控制与决策
Control and Decision

基于改进近端策略优化算法的草酸钴合成过程优化

贾润达, 宁文彬, 何大阔, 褚菲, 王福利

引用本文:
贾润达,宁文彬,何大阔,褚菲,王福利. 基于改进近端策略优化算法的草酸钴合成过程优化[J]. 控制与决策, 2023, 38(11): 3075−3082.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.2280

基于近端强化学习的股价预测方法

Method of stock prices forecast based on proximal reinforcement learning

控制与决策. 2021, 36(4): 967−973 https://doi.org/10.13195/j.kzyjc.2019.1245

基于改进近端策略优化算法的草酸钴合成过程优化

贾润达[1,2†], 宁文彬[1], 何大阔[1,2], 褚 菲[3,4], 王福利[1,2]

(1. 东北大学 信息科学与工程学院，沈阳 110819；2. 东北大学 流程工业综合自动化国家重点实验室，沈阳 110819；3. 中国矿业大学 信息与控制工程学院，江苏 徐州 221116；4. 中国矿业大学 地下空间智能控制教育部工程研究中心，江苏 徐州 221116)

摘 要: 金属钴被广泛用于电池和金属复合材料,草酸钴合成过程是影响产品质量的关键工序. 针对草酸钴平均粒径的优化问题,提出一种基于改进的近端策略优化(PPO)算法的草酸钴合成过程优化方法. 首先,根据草酸钴合成过程的优化目标及约束条件设计相应的奖励函数,通过建立过程的马尔科夫决策模型,将优化问题纳入强化学习框架;其次,针对策略网络在训练过程中出现的梯度消失问题,提出将残差网络作为PPO算法的策略网络;最后,针对过程连续状态空间导致PPO算法陷入局部最优策略问题,利用交错模仿学习对初始策略进行改进. 将所提出的方法与传统PPO算法进行比较,改进的PPO算法在满足约束条件的同时,具有更好的优化效果和收敛性.

关键词: 强化学习；近端策略优化；草酸钴合成过程；残差网络；交错模仿学习；间歇过程

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.2280

引用格式: 贾润达,宁文彬,何大阔,等. 基于改进近端策略优化算法的草酸钴合成过程优化[J]. 控制与决策, 2023, 38(11): 3075-3082.

Optimization of cobalt oxalate synthesis process based on modified proximal policy optimization algorithm

JIA Run-da[1,2†], *NING Wen-bin*[1], *HE Da-kuo*[1,2], *CHU Fei*[3,4], *WANG Fu-li*[1,2]

(1. College of Information Science and Engineering，Northeastern University，Shenyang 110819，China；2. State Key Laboratory of Synthetical Automation for Process Industries，Northeastern University，Shenyang 110819，China；3. College of Information and Control Engineering，China University of Mining and Technology，Xuzhou 221116，China；4. Engineering Research Center of Ministry of Education for Intelligent Control of Underground Space，China University of Mining and Technology，Xuzhou 221116，China)

Abstract: Cobalt is widely used in batteries and metal composite materials, and the cobalt oxalate synthesis process is a key process that affects product quality. To optimize the average particle size of cobalt oxalate, this work presents an optimization method for the cobalt oxalate synthesis process based on the modified proximal policy optimization (MPPO) algorithm. First, the reward function is designed according to the optimization objectives and constraints of the cobalt oxalate synthesis process, and the optimization problem is incorporated into the reinforcement learning framework by establishing the Markov decision model of the process. Secondly, to deal with the problem of gradient disappearance in the training process of the policy network, the residual network is proposed as the policy network of the PPO algorithm. Finally, to avoid the PPO algorithm falling into the local optimal strategy due to the continuous state space of the process, the initial strategy is improved by interleaved imitation learning. The proposed method is compared with the traditional PPO algorithm, and the modified PPO algorithm has a better optimization effect and convergence while satisfying the constraints.

Keywords: reinforcement learning；proximal policy optimization；cobalt oxalate synthesis process；residual network；interleaved imitation learning；batch processes

0 引 言

金属钴被广泛用于碱性可充电电池和具有较强耐热性能和防腐性能的金属复合材料当中. 在钴湿法冶金工业流程中,通过煅烧草酸钴得到氧化钴,然后利用氢气将氧化钴还原为金属钴. 因此,草酸钴合成过程的质量在很大程度上影响着金属钴的品

收稿日期: 2021-12-30；**录用日期:** 2022-05-31.

基金项目: 国家自然科学基金项目(61873049, 61733003, 62173078, 61973304).

†通讯作者. E-mail: jiarunda@ise.neu.edu.cn.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

质[1]. 草酸钴合成过程是一种小批量的间歇过程,草酸钴晶体的品质通常通过其平均粒径的大小进行评价,例如:较大的平均粒径可以有效缩短过滤和洗涤的时间. 因此,草酸钴合成过程的优化目标可以是最大化平均粒径,并同时尽量减小批次间平均粒径的变化幅度[2].

传统方法在对草酸钴合成过程进行优化时,一般需要基于过程数据建立统计学模型. 常玉清等[3]结合过程数据建立了最小二乘支持向量机与草酸钴机理模型相结合的混合模型;黄碧璇等[4]建立了多向偏最小二乘回归模型;Chu等[5]提出了基于迁移学习的联合输出偏最小二乘数据建模方法. 草酸钴合成过程具有复杂的非线性动力学特性,仅依赖历史数据和机理模型建立一个准确的过程模型是十分困难的,因此基于模型的优化方法面临着模型失配的问题[6]. 此外,大多数过程优化问题存在各种不确定扰动,如:入料温度、浓度的不确定性等[7]. 因此,基于确定性模型的优化结果常常是次优的,且可能导致违反生产过程的约束条件.

