自动化学报
Acta Automatica Sinica
ISSN 0254-4156,CN 11-2109/TP

《自动化学报》网络首发论文

题目：基于深度强化学习的制造过程 Run-to-Run 控制
作者：郭鹏，余建波
DOI：10.16383/j.aas.c190546
收稿日期：2019-07-23
网络首发日期：2020-03-10
引用格式：郭鹏，余建波．基于深度强化学习的制造过程 Run-to-Run 控制[J/OL]．自动化学报. https://doi.org/10.16383/j.aas.c190546

自 动 化 学 报
ACTA AUTOMATICA SINICA

基于深度强化学习的制造过程 Run-to-Run 控制

郭鹏 [1] 余建波*[1]

摘要 针对化学机械研磨(Chemical mechanical planarization, CMP)过程中材料去除率逐渐下降、目标去除率变化等异常现象，本文提出了一种基于深度强化学习的 CMP 过程 Run-to-Run 控制模型，来进行 CMP 过程的在线学习优化与自适应控制. 首先，提出了基于无模型的深度强化学习的 Run-to-Run 控制模型，避免了传统方法中存在的系统识别和精确建模的困难. 其次，对深度强化学习的策略网络结构进行了改进，通过显式地将控制策略分为线性部分与非线性部分，提高了深度强化学习在 Run-to-Run 控制中的控制效果与鲁棒性. 最后通过仿真实验表明，本文提出的控制模型能够通过控制 CMP 过程参数来自适应补偿 CMP 过程的异常现象，有效降低批次间的产品质量波动.

关键词 化学机械研磨，过程控制，批次控制，深度强化学习，深度学习
引用格式 郭鹏，余建波. 基于深度强化学习的制造过程 Run-to-Run 控制. 自动化学报
DOI 10.16383/j.aas.c190546

Run-to-Run Control of Manufacturing Process based on Deep Reinforcement Learning

Guo Peng[1] Yu Jian-Bo*[1]

Abstract Aiming at dealing with the anomalies in chemical mechanical planarization (CMP), a Run-to-Run control model of CMP process based on deep reinforcement learning is proposed to optimize and adaptively control CMP process. Firstly, a Run-to-Run control model based on model-free deep reinforcement learning is proposed, which avoids the difficulties of system identification and accurate modeling in traditional methods. Secondly, the strategy network structure of deep reinforcement learning is improved. By explicitly dividing the control strategy into linear part and non-linear part, the control result and robustness of deep reinforcement learning in Run-to-Run control are improved. Finally, the simulation results show that the proposed control model can compensate the anomalies of CMP process by controlling the parameters of CMP process, and effectively reduce the fluctuation of product quality between batches.

Key words chemical mechanical planarization, process control, Run-to-Run control, deep reinforcement learning, deep learning
Citation Guo Peng, Yu Jian-Bo. Run-to-Run Control based on Manufacturing Process based on Deep Reinforcement Learning. *Acta Automatica Sinica*

化学机械研磨，或称为化学机械平坦化(Chemical mechanical planarization, CMP)是半导体制造中重要的过程之一[1]. CMP 的目标是将圆晶上的介电层与金属层磨平，使其全局平坦化，进而达到立体布线或多层布线、提升配线密度同时降低缺陷密度，减小批次间产品质量差异和减少再工次数的目的[2]. 由于缺少在线测量手段，无法对 CMP 过程进行实时控制，因此 CMP 过程往往采用批间控制的方式来进行加工制造[3]. 在 CMP 过程控制领域存在以下挑战：(1)由于缺乏实时传感器(in-situ sensor)，难以对 CMP 过程进行实时控制；(2)由于设备老化或环境因素的影响，产品质量不可避免地存在渐进漂移(Drift)和突变漂移(Shift)现象；(3)半导体生产的多品种、小批量特性使得 CMP

收稿日期 2019-07-23 录用日期 2020-02-07
Manuscript received July 23, 2019; accepted February 7, 2020
国家自然科学基金(71777173)，中央高校基本业务经费项目资助
Supported by National Natural Science Foundation of China (71777173), Fundamental Research Funds for the Central Universities
1. 同济大学机械与能源工程学院 上海 201804
1. School of Mechanical Engineering, Tongji University, Shanghai, 201804

过程需要快速响应制造目标的改变；(4)CMP 过程具有高度非线性.

在半导体生产过程中实施先进控制往往受到以上提及的几种挑战的制约. Run-to-Run(批次控制，R2R)控制正是为了克服这些制约提出的. R2R 控制是优化控制和实时控制的统一，它结合了统计过程控制和反馈控制，通过对前面批次制程中产品特性的检测数据进行统计分析，自动地改进生产过程的制程方案，达到优化控制并提高产品质量的目的. R2R 控制是 CMP 过程控制的主要方法，Run 可以是单个圆晶或一批圆晶. 根据控制算法的不同, CMP 过程的 R2R 控制器通常可以分为指数加权移动平均(EWMA)类、模型预测控制(model predictive control, MPC)类和智能控制类 3 种[3]。

多数批次过程是具有较明显的非线性的. 为了降低在线计算的复杂度，基于迭代学习控制的批次过程算法大多用一个线性时变模型来近似非线性系统[4]. 文献[5-6]提出了基于线性模型的 EWMA 和 d-EWMA 控制器，用于补偿 CMP 过程的光滑漂移和扰动，但是无法处理过程突变和严重的漂移. 文献[7]提出的 PCC(predictor corrector control)是 EWMA 控制器的扩展. 不同于 EWMA, PCC 对噪声和过程漂移进行分别处理，提高了目标跟踪性能和优化控制效果. 文献[8]提出了基于 MPC 的 Run-to-Run 控制方法，能够提供直接的多变量控制并且方便地处理输入、输出约束. 文献[9]提出了 OAQC(optimizing adaptive quality controller)控制器，对于非线性过程具有更好的性能. 它主要包含递归最小二乘估计器和非线性约束优化器 2 部分.

随着智能控制器在工业过程控制中的应用，文献[10-12]提出了基于神经网络的 CMP 过程控制器. 文献[2]提出了基于径向基神经网络和微粒群算法的 Run-to-Run 预测控制器. 文献[13]提出了一种基于灰色模型和克隆选择免疫算法的 Run-to-Run 预测控制器.

以上基于模型(model-based)的方法往往需要对 CMP 过程进行建模，如果所建立的系统模型与实际系统模型之间存在较大偏差，那么势必造成基于模型方法的控制结果与实际系统的控制结果存在较大偏差，因此其控制效果很大程度上依赖于所建立模型的准确性. 然而 CMP 过程的非线性和时变特性使得其过程模型难以精确拟合，制约了基于模型方法的控制效果. 无模型(model-free)的方法(如强化学习)通过与环境进行交互，在线地学习控制策略，避免了系统模型的拟合，提高了制造过程控制的控制效果. 文献[14,15,18]将强化学习应用于过程控制领域，取得了比传统方法更好的控制结果与控制精度.

深度学习最早由 Hinton[19]等人提出，指基于样本数据通过一定的训练方法得到包含多个层级的深度网络结构的机器学习过程，深度学习通过组合低层特征形成更加抽象的高层表示，以发现数据的分布式特征表示[20]. 深度学习所得到的深度网络结构包含大量的神经元，每个神经元与大量其他神经元相连接，神经元间的权值在学习过程中修改并决定网络的功能，深度网络就是深层次的神经网络，即深度神经网络(deep neural networks, DNN). 深度神经网络是由多个单层非线性网络叠加而成的，可完成复杂的函数逼近，获得输入数据的分布表示.

深度强化学习[21]将深度学习的感知能力与强化学习的决策能力相结合，是一种端对端(end-to-end)的感知与控制系统，具有很强的通用性. 其学习过程可以描述为：(1)在每个时刻 agent 与环境交互得到一个高维度的观察，并利用深度学习方法来感知观察以得到抽象、具体的状态特征表示；(2)基于预期回报来评价各动作的价值函数，并通过某种策略将当前状态映射为相应的动作. (3)环境对此动作做出反应，并得到下一个观察. 通过不断循环以上过程，最终可以得到实现目标的最优策略. 通过结合深度学习与强化学习，能够将强化学习拓展到具有高维状态空间与动作空间的问题. 目前深度强化学习已经广泛应用于游戏[22-23]、机器人[24-25]、对话系统[26-27]、自动驾驶[28-29]等领域. 基于值函数和策略梯度的深度强化学习是核心的

基础方法和研究重点. 基于值函数的深度强化学习使用 DNN 逼近奖励值函数. 类似地, 基于策略梯度的深度强化学习用 DNN 逼近策略并利用策略梯度方法求得最优策略. 深度 *Q* 网络[23](Deep Q Network，DQN)和基于 DQN 的各种改进方法[30-31]是基于值函数的深度强化学习的最主要方法, 而深度确定性策略梯度[32](Deep Deterministic Policy Gradient，DDPG)、信赖域策略优化[33](Trust Region Policy Optimization, TRPO)和异步的优势行动者评论家算法[34](Asynchronous Advantage Actor-Critic， A3C)是基于策略梯度的深度强化学习最主要的方法.

