的回归模型。

2　基于 SDAE 的发酵过程回归建模

发酵过程的本质是微生物在生命周期内的一系列代谢活动，其一般分为 4 个生长阶段，即迟滞期、对数生长期、稳定期和凋亡期，而细胞生长代谢状态会在不同生长阶段随着自身特性和培养环境的变化而变化，具有显著的非线性过程特征。另外，由于发酵过程中物理或化学参数和生物学参数的测量形式不同，会产生大量的无标签数据，其中很可能蕴含丰富的过程信息。同时，用于监测发酵过程物理和化学参数的传感器会出现性能失准，如传感器探头老化、探头敏感部位被反应液堵塞等现象，导致过程数据中含有噪声。

2.1　SDAE 方法

2.1.1　自编码器

AE 本质是一个试图还原初始输入的系统，它的神经网络结构形式如图 1，由输入层、隐含层和输出层组成。训练过程使用无标签数据，使得 AE 能去充分挖掘无标签数据中的过程信息。

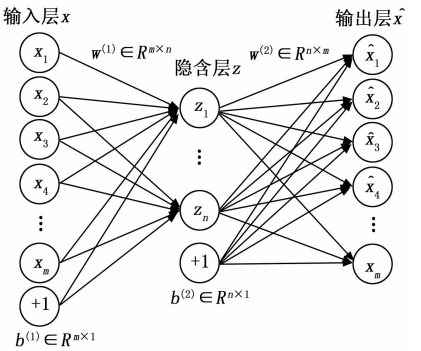

图 1　自编码器（神经网络形式）

输入向量 x 经过编码可以获得隐含层向量 z，该过程表示如下：

$$\boldsymbol{z} = f(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}) \tag{1}$$

其中：$x \in \boldsymbol{R}^{m\times 1}$ 为输入向量，$\boldsymbol{W}^{(1)} \in \boldsymbol{R}^{m\times n}$ 为权值矩阵，$\boldsymbol{b}^{(1)} \in \boldsymbol{R}^{m\times 1}$ 为输入偏置，$\boldsymbol{z} \in \boldsymbol{R}^{m\times 1}$ 为隐含层向量，$f(\cdot)$ 是激活函数。

隐含层向量 z 经过解码可以获得重构输入向量 $\hat{x}$，函数表示如下：

$$\hat{\boldsymbol{x}} = g(\boldsymbol{W}^{(2)}\boldsymbol{z} + \boldsymbol{b}^{(2)}) \tag{2}$$

其中：$\hat{\boldsymbol{x}} \in \mathbf{R}^{n\times 1}$ 为重构输入向量，$\boldsymbol{b}^{(1)} \in \boldsymbol{R}^{n\times 1}$ 为输出偏置，重构误差可表示为：

$$L = \left\| x - g(f(x)) \right\|^2 \tag{3}$$

定义代价函数为：

$$J(\boldsymbol{W}^{(1)},\boldsymbol{W}^{(2)},\boldsymbol{b}^{(1)},\boldsymbol{b}^{(2)}) = \frac{1}{N}\sum_{i=1}^{N}\left\| \boldsymbol{x}^{(i)} - g(f(\boldsymbol{x}^{(i)})) \right\|^2 \tag{4}$$

其中：N 为样本的数量，$\boldsymbol{x}^{(i)}$ 代表第 i 个样本，最优 $(\boldsymbol{W}^{(1)},\boldsymbol{W}^{(2)},\boldsymbol{b}^{(1)},\boldsymbol{b}^{(2)})$ 可以通过误差反向传播算法得到。