为了克服上述问题,基于强化学习(reinforcement learning, RL)的优化控制方法在间歇过程中得到了广泛的应用. 在基于强化学习的方法中,智能体直接与真实环境进行交互,并基于交互数据学习最优策略[8]. Petsagkourakis等[9]提出了基于策略梯度(policy gradient)方法的间歇过程两阶段优化框架,用于解决复杂生化过程优化中的模型失配问题;Lillicrap等[10]和Yoo等[11]提出了一种改进的深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法,该算法针对过程的不同阶段设计了相应的价值函数和奖励函数,保证了智能体能够更加稳定、高效地学习最优策略;Ma等[12]针对半间歇过程的优化问题,提出了基于异步优势动作评价(asynchronous advantage actor-critic, A3C)算法的间歇过程优化算法框架,该算法利用多个智能体与环境交互产生过程数据,提高了策略网络训练的稳定性.

虽然强化学习算法能够在过程模型未知的情况下对间歇过程进行优化,但将其应用于草酸钴合成过程优化仍存在一些不足. 首先,草酸钴合成过程作为典型的结晶过程和间歇过程,其动力学特性复杂. 到目前为止,大多数基于强化学习的间歇过程的优化控制研究在拟合控制策略时都采用浅层人工神经网络(artificial neural network, ANN). 虽然3层人工神经网络已被证明能以高精度有效逼近任何一个非线性过程[13],但对于具有强非线性的草酸钴合成过程,采用深层网络才能更好地拟合控制策略[14]. 另外,上述的基于强化学习的间歇过程优化算法,在策略更新时难以选择合适的步长,导致新旧策略的变化差异过大,算法最终难以取得良好的训练效果.

针对以上问题,本文提出基于近端策略优化方法的草酸钴合成过程优化算法. 该算法在策略更新时引入KL散度(Kullback-Leibler divergence)和重要性采样方法,并使用改进的梯度裁剪方法,解决策略更新时梯度步长难以选择的问题. 同时该算法使用深层神经网络作为策略网络,可有效解决浅层ANN作为策略网络模型时,对最优策略的拟合能力不足的问题. 同时,为了解决深层网络在算法训练过程中出现的梯度消失问题,本文提出基于残差网络结构的策略网络模型.

综上所述,本文首先将草酸钴合成过程建模为马尔可夫决策过程,然后针对过程特性设计有效的奖励函数,最后使用改进的近端策略优化算法对其进行优化. 针对草酸钴合成过程具有高维连续的状态和动作空间,导致策略在训练初期容易收敛至局部最优的问题,本文提出交错模仿学习的预训练方法,使智能体与环境交互初期就拥有一个较优的策略. 通过将改进的近端策略优化(modified proximal policy optimization, MPPO)算法应用于草酸钴合成过程仿真平台,验证算法的有效性,同时进一步提高草酸钴合成过程的产品质量.

1 过程描述

草酸钴的生产过程是典型的批次结晶过程,因此其反应过程较为复杂[4]. 而合成草酸钴最关键的一步是草酸铵与氯化钴溶液在结晶器中进行化学反应,其反应的核心化学方程式如下:

$$CoCl_2 + (NH_4)_2C_2O_4 \rightarrow CoC_2O_4 \downarrow + 2NH_4Cl. \quad (1)$$

草酸钴的合成过程如图1所示. 首先,将草酸铵溶液存储在溶解器中;然后,将固定体积的氯化钴溶液加入到结晶器中,并将其加热至适宜温度;最后,将草酸铵溶液逐步加入至配备连续搅拌功能的结晶器中与氯化钴进行结晶反应. 为了保证结晶器内温度恒定,一般使用带有PI控制器的加热夹套进行温度控制. 在生产过程中,一般将结晶器中的温度和搅拌速率设为恒定值,将草酸铵的进料速率设置为唯一的操作变量,草酸钴晶体的平均粒径作为生产指标[5].

采用常微分方程组描述草酸钴合成过程的动态模型如下:

$$\frac{\mathrm{d}V}{\mathrm{d}t} = F_B; \quad (2)$$

$$\frac{\mathrm{d}\mu_0}{\mathrm{d}t} = B - \frac{F_B\mu_0}{V}; \tag{3}$$

$$\frac{\mathrm{d}\mu_j}{\mathrm{d}t} = jG\mu_{j-1} - \frac{F_B\mu_j}{V},\ j = 1,2,3; \tag{4}$$

$$\frac{\mathrm{d}C}{\mathrm{d}t} = \frac{F_BC_BV_{A0}}{(V_{A0}+F_Bt)^2} - 3\rho_ck_vG\mu_2 - \frac{F_BC}{V}. \tag{5}$$

其中: V 表示悬浮液体积, F_B 表示草酸铵进料速率, μ_j 表示草酸钴粒度分布函数的 j 阶矩, B 表示结晶成核速率, G 表示生长速率, C 表示溶液浓度, C_B 表示草酸铵溶液浓度, V_{A0} 表示氯化钴的初始体积, ρ_c 表示晶体密度, k_v 表示体积形状因子.