本文提出了基于深度强化学习的 CMP 过程 R2R 控制算法, 主要的创新点如下：(1)通过结合 R2R 控制思想, 将深度强化学习框架应用于 CMP 过程控制领域, 构建了基于深度强化学习的 R2R 控制器, 其将无模型的方法与 R2R 控制进行有效结合, 拓展了 CMP 过程控制方法；(2)通过将深度强化学习中的策略网络分为线性与非线性部分, 以及对成本函数进行设计, 提高了深度强化学习在 CMP 过程控制中的控制效果与鲁棒性；(3)本文所建立的无模型的基于深度强化学习的 R2R 控制器相比于基于模型的 R2R 控制器, 实现了在不同制造环境下的在线自适应决策, 能够快速响应 CMP 过程的生产波动, 如 CMP 过程因子改变、目标去除率变化等.

1 理论背景

1.1 马尔科夫决策过程

强化学习(Reinforcement Learning, RL)问题一般由马尔可夫决策过程(Markov Decision Process, MDP)进行建模. 通常将 MDP 定义成一个四元组 $(\mathbf{S},\mathbf{A},r,p)$，其中：(1) $\mathbf{S}$ 为所有系统状态集合. $s_t\in\mathbf{S}$ 表示智能体(agent)在时刻 *t* 的系统状态；(2) $\mathbf{A}$ 为动作集合. $a_t\in\mathbf{A}$ 表示 agent 在时刻 *t* 所采取的动作；(3) r 为回报函数. $r(s_t,a_t)$ 表示在状态 s_t 下采取动作 a_t 后的奖励值；(4) p 为状态转移概率分布函数. $p(s_{t+1}|s_t,a_t)$ 表示在状态 s_t 下采取动作 a_t 后转移到下一状态 s_{t+1} 的概率.

所谓强化学习是指从环境状态到动作映射的学习, 以使动作从环境中获得的累积奖赏值最大[16]. RL 的优势在于可以求解先验信息较少或不需要先验信息的复杂优化决策问题, 因此其在解决模型未知的复杂非线性系统最优控制问题中可以起到十分重要的作用[17]. 在强化学习中, 定义策略 $\pi:\mathbf{S}\to\mathbf{A}$ 为状态空间到动作空间的一个映射. 在每个离散步长 *t*, agent 在当前状态 s_t 下根据策略 π 采取动作 a_t，接收到回报值 $\mathrm{r}(s_t,a_t)$ 并转移到下一状态 s_t. 定义 R_t 为从 *t* 时刻开始到 *T* 时刻情节(episode)结束时的累积回报值：

$$R_t=\sum_{i=t}^{T}\gamma^{i-t}r(s_i,a_i) \tag{1}$$

其中，$\gamma\in[0,1]$ 为折扣率, 用来确定短期回报的优先程度.

1.2 深度强化学习

文献[32]利用 DQN 扩展 *Q* 学习算法的思路对确定性策略梯度[35](Deterministic policy gradient, DPG)方法进行改造, 提出了一种基于行动者评论家(Actor-Critic, AC)框架的深度确定性策略梯度(Deep deterministic policy gradient, DDPG)算法.

DDPG 使用了两个深度神经网络来表示确定性策略 $a=\pi_\phi(s)$ 和值函数 $Q_\theta(s,a)$，参数分别为 ϕ 和 θ. 其中, 策略网络用来更新策略, 对应于行动者；值网络用来逼近值函数, 并为策略网络的更新提供梯度信息, 对应于评论家. DDPG 的目标是寻找到一个最优策略 π_ϕ 来最大化期望回报值 $J(\phi)=\mathrm{E}_{s_i\sim p_\pi,a_i\sim\pi}[R_0]$，通过梯度 $\nabla_\phi J(\phi)$ 进行策略网络的参数更新：

$$\nabla_\phi J(\phi)=\mathrm{E}_{s\sim p_\pi}\left[\nabla_a Q^\pi(s,a)\big|_{a=\pi(s)}\nabla_\phi\pi_\phi(s)\right] \tag{2}$$

式中，$Q^{\pi}\left(s,a\right)=\mathrm{E}_{s_i\sim p_{\pi},a_i\sim\pi}\left[R_t\left|s,a\right.\right]$表示在遵循策略$\pi$情况下，在状态$s$采取动作$a$后的期望回报值.

根据 DQN 中更新值网络的方法来更新评论家网络，即最小化损失函数$L\left(\theta\right)$：

$$L\left(\theta\right)=\mathrm{E}_{s_t,a_t,r\left(s_t,a_t\right),s_{t+1}}\left[\left(y_t-Q_{\theta}\left(s_t,a_t\right)\right)^2\right] \tag{3}$$

来进行值网络参数的更新. 其中，$y_t=r\left(s_t,a_t\right)+\gamma Q_{\theta'}\left(s_{t+1},a_{t+1}\right)$，$a_{t+1}\sim\pi_{\phi'}\left(s_{t+1}\right)$，$\phi'$和$\theta'$分别表示目标策略网络和目标值网络. DDPG 使用经验回放机制[34]获得训练样本，通过值网络将 Q 值函数关于动作的梯度信息传递到策略网络，并依据式(2)沿着提升 Q 值的方向进行策略网络的参数更新.

1.3 分布强化学习

文献 [36] 提出了分布强化学习 (Distributional Reinforcement learning)，通过学习Q值分布来代替原来的期望Q值，提升了算法能力. Q 值分布$Z^{\pi}\left(s,a\right)$的分布贝尔曼方程为：

$$\begin{aligned}&Z^{\pi}\left(s,a\right)\overset{D}{=}r\left(s,a\right)+\gamma Z^{\pi}\left(s',a'\right)\\&s'\sim p\left(\cdot\left|s,a\right.\right)\\&a'\sim\pi\left(\cdot\left|s\right.\right)\end{aligned} \tag{4}$$

其中，$X\overset{D}{=}Y$表示两个随机变量X与Y符合相同的概率分布. 分布贝尔曼最优性算子为：

$$\begin{aligned}&TZ\left(s,a\right)=r\left(s,a\right)+\\&\gamma Z\left(s',\arg\max_{a'}\mathrm{E}_{p,r}\left[Z\left(s',a'\right)\right]\right)\\&s'\sim p\left(\cdot\left|s,a\right.\right)\end{aligned} \tag{5}$$

当进行决策时，依然根据Q值分布的期望值进行动作选择.

文献[37]通过一系列分位数(quantiles)来进行$Z\left(s,a\right)$的近似. Z 的分布由 N 个支撑分位数的均匀混合来表示：

$$Z_{\varphi}\left(s,a\right)=\frac{1}{N}\sum_{i=1}^{N}\delta_{q_i\left(s,a;\varphi\right)} \tag{6}$$

其中，δ_x表示在$x\in\mathbf{R}$处的一个 Dirac 函数，即通过N个 Dirac 函数的混合来表示Q值分布，并且每一个q_i表示对应于分位数水平的分位数估计$\hat{\tau}_i=\frac{\tau_{i-1}+\tau_i}{2}$，其中$\tau_i=\frac{i}{N}\left(0\le i\le N\right)$. 这种分布近似被称为分位数近似，通过 Huber 分位数回归损失[38]来对分位数估计进行训练：

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left[\rho_{\hat{\tau}_i}^{\kappa}\left(y_{t,i'}-q_i\left(s_t,a_t\right)\right)\right] \tag{7}$$

其中，$y_{t,i'}=r\left(s_t,a_t\right)+\gamma q_{i'}\left(s_{t+1},\arg\max_{a'}\sum_{i=1}^{N}q_i\left(s_{t+1},a'\right)\right)$并且$\rho_{\hat{\tau}_i}^{\kappa}\left(x\right)=\left|\hat{\tau}_i-\mathrm{I}\left\{x<0\right\}\right|L_{\kappa}\left(x\right)$，其中I为指示函数并且$L_{\kappa}$为 Huber 损失：

$$L_{\kappa}\left(x\right)=\begin{cases}\dfrac{1}{2}x^2 & if\quad x\le\kappa\\ \kappa\left(\left|x\right|-\dfrac{1}{2}\kappa\right) & otherwise\end{cases} \tag{8}$$

文献[39]将基于分位数回归的分布强化学习与DDPG相结合，提出了分位数回归深度确定性策略梯度(Quantile Regression DDPG, QR-DDPG)。

2 基于深度强化学习的 CMP 过程 Run-to-Run 控制模型

2.1 CMP 过程模型

典型旋转 CMP 系统如图 1 所示[10]，圆晶由圆晶载具固定并压在研磨垫上，通过载具与研磨台的相对旋转来进行研磨加工. 在研磨过程中，不断地注入研磨液，经由化学侵蚀和机械移除的双重作用下，达到圆晶平坦化的目的.

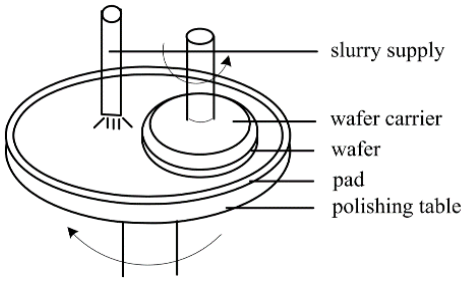

图 1　旋转 CMP 系统示意图

Fig.1　Schematic of a typical rotational CMP system

CMP 过程的一个重要被控量是材料去除率(Material Removal Rate, MRR), 这里 MRR 是指每批次圆晶研磨膜厚度的变化速率. 为了刻画在实际 CMP 过程中由于研磨垫老化和其它因素引起的 MRR 的 shift 和 drift 异常, 根据 Preston 方程, 本文使用式(9)所示的非线性过程模型[40]来进行 CMP 过程的仿真.

$$y = K_P pv + K_E \varepsilon - K_D n \tag{9}$$

其中, y 为去除率, K_P 为过程系数, p 为下压力, v 为研磨垫与晶圆之间的相对速度, K_E 为白噪声的比例系数, ε 为白噪声, K_D 为衰减率, n 为过程的 run 数.

