冯子豪, 万会龙, 林江豪, 等. 基于强化学习的注塑工艺参数自动调优方法及应用[J]. 广东工业大学学报, 2025, 42(2): 59–69. doi: 10.12052/gdutxb.240142.
Feng Zihao, Wan Huilong, Lin Jianghao, *et al*. Reinforcement learning-based automatic tuning method and application of injection molding process parameters[J]. Journal of Guangdong University of Technology, 2025, 42(2): 59–69. doi: 10.12052/gdutxb.240142.

基于强化学习的注塑工艺参数自动调优方法及应用

冯子豪，万会龙，林江豪，任志刚
（广东工业大学　自动化学院 , 广东　广州 510006）

摘要: 注塑工艺参数的优化在当代制造业中至关重要，它不仅影响产品质量，还决定了生产成本和效率。传统的人工调优方法依赖试错，导致生产周期延长和成本上升。为此，本文首次将强化学习(Reinforcement Learning, RL)技术应用于注塑工艺参数调优，提出了一种基于RL的新型注塑工艺参数自动调优算法，旨在通过智能化手段优化注塑过程。首先，本文将注塑工艺参数调优问题建模为顺序决策问题，并设计了一个定制化的马尔可夫决策过程模型。然后，提出了一种无模型的RL方法，基于Q-learning算法实现了注塑工艺参数的自动选择。与传统方法相比，该算法能在动态变化的生产环境中自动探索并优化工艺参数配置。最后，通过实验验证了所提方法的可行性和有效性，展示了较好的应用潜力。

关键词: 注塑成型；强化学习；Q-learning；智能制造；知识自动化
中图分类号: TP271　　**文献标志码:** A　　**文章编号:** 1007–7162(2025)02–0059–11

Reinforcement Learning-based Automatic Tuning Method and Application of Injection Molding Process Parameters

Feng Zihao, Wan Huilong, Lin Jianghao, Ren Zhigang
(School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: The optimization of injection molding process parameters is crucial in contemporary manufacturing, as it affects not only product quality but also production costs and efficiency. Traditional manual tuning methods rely on trial and error, leading to extended production cycles and increased costs. To address this, reinforcement learning (RL) technology is applied, for the first time, to the tuning of injection molding process parameters, proposing a novel automatic tuning algorithm based on RL for injection molding process parameters. The paper first models the injection molding process parameter tuning problem as a sequential decision-making problem and designs a customized Markov Decision Process model. Subsequently, a model-free RL method is proposed, implementing the automatic selection of injection molding process parameters based on the Q-learning algorithm. Compared with traditional methods, this algorithm can automatically explore and optimize process parameter configurations in dynamically changing production environments. Finally, the feasibility and effectiveness of the proposed method are experimentally validated, demonstrating significant application potential.

Key words: injection molding; reinforcement learning; Q-learning; intelligent manufacturing; knowledge automation

注塑成型是现代离散制造业中广泛使用的一种塑料成型技术，它通过加热熔化热塑性或热固性塑料，将其注入模具，冷却凝固后，获得所需形状和尺寸的零件或产品[1-4]。注塑成型具有生产效率高、产品质量好、成型范围广、资源利用率高等优点，其产品涉及汽车、家用电器、医疗器械、电子产品等多个领域。

注塑成型工艺(Injection Molding Process, IMP)本身错综复杂，涉及多个参数的精确控制。影响注塑

收稿日期： 2024-11-16　　**录用日期：** 2025-02-14
基金项目： 国家自然科学基金资助项目(62073088)；广东省基础与应用基础研究基金资助项目(2024A1515011768)
作者简介： 冯子豪(1999–)，男，硕士研究生，主要研究方向为强化学习和工艺参数寻优，E-mail：849612040@qq.com
通信作者： 任志刚(1987–)，男，副教授，硕士生导师，主要研究方向为复杂工业过程优化与控制、工业智能，E-mail：renzhigang@gdut.edu.cn

产品质量的因素一般可分为3个方面：机器参数、工艺参数和质量指标。其中，注塑过程中的工艺参数(如模具温度、熔体温度、注塑速度、填充时间、保压时间等) 是影响注塑产品质量的关键因素。获得并保持优化工艺参数是提高注塑成型性能和降低成本的重要手段。在传统的IMP 中，初始工艺参数设置受材料和机器特性的影响。因此，每个注塑成型周期完成后，都必须对产品质量进行检查和评估。如果产品不符合质量标准，就会被剔除，然后在下一个生产周期重新调整工艺参数。经过反复迭代后，工艺就能得到有效控制并按比例进行生产。生产成本可能会随着产品的复杂程度而增加，当工艺人员的技能和专业知识存在差异时，工艺参数设置的质量也可能会出现差异，从而不可避免地导致生产周期的延长和生产成本的增加。由于产品的多样性，对IMP的优化要求越来越高。IMP参数调整(IMP Parameter Tuning, IMPPT) 是指根据不同的塑料材料、模具结构和产品要求，选择和调整注塑机的最佳模具温度、熔体温度、注射速度、充填时间、保压时间等工艺参数的过程。IMPPT的任务对注塑工程师的专业技术要求很高，既需要丰富的理论知识，又需要实践经验。目前，注塑成型工艺参数的设计在很大程度上仍依赖于现场工作人员的经验和试错。然而，这些启发式方法并不能保证获得最佳条件，而且往往需要漫长的微调过程。因此，需要一种系统的工艺优化方法来改善这种情况。

近年来，优化注塑成型工艺参数已成为制造领域的一个研究重点。研究人员在这一领域进行了广泛的研究，旨在提高产品质量、降低生产成本和增强竞争力[3-14]。例如，Shen等[7]提出了一种结合人工神经网络和遗传算法(Genetic Algorithm, GA) 的方法来优化注塑成型过程。采用反向传播神经网络模型来映射工艺条件与模塑件质量指标之间的复杂非线性关系，并利用GA进行工艺条件优化。该方法应用于工业部件，以改善体积收缩变化的质量指标。Xu[10]通过有限元分析研究了注塑过程中因翘曲引起的残余应力和机械性能。使用反向传播神经网络模型来映射工艺参数与产品性能之间的非线性关系，并用粒子群优化(Particle Swarm Optimization, PSO) 算法优化这些参数，从而显著提高机械性能。以聚碳酸酯车窗为案例，确定了在冲击载荷下最小化车窗内最大冯米塞斯应力的最佳工艺参数。Alam等[13]利用响应面法和遗传算法，通过优化折叠产品的折角、周期长度和座位与踏板之间的高度这三个设计参数，减少折叠时间。通过中央复合设计进行实验，建立了3D模型和数学方程，旨在优化设计参数，以提高产品开发的效率。Ribeiro[15]比较了c-SVM (Support Vector Machine, SVM) 和v-SVM分类器与径向基函数(Radial Basis Function, RBF) 神经网络在塑料注塑机产品故障数据中的表现，通过超参数优化来选择合适的监控条件，研究目标是监控过程数据并快速响应意外的工艺干扰。实验结果表明SVM，特别是大间隔分类器，相较于RBF神经网络在泛化能力和性能上有所提升，提高了模型在实际应用中的有效性。尽管上述方法在一定程度上取得了改进，但仍然存在局限性。经验法过度依赖于操作员的专业技能和知识；上述的数学模型方法和生物启发的算法通常计算资源消耗大，在处理高维和非线性参数空间时较为吃力；人工神经网络与机器学习的方法需要大量的标记数据进行训练。这些模型方法缺乏实时适应性和持续学习能力，难以应对瞬息万变的生产环境和实时调整的需求。