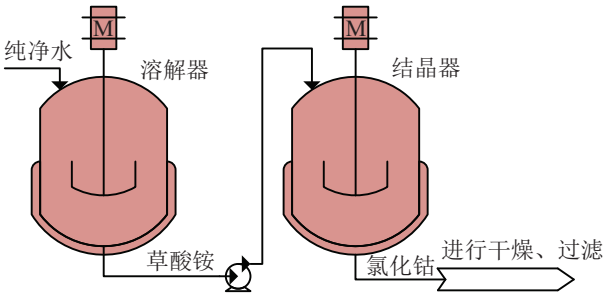

图 1　草酸钴生产过程

平均结晶尺寸是草酸钴合成过程中最重要的生产指标, 因此本文的优化目标是最大化草酸钴晶体的平均粒径 $\mathrm{Ln} = \dfrac{\mu_1(t_{\mathrm{end}})}{\mu_0(t_{\mathrm{end}})}$. 根据过程模型与优化目标建立如下动态优化问题:

$$\begin{aligned} &\max_{\pi(\cdot)} \mathrm{Ln};\\ &\text{s.t. } C_{\max} \geqslant \int_0^{t_{\mathrm{end}}} F_B(t)C\mathrm{d}t \geqslant C_{\min},\\ &\qquad F_{\min} \leqslant F_B(t) \leqslant F_{\max}. \end{aligned} \tag{6}$$

其中: $C_{\max}$ 和 $C_{\min}$ 表示结晶器内草酸铵添加量的上限和下限, $F_{\min}$ 和 $F_{\max}$ 表示草酸铵进料速率的极大和极小值.

2　改进的近端策略优化算法

本节将提出一种 MPPO 算法. 草酸钴合成过程属于结晶过程和间歇过程, 具有非线性系统动力学特性, 并且实际生产过程中存在各种不确定性扰动, 使得其控制策略更加难以优化. 传统的 PPO 算法通过对新旧策略的比值加以限制, 改善了策略参数更新过程中步长难以选择的问题. 但由于采用浅层神经网络作为策略网络拟合最优策略, 该算法存在着学习能力不足的缺点. 因此本文将使用深层神经网络作为策略网络, 用以提高智能体学习最优策略的能力, 并针对深层网络训练过程中出现的梯度消失问题, 提出将残差网络结构与传统 PPO 算法相结合, 在提升智能体学习能力的基础上进一步提高强化学习算法的收敛性能. 为解决强化学习算法前期难以训练的问题, 本节还提出交错模仿学习的预训练策略, 使智能体在与环境交互初期就具有良好的初始策略.

2.1　残差网络架构的策略网络

2.1.1　残差网络结构

草酸钴合成过程是复杂的化工过程, 其非线性特征显著, 且反应过程容易受到外部环境的干扰, 因此使用常见的 3 层全连接神经网络很难学习到最优的控制策略. 随着策略网络深度的增加, 网络对于各类复杂模型的拟合能力大大加强, 因此, 在草酸钴合成过程优化中可以采用深层网络拟合控制策略. 但深层神经网络也存在一些问题, 神经网络的优化一般采用梯度反向传播原理, 随着网络层数的增加, 网络出现了梯度消失的问题. 因此仅仅增加网络的层数, 甚至可能导致网络性能下降[15].

考虑如图 2(a) 所示的网络结构, 在忽略激活函数的情况下, 该网络的输入为 x_l, 输出为 x_{l+1}, 其网络的映射关系为 $Y(x_l)$. 当对该网络进行优化时, 需要将隐藏层的权值优化至最优. 针对该优化问题, 本文考虑另外一种网络结构 (如图 2(b) 所示), 网络的输入为 x_l, 输出为 x_{l+1}, 该网络的映射关系为 $W(x_l) + x_l$. 相比于图 2(a), 网络的映射关系 $Y(x_l)$ 被重构为 $W(x_l) + x_l$. 在极端情况下, 如果两个网络的最优映射关系为恒等映射 $Y(x_l) = x_l$, 则相比于将原始映射 $Y(x_l)$ 拟合为恒等映射, 将残差映射 $W(x_l)$ 拟合至 0 要更加容易实现.

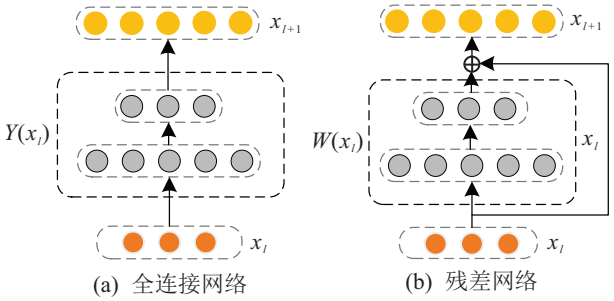

图 2　不同神经网络结构

在实际情况中, 恒等映射不一定是最优映射关系, 但残差网络的构造表明, 当网络需要拟合的最优映射关系更接近恒等映射时, 残差网络更容易被优化至最优映射关系.

$W(x_l) + x_l$ 的映射关系可以通过具有跳跃连接 (skip connection) 形式的前馈神经网络实现, 这种实现方式可以将网络的输入直接叠加到网络深层的输出中, 因此不增加额外的参数, 也不增加计算复杂度. 在该网络结构中, 传统的反向梯度传播方法仍然适用, 且不需要对求解器进行额外的修改.

2.1.2 残差网络的反向梯度传播

通过叠加上节所述的残差块网络结构,可以得到残差网络架构的深层神经网络,其信息传递表示为如下形式:

$$y_l = x_l + W(x_l, \omega_l), \tag{7}$$

$$x_{l+1} = h(y_l). \tag{8}$$

其中:x_l表示第l层残差块的输入特征;w_l表示第l层残差块的权值;W表示残差块映射;函数y_l表示残差块映射和残差块输入特征的相加操作;h表示相加后的操作,在式(8)中h为ReLU激活函数.