本文以 CMP 过程模型的下压力 p 和相对速度 v 作为输入变量, 以材料去除率 MRR 作为输出变量, 通过基于深度强化学习的 CMP 过程 R2R 控制模型来自适应地控制输入变量, 以补偿 CMP 过程中出现的各种异常. 本文考虑了 CMP 过程中的 5 种异常: (1)CMP 过程发生 drift 现象;(2)CMP 过程因子变化; (3)目标 MRR 变化; (4)CMP 过程发生 shift 现象; (5)输入输出存在噪声.

2.2 R2R 控制模型

本节提出了基于深度强化学习的 CMP 过程 R2R 控制模型, 包括基于 DDPG 的 R2R 控制模型、基于 QR-DDPG 的 R2R 控制模型和基于 SCN-DDPG 的 R2R 控制模型.

1)　**基于 DDPG 的 R2R 控制模型**

通过将 DDPG 引入 R2R 控制中, 构成了基于 DDPG 的 R2R 控制模型, 以下压力 p 和相对速度 v 作为控制动作, 通过自适应地控制输入参数来补偿 CMP 过程中的各种异常. 状态空间、成本函数等模型细节与基于 SCN-DDPG 的 R2R 控制模型相同.

2)　**基于 QR-DDPG 的 R2R 控制模型**

通过将 QR-DDPG 引入 R2R 控制中, 构成了基于 QR-DDPG 的 R2R 控制模型. 分布强化学习通过学习 Q 值分布提高了 Q 值的估计精度, 能够进一步提升深度强化学习在 R2R 控制中的控制效果. 状态空间、成本函数等模型细节与基于 SCN-DDPG 的 R2R 控制模型相同.

3)　**基于 SCN-DDPG 的 R2R 控制模型**

受传统非线性控制理论启发, 文献[41]提出了结构化控制网络(Structured Control Network, SCN). 文献将 AC 框架中的策略网络分为两个部分: 非线性部分与线性部分. 通过将两个部分的动作值相加得到最终动作:

$$\pi_\phi(s) = \pi^n(s) + \pi^l(s) \tag{10}$$

式中, 线性项 $\pi^l(s) = K \cdot s + b$, K 与 b 为线性控制增益矩阵与偏置项. 非线性项 $\pi^n(s)$ 为一个全连接多层神经网络, 并去除输出层的偏置项.

结构化网络通过将策略网络分为线性模块以及非线性模块, 显式地将策略划分为全局部分和局部部分. 这种简单的结构变化能够有效地提升深度强化学习的性能, 在机器人控制以及视频游戏等领域均取得了比原网络结构更加优异的表现. 本文通过将结构化网络引入到 DDPG 算法中, 构成了 SCN-DDPG(Structured Control Network DDPG)算法。

文献[35]从数学上证明了确定性策略梯度是动作值函数的期望梯度, 从而从理论上保证了确定性策略梯度算法的收敛性. 同时作者通过实验表明在高维动作空间任务中, 确定性策略梯度算法的学习效率与控制效果均优于基于随机策略的算法. 文献[31]将确定性策略梯度算法与深度学习相结合建立了 DDPG 算法, 实现了在高维连续动作空间的端到端的策略学习, 并通过实验证明了 DDPG 在多个仿真物理控制任务中的有效性. 因此本文所提出的 SCN-DDPG 算法的收敛性具有相关的理论基础.

表 1 基于 SCN-DDPG 的压边力控制算法

Table 1 SCN-DDPG based blank holder force control algorithm

算法 1 SCN-DDPG

1: Initialize critic network Q_θ and actor network π_ϕ with parameters θ and ϕ.

2: $\theta' \leftarrow \theta$, $\phi' \leftarrow \phi$. // Initialize target network.

3: Initialize replay buffer B.

4: For episode=1, *M* do

5: Initialize CMP process model state s_1

6: For *t*=1, *T* do

7: $a_t \leftarrow \pi_\phi(s_t)+\varepsilon, \varepsilon \sim \mathrm{N}(0,\sigma)$ //select action a_t with exploration noise ε

8: Execute a_t in CMP process model, output s_{t+1} and r_t

9: $(s_t, a_t, r(s_t,a_t), s_{t+1}) \rightarrow B$ //store a transition to replay buffer

10: Sample a minibatch of *N* transitions $\left(s_t^i, a_t^i, r\left(s_t^i, a_t^i\right), s_{t+1}^i\right)_{i=1,\ldots,N}$ from B

11: $y_t^j \leftarrow r\left(s_t^j, a_t^j\right) + \gamma Q_{\theta_i'}\left(s_{t+1}^j, \pi_{\phi'}\left(s_{t+1}^j\right)\right)$ //compute target values for each sample transition

13: Update θ with gradient $\nabla_\theta L(\theta) = N^{-1}\sum_j \left(y_t^j - Q_{\theta_i}\left(s_t^j, a_t^j\right)\right)\nabla_\theta Q_\theta\left(s_t^j, a_t^j\right)$

15: Update ϕ with gradient $\nabla_\phi J(\phi) = N^{-1}\sum_j \nabla_{a_t^j} Q_{\theta_1}\left(s_t^j, a_t^j\right)\Big|_{a_t^j=\pi_\phi\left(s_t^j\right)} \nabla_\phi \pi_\phi\left(s_t^j\right)$

16: $\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i', \phi' \leftarrow \tau\phi + (1-\tau)\phi'$ //update target networks.

18: End for

19: End for

如表 1 所示，本文提出的 SCN-DDPG 算法在进行环境与智能体的交互之前，首先进行回放经验池、神经网络的参数的初始化. 在每个 episode 开始时，先初始化系统状态 s_1. 在 episode 开始后的每个控制步长 *t*，将当前时刻的状态 s_t 输入策略网络并得到动作 a_t，执行动作后得到下一时刻的状态 s_{t+1} 以及回报 $r(s_t,a_t)$. 然后将一个完整的转移经验 $(s_t, a_t, r(s_t,a_t), s_{t+1})$ 加入到回放经验池 B 中，并从回放经验池中采样出 *N* 个转移经验的 mini-batch. 最后，根据采样出的 mini-batch 来进行价值与策略网络及其目标网络的参数更新. 本文所使用的深度强化学习中的深度不仅仅体现在使用 5 层的深度网络来拟合价值函数与策略函数，还体现在深度强化学习算法中使用到的参数训练技术，如目标网络、经验回放等. 本文使用了 5 层的前馈神经网络来进行价值函数与策略函数的拟合，其中神经网络的层数由针对网络层数的敏感性实验确定.

本文提出的基于 SCN-DDPG 的 CMP 过程 R2R 控制模型如图 2 所示：

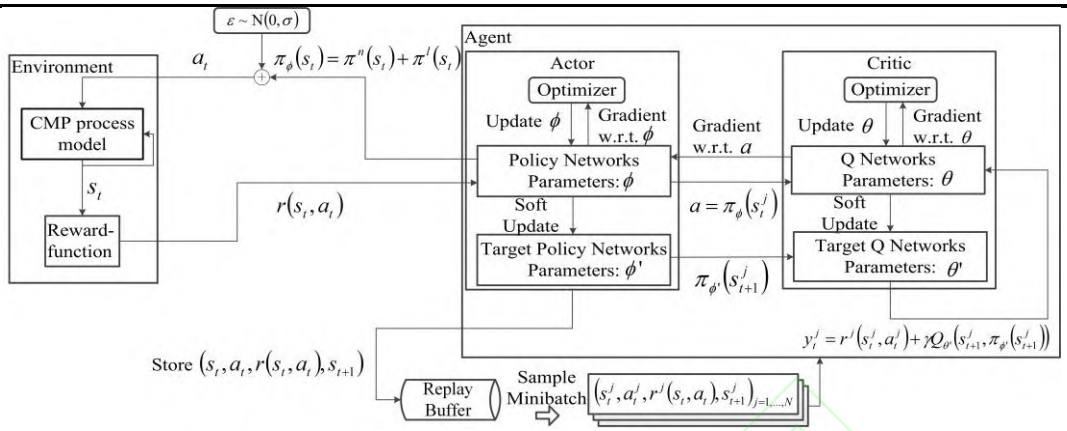

图 2 基于 SCN-DDPG 的 CMP 过程 R2R 控制模型

Fig.2 R2R control model of CMP process based on SCN-DDPG

控制模型主要由环境与智能体两部分组成. 其中, 环境由 CMP 过程模型与成本函数组成; 智能体由价值网络 Q_θ 以及策略网络 π_ϕ 组成. 环境接受到动作 a, 根据前一 run 系统状态得到当前的回报 r 与观察值 s 并将其输入智能体, 智能体输出下一 run 动作, 开始下一次交互. 智能体在与环境进行交互的过程中利用深度强化学习算法不断地更新网络参数, 最终学习到一个最优的 CMP 过程控制策略.