针对以上问题，本文提出了一种基于强化学习(Reinforcement Learning, RL) 的注塑工艺参数调优方法。RL是一种交互式学习的方法，具有实时适应能力、高维处理能力、对不断变化的环境的适应能力和持续学习能力。这种方法通过智能体与环境交互，反复实验和学习来优化决策，非常适合解决多变量的复杂问题。近年来，RL在游戏[16-17]、机器人控制[18-21]、推荐系统[22-23]等领域取得了较多成果[24-29]，但在注塑工艺参数优化方面的应用有限。本研究旨在将RL技术应用于注塑工艺参数的自动调优，以实现工艺参数决策的自动化，同时减轻操作员负担并降低人为错误风险，对提升注塑工业生产效率和产品质量具有重要意义。为此，本文首先将注塑成型工艺参数调优问题建模为一个顺序决策问题，并构建了定制化的马尔可夫决策过程(Markov Decision Process, MDP) 模型。接着，提出了一种定制的无模型、基于非策略Q-learning 的RL方法来自动选择注塑工艺参数。最后，实验结果表明，该方法在实现高质量注塑产品成果方面的可行性和有效性。本研究首次将基于RL的技术应用于注塑工艺参数自动调优领域，为将先进的机器学习技术集成到注塑成型工艺优化的自动化中奠定了基础，为未来提高注塑生产效率和产品质量开辟了新的可能性。

1 注塑成型参数优化问题

IMP工艺参数的优化一直是制造业关注的关键问题，它包括但不限于对料筒和模具的温度控制、注

射压力、夹紧压力、背压、注射速度、冷却时间、保压时间等设置。优化这些参数的目的是提高产品质量、降低生产成本、提高生产效率等。

在注塑工艺领域，主要目标通常是确定工艺参数的最佳组合，以达到注塑产品的期望质量标准。注塑成型工艺的固有参数包括注射速度v(cm/s)、保压时间t(s)和注射压力P(MPa)，对工艺的效率和产品的质量都有直接的影响。因此，工程师和研究人员努力确定工艺参数的最佳组合，以尽量减少成型产品的缺陷，确保缺陷在可接受的参数范围，最大程度降低产品的缺陷率。本文研究侧重于确定IMP工艺参数的最佳组合，特别是v、t和P。因此，本文提出的最小化目标函数为

$$\min_{v,t,P}\left|D(v,t,P)-D_{\text{target}}(v,t,P)\right| \tag{1}$$

式中：$D(v,t,P)$是实际成型产品的质量指标，而$D_{\text{target}}(v,t,P)$是成型产品的期望质量指标或者合格标准。$D(v,t,P)$是一个复杂的数学函数，取决于v、t和P，可以由实际生产数据和经验推导获得。为了保证工艺参数的实用性、有效性和安全性，本文引入式(2)的约束条件。

$$\begin{cases} v_{\min}\leqslant v\leqslant v_{\max} \\ t_{\min}\leqslant t\leqslant t_{\max} \\ P_{\min}\leqslant P\leqslant P_{\max} \end{cases} \tag{2}$$

式中：$v_{\min}$和$v_{\max}$为注射速度允许的最小值和最大值。$t_{\min}$和$t_{\max}$为保压时间的最小与最大值。$P_{\min}$和$P_{\max}$为注射压力的最小与最大值，其在实际生产过程中根据设备不同在保证安全性的情况下会有所变化，这些约束保证了工艺参数保持在合理的范围内。

本文简化了注塑成型参数的设置过程，以促进未来的研究工作。当前的研究假设注塑成型产品质量指标仅与3个关键参数相关即v、t和P，假设其他参数配置合理，在未来的研究中将探索更复杂的工艺参数配置。本文所提方法的目标是确定约束条件内v、t和P的最佳组合，以获得满足要求的成型产品。具体而言，目标是解决式(3)的优化问题。

$$\min_{v,t,P}\left|D(v,t,P)-D_{\text{target}}(v,t,P)\right|,\quad \text{s.t.}\begin{cases} v_{\min}\leqslant v\leqslant v_{\max} \\ t_{\min}\leqslant t\leqslant t_{\max} \\ P_{\min}\leqslant P\leqslant P_{\max} \end{cases} \tag{3}$$

2　基于强化学习的过程参数选择

2.1　Q-learning 算法原理

在优化IMP参数的背景下，本文采用Q-learning[30]作为基础的方法，解决注塑工艺参数的自动调整和优化问题。

Q-learning是一种无模型、无策略的RL算法，智能体通过和环境进行不断交互学习得到最佳行动的选择策略。它通过估计每个状态−动作对的价值(Q值)，并根据与环境交互过程中获得的奖励和经验迭代更新Q值来实现这一目标。Q-learning是一种成熟的RL算法，在优化和控制问题中得到了广泛的应用，特别适合用于具有离散动作空间与状态空间的场景，因此也是离散IMP参数优化问题的最佳选择。在Q-learning中，Q值表示为$Q(\boldsymbol{s},\boldsymbol{a})$，表示智能体在$\boldsymbol{s}$状态下选择$\boldsymbol{a}$动作同时遵循最优策略所能获得的预期累计奖励价值。

在Q-learning算法中，其关键概念为状态、动作、奖励、Q值、ε-贪婪策略。状态表示环境中的特定条件或情况，通常用$\boldsymbol{s}$表示。状态可以是离散或者连续的，这取决于问题的性质，本文采用离散状态。动作表示智能体可以采取的策略，通常用$\boldsymbol{a}$来表示，与状态类似，动作可以是离散的或者连续的，取决于问题的性质，本文采用离散动作。奖励是通过在状态$\boldsymbol{s}$情况下采取特定的动作$\boldsymbol{a}$，根据设计的奖励机制获得数值反馈，通常用r表示，用于评估该状态下该动作的价值，以反馈其可取性，从而使智能体累积奖励最大化。价值函数，记作$Q(\boldsymbol{s},\boldsymbol{a})$，表示当前$\boldsymbol{s}$状态下采取动作$\boldsymbol{a}$的预期累计奖励。$Q$值是Q-learning算法的核心，Q-learning本质上就是基于$Q(\boldsymbol{s},\boldsymbol{a})$来进行迭代学习的，它近似于状态−动作对的长期预期累计奖励。Q值更新的核心公式[31]为

$$Q^{\text{new}}(\boldsymbol{s}_k,\boldsymbol{a}_k)=Q(\boldsymbol{s}_k,\boldsymbol{a}_k)+\alpha\left[r_k+\gamma\max_{\boldsymbol{a}}Q(\boldsymbol{s}_{k+1},\boldsymbol{a})-Q(\boldsymbol{s}_k,\boldsymbol{a}_k)\right] \tag{4}$$