假设从y_l到x_{l+1}为恒等映射,推导可得

$$x_L = x_l + \sum_{i=l}^{L-1} W(x_i, w_i). \tag{9}$$

其中:x_l表示经过多个残差块网络映射后的输出特征,从该形式可知,堆叠了任意多个残差块网络的深层网络的输出,可以通过将浅层输入特征x_l和若干个残差块映射相加得到[16].

就神经网络反向传播的角度而言,具有残差块结构的深层神经网络具有非常好的传播特性,由反向传播的链式法则可得

$$\frac{\partial \xi}{\partial x_l} = \frac{\partial \xi}{\partial x_L}\frac{\partial x_L}{\partial x_l} = \frac{\partial \xi}{\partial x_L}\Big(1 + \frac{\partial}{\partial x_l}\sum_{i=l}^{L-1} W(x_i, w_i)\Big), \tag{10}$$

其中ξ为损失函数. 式(10)表明反向传播的梯度被分解为两项,其中一项能够直接传播梯度,且不涉及任何隐藏层,因此即使隐藏层部分的权值是任意小的,传递至网络浅层的梯度也不会消失,这保证了反向传播的梯度能够直接优化网络浅层部分的权值. 因此具有残差网络结构的深层神经网络能够有效避免梯度消失的问题,且相比于一般的全连接结构的深层网络更易训练.

为了使策略网络能够更好地拟合最优控制策略,在基于近端策略优化算法的草酸钴合成过程的优化控制中,本文将采用残差网络结构的深层神经网络作为策略网络.

2.2 交错模仿学习

草酸钴合成过程具有高维连续的状态和动作空间,在基于强化学习算法的优化过程中,如果单纯依靠智能体在环境中搜索,则找到多步决策的最优控制策略是十分困难的. 而利用专家提供的轨迹对智能体进行预训练,是解决上述问题的常规方法,该方法被称为直接模仿学习[17]. 但直接模仿学习方法仍存在不足,由于专家提供的都是较优轨迹,在预训练过程中,智能体对环境的探索并不充分,得到的是对于局部最优策略的有偏估计. 在智能体与真实环境交互过程中首次遇到某种状态时,容易给出非常差的动作,导致策略收敛至较差的策略.

为了解决直接模仿学习在预训练阶段的有偏估计问题,本文将智能体训练的模仿学习阶段与环境交互阶段融合为一个阶段,即进行交错式的模仿学习. 本文通过设置n个初始批量的专家轨迹和衰减率η,控制模仿学习和智能体探索真实环境的训练比例,避免直接模仿学习的缺点.

3 基于MPPO算法的草酸钴合成过程优化

3.1 强化学习问题构建

在本研究中,假设草酸钴合成过程的系统动力学特性符合一个未知的概率分布,考虑到该反应过程存在一定的外部扰动,系统描述为

$$x_{k+1} = f(x_k, u_k) + e_k. \tag{11}$$

其中:k表示离散时间;x_k表示状态向量;u_k表示输入向量;e_k表示反应过程的外部扰动向量,假设该扰动符合高斯分布$e_k \sim N(0, \sigma^2)$;$f(\cdot)$表示系统的非线性动力学特性.

为了将草酸钴合成过程纳入强化学习框架进行优化控制,首先将其建模为马尔可夫决策过程. 该过程的几个核心要素如下:状态(states)、动作(actions)、奖励(reward)和策略(policy).

将马尔可夫决策过程的状态和动作定义为草酸钴合成过程中的状态变量和操作变量,具体表示为

$$S = [C, V, \mu_0, \mu_1, \mu_2, \mu_3],\ A = [F_B]. \tag{12}$$

其中:S和A分别表示马尔可夫决策过程的状态和动作,其余各变量与第1节中对应变量意义相同. 奖励函数的具体形式将在下节给出,策略由本文提出的残差网络架构的深层神经网络进行建模.

3.2 奖励函数设置

在训练过程中的每一个离散时刻,环境会向智能体反馈即时奖励,奖励函数的设置在很大程度上影响智能体的学习. 在本研究中,最终目标是最大化草酸钴合成过程的终端产品质量,即草酸钴晶体的平均粒径,为了最有效地激励智能体寻找最优动作,以实现平均粒径的最大化,本文将奖励函数设置为3个部分的累加,即

$$r_t = r_{t1} + r_{t2} + r_{t3}. \tag{13}$$

首先,由于草酸钴合成过程的评价指标为最终时刻草酸钴的平均粒径,该部分奖励函数设置为如下稀疏形式:

$$r_{t1}=\begin{cases}\alpha_1\cdot Y_{\rm Ln},\ t=t_{\rm end}\,;\\ -0.1,\ t<t_{\rm end}\,.\end{cases}\tag{14}$$

其中: α_1 表示奖励系数, $Y_{\rm Ln}$ 表示一个批次结束时草酸钴的平均粒径.

为了同时满足优化问题的多个约束条件, 本文设置了两种针对不同约束条件的奖励函数, 其中为了满足下界约束条件设计的奖励函数如下:

$$r_{t2}=\begin{cases}\alpha_2\Big(\displaystyle\sum_{T_i=0}^{T_i=t_{\rm end}}C-C_{\min}\Big),\ t=t_{\rm end}\,;\\ 0,\ t<t_{\rm end}\,.\end{cases}\tag{15}$$

其中 α_2 表示奖励系数. 由于该优化问题的约束为操作变量在整个批次的累加量, 只能在每个批次结束时才能判断是否违反了下界约束. 在一个批次的终端时刻, 如果操作变量累计值小于约束的下界, 则环境会给智能体施加一个惩罚, 累计值偏离下界越大, 惩罚越大. 当草酸铵累加量在下限值以上时, 该奖励为正值. 由于 MPPO 算法中优势函数的计算与该批次中每一个时刻的奖励有关, 终端惩罚项会影响到全局策略的更新.