本文将历史 run 的 MRR, 目标 MRR, 以及上一 run 的控制参数 p, v 作为系统状态. 第 t 个 run 的系统状态 s_t 由 8 维数据组成, 分别为 5 维历史去除率、1 维目标去除率、1 维控制参数 p 以及 1 维控制参数 v 组成, 系统状态定义为 $s_t=\left(y_{t-5},y_{t-4},y_{t-3},y_{t-2},y_{t-1},T,p_{t-1},v_{t-1}\right)$. 状态空间是定义为 $\mathbf{S}=\{s_1,s_2,s_3,...,s_i\}$ 的连续空间. 第 t 个 run 的系统动作 a_t 由 2 维数据组成, 分别由 1 维控制参数 p 以及 1 维控制参数 v 组成, 系统动作定义为 $a_t=\left(p_t,v_t\right)$. 动作空间是定义为 $\mathbf{A}=\{a_1,a_2,a_3,...,a_i\}$ 的连续空间. 成本函数 $r_t=\left|y_t-T\right|+w_1\left(w_2\left|p_t-p_{t-1}\right|+w_3\left|v_t-v_{t-1}\right|\right)$ 由两部分组成, $\left|y_t-T\right|$ 表示 MRR 偏离目标值的成本项, $\left|p_t-p_{t-1}\right|$ 与 $\left|v_t-v_{t-1}\right|$ 限制了当前 run 控制参数与上一 run 控制参数的偏离程度, 保证控制参数不会发生剧烈变化. $w_i, i=1,2,3$ 表示重要性权重. 本文依据控制参数 p 与参数 v 的变化范围间的数量级进行重要性权重 w_2 与 w_3 的选择. p 的变化范围大概是[2, 8], v 的变化范围大概是[25, 55]. 它们之间大概相差了 5 倍, 因此选择 $w_2=1$, $w_3=0.2$, 保证控制参数 p 和 v 的偏离成本项基本保持在同一数量级. 至于重要性权重 w_1 的确定, 则是在粗略估计目标 MRR 偏离项与控制参数偏离项数量级的基础上, 通过对比实验确定 w_1=0.5.

3 实验验证与分析

为了验证本文所提出的深度强化学习算法 SCN-DDPG 在 CMP 过程 R2R 控制中的有效性, 根据式(9)进行 CMP 过程的计算机仿真实验, 并与传统控制方法 EWMA[5]、PCC[7], 基于神经网络的预测控制方法 NNPR2R[2]以及深度强化学习算法 DDPG、QR-DDPG 进行控制效果的比较分析. 本文设计了 5 种不同的制造异常波动环境: (1)CMP 过程发生 drift 现象;(2)CMP 过程因子变化; (3)目标 MRR 变化; (4)CMP 过程发生 shift 现象; (5)输入输出存在噪声, 并针对不同的制造环境进行实验来验证所提出的基于 DRL 的 R2R 控制器的有效性. 在每个实验中, 使用控制结果的 MRR 与目标

MRR 之间的均方根误差作为定量评价指标, 比较了各R2R控制方法之间的优劣. 在本文的实验中, 基于 DRL 的 R2R 控制器在某一制造环境下学习到策略网络并不会直接用于其他制造环境下的 CMP 过程控制, 而是在各个特定制造环境下对策略网络进行重新训练, 探究基于 DRL 的 R2R 控制器在不同制造异常波动下的鲁棒性.

3.1 SCN-DDPG 控制效果验证

在本节实验中, 每个 episode 为 600 run, 保持固定的控制目标 T 为 30 A/s, 仿真模型的系数 K_P=0.15, K_E=0.05, K_D=0.01. 本文针对基于深度强化学习的 R2R 控制器中的神经网络层数进行了敏感性实验, 分析在不同网络层数下, 基于深度强化学习的 R2R 控制器在 CMP 过程控制中的控制效果, 实验结果表明当网络层数为 4 或 5 层时, 基于深度强化学习的 R2R 控制器的控制效果与训练效率较理想.

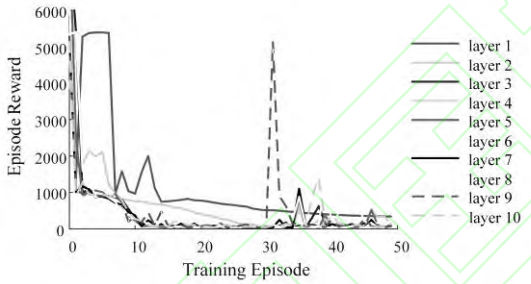

图 3 不同网络层数下 DDPG 训练过程回报值变化

Fig.3 Reward of DDPG during training process under different network layers

在不同网络层数下, 基于 DDPG 的 R2R 控制器随训练过程的回报值变化如图 3 所示. 由图 3 可知, 随着网络层数的增加, 回报值的收敛逐渐变快. 其中, 当网络层数为 1 层与 2 层时, 收敛速度明显小于深层神经网络. 从各层数的最终收敛水平来看, 浅层网络(如 1 层、2 层和 3 层)的最终获得的回报值的水平大于深层网络. 然而当网络层数逐渐加深, 回报值的收敛速度与最终水平并没有得到进一步优化, 降低了训练过程的效率, 因此综合考虑训练效率与控制效果, 网络层数选择 4 层或 5 层是较为合理的.

因此各深度强化学习算法的网络结构均设为 5 层, 隐藏层的节点数为 50, 其中 QR-DDPG 的分位数数目为 50. 由于本文所建立的基于深度强化学习的 R2R 控制器是无模型的, 假设关于系统的模型信息无法事先获取, 因此只能采用随机初始化的方法来对策略网络的参数进行初始化, 无法根据系统的模型信息或者其他理论知识来对参数进行初始化.

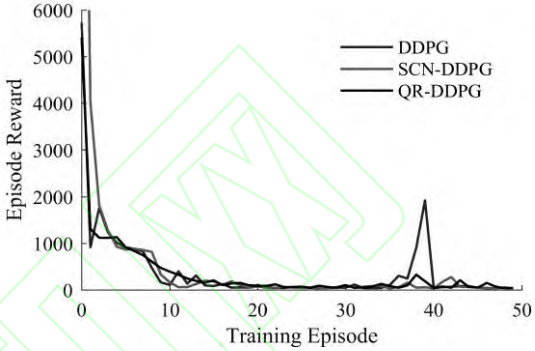

图 4 各深度强化学习训练过程回报值变化

Fig.4 Reward during training process

训练过程中, 每个训练 episode 结束后对算法进行评估, 即去除动作探索噪声后进行一次完整的 episode 控制, 并取 10 次评估的回报值平均值作为 episode reward. DDPG 与 SCN-DDPG、QR-DDPG 的训练过程 reward 变化如图 4 所示. 从回报值的总体变化趋势上看, DDPG 与 SCN-DDPG 的收敛速度基本相同.

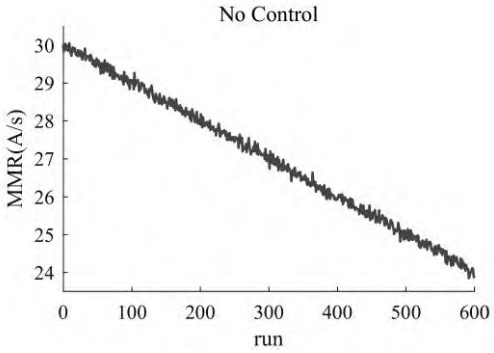

图 5 无控制下的 MRR 变化

Fig.5 MRR change under no control

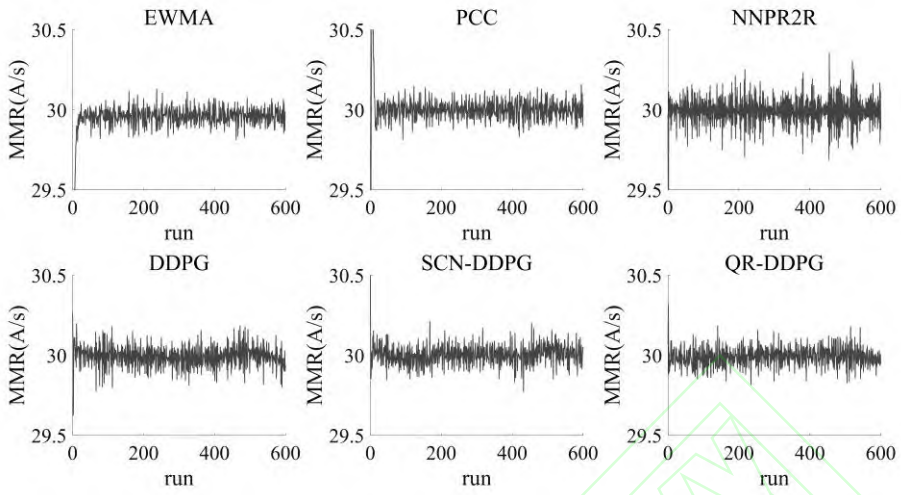

图 6 各方法控制下的 MRR 变化

Fig.6 MRR change under different methods

图 5 展示了无控制下的 MRR 变化情况, 可知 MRR 随加工 run 数的增加逐渐减小,同时其自身存在波动. 取各深度强化学习算法训练过程中获得的最好 MRR 控制效果与相应的最优控制参数轨迹进行结果分析. 图 6 展示了各算法控制下的 MRR 变化, 可知各算法均补偿了 MMR 的下降现象, 使得 MRR 维持在一定水平上. 本文使用均方根误差 (Root Mean Squared Error, RMSE)作为控制效果的评价指标, 图 6 中各方法控制下的 RMSE 对比如表 2 所示.