式中：α表示学习率，控制每次更新的步长；γ表示折扣因子，用来权衡当前价值函数与未来价值函数；k表示迭代次数；r_k表示当前状态下执行当前动作通过奖励机制所获得的即时奖励；$\boldsymbol{s}_k$和$\boldsymbol{a}_k$表示当前的状态和动作，$\boldsymbol{s}_{k+1}$表示状态$\boldsymbol{s}_k$执行动作$\boldsymbol{a}_k$后的状态。$\max_{\boldsymbol{a}}Q(\boldsymbol{s}_{k+1},\boldsymbol{a})$是指选择下一个状态中对应动作价值最高的价值函数。此外还有ε-贪婪策略，用来权衡智能体的探索与利用，Q-learning算法通常采用ε-贪婪策略。具体来说，智能体有ε的概率随机选择动作，并以概率$(1-\varepsilon)$选择Q值最优的动作，利用这种策略可以驱动智能体倾向选择最优动作的同时一定程度上避免陷入局部最优解。

Q-learning有收敛的特性，理论上经过无限次迭

代后，Q值函数将收敛到真正最优的Q值，一旦收敛就可以通过每个状态中选择最优Q值的动作来推导出最优策略，这种收敛特性巩固了它在RL任务中的有效性，也确保在智能体持续探索和与环境交互学习过程中，Q值逐渐接近其真实值，从而得到获得最高累积奖励的最优策略。Q-learning的收敛性为其应用于各种决策和优化问题，包括IMP参数优化等提供了理论基础。

2.2 基于Q-learning的训练数据集的个性化定制

本节具体描述了在IMP框架中各种约束条件下，如何设计Q-learning算法以获得最佳的注塑工业参数组合。为了在Q-learning学习框架中解决IMPPT问题，将IMPPT问题初始化为MDP是至关重要的。MDP是一种概率数学框架，用于模拟智能体和环境之间的交互。通常MDP包含状态、动作、转换函数和奖励4个主要组件，本文将定义各组件，同时将IMPPT问题转化为MDP问题。

2.2.1 状态空间定义

在注塑问题中，状态表示工艺参数的特定组合。例如，状态可以包括v、t和P。在这项工作中，本文使用状态向量$\boldsymbol{s}$来表示当前状态。$\boldsymbol{S}$表示状态空间，包括过程参数的所有可能组合，通常是连续或者离散的值，可以表示为$\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_N\}$，$N$表示状态数。每个状态$\boldsymbol{s}$代表一个唯一的参数组合，如$\boldsymbol{s}_1=\{v_1,t_1,P_1\}$，其中$v_1$、$t_1$、$P_1$分别为状态$\boldsymbol{s}_1$对应的注射速度、保压时间和注射压力的值。最后，本文定义状态空间为

$$\begin{aligned}\boldsymbol{S}=\{&(v,t,P)|v\in[v_{\min},v_{\max}],t\in[t_{\min},t_{\max}],\\&P\in[P_{\min},P_{\max}]\}\end{aligned}\tag{5}$$

2.2.2 动作空间定义

动作表示对应过程控制中的参数调整，记为向量$\boldsymbol{a}$。在注塑成型问题中，动作可以调整注塑时间、保压时间和注射压力，表示为Δv、Δt和ΔP。动作空间记为$\boldsymbol{A}=\{\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_Z\}$，$Z$为动作数，每个动作$\boldsymbol{a}$表示一种修改参数的方式，如$\boldsymbol{a}_1=\{\Delta v_1,\Delta t_1,\Delta P_1\}$表示改变注射速度、保压时间和注射压力的值，$\boldsymbol{A}$表示所有可能动作的集合，通常是连续或者离散的值。最后本文的动作空间定义为

$$\begin{aligned}\boldsymbol{A}=\{&(\Delta v,\Delta t,\Delta P)|\Delta v\in[\Delta v_{\min},\Delta v_{\max}],\\&\Delta t\in[\Delta t_{\min},\Delta t_{\max}],\Delta P\in[\Delta P_{\min},\Delta P_{\max}]\}\end{aligned}\tag{6}$$

2.2.3 奖励函数定义

奖励函数用于衡量每一步采取的动作对注塑产品质量的影响。为了最大限度地降低缺陷率，提高产品质量，奖励函数的设计至关重要。在注塑问题中可以根据实际质量指数和目标质量指数的误差来定义奖励函数。本文的奖励函数设计为

$$R(\boldsymbol{s},\boldsymbol{a})=\begin{cases}hw_1+\dfrac{(E_{k-1}-E_k)}{\mathrm{tol}_{\mathrm{sa}}}w_2, & E_k\leqslant\mathrm{tol}_{\mathrm{ba}}\\ w_1+\dfrac{(E_{k-1}-E_k)}{\mathrm{tol}_{\mathrm{sa}}}w_2, & \mathrm{tol}_{\mathrm{ba}}<E_k\leqslant\mathrm{tol}_{\mathrm{sa}}\\ -w_1+\dfrac{(E_{k-1}-E_k)}{\mathrm{tol}_{\mathrm{sa}}}w_2, & E_k>\mathrm{tol}_{\mathrm{sa}}\end{cases}\tag{7}$$

其中

$$E=\left|D(v,t,P)-D_{\mathrm{target}}(v,t,P)\right|\tag{8}$$

式(7)、(8)中：E表示注塑产品在特定状态下的质量指标和目标质量指标之间的误差；E_k表示注塑产品在当前状态下的质量指标和目标质量指标之间的误差；E_{k-1}表示上一个状态下的质量指标和目标质量指标的差值；$R(\boldsymbol{s},\boldsymbol{a})$表示在状态$\boldsymbol{s}$下采取动作$\boldsymbol{a}$的奖励；$\mathrm{tol}_{\mathrm{ba}}$和$\mathrm{tol}_{\mathrm{sa}}$是两个容差参数，$\mathrm{tol}_{\mathrm{sa}}$代表可以接受的质量指数误差范围，$\mathrm{tol}_{\mathrm{ba}}$确保每次迭代的收敛性，当误差低于这个范围时，算法训练会提前终止；h是一个大于1的常数，当算法提前终止时会给出相对更大的正奖励；w_1和w_2是平衡相对奖励和绝对奖励的权重参数。

奖励函数也可以定义为多个目标。例如最小化收缩率和获得期望的产品质量，可以根据式(7)分别计算奖励r_1和r_2，最后总奖励可以定义为$r=\lambda r_1+\mu r_2$，其中λ和μ是平衡2个目标参数的加权参数。

2.2.4 转移概率

转移概率定义了智能体在采取特定动作后如何从一种状态过渡到另外一种状态。在注塑问题中，本文假设注塑过程状态是确定的，表示注塑智能体在执行一个动作后直接从状态$\boldsymbol{s}_k$转移到新状态$\boldsymbol{s}_{k+1}$。

2.2.5 目标函数

本文的目标是找到一组最优的IMP参数组合，使得累计奖励最大化。因此目标函数可以定义为总奖励的最大化：

$$\max J(\theta)=\sum_{k=0}^{\mathrm{epochs}}\gamma^k R(\boldsymbol{s}_k,\boldsymbol{a}_k)\tag{9}$$