另外, 为满足操作变量累加值的上界约束, 设计的奖励函数形式如下:

$$r_{t3}=\begin{cases}\alpha_3\Big(C_{\max}-\displaystyle\sum_{T_i=0}^{T_i=t}C\Big),\ \displaystyle\sum_{T_i=0}^{T_i=t}C>C_{\max};\\ 0,\ \displaystyle\sum_{T_i=0}^{T_i=t}C\leqslant C_{\max}.\end{cases}\tag{16}$$

其中: α_3 表示奖励系数, 在每一批次的草酸钴合成过程当中, 截止到当前时刻的草酸铵累加量将会被记录下来, 并判断该值是否大于上界值. 该奖励函数的作用是, 累加值超出上界值越大, 惩罚越大.

以上 3 种奖励函数的设置保证了强化学习算法能够对智能体进行有效训练, 并保证了在与环境交互过程中, 能够满足生产过程的约束条件, 同时保障了草酸钴合成过程控制优化的经济性能.

3.3　基于 MPPO 算法的草酸钴合成过程优化

本文提出的基于残差网络架构的 PPO 算法, 是一种基于策略梯度方法的强化学习算法. 在该算法中, 为了更好地拟合控制策略, 并提高深层网络的学习能力, 策略采用具有残差网络结构的深层神经网络进行建模, 策略 π 根据观测信息 S_t 选择一个动作 A_t, 然后利用环境给出的奖励 R_t 直接对被选择的动作的概率进行修正, 奖励值高的动作下一次被选中的概率会增加, 奖励值为负值的动作下次被选中的概率会减小. 在策略梯度方法中, 策略参数一般为神经网络的权值 θ, 通过对策略参数进行梯度上升, 进而对当前策略进行更新.

$$\theta\leftarrow\theta+\alpha\nabla_\theta J(\theta),\tag{17}$$

其中 α 表示学习率. 使用策略梯度定理, 策略梯度 $\nabla_\theta J(\theta)$ 有几种不同的表示方法, 其中最常用的梯度估计量形式[18] 如下:

$$\hat{g}=E_t[\nabla_\theta\log\pi_\theta(a_t\mid s_t)N_t].\tag{18}$$

其中: π_θ 表示当前策略; N_t 表示 t 时刻的优势函数, 即策略更新后相比于旧策略的优势.

在基于策略梯度的强化学习算法中, 策略更新步长的选取是一个至关重要的问题, 若选取步长太长, 则策略很容易发散, 若步长太短, 则策略收敛速度很慢. 为了解决在策略更新时的步长选取问题, PPO 算法在策略更新时引入了一个新旧策略的比值约束

$$J^{\rm clip}(\theta)=E_t,$$
$$[\min(r_t(\theta)N_t,{\rm clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)N_t)].\tag{19}$$

其中: $r_t(\theta)$ 表示 t 时刻的策略比值, ε 表示裁剪比. 策略比 $r_t(\theta)$ 如下所示:

$$r_t(\theta)=\frac{\pi_\theta(a_t\,|\,s_t)}{\pi_{\theta{\rm old}}(a_t\,|\,S_t)}.\tag{20}$$

为了减小计算优势函数的估计时的方差, PPO 算法采用了 Actor-Critic(A-C) 结构, 优势函数的计算[19] 如下:

$$\begin{aligned}N_t=&-V(s_t)+r_t+\gamma r_{t+1}+\ldots+\\&\gamma^{T-t+1}r_{T+1}+\gamma^{T-t}V(s_T),\end{aligned}\tag{21}$$

其中 $V(s_t)$ 表示状态 s_t 的状态值. Critic 网络的目标函数[20] 为

$$J(\omega)=E[N_t^2].\tag{22}$$

综上, 将改进后的 PPO 算法应用于草酸钴合成过程, 其流程示意如图 3 所示.

4　实验验证

本节将对基于改进近端策略优化的草酸钴合成过程优化算法进行仿真验证.

4.1　训练结果

本节将提出的 MPPO 算法用于草酸钴合成过程的优化控制. 基于残差结构的策略网络包含输入层、4 层隐藏层以及输出层, 其中网络的第 2 层 ~ 第 5 层使用跳跃连接的方式构造残差网络结构; 值函数网络为全连接结构的人工神经网络, 包含输入层、2 层隐藏层及输出层. 为了模拟实际过程中观测状态受到的测量误差的干扰, 在每个采样时刻为观测状态加入当前状态值的 ±5 % 的高斯噪声.