表 2 各方法控制下的 RMSE

Table 2 RMSE under different methods

methods	RMSE
EWMA	0.1630
PCC	0.1385
NNPR2R	0.1021
DDPG	0.0707
SCN-DDPG	0.0671
QR-DDPG	0.0627

根据图 6 与表 2 可知, 深度强化学习算法的控制效果总体上优于传统控制方法以及基于神经网络的预测控制方法. 传统算法中, PCC 的控制效果优于 EWMA, 这得益于 PCC 对噪声和过程漂移的分别处理对渐进性能带来的提升, 使得其对过程模型的拟合更加准确, 相对于 EWMA 将控制效果提升了 2.45%. NNPR2R 中使用了神经网络来进行 CMP 过程的拟合, 相比于传统方法中使用的线性模型, 提升了模型拟合的精确度, 因此获得了比传统方法更好的控制效果, 相对于 EWMA 将控制效果提升了 6.09%. 而无模型的深度强化学习避免了不精确模型拟合对控制带来的影响, 进一步提升了控制效果, DDPG 相对于 EWMA 将控制效果提升了 56.62%. 其中, QR-DDPG 获得最佳的控制效果, 相对于 EWMA 将控制效果提升了 61.53%. SCN-DDPG 获得了比 DDPG 更佳的控制效果, 相对于 EWMA 将控制效果提升了 58.83%.

EWMA、PCC、NNPR2R 等均为基于模型的 R2R 控制方法. 其中, EWMA 与 PCC 均使用了一个时变线性模型来进行 CMP 过程模型的拟合. 在控制过程中, 不断地利用历史加工数据来进行过程模型的更新, 然后根据更新后的模型进行控制参数值的计算. NNPR2R 则使用了一个神经网络来拟合 CMP 过程模型, 相对于 EWMA 与 PCC 提高了模型拟合的精确性, 然后使用 PSO 算法滚动优化求取控制律. 这三种 CMP 过程控制方法均需要对 CMP 过程进行建模. CMP 过程包括复杂的化学反应和机械抛光, 无法建立精确的机理模型, 是一个典型的非线性动态过程[3]. 针对 CMP 过程, 文献[42-44]设计了多种基于模型的 R2R 控制器来进行 CMP

过程控制. 然而基于模型的R2R控制方法往往受到所建立过程模型的精确性的制约, 无法获得最优的控制效果. 本文所提出的基于深度强化学习的 R2R 控制器利用无模型的深度强化学习算法来进行控制策略的学习, 避免了系统模型的拟合, 因此能够获得比传统 R2R 控制方法更优的控制效果. 在 SCN-DDPG 中, 线性与非线性网络结构的引入使得策略网络能够同时学习局部策略与全局策略, 进一步提升了控制效果. 在 QR-DDPG 中, 对 Q 值分布的学习提升了值网络对 Q 值估计的精确程度, 从而获得最佳的控制效果.

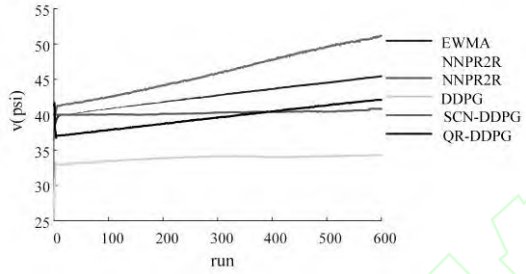

图 7 各方法控制下的 v 的变化轨迹

Fig.7 Trajectory of v under different methods

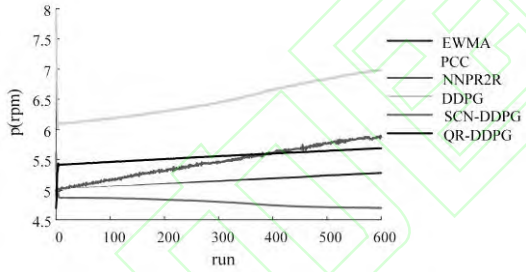

图 8 各方法控制下的 p 的变化轨迹

Fig.8 Trajectory of p under different methods

图 7 与 8 分别展示了各方法控制下的控制变量 v 与 p 的变化轨迹. 总体上, 各算法都学习到了通过提高过程参数来补偿 MRR 的下降现象, 其中深度强化学习算法的控制策略反应更加迅速, 能够快速稳定到合理的参数水平, 而传统方法则需要经过一段时间才能稳定到合理的参数水平. 从实验结果上来看, p 与 v 的变化随着 run 数的增加呈现出增加或下降的趋势, 但并不是严格的线性关系. 由于本文所建立的 CMP 过程的仿真模型非线性程度不是很强, 并且成本函数中存在对于参数剧烈变化的惩罚项, 使得参数 p 和 v 的变化较为平缓, 所以呈现一种控制参数与 run 数成类似线性的关系, 但这并不影响本文提出的 R2R 控制器所学习到的控制策略的合理性.

3.2 环境改变下的鲁棒性分析

在 CMP 过程中, 研磨垫保养会导致加工环境的变化. 为了模拟这种变化, 本节使用 $0.7K_p$ 替换原来的过程因子, 探究算法在加工环境改变下的鲁棒性.

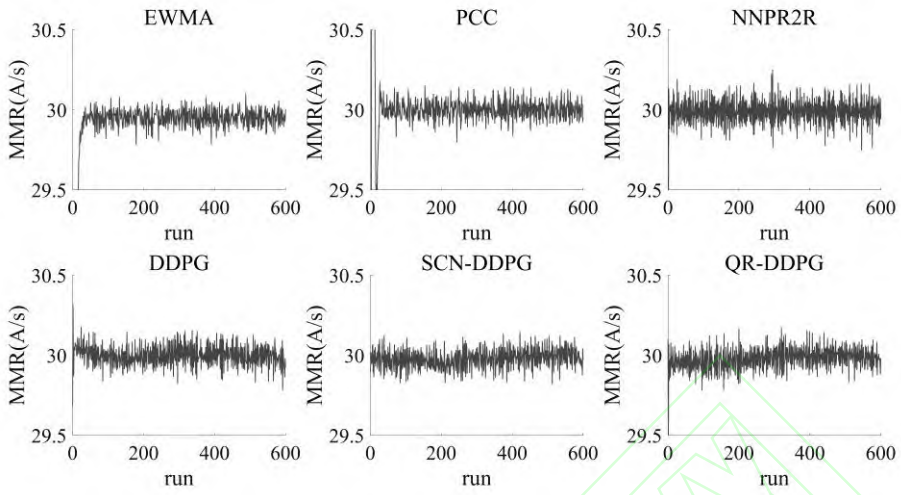

图 9 环境改变情况下的各方法控制下的 MRR 变化

Fig.9 MRR change under different methods in environment change situation

取各深度强化学习算法训练过程中获得的最好 MRR 控制效果与相应的最优控制参数轨迹来进行结果分析. 图 9 展示了在环境改变情况下的各算法控制下的 MMR 变化情况, 各方法控制下的 RMSE 对比如表 3 所示.

表 3 环境改变情况下的各方法控制下的 RMSE

Table 3 RMSE under different methods in environment change situation

methods	RMSE
EWMA	0.1630
PCC	0.1385
NNPR2R	0.1021
DDPG	0.0707
SCN-DDPG	0.0671
QR-DDPG	0.0627

根据图 9 与表 3 可知, 深度强化学习算法的控制效果总体上优于传统控制方法, 说明深度强化学习算法对于环境变化的鲁棒性优于传统控制方法. 由表 3 可知, 基于深度强化学习的 R2R 控制器总体上获得了比基于模型的各 R2R 方法更小的 RMSE. SCN-DDPG 获得了最低的 RMSE, EWMA 获得了最高的 RMSE. 在深度强化学习方法中, SCN-DDPG 取得了最好控制效果.