AE 通过神经网络来学习每个样本的唯一抽象表示，但是当神经网络的参数复杂到一定程度时 AE 很容易存在过拟

合的风险。

2.1.2 降噪自编码器

DAE 是先对输入向量随机地加入噪声，然后对其进行编码、解码，使提取到的特征具有一定的鲁棒性，其基本结构如图 2 所示。

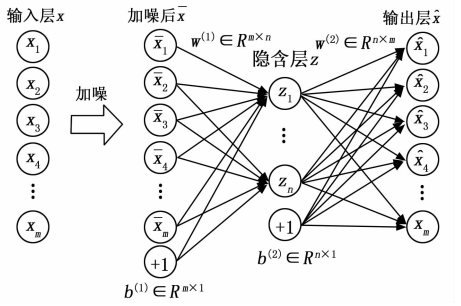

图 2 降噪自编码器

图中，x 为原始输入向量，$\bar{x}$ 为加入噪声后的输入数据，$\hat{x}$ 是重构的输入向量，重构误差表示为：

$$L = \left\| x - g(f(\bar{x})) \right\|^2 \tag{5}$$

定义代价函数为：

$$J(\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}) = \frac{1}{N} \sum_{i=1}^{N} \left\| x^{(i)} - g(f(\bar{x}^{(i)})) \right\|^2 \tag{6}$$

目前加噪声的方式分为两种，一种是添加服从特定分布的随机噪声，另一种是随机将特定比例的输入节点置为零。

2.1.3 栈式降噪自编码器

将多个 DAE 逐层堆叠构成 SDAE，其深度模型结构如图 3 所示。

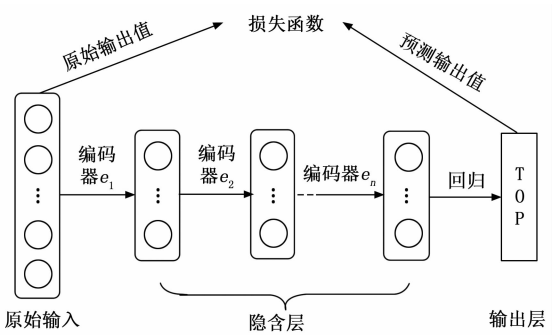

图 3 栈式降噪自编码器

深度模型具有更强大的近似复杂函数的能力，且经过多层提取得到的特征更具有表示性。首先对网络前 n 层采用逐层贪婪学习算法进行无监督预训练，即使用 DAE 算法训练第一层编码器，记录该层参数，将第一层得到的隐含层输出作为第二层输入，训练第二层编码器后继续记录参数，直到第 n 层编码器训练完毕，然后将前 n 层记录好的参数作为整体网络的初始参数，最后对整体网络进行有监督地微调。

2.2 算法流程

基于 SDAE 的发酵过程回归建模流程描述如下：

1）采集发酵过程样本数据。

2）对过程数据进行数据预处理。将数据分为预训练集、微调集、验证集和测试集四部分。

3）利用预训练集、微调集和验证集建立 SDAE 模型。预训练集用来对 SDAE 进行逐层贪婪训练，获得初始参数。随后加一层神经网络作为输出层构成 SDAE－NN，使用微调集对 SDAE－NN 的参数进行微调。验证集是在微调过程中监控模型性能，可以有效防止过拟合。

4）利用测试集评估模型性能。

3 青霉素发酵过程建模仿真分析

青霉素发酵过程是已知的用于分批补料反应器建模的基准工艺[18]。青霉素发酵过程仿真平台 PenSim 以 Birol 机理模型为内核[19]，可以在不同的操作模式下运行，得到了广泛的应用。

3.1 数据产生、预处理与划分

3.1.1 PenSim 产生数据

青霉素发酵过程在不同批次间会存在特性差异，本文则利用计算机模拟来随机设定 PenSim 平台的初始条件，共产生 50 批青霉素发酵过程数据，其中每批发酵总时长均为 400 h，采样间隔为 0.5 h。

青霉素发酵过程模型输入变量中的通风率和搅拌功率可以调控溶氧，底物流加速率用于控制基质浓度，底物流加温度、发酵罐温度和 pH 则影响发酵液的物理性质。输出变量中的青霉素浓度是实际发酵生产水平的主要体现，而基质浓度会影响生产效率，过低会导致菌体营养不良，但过高又会使得菌体耗氧增加，降低青霉素产率。