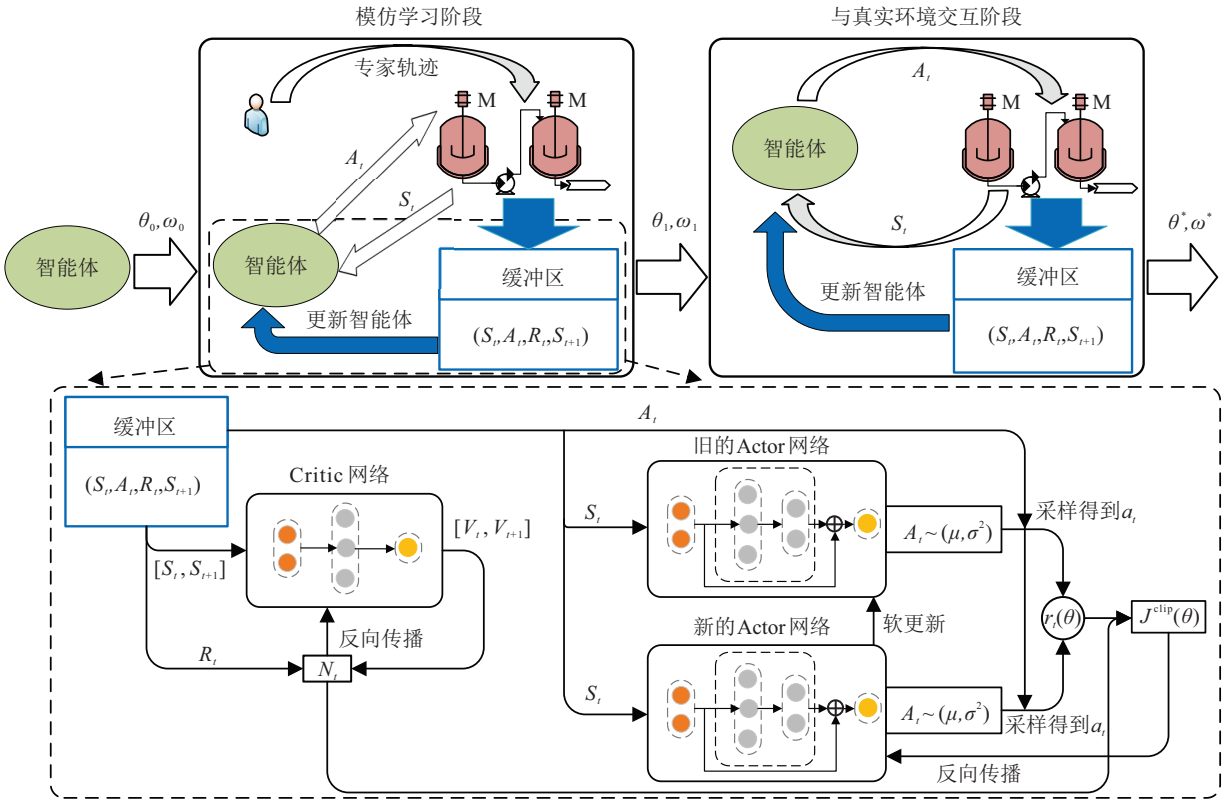

图 3 基于改进近端策略优化算法的草酸钴过程优化

图4为采用残差网络架构的PPO算法与环境交互的奖励曲线,图4中:横轴表示训练轮次,其中每轮次包含8个批次的交互过程;纵轴表示每一轮次中所有批次奖励的平均值. 可以看出,由于对智能体进行了预训练,在与环境交互时取得了较好的初始值. 在交互至200轮次左右时,出现了奖励曲线迅速上升的趋势,推测是由于智能体与环境交互一定次数后,策略网络从局部最优处向全局最优处收敛. 在智能体训练过程中,将50个轮次内的平均奖励停止上升作为智能体训练终止的判断条件,防止其达到过拟合状态. 随着训练轮次的增加,每一轮次的平均奖励逐渐上升,最终收敛到最优值.

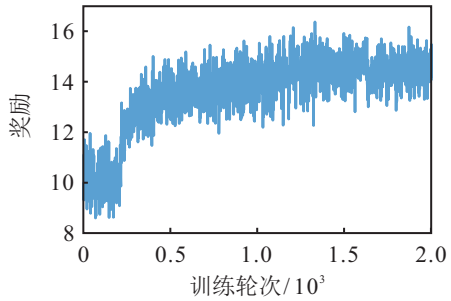

图 4 智能体与环境交互的奖励

图5为训练至不同轮次时的动作序列曲线,随着训练的进行,动作序列逐渐向最优动作序列收敛. 由于使用了改进的奖励函数,每个时刻的草酸铵流量都在约束范围以内,且总体的草酸铵累加量也满足约束上限和下限,保证了生产过程的平稳进行. 对智能体的训练完成后,对最优策略网络进行测试,得到图6所示的草酸钴平均粒径的变化曲线.

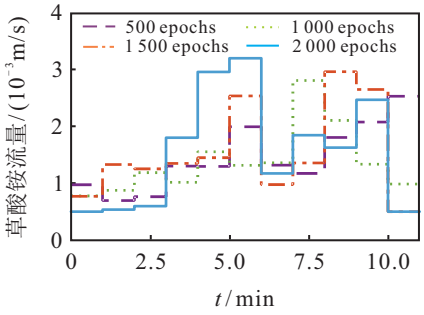

图 5 动作序列

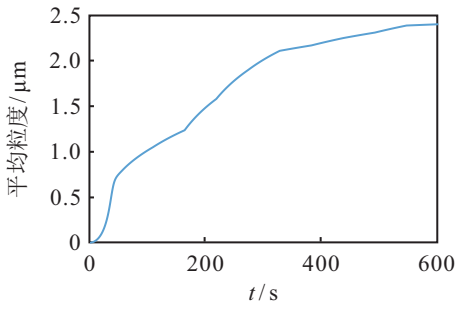

图 6 平均粒径变化

由图6可以看到,在最优动作序列下草酸铵与氯化钴经过11 min的反应,最终得到的草酸钴晶体的平均粒径达到2.40 μm. 最优动作序列如图5所示,在反应过程中,每隔1 min改变一次草酸铵的流量. 输

入的草酸铵流量的上限为 $4\times10^{-3}/(\mathrm{m}^3/\mathrm{s})$, 下限为 $5\times10^{-4}/(\mathrm{m}^3/\mathrm{s})$, 可以看出, 动作在任意时刻均满足约束. 由于对奖励函数的改进, 草酸铵流量的变化总体上比较平稳, 便于实际生产中进行操作. 草酸铵的累积量在约束范围以内, 在最优动作序列下, 草酸铵消耗累积量为 1 745 mol.