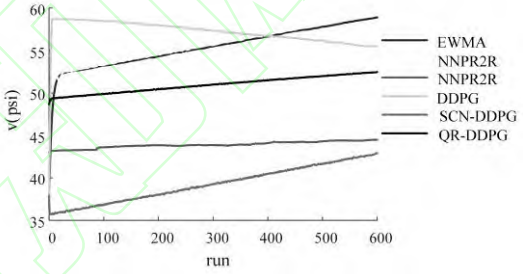

图 10 环境改变情况下的各方法控制下的 v 的变化轨迹

Fig.10 Trajectory of v under different methods in environment change situation

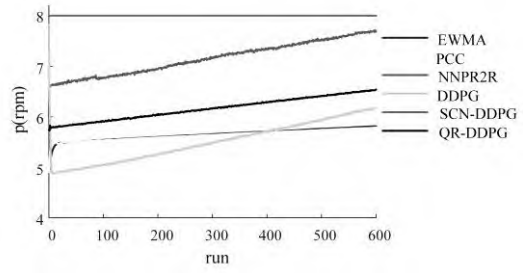

图 11 环境改变情况下的各方法控制下的 p 变化轨迹

Fig.11 Trajectory of p under different methods in environment change situation

图 10 与图 11 分别展示了环境改变情况下的各方法控制下的控制变量 v 与 p 的变化轨迹。

3.3 控制目标变化下的鲁棒性分析

在 CMP 过程中，多品种小批量的制造模式会导致目标 MRR 改变的情况发生。本节探究算法在控制目标变化下的鲁棒性。

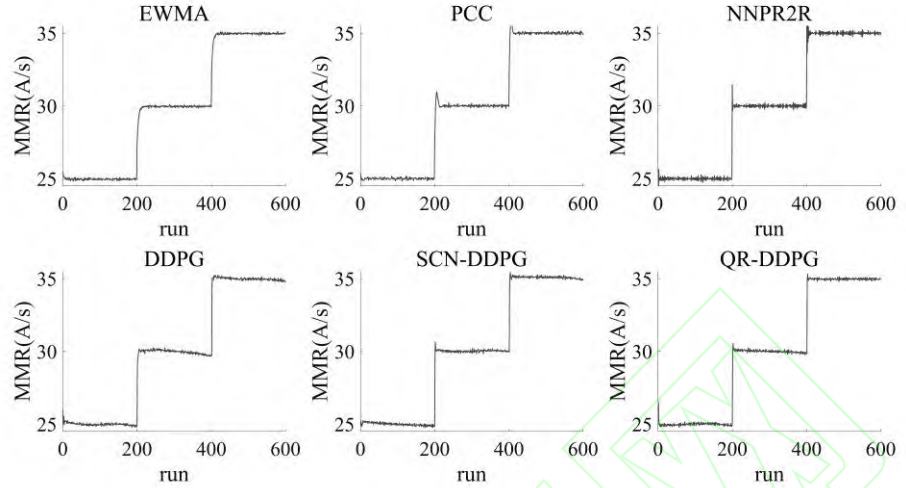

图 12 控制目标变化情况下各方法控制下的 MRR 变化

Fig.12 MRR change under different methods in target change situation

图 12 展示了在环境改变情况下的各算法控制下的 MMR 变化情况，其中 0 到 200 run 的控制目标为 25 A/s, 200 到 400 run 的控制目标为 30 A/s, 400 到 600 run 的控制目标为 35 A/s. 各方法控制下的 RMSE 对比如表 4 所示.

表 4 控制目标变化情况下各方法控制下的 RMSE

Table 4 RMSE under different methods in target change situation

methods	RMSE
EWMA	0.2338
PCC	0.2010
NNPR2R	0.1999
DDPG	0.1292
SCN-DDPG	0.1013
QR-DDPG	0.1093

由图12以及表4可知，深度强化学习在控制目标变化下的控制效果均优于传统方法. 由表 4 可知，基于深度强化学习的 R2R 控制器总体上获得了比基于模型的各 R2R 方法更小的 RMSE. SCN-DDPG 获得了最低的 RMSE，EWMA 获得了最高的 RMSE. 其中 SCN-DDPG 获得最优控制效果，对目标变化的鲁棒性最好.

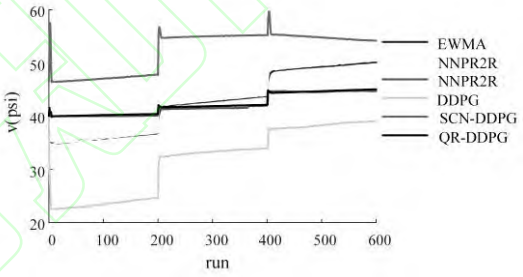

图 13 控制目标变化情况下各方法控制下的 v 变化轨迹

Fig.13 Trajectory of v under different methods in target change situation

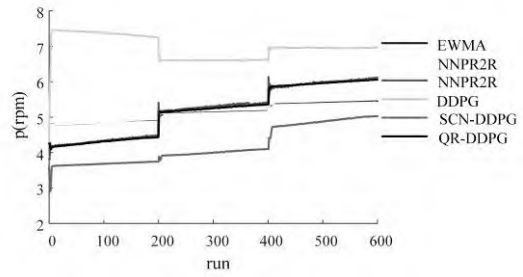

图 14 控制目标变化情况下各方法控制下的 p 的变化轨迹

Fig.14 Trajectory of p under different methods in target change situation

图 13 与 14 分别展示了环境改变情况下的各方法控制下的控制变量 v 与 p 的变化轨

迹. 从总体上看, 各算法均学习到了控制参数的阶跃策略来处理控制目标改变的情况, 其中深度强化学习控制策略的阶越速度比传统方法更加迅速, 能够使得 MRR 快速稳定到控制目标水平.

3.4 输入输出噪声下的鲁棒性分析

本节探究算法在输入、输出噪声情况下的鲁棒性. 输入输出噪声均为方差为 0.1 的白噪声.

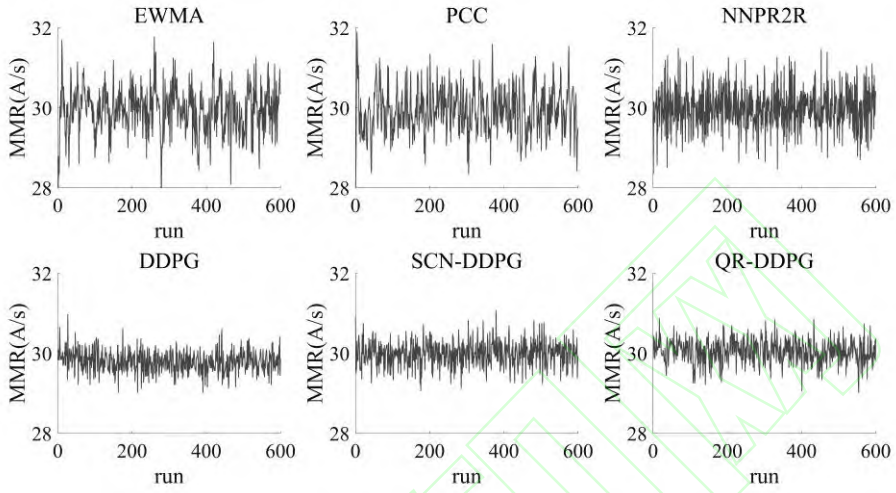

图 15　输入噪声情况下各方法控制下的 MRR 变化

Fig.15 MRR change under different methods in input noise situation

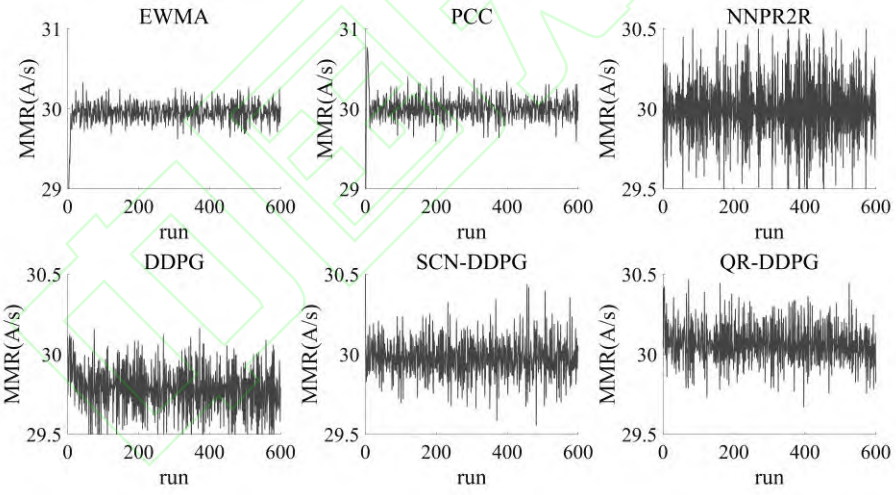

图 16　输出噪声情况下各方法控制下的 MRR 变化

Fig.16 MRR change under different methods in output noise situation

图 15 与 16 分别展示了在输入噪声与输出噪声情况下的各算法控制下的 MRR 变化情况, 各方法控制下的 RMSE 对比如表 5 所示.

表 5　噪声情况下各方法控制下的 RMSE

Tab.5 RMSE under different methods in noise situation

methods	input noise	output noise
EWMA	0.5859	0.1918
PCC	0.5737	0.1801
NNPR2R	0.5252	0.2308
DDPG	0.3387	0.2634
SCN-DDPG	0.3094	0.1311
QR-DDPG	0.2984	0.1464

在输入噪声情况下，由表 5 可知，基于深度强化学习的 R2R 控制器总体上获得了比基于模型的各 R2R 方法更小的 RMSE. QR-DDPG 获得了最低的 RMSE，EWMA 获得了最高的 RMSE. 深度强化学习方法均优于传统方法和 NNPR2R. 在输出噪声情况下，基于深度强化学习的 R2R 控制器总体上获得了比基于模型的各 R2R 方法更小的 RMSE. SCN-DDPG 获得了最低的 RMSE，EWMA 获得了最高的 RMSE. NNPR2R 以及 DDPG 的控制效果最差，这是由于它们的输入状态包含了历史 run 的 MRR，受 MRR 噪声的影响容易产生不准确的决策动作，而 SCN-DDPG 与 QR-DDPG 对于 MRR 噪声的鲁棒性较好，控制效果最优.