3.1.2 数据预处理

青霉素发酵过程数据通常具有不同的量纲，这会对建模产生不利影响并减缓算法收敛速度，因此需要将数据归一化。本文使用 Z－score 标准化方法，其变换形式为：

$$X'_i(t) = \frac{X_i(t) - \bar{x_i}}{\sigma_i} \tag{7}$$

其中：$X_i(t)$ 为原始数据，$X'_i(t)$ 是归一化后的数据，$\bar{x_i}$ 是第 i 个变量的均值，σ_i 是第 i 个变量的标准差。

3.1.3 数据划分

划分数据的方式是影响青霉素发酵过程模型性能的一个重要因素。一方面，训练集应包含多样的过程信息，否则模型将学习不到训练集中不存在的信息，这将会对模型预测性能产生影响。另一方面，测试数据集中不应覆盖近似的过程信息，否则模型性能会随着过程特性变化而出现显著差异。

本文首先将生成的 50 批数据中，40 批用作训练集，5 批用作验证集，5 批用作测试集，然后随机删去训练集中 40％数据点的目标向量，从而将训练数据进一步分为预训练集（无标签数据）和微调集（标签数据）。预训练集来用于进行无监督的预训练，微调集是用于进行有监督的微调，验证集既用于选择模型中的超参数，也可以在训练过程中监控模型性能变化从而避免过拟合，测试集是用于测试模型泛化性能。

3.2　仿真结果分析

本文采用均方误差 σ_{RMSE} 和最大绝对值误差 σ_{MAXE} 评价指标来分析回归模型的估计性能，指标定义如下：

$$\sigma_{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2},i=1,2,3,\ldots,n \quad (8)$$

$$\sigma_{MAXE}=\max(|Y_i-\hat{Y}_i|),i=1,2,3,\ldots,n \quad (9)$$

式中，N 是样本数，Y_i 是真实值，$\hat{Y}_i$ 是模型预测值，σ_{RMSE} 和 σ_{MAXE} 值越小说明模型预测性能越好。

通过实验分析不同 SDAE 模型在验证集上的性能，确定 SDAE－NN 网络结构为 6－5－5－4－3－2。为验证 SDAE 方法的有效性，将该方法与传统多层 ANN 与 SAE (Stacked Autoencoder) 进行比较，3 种方法采用的网络结构保持一致，传统 ANN 采用随机初始化权值的策略进行模型训练，SAE 相较于 SDAE 则是在逐层贪婪预训练阶段中没有对输入数据添加随机噪声。

为模拟现实过程中存在一些仪器和测量噪声的情况，在保证样本数据不失真的情况下，把原始样本数据中的过程变量加入 5%～10%的高斯噪声，得到含有噪声的样本数据。实验过程中利用原始样本和加噪样本两组数据，来分别测试模型性能。