4.2　对比实验

为了验证残差网络结构对强化学习算法性能的影响, 本节将对比使用残差网络结构和未使用的深层网络对草酸钴过程进行优化的结果, 其中每个训练轮次包含 40 个批次的交互过程. 仿真结果如图 7、图 8 所示.

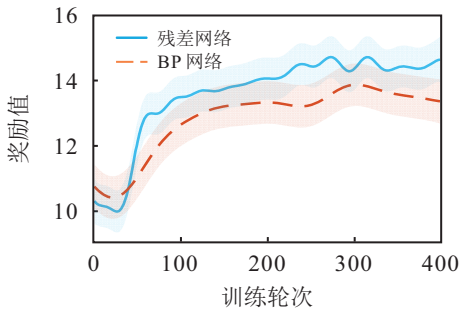

图 7　不同结构下 6 层深度网络的奖励

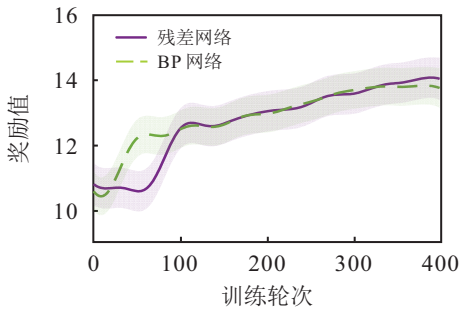

图 8　不同结构下 4 层深度网络的奖励

从图 7 和图 8 可以看出, 不同网络结构和网络深度下的 PPO 算法的奖励曲线都在逐渐上升, 且逐渐达到收敛. 这是因为传统的 PPO 算法也使用了改进的策略梯度裁剪方法, 虽然对最优策略拟合能力不足, 但在策略更新过程中, 新旧策略的差距不会过大, 也可以保证策略网络的性能逐步上升.

首先, 就 6 层网络的奖励曲线对比而言, 采用残差网络结构的策略网络的奖励曲线明显高于采用全连接 BP 网络的策略网络的奖励曲线, 且具有较快的收敛速度和较大的奖励值. 这是由于全连接结构的策略网络, 梯度在反向传播时可能会由于隐藏层权值过小而导致传播至浅层的梯度几乎为 0, 使得策略网络无法更好地学习. 而残差网络结构的策略网络在网络浅层和网络深层之间有一条快捷通道, 梯度将从残差块网络和快捷通道两部分向浅层传递, 使得梯度必然传递至浅层, 进而使策略网络达到更好的训练效果. 其次, 通过对较浅的 4 层策略网络的性能进行对比发现, 使用残差网络结构的策略网络和全连接结构的策略网络的收敛性能并无明显差别. 这是由于对于较浅的网络而言, 梯度在反向传播时较为容易传导至浅层, 因此在加入快捷通道后, 对于性能的改善并不明显. 通过观察发现, 尽管残差网络结构的策略网络在参数数量上相比于全连接网络并没有增加, 但从局部最优策略向全局最优策略更新时, 速度要慢于全连接结构的浅层策略网络.

4 种网络的特征记录如表 1 所示.

表 1　不同网络结构性能对比

网络层数	网络结构	每轮次最高奖励	训练轮次
6	残差网络	15.21	400
6	BP 网络	14.50	400
4	残差网络	14.60	400
4	BP 网络	14.43	400

由表 1 可知, 在深层网络中使用残差网络结构能够明显提升策略网络的性能. 对比未使用残差网络结构的浅层网络和深层网络, 由于深层网络的梯度消失和退化现象, 尽管网络的层数加深了, 但并不能提升网络的性能.

综上所述, 将本文提出的 MPPO 算法应用于草酸钴合成过程的控制优化中, 取得了较好的控制效果. 首先, 在训练初始阶段, 使用交错模仿学习的预训练策略, 更有利于实现智能体在学习与探索之间的平衡. 对不同网络结构的强化学习算法进行了实验对比, 本文提出的基于残差网络结构的 PPO 算法相比于传统的 PPO 算法, 能够更好地发挥该算法在复杂非线性系统中的作用, 得到更加优良的优化效果.

5　结　论

本文针对草酸钴合成过程具有复杂系统动力学特性, 难以建立其精确的系统模型, 导致难以对该过程进行有效优化的问题, 提出了一种基于改进的近端策略优化的草酸钴合成过程优化算法. 该方法首先将草酸钴合成过程建模为马尔可夫决策过程, 确立强化学习问题中的各部分关键变量; 其次, 为了提高策略网络对最优策略的拟合能力, 并解决深层网络训练过程中的梯度消失等问题, 提出了基于残差网络架构的 PPO 算法; 为了解决使用专家轨迹进行预训练时导致控制策略陷入局部最优的问题, 本文还提出了交错模仿学习的预训练方法; 最后, 在仿真平台上进行了实验验证, 与传统的 PPO 算法相比, 所提出的 MPPO 算法不仅具有更好的收敛性能, 而且学习到的最优策略具有更好的优化效果.