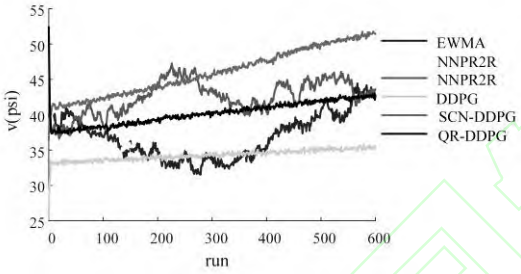

图 17 输入噪声下各方法控制下的 v 的变化轨迹

Fig.17 Trajectory of v under different methods in input noise situation

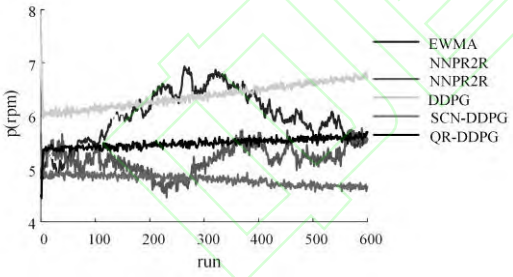

图 18 输入噪声下各方法控制下的 p 的变化轨迹

Fig.18 Trajectory of p under different methods in input noise situation

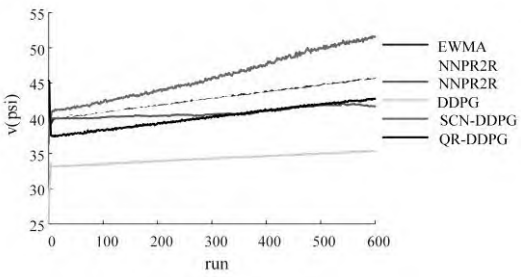

图 19 输出噪声下各方法控制下 v 的变化轨迹

Fig.19 Trajectory of v under different methods in output noise situation

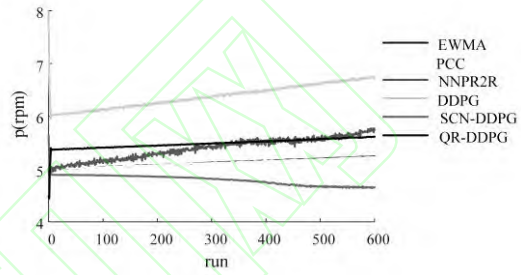

图 20 输出噪声下各方法控制下 p 的变化轨迹

Fig.20 Trajectory of p under different methods in output noise situation

图 17-20 分别展示了输入输出噪声情况下的各方法控制下的控制变量 v 与 p 的变化轨迹. 从图 17、18 也可以看出，深度强化学习在输入噪声下依然能够得到稳定的控制变量轨迹.

3.5 MMR 过程发生 shift 下的鲁棒性分析

在 CMP 过程中，当研磨垫磨损到一定程度后，操作人员一般会使用新的研磨垫进行替换，这就势必造成加工环境的剧烈改变，主要表现为 MRR 的 shift 现象. 本节探究算法在 CMP 过程发生 shift 情况下的鲁棒性，假设 shift 在 300 run 时进行替换.

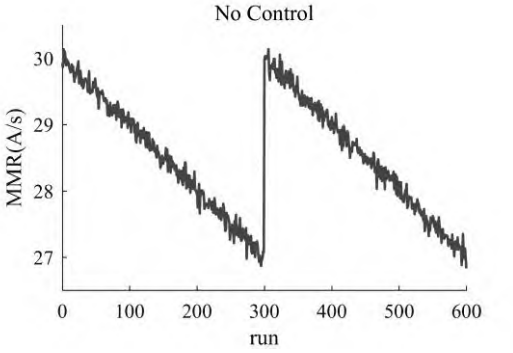

图 21　CMP 过程发生 shift 下无控制下的 MRR 变化

Fig.21　MRR change under no control in shift situation

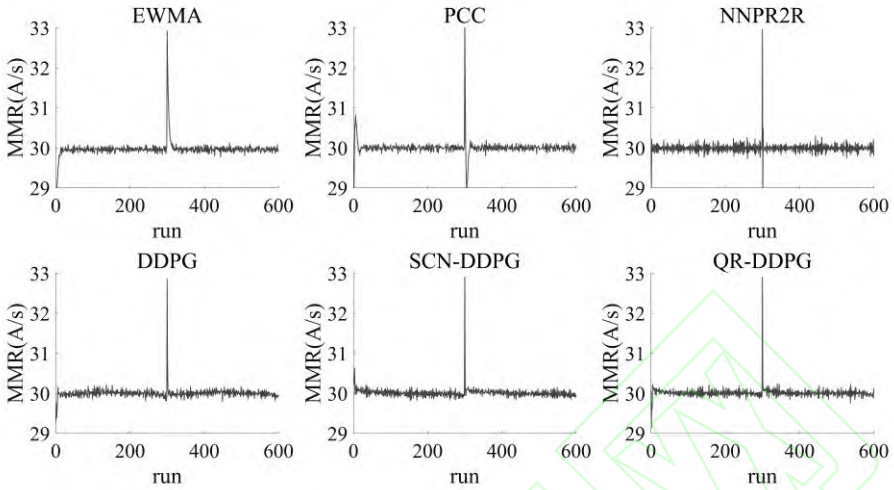

图 22　CMP 过程发生 shift 下各方法控制下的 MRR 变化

Fig.22　MRR change under different methods in shift situation

图 21 展示了 CMP 过程发生 shift 下无控制下的 MRR 变化，在第 300 run, MRR 产生了一个突变. 图 22 展示了 CMP 过程发生 shift 下各方法控制下的 MRR 变化，各算法均能将从突变中将 MRR 拉回到受控状态. 图 22 中各方法控制下的 RMSE 对比如表 6 所示. 由表 6 可知，基于深度强化学习的 R2R 控制器总体上获得了比基于模型的各 R2R 方法更小的 RMSE. SCN-DDPG 获得了最低的 RMSE, EWMA 获得了最高的 RMSE.

表 6　CMP 过程发生 shift 情况下各方法控制下的 RMSE

Table 6　RMSE under different methods in shift situation

methods	RMSE
EWMA	0.2454
PCC	0.2140
NNPR2R	0.1377
DDPG	0.1391
SCN-DDPG	0.1321
QR-DDPG	0.1493

由图 22 可知深度强化学习算法能够较快的从 shift 情况中调整到受控状态，其中 SCN-DDPG 的控制效果最好.

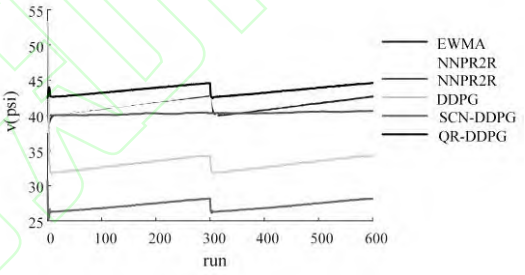

图 23　CMP 过程发生 shift 下各方法控制下 v 的变化轨迹

Fig.23　Trajectory of v under different methods in shift situation

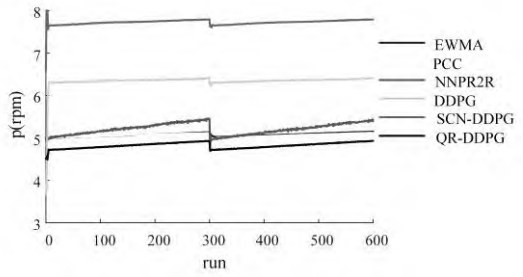

图 24　CMP 过程发生 shift 下各方法控制下 p 的变化轨迹

Fig.24　Trajectory of p under different methods in shift situation

图 23 和 24 分别展示了 CMP 过程发生 shift 下的各方法控制下的控制变量 v 与 p 的

变化轨迹. 从总体上看, 各算法均学习到了控制参数的阶跃策略来补偿 MRR 的 shift 现象, 其中深度强化学习的控制策略的阶跃速度比传统方法更快, 能够更快速地补偿 MRR 的 shift 现象.

通过以上各制造环境下对于基于 DRL 的 R2R 控制器与基于模型的 R2R 控制器的控制效果的比较分析可知, 基于 DRL 的 R2R 控制器在控制效果以及不同制造环境的鲁棒性下均优于基于模型的 R2R 控制器.

4 结论

1） 本文提出了基于深度强化学习的 CMP 过程 R2R 控制方法, 构建了基于深度强化学习的 CMP 过程 R2R 控制模型, 相较于传统方法和基于神经网络的预测控制方法, 显著提高了控制效果.

2） 将控制策略分为线性部分与非线性部分, 改进了策略网络的结构. 实验表明这种改进能够提高深度强化学习在 CMP 控制任务中的控制效果以及在不同加工环境下算法的鲁棒性.

3） 本文所提出的控制模型能够进一步地拓展到其他过程控制问题中, 并进一步地改进深度强化学习以提高其在工业过程控制问题中的有效性. 比如探究如何将 DRL 学习到的参数控制策略迁移到其他制造环境下, 避免对策略网络的重新训练以及研究基于深度强化学习的 R2R 控制器在更为复杂的制造环境下(如 shift 与 drift 同时存在)的自适应性.