式中：$J(\theta)$为目标函数，θ为IMP参数，epochs为迭代次数，γ为折扣因子。

2.2.6 Q表设计

本文使用Q表存储Q值，Q表中的每个表项$Q(s,a)$表示在状态$\boldsymbol{s}$中采取行动$\boldsymbol{a}$后通过式(4) 计算得到的Q值，同时Q表使用式(4) 中的更新规则进行

更新。这种方法使智能体能够通过与环境的交互学习来得到最佳的注塑工艺参数组合，以获得所期望的产品质量。该方法的关键在于定义结构良好的状态空间、动作空间、奖励函数，并选择合适的Q-learning算法进行训练。图1展示了状态空间、动作空间和Q表的设计。最终，利用学习到的Q值来选择最优的参数组合，从而加强注塑过程的控制和优化。

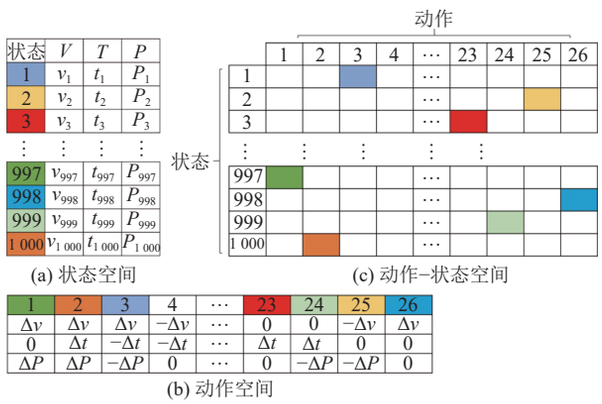


图 1　离散空间设计

Fig.1　Discrete space design

假定环境为注塑成型系统，该系统通过注塑数字孪生系统进行模拟或在实际机器上实现。基于Q-learning的优化算法作为注射成型执行器的智能体在状态空间$\boldsymbol{S}$内运行，状态空间$\boldsymbol{S}$由(v,t,P)的离散域组成。在k次迭代下，如果智能体占据的状态为$\boldsymbol{s}_k$，那么它可以在动作空间$\boldsymbol{A}$执行动作$\boldsymbol{a}$过渡到新的状态$\boldsymbol{s}_{k+1}$。在这种情况下，一个动作需要在(v,t,P)上同时进行调整，并且可以表现为26种不同的组合(不包括不变的情况)，如图1(b) 所示。随后，根据在状态下评估的质量指标$D(v,t,P)$与期望的产品质量指标$D_{\text{target}}(v,t,P)$之间的差值，对智能体的行为按照式(7)进行奖励或惩罚$(R(\boldsymbol{s},\boldsymbol{a}))$，使智能体以最大化累积奖励为目标遍历状态空间。每个状态–动作对的价值通过存储在Q表中的Q值来量化，该Q值作为状态–动作空间的综合表示，如图1(c) 所示。

2.3　训练及学习过程

Q-learning的学习过程包括智能体与环境的交互、动作选择策略(通常用ε-贪婪策略) 和Q值的更新，如图2所示。本文用强制ε-贪婪策略代替了传统ε-贪婪策略算法。强制ε-贪婪策略是传统ε-贪婪策略算法的扩展，它规定了在给定状态下每个动作必须执行的最小次数，当达到某个状态时，强制ε-贪婪策略算法将随机选择未达到最小执行次数的动作，如果某个状态下每个动作都执行了最小次数，那么强制ε-贪婪策略算法将退化为传统ε-贪婪策略算法。这种方法可以使得智能体更加全面地对环境进行探索，从而减少陷入局部最优的可能性。

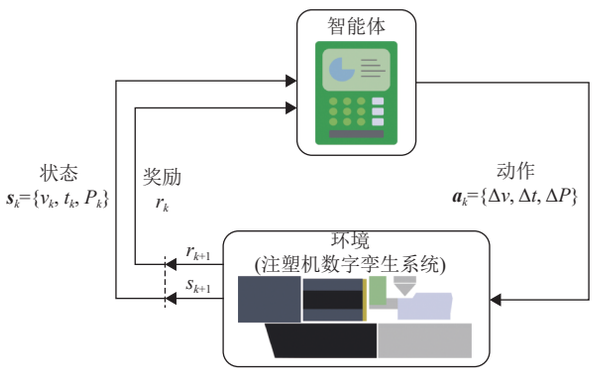


图 2　基于强化学习的IMP参数优化框架

Fig.2　Reinforcement learning-based IMP parameter optimization framework

本文使用定制的Q-learning算法，通过与环境的交互来训练智能体，以学习最优策略为目标。注塑工艺参数选择的学习过程如算法1所示。该算法从初始化超参数开始，包括了ε-贪婪策略、学习率α、折扣因子γ和动作执行空间的离散化。对于期望D_{target}，通过设置固定的步数、最大允许的迭代次数(n) 以及2个容差参数tol_{sa}和tol_{ba}来管理训练。基于tol_{sa}和tol_{ba}，智能体所采取的行动会得到奖励或惩罚(r_k)。根据式(7)计算奖励。训练完成后，可以通过选择具有最高Q值的状态–动作来获得最优参数组合。

算法1　基于Q-learning的参数优化算法

(1) 输入：$Q(\boldsymbol{s},\boldsymbol{a})=0,\forall \boldsymbol{s}\in\boldsymbol{S},\forall \boldsymbol{a}\in\boldsymbol{A}$，设置训练参数$\alpha,\gamma,\varepsilon,\text{tol}_{\text{sa}},\text{tol}_{\text{ba}},\text{episodes},n,k,w_1,w_2$。

(2) 初始化环境：随机初始化$\boldsymbol{s}_k(v_k,t_k,P_k)$，$E_{k-1}=0$，epochs=0，计算初始$E_k$。

(3) 迭代过程：每一轮episode执行以下操作直至$E_k\leqslant\text{tol}_{\text{ba}}$或 epochs$\geqslant n$；

(a) 选择动作$\boldsymbol{a}_k$：按照强制ε-贪婪策略；

(b) 执行动作$\boldsymbol{a}_k$，将状态$\boldsymbol{s}_k$转换为$\boldsymbol{s}_{k+1}$并计算E_{k+1}；

(c) 奖励计算：若$E_{k+1}\leqslant\text{tol}_{\text{ba}}$，设定奖励$r_k$：

$$r_k=hw_1+\frac{(E_{k+1}-E_k)}{\text{tol}_{\text{sa}}}w_2$$

若$\text{tol}_{\text{ba}}<E_k\leqslant\text{tol}_{\text{sa}}$，设定奖励$r_k$：

$$r_k=w_1+\frac{(E_{k+1}-E_k)}{\text{tol}_{\text{sa}}}w_2$$

否则设定奖励r_k：

$$r_k=-w_1+\frac{(E_{k+1}-E_k)}{\text{tol}_{\text{sa}}}w_2$$

(d) $E_{k-1}=E_k$；

(e) 根据Q表更新规则更新Q表：

$$Q^{\text{new}}(\boldsymbol{s}_k,\boldsymbol{a}_k)=Q(\boldsymbol{s}_k,\boldsymbol{a}_k)+\alpha\left[r_k+\gamma\max_{\boldsymbol{a}}Q(\boldsymbol{s}_{k+1},\boldsymbol{a})-Q(\boldsymbol{s}_k,\boldsymbol{a}_k)\right]$$