以其中一批青霉素浓度为例，3 种模型的预测性能见图 4（上为原始样本，下为加噪样本），表 1 则定量地列出两组样本在各个模型测试集上的性能指标。

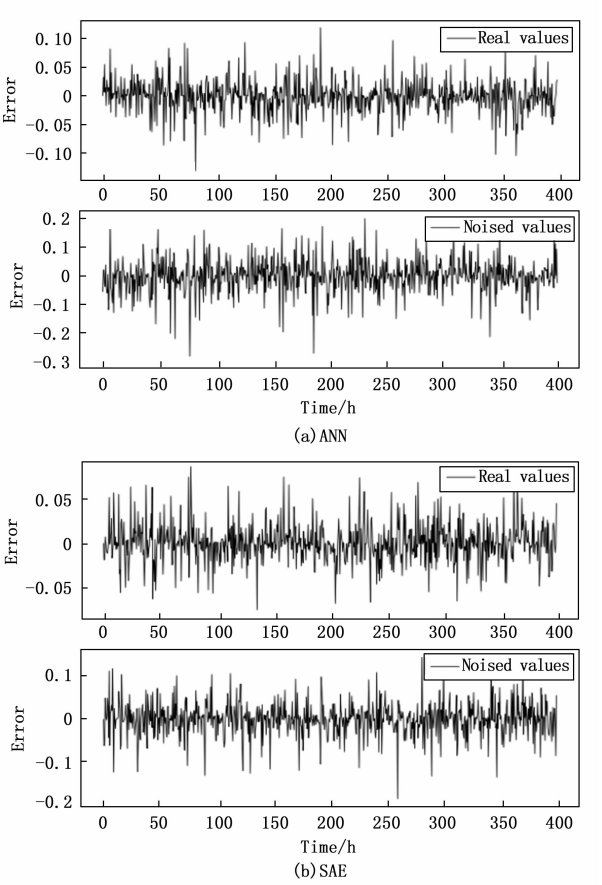

图 4　3 种模型的测试集预测性能

结合图 4 和表 1 分析可得，一方面，对于原始样本或加噪样本，SDAE 的均方误差 σ_{RMSE} 和最大绝对值误差 σ_{MAXE} 都是最小的，另一方面，对原始样本加噪后，虽然 ANN、SAE、SDAE 模型的泛化能力都有所下降，但 SDAE 模型对于加噪样本仍然具有很好的预测性能，由此说明基于 SDAE 的发酵过程回归模型不仅具有更强的非线性拟合能力，而且对于含噪声数据具有更好的泛化性。

表 1　3 种模型的性能指标

模型	样本类型	σ_{RMSE}	σ_{MAXE}
ANN	原始样本	0.030	0.142
	加噪样本	0.052	0.269
SAE	原始样本	0.021	0.093
	加噪样本	0.045	0.192
SDAE	原始样本	0.011	0.057
	加噪样本	0.021	0.096

4　结束语

本文提出基于 SDAE 的发酵过程回归建模方法，该策略的多层神经网络结构可以有效拟合发酵过程的非线性，而且半监督的学习策略能够充分挖掘发酵过程的无标签数据信息，同时能够提取出发酵过程数据中深层次的鲁棒特征，使模型具有一定的噪声适应性，进而提升模型的泛化性能。最后利用 PenSim 仿真数据进行多组对比试验，通过 σ_{RMSE} 和 σ_{MAXE} 两项性能指标说明 SDAE 与 ANN、SAE 模型相比，预测性能更好，这对发酵过程的生物学参数在线监测、控制、优化有重要的理论和应用价值。

参考文献：

[1] ZHU X, REHMAN K U, WANG B, et al. Modern Soft－Sensing Modeling Methods for Fermentation Processes [J]. Sensors, 2020, 20 (6): 1771.

[2] FORMENTI L R, NORREGAARD A, BOLIC A, et al. Challenges in Industrial Fermentation Technology Research [J]. Biotechnology Journal, 2014, 9 (6): 727－738.

[3] TRELEA I C, TITICA M, LANDAUD S, et al. Predictive Modelling of Brewing Fermentation: from Knowledge－based to Black－box Models [J]. Mathematics and Computers in Simulation, 2001, 56 (4－5): 405－424.

[4] 张玲玲. 基于混合模型的发酵过程优化研究 [D]. 沈阳：东北大学，2014.

[5] DACH J, KOSZELA K, BONIECKI P, et al. The Use of Neural Modelling to Estimate the Methane Production from Slurry Fermentation Processes [J]. Renewable and Sustainable Energy Reviews, 2016, 56: 603－610.

[6] ZHONG W, PI D, SUN Y. SVM Based Soft Sensor for Antibiotic Fermentation Process [A]. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme－System Security and Assurance [C]. IEEE, 2003, 1: 160－165.

(下转第 155 页)