References

1. Limanond S, Si J, Tsakalis K. Monitoring and control of semiconductor manufacturing processes. *Control Systems IEEE*, 1998, **18**(6): 46-58
2. Wang Liang, Hu Jing-Tao. Neural network based intelligent r2r predictive control to cmp process. *Semiconductor Technology*, 2012, **37**(4): 305-311 (in Chinese)
(王亮, 胡静涛. 基于神经网络的 CMP 过程智能 R2R 预测控制. 半导体技术, 2012, **37**(4): 305-311)
3. Wang Shu-Qing, Zhang Xue-Peng, Chen Liang. Review on run-to-run control in semiconductor manufacturing. *Journal of Zhejiang University*(*Engineer Science*), 2008, **42**(8): 1393-1398 (in Chinese)
(王树青, 张学鹏, 陈良. 半导体生产过程的 Run-to-Run 控制技术综述. 浙江大学学报(工学版), 2008, **42**(8): 1393-1398)
4. Lu Jing-Yi, Cao Zhi-Xing, Gao Fu-Rong. Batch process control—overview and outlook. *Acta Automatica Sinica*, 2017, **43**(6): 933-943 (in Chinese)
(卢静宜, 曹志兴, 高福荣. 批次过程控制—回顾与展望. 自动化学报, 2017, **43**(6): 933~943)
5. Boning D S, Moyne W P, Smith T H, Moyne J, Telfeyan R, Hurwitz A, et al. Run by run control of chemical-mechanical polishing. *IEEE Transactions on Components, Packaging & Manufacturing, Technology, Part C*, 1996, **19**(4): 307-314
6. Chen A, Guo R S. Age-based double EWMA controller and its application to CMP processes. *IEEE Transactions on Semiconductor Manufacturing*, 2001, **14**(1): 11-19
7. Yi J, Sang W S, Zhao E. A run-to-run film thickness control of chemical-mechanical planarization processes. In: Proceedings of 2005 American Control Conference. Portland, USA: IEEE, 2005. 4231-4236
8. Bode C A, Ko B S, Edgar T F. Run-to-run control and performance monitoring of overlay in semiconductor manufacturing. *Control Engineering Practice*, 2004, **12**(7): 893-900
9. Del Castillo E, Yeh J Y. An adaptive run-to-run optimizing controller for linear and nonlinear semiconductor processes. *IEEE Transactions on Semiconductor Manufacturing*, 1998, **11**(2): 285-295.
10. Yi J, Sheng Y, Xu C S. Neural network based uniformity profile control of linear chemical-mechanical planarization. *IEEE Transactions on Semiconductor Manufacturing*, 2003, **16**(4): 609 - 620
11. Wang G J, Yu C H. Developing a neural network-based run-to-run process controller for chemical-mechanical planarization. *International Journal of Advanced Manufacturing Technology*, 2006, **28**(9-10): 899-908.
12. Wang G J, Lin B S, Chang K J. In-situ neural network process controller for copper chemical mechanical polishing. *International Journal of Advanced Manufacturing Technology*, 2007, **32**(1-2): 42-54
13. Wang Liang, Hu Jing-Tao. Grey model based immune predictive R2R control of CMP process. Chinese Journal of Scientific Instrument, 2012, **33**(2): 306-314 (in Chinese)
(王亮, 胡静涛. 基于灰色模型的 CMP 过程免疫预测 R2R 控制. 仪器仪表学报, 2012, **33**(2): 306-314)
14. Shah H, Gopal M. Model-Free Predictive Control of Nonlinear Processes Based on Reinforcement Learning. *IFAC-PapersOnline*, 2016, **49**(1): 89-94
15. Hafner R, Riedmiller M. Reinforcement learning in

feedback control. *Machine Learning*, 2011, **84**(1-2): 137-169

16. Gao Yang, Chen Shi-Fu, Lu Xin. Research on reinforcement learning technology: a interview. *Acta Automatica Sinica*, 2004, **30**(1): 86-100 (in Chinese)
(高阳, 陈世福, 陆鑫. 强化学习研究综述. 自动化学报, 2004, **30**(1): 86-100)

17. Liu De-Rong, Li Hong-Liang, Wang Ding. Data-based self-learning optimal control: research progress and prospects. *Acta Automatica Sinica*, 2013, **39**(11): 1858-1870 (in Chinese)
(刘德荣, 李宏亮, 王鼎. 基于数据的自学习优化控制: 研究进展与展望. 自动化学报, 2013, **39**(11): 1858-1870)

18. Spielberg S P K, Gopaluni R B, Loewen P D. Deep reinforcement learning approaches for process control. In: Proceedings of 2017 6th international symposium on advanced control of industrial process (AdCONIP). Taipei, China: IEEE, 2017. 28-31

19. Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527-1554

20. Bengio Y, Delalleau O. On the expressive power of deep architectures. In: Proceedings of International Conference on Algorithmic Learning Theory. Heidelberg, Berlin: Springer, 2011. 18-36

21. Liu Quan, Zhai Jian-Wei, Zhang Zong-Chang. A Survey on Deep Reinforcement Learning. *Chinese Journal of Computers*, 2018, **41**(1): 1-27 (in Chinese)
(刘全, 翟建伟, 章宗长,等. 深度强化学习综述. 计算机学报, 2018, **41**(1):1-27)

22. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. In: Proceedings of the Work-shops at the 26th Neural Information Processing Systems 2013. Lake Tahoe, USA: Curran Associates, 2013. 201-220

23. Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529-533

24. Zhang M, Geng X, Bruce J, Caluwaerts K. Deep reinforcement learning for tensegrity robot locomotion. In: Proceedings of the IEEE International Conference on Robotics and Automation. Singapore: IEEE, 2017. 634-641

25. Gu S, Holly E, Lillicrap T, Levine S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In: Proceedings of the IEEE International Conference on Robotics and Automation. Singapore: IEEE, 2017. 3389-3396.

26. Cuayahuitl H, Yu S, Williamson A, Carse J. Scaling up deep reinforcement learning for multi-domain dialogue systems. In: Proceedings of the International Joint Conference on Neural Networks. Anchorage, USA: IEEE, 2017. 3339-3346

27. Zhao T, Eskenazi M. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Los Angeles, USA: ACM, 2016. 1-10

28. Xiong X, Wang J, Zhang F, Li K. Combining deep reinforcement learning and safety based control for autonomous driving[Online], available: https://arxiv.xilesou.top/abs/1612.00147, December 1, 2016

29. Sallab A E L, Abdou M, Perot E, Senthil Y. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017, **2017**(19): 70-76

30. Hasselt H V. Double Q-learning. In: Proceedings of the 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, 2010. 2613-2621

31. Wang Z, Sehaul T, Hessel M, Hasselt H V, Lanctot M, Freitas N D. Dueling network architectures for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning. New York, USA: ACM, 2016. 1995-2003

32. Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. Computer Science, 2015, 8(6):A187.

33. Schulman J, Levine S, Moritz P, Abbeel P. Trust region policy optimization. In: Proceedings of the International Conference on Machine Learning. Lugano, Switzerland: ACM, 2015. 1889-1897

34. Mnih V, Badia A P, Mirza M, Graves A, Harley T, Lillicrap T P, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the International Conference on Machine Learning. New York, USA: ACM, 2016. 1928-1937

35. Silver D, Lever G, Heess N, Degris D, Wierstra D, Riedmiller. Deterministic policy gradient algorithms. In: Proceedings of the International Conference on Machine Learning. Beijing, China: ACM, 2014: 387-396.

36. Bellemare M G, Dabney W, Munos R. A distributional perspective on reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: ACM, 2017. 449-458

37. Dabney W, Rowland M, Bellemare M G, Munos R. Distributional Reinforcement Learning with Quantile Regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. Louisiana, USA: AAAI, 2018. 2892-2901

38. Huber, Peter J. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 1964, **35**(1):73-101.

39. Zhang S, Yao H. QUOTA: The Quantile Option

Architecture for Reinforcement Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019. 5797-5804

40 . Wang G J, Wang M T, Yang F C, Huang M L, Loh K N, Chen J. Grey forecasting run-to-run control system in copper chemical mechanical polishing. *International Journal of Advanced Manufacturing Technology*, 2009, **41**(1-2): 48-56

41 . Srouji M, Zhang J, Salakhutdinov R . Structured Control Nets for Deep Reinforcement Learning [Online]. Available: https://arxiv.xilesou.top/abs/1802.08311, February 22, 2018

42 . Boning D S, Moyne W P, Smith T H, Moyne J, Telfeyan R, Hurwitz A, et al . Run by run control of chemical mechanical polishing . *IEEE Transactions on Semiconductor Manufacturing*, 1996, 19(4): 307-314

43 . Edgar T F, Campbell W J, Bode C . Model-based control in microelectronics manufacturing. In: Proceedings of 38th Conference on Decision and Control .Phoenix, USA: IEEE, 1999. 4185-4191

44 . Zhang C, Deng H, Baras J S . Run-to-run control methods based on the DHOBE algorithm . Automatica, 2003, 39(1): 35-45

郭鹏 同济大学机械与能源工程学院研究生. 2017 年获南京理工大学机械工程学院学士学位. 主要研究方向为生产过程控制与深度强化学习.
E-mail: guopeng19940821@163.com
(**GUO Peng** Postgraduate candidate at the Mechanical and Energy Engineering, Tongji University. He received his bachelor degree from Nan Jing University of Science and Technology. His research interests include production system control and deep reinforcement learning)

余建波 通讯作者, 同济大学机械与能源工程学院教授. 2009 年获上海交通大学机械工程学院博士学位. 主要研究方向为: 机器学习, 深度学习, 智能质量管控, 过程控制, 视觉检测与识别.
E-mail: jbyu@tongji.edu.cn
(**YU Jian-Bo** Professor at the Mechanical and Engineering, Tongji University. He received his Ph. D. degree from Shanghai Jiaotong University. His research interests include machine learning, deep learning, intelligent quality control, process control, visual inspection and identific