(f) 更新状态：$v_k = v_{k+1}, t_k = t_{k+1}, P_k = P_{k+1}$；

(g) epochs = epochs + 1。

(4) 选择最佳参数组合：

(a) $Q_{max} = \max(Q(s,a))$；

(b) 根据Q_{max}得到对应的最优组合(s_{opt}, a_{opt})；

(c) 根据(s_{opt}, a_{opt})得到最优组合参数$(v_{opt}, t_{opt}, P_{opt})$。

图3展示了本文研究的整体框架。v_i、t_i、P_i表示在状态空间随机初始化的初始状态，v_k、t_k、P_k表示当前参数状态。符号v_o、t_o、P_o表示最终输出的最优参数集。D_i是从拟合网络中得到的初始产品质量指标，而D_k表示当前输入到模型中计算的质量指标。a_k对应于当前的行动选择，r_k是针对当前状态计算的奖励。

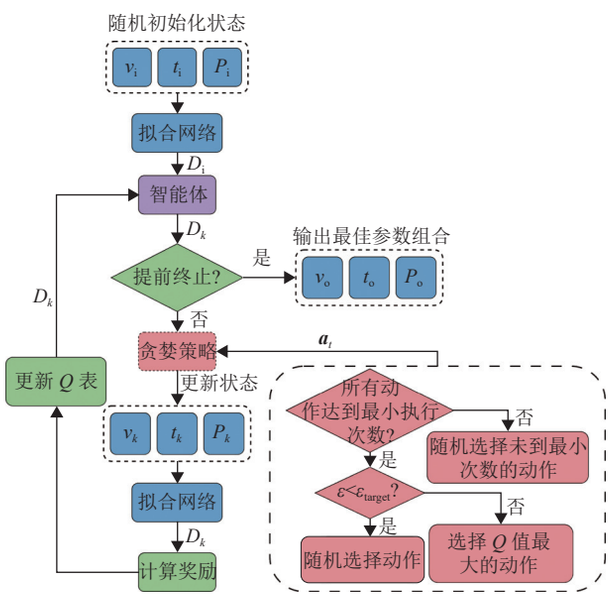


图 3 基于Q-learning的注塑参数优化框架流程图

Fig.3 The overall framework flow chart of injection molding parameter optimization based on Q-learning

3 实验结果与分析

本文提出的算法训练与仿真实验均运行在配置为Intel Core i7-10700，32 GB内存的Windows计算机平台上，使用Python/PyTorch编写基于Q-learning的IMP参数优化框架。本文用于训练Q-learning算法的注塑数据集主要通过注塑数字孪生系统进行采集，该系统连接到BORCHE品牌注塑机，型号为Bi260。机器与系统通过OPCUA(Open Platform Communications Unified Architecture) 协议进行数据交换。数据采集时间为2周，对应生产产品的材料是聚丙烯(Polypropylene, PP)，具体注塑产品为玻璃压块，如图4所示，该产品的质量指标为质量在100 g上下误差1%。

研究目标是确定一组最佳的工艺参数，包括v、t和P，以达到预期的产品质量。具体来说，研究目标是解决式(10)的优化问题。

$$\min_{v,t,P} \left| M(v,t,P) - M_{target}(v,t,P) \right|,$$
$$\text{s.t.} \begin{cases} v_{min} \leqslant v \leqslant v_{max} \\ t_{min} \leqslant t \leqslant t_{max} \\ P_{min} \leqslant P \leqslant P_{max} \end{cases} \tag{10}$$

式中：$M(v,t,P)$表示当前产品质量，$M_{target}(v,t,P)$为产品的目标质量。假设IMP中其他参数设置已经提前预设好。

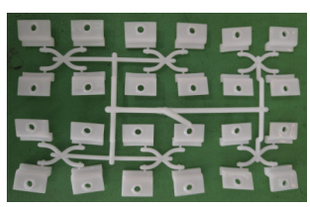


图 4 聚丙烯玻璃压块

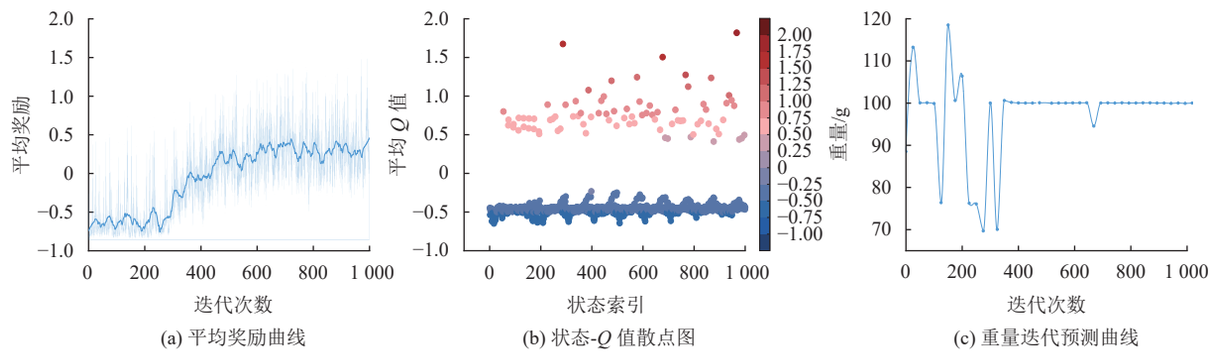
Fig.4 Polypropylene glass press block

3.1 Q-learning的有效性实验

本实验利用本文设计的Q-learning算法对离散化为1 000个状态的状态空间进行仿真实验。目标是利用本文设计的Q-learning 算法框架去寻找最优的参数组合(v,t,P)，以实现期望的注塑产品目标质量$M_{target}(v,t,P)$(100 g)。为此，本文将待优化的工艺参数(v,t,P) 各自离散为10个状态。在超参数$\alpha = 0.25$、$\gamma = 0.25$和$\varepsilon = 0.25$的情况下，经过算法的迭代，得到了较为满意的实验结果，如图5所示。

图5(a)显示了注塑产品目标质量的参数优化实验在1 000轮结束时的结果。可以观察到，在约前250轮中，由于采用了强制ε-贪婪策略，Q-learning算法处于纯探索状态，250轮以后平均奖励逐渐上升，并经过一定迭代次数后逐渐收敛。图5(b)展示了每个状态在算法过程中累计Q值的散点图，颜色越红代表Q值越大，越蓝代表Q值越小。将散点图映射到状态空间，可以直观地观察到最优的参数组合。如表1所示，当v = 10 cm/s、t = 12.0 s、P = 8.5 MPa时，得到了最接近目标质量(100 g) 的最优参数组合，其预测质量为100.007 g，误差百分比为0.007%。同时，还观察到其他具有较高Q值且满足需求的参数组合，例如v = 7 cm/s、t = 11.0 s、P = 8.5 MPa，其预测质量为100.322 g，误差百分比为0.322%，以及v = 5 cm/s、t = 12.0 s、P = 8.5 MPa，其预测质量为100.214 g，误差百分比为0.214%。图5(c) 展示了算法的探索过程。前期曲线反映了算法的探索阶段，在强制ε-贪婪策略