虽然使用MPPO能够对草酸钴合成过程进行优化控制,但目前仍存在一些问题. 强化学习的安全性问题一直以来受到广泛的关注,在策略训练过程中,如果智能体给出的动作导致状态违反了约束,则环境一般会反馈给智能体一个较大的惩罚,从而激励智能体在相同状态下不要给出导致约束违反的行为. 但这种形式并不能保证足够的安全性,未来的研究将集中在如何保证强化学习方法能够始终满足过程约束条件.

参考文献(References)

[1] Jia R D, Mao Z Z, Wang F L, et al. Batch-to-batch optimization of cobalt oxalate synthesis process using modifier-adaptation strategy with latent variable model[J]. Chemometrics and Intelligent Laboratory Systems, 2015, 140: 73-85.

[2] Jia R D, Zhang S L, You F Q. Transfer learning for end-product quality prediction of batch processes using domain-adaption joint-Y PLS[J]. Computers & Chemical Engineering, 2020, 140: 106943.

[3] 常玉清, 梁倩, 王姝, 等. 湿法冶金草酸钴粒度分布建模与优化研究[J]. 东北大学学报: 自然科学版, 2012, 33(11): 1533-1537.
(Chang Y Q, Liang Q, Wang S, et al. Modeling and optimization of particle size distribution of cobalt oxalate in hydrometallurgy[J]. Journal of Northeastern University: Natural Science, 2012, 33(11): 1533-1537.)

[4] 黄碧璇, 毛志忠, 贾润达. 草酸钴合成过程批次间自适应优化[J]. 控制理论与应用, 2016, 33(2): 189-195.
(Huang B X, Mao Z Z, Jia R D. A batch-to-batch adaptive optimization for the cobalt oxalate synthesis process[J]. Control Theory & Applications, 2016, 33(2): 189-195.)

[5] 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.

[6] 栾小丽, 刘晓凤, 刘飞. 基于互信息操作变量曲线参数化的间歇过程批内修正优化[J]. 控制与决策, 2021, 36(1): 234-240.
(Luan X L, Liu X F, Liu F. Intra-batch correction optimization of batch process with manipulated variable trajectory parameterization based on mutual information[J]. Control and Decision, 2021, 36(1): 234-240.)

[7] Yoo H, Byun H E, Han D, et al. Reinforcement learning for batch process control: Review and perspectives[J]. Annual Reviews in Control, 2021, 52: 108-119.

[8] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. Cambridge: MIT Press, 2018: 47-54.

[9] Petsagkourakis P, Sandoval I O, Bradford E, et al. Reinforcement learning for batch bioprocess optimization[J]. Computers & Chemical Engineering, 2020, 133: 106649.

[10] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J/OL]. 2015, arXiv: 1509. 02971.

[11] Yoo H, Kim B, Kim J W, et al. Reinforcement learning based optimal control of batch processes using Monte-Carlo deep deterministic policy gradient with phase segmentation[J]. Computers & Chemical Engineering, 2021, 144: 107133.

[12] Ma Y, Noreña-Caro D A, Adams A J, et al. Machine-learning-based simulation and fed-batch control of cyanobacterial-phycocyanin production in Plectonema by artificial neural network and deep reinforcement learning[J]. Computers & Chemical Engineering, 2020, 142: 107016.

[13] 焦焕炎, 冯浩东, 魏东, 等. 基于强化学习的地铁站空调系统节能控制[J]. 控制与决策, 2022, 37(12): 3139-3148.
(Jiao H Y, Feng H D, Wei D, et al. Energy saving control for subway station air conditioning systems based on reinforcement learning[J]. Control and Decision, 2022, 37(12): 3139-3148.)

[14] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[15] He K M, Sun J. Convolutional neural networks at constrained time cost[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 5353-5360.

[16] He K M, Zhang X Y, Ren S Q, et al. Identity mappings in deep residual networks[C]. European Conference on Computer Vision. Cham, 2016: 630-645.

[17] 王思鹏, 杜昌平, 郑耀. 基于强化学习的扑翼飞行器路径规划算法[J]. 控制与决策, 2022, 37(4): 851-860.
(Wang S P, Du C P, Zheng Y. Local planner for flapping wing micro aerial vehicle based on deep reinforcement learning[J]. Control and Decision, 2022, 37(4): 851-860.)

[18] Byun H E, Kim B, Lee J H. Robust dual control of batch processes with parametric uncertainty using proximal policy optimization[C]. The 59th IEEE Conference on Decision and Control. Jeju, 2021: 3016-3021.

[19] Schulman J, Moritz P, Levine S, et al. High-dimensional continuous control using generalized advantage estimation[J/OL]. 2015, arXiv: 1506. 02438.

[20] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J/OL]. 2017, arXiv: 1707. 06347.

作者简介

贾润达(1981−), 男, 副教授, 博士生导师, 从事机器学习、工业人工智能等研究, E-mail: jiarunda@ise.neu.edu.cn;

宁文彬(1998−), 男, 硕士生, 从事间歇过程优化控制、强化学习等研究, E-mail: nwb0718@163.com;

何大阔(1975−), 男, 教授, 博士生导师, 从事复杂工业生产全流程智能建模与优化控制等研究, E-mail: hedakuo@ise.neu.edu.cn;

褚菲(1984−), 男, 副教授, 博士生导师, 从事流程工业过程智能建模、控制与优化等研究, E-mail: chufeizhufei@sina.com;

王福利(1957−), 男, 教授, 博士生导师, 从事复杂工业过程的建模、控制与优化、过程检测与故障诊断等研究, E-mail: flwang@mail.neu.edu.cn.