的驱使下，智能体在目标周围波动，有一定的试错过程。随着经验的积累，算法模型逐渐稳定于一个满足要求的结果。此外在模型训练结束后，根据实验统计，在该环境下针对质量问题，平均计算时间约为19 ms。综上所述，Q-learning算法用于解决注塑工艺产品的目标质量预测参数组合寻优问题是具有逻辑性且有效的。



(a) 平均奖励曲线　(b) 状态-Q 值散点图　(c) 重量迭代预测曲线

图 5　基于Q-learning最优实验结果图

Fig.5　Q-learning based optimal experimental results graphs

表 1　针对质量的最优参数组合

Table 1　Optimized parameter combinations for weight

v/(cm·s^{-1})	t/s	P/MPa	M_{target}/g	M/g
10	12.0	8.5	100	100.007

3.2　超参数对算法的影响实验

在Q-learning算法中，主要有4个超参数对算法的行为有影响分别是ε参数、折扣因子γ、学习率α和迭代次数。ε控制了算法在探索和利用之间的平衡。γ决定了未来奖励的重要性。α控制了智能体在更新Q值时对新奖励的重视程度。迭代次数决定了算法运行的计算成本。

本节将介绍每个参数对算法行为的影响。在所有情况下，默认参数ε、γ和α均为0.25，迭代次数为1 000，除非特别指定。

3.2.1　折扣因子γ的影响实验

折扣因子γ在强化学习中用于权衡未来奖励与即时奖励，通常取值在0到1之间。当γ值越大时，智能体更加关注未来奖励，更愿意采取可以在未来获得更大奖励的策略。相反，当γ越小时，智能体越贪婪，更倾向于选择即时奖励较大的策略。本文在实验中测试了不同γ值下的实验结果，如图6所示。

从图6(a)可以看出，在不同的γ值下，收缩率预测的平均奖励的收敛曲线变化并不明显。这表明γ与平均奖励的增减趋势没有明显的关联性。在图6(b) ~(e)中，可以观察到随着γ越大，Q值分布更加扩散，中间值比例增多，这表明更大的γ值会使智能体更倾向于选择质量价值更高的状态。特别是在$\gamma=1$的情况下，Q值远大于其他3个γ值，这表明γ取较大值时，Q值的更新受未来状态的价值影响较大，同时导致Q值相对较高。实验结果显示，在γ别为0.25、0.5、0.75和1时，平均质量误差分别为0.007%、0.067%、0.624%和0.510%，平均质量分别为100.007 g、100.067 g、100.624 g和100.510 g。其中，γ值为0.25时的最优参数组合表现出最小的质量误差，其平均预测质量M为100.007 g，最接近目标质量(100 g)。因此，选择γ值为0.25 是合理的。

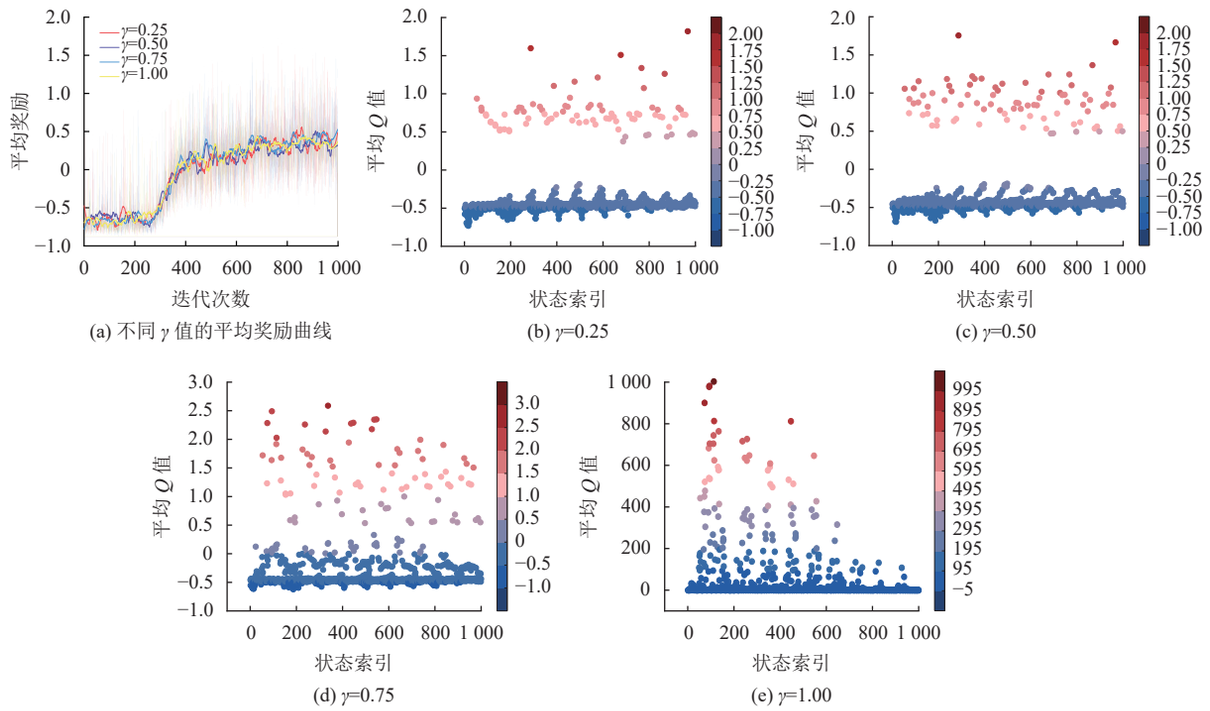
3.2.2　学习率α的影响实验

学习率α定义了旧Q值对新Q值的权重，其值通常设置在0到1之间，其数值的大小直接影响着旧Q值对新信息的学习程度。较大的α值意味着在Q值更新中智能体更加重视即时信息，使得算法更具灵活性，可以更敏锐地响应环境的变化。相反，较小的α值则表明算法更倾向于依赖先验知识，智能体更为保守。图7展示了不同学习率α值下的实验结果。

从图7(a)中可以观察到，不同α值对于目标质量预测的平均奖励曲线的收敛变化影响不明显。从图7(b) ~(e)中可以看出，随着α值的增大，Q值在状态空间中呈现更为扩散的趋势，这表明较大的α导致智能体更依赖于当前状态的质量信息，从而过快地遗忘了历史经验。实验结果表明，α值为 0.25 时，获得了接近目标质量的结果，平均预测质量M为100.021 g，平均误差为0.021%。相比之下，α值为0.5、0.75和1时的实验误差结果分别为0.022%、0.046%、0.062%，平均质量分别100.022 g、100.046 g和100.062 g，α值为0.25 时算法展现了最优的性能。因此，选择α值为0.25 是合理的。

3.2.3　超参数ε的影响实验

超参数ε用于权衡算法的探索与利用。较高的ε值会使智能体更倾向于探索新的动作而不是根据

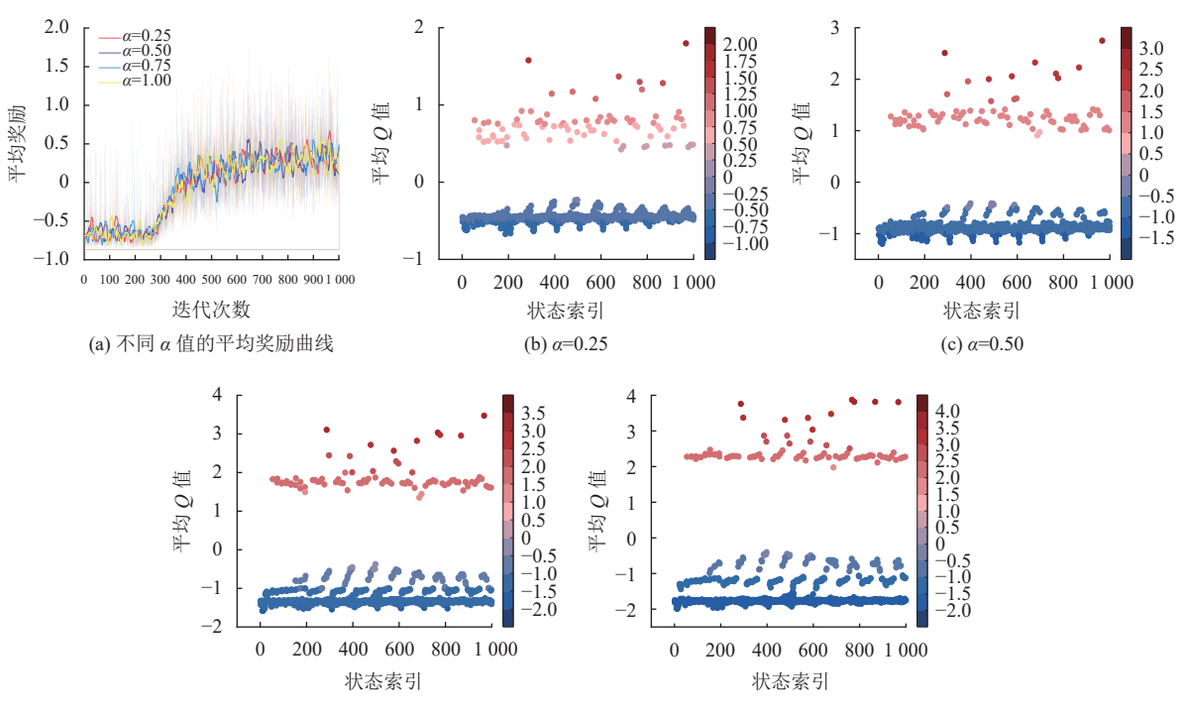


(a) 不同 γ 值的平均奖励曲线 (b) γ=0.25 (c) γ=0.50

(d) γ=0.75 (e) γ=1.00

图 6 不同γ值下的实验结果

Fig.6 Comparison of experimental results at differentγvalues



(a) 不同 α 值的平均奖励曲线 (b) α=0.25 (c) α=0.50

(d) α=0.75 (e) α=1.00

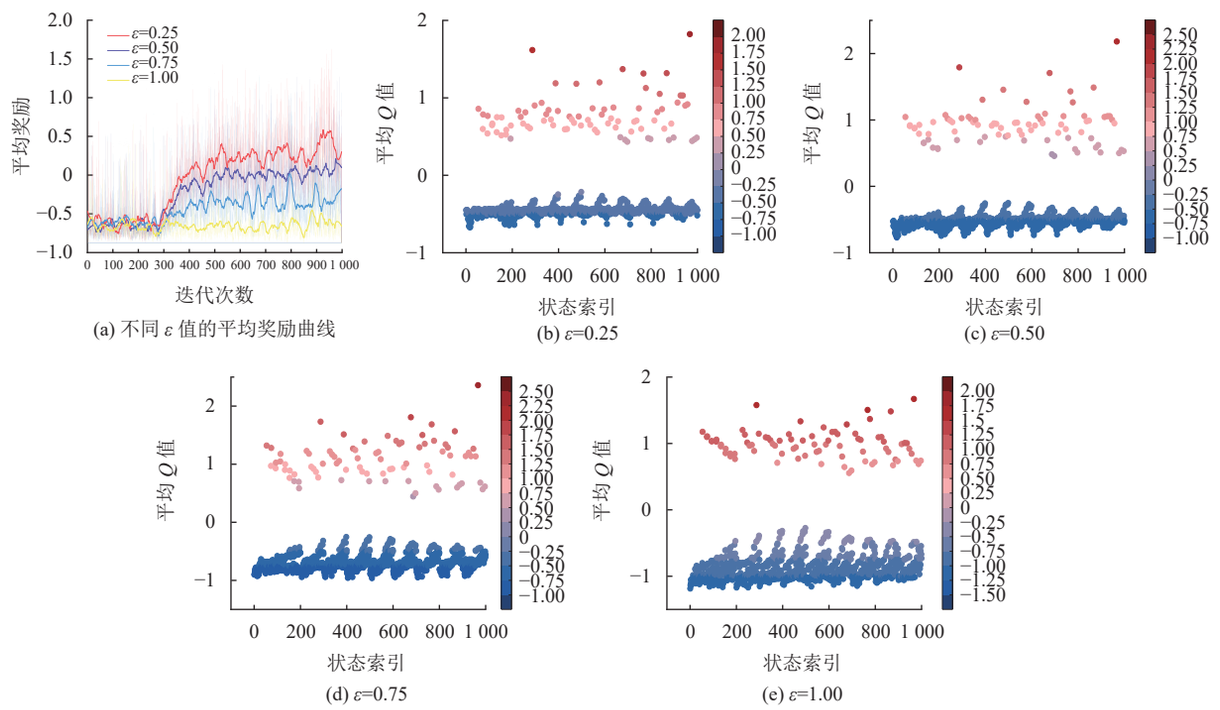
图 7 不同α值下的实验结果

Fig.7 Comparison of experimental results at differentαvalues

经验选择最佳动作，从而更好地估计多个动作的期望奖励，但这可能会使得智能体失去选择最优动作的机会。相反，较低的ε值会使得智能体更倾向于选择已知的最优动作，但这可能会导致算法陷入局部最优，从而失去探索更好状态的机会。图8展示了不同超参数ε值下的实验结果。

从图8(a)中可以观察到，ε值越大，平均奖励曲线奖励值越小，增长趋势和收敛速度越慢。相反，ε值越

小，曲线收敛速度和增长趋势越快。这是因为较低的ε值使得算法更倾向于选择已知的最接近目标质量的状态。在ε值为1的情况下，平均奖励曲线呈现相对平坦的趋势，说明纯探索策略在大部分情况下难以使算法选择到更高价值的状态，除此之外其他策略在不同程度上利用了过去的经验，因此有机会选择到更高价值的状态。从图8(b) ~(e)可以观察到，随着ε值的增加，不同状态的Q值更为分散，表明较大的ε值能够更全面地估计每个质量状态的期望值，相对地，对于选择较高价值的状态的倾向就不那么明显。在ε分别为0.25、0.5、0.75 和1时，平均质量误差分别为 0.007%、0.076%、0.029%和0.043%，预测质量分别为100.007 g、100.076 g、100.029 g和100.043 g。当ε = 0.25时，收敛速度最快，增长趋势最明显，且平均质量最接近目标质量，因此选择将ε值设定为0.25。



(a) 不同 ε 值的平均奖励曲线　(b) ε=0.25　(c) ε=0.50　(d) ε=0.75　(e) ε=1.00

图 8　不同ε值下的实验结果

Fig.8　Comparison of experimental results at differentεvalues

3.2.4　迭代次数的影响实验

迭代次数对于Q-learning模型的效果也具有一定的影响。由于采用了强制ε-贪婪策略，模型在初始阶段约为250次迭代中处于纯探索状态，而在经过约250次迭代后，动作选择策略转变为传统的ε-贪婪策略。因此，只有在经过一定数量的迭代后才能观察到迭代次数对算法效果的影响。在不考虑算法计算成本的情况下，无法确保当前得到的结果是最优的迭代次数，只能通过对实验结果的分析以及观察每个状态的Q值分布情况，在有限的迭代次数内选择一个相对合适的值。图9展示了不同迭代次数下的实验结果。

从图9(a)~(e)中可以观察到，在750 轮迭代后，每个状态的Q值分布趋于一致，没有显著的区别，证明在1 000次的训练迭代后，算法模型已经趋于相对稳定的状态。在迭代次数大于1 000后，最终预测质量的细微差别可能存在一些偶然性，可以忽略。实验发现在训练周期为250时，算法得到的最优质量误差为0.341%，预测质量为100.341 g，满足了需求，证明在较少的迭代次数内，该算法模型能够找到满足质量目标的参数组合。500次迭代后，结果得到显著提升，迭代次数为500、750和1 000时，预测质量平均误差分别为0.036%、0.048%和0.007%，预测质量分别为100.036 g、100.048 g和100.007 g，迭代次数为1 250和1 500 时，预测质量平均误差为0.034%，预测质量为100.034 g。实验证明，在1 000次迭代后，算法已经取得了令人满意的结果，进一步增加迭代次数并不会显著改善性能，还可能增加计算成本。这表明，本实验的Q-learning 算法框架在不同的环境下，合理的迭代次数可以在保证性能的同时有效地减少计算资源的开销。

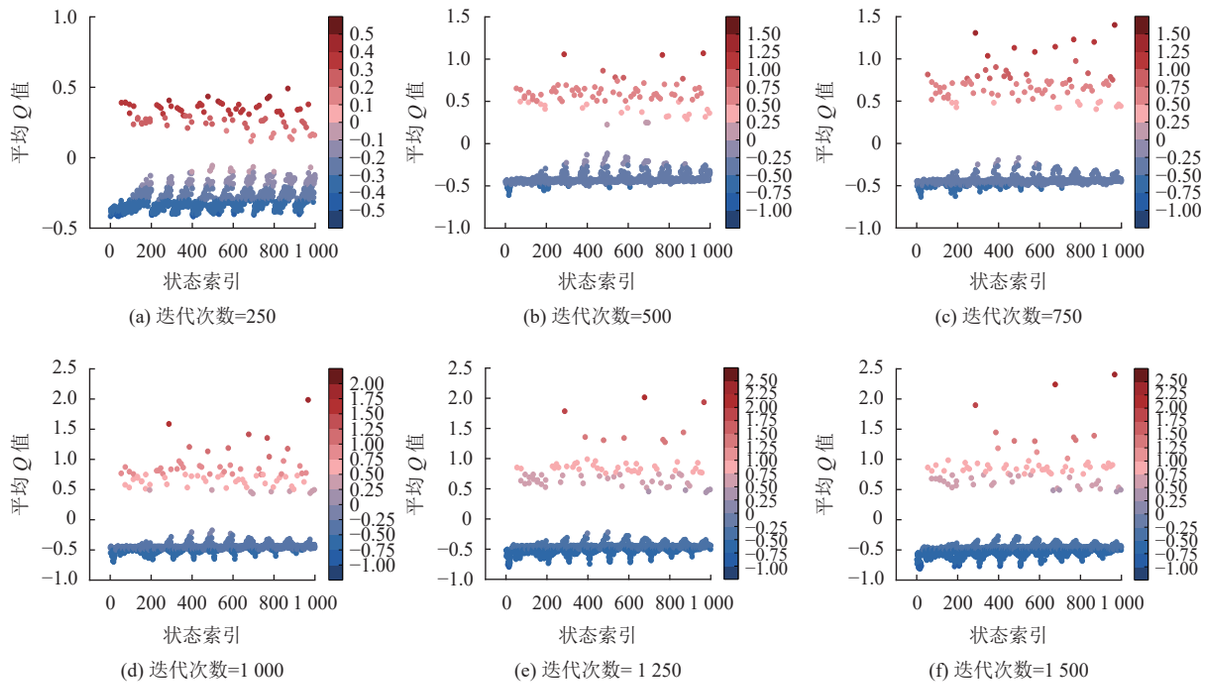


图 9 不同迭代次数下实验结果比较

Fig.9 Comparison of experimental results at different epochs

4 结论

本文首次探索了基于Q-learning的RL算法在IMPPT中的应用，提出了基于RL的自动组合IMP参数的方法。实验结果表明，基于Q-learning的策略在工艺参数选择上具有有效性。因此，RL方法为解决IMPPT中固有的计算复杂性提供了一种有前景的解决方案。

鉴于RL的最新进展，未来研究可以关注以下几个方向。首先，离散参数空间的使用限制了对设计空间内特定组合的访问，这种限制可能会导致有价值的组合丢失。为此，可采用近似策略优化等方法[32]，扩展至连续参数空间。值得注意的是，对于有效的离散化采样空间仍然很有价值，尤其是在处理有限信息时。其次，通过更合理的初始Q表设置，融合专家知识，能够加速优化过程。在实际应用中，制造商可通过赋予高初始Q值的参数组合来引导搜索，从而加速最优解的寻找。然而，当状态或动作空间非常大时，Q表的维护和更新成本过高。为此，未来研究可以考虑采用DQN(Deep Q-Network) 方法[33]，通过神经网络取代传统Q表，从而提高高维空间的处理效率。同时，还可以应用多智能体强化学习[34]，将每个参数视为一个独立的智能体，并赋予其独立的Q表，进一步解决高维问题。

参考文献：

[1] ZHENG R, TANNER R I, FAN X J. Injection molding: integration of theory and modeling methods[M]. Berlin: Springer Science & Business Media, 2011.

[2] ZHOU H M, HU Z X , LI D Q.Mathematical models for the filling and packing simulation[J/OL]. Computer Modeling for Injection Molding: Simulation, Optimization, and Control(2013-02-22)[2024-11-01]. https://onlinelibrary.wiley.com/doi/10.1002/9781118444887.ch3.

[3] 余世浩, 何星明, 张国英. 基于响应面模型和NSGA-Ⅱ算法的注塑成型工艺优化[J]. 塑性工程学报, 2014, 21(3): 15-19.
YU S H, HE X M, Z G Y. Optimization of injection molding process based on response surface model and NSGA-II algorithm[J]. Journal of Plasticity Engineering, 2014, 21(3): 15-19.

[4] KHOSRAVANI M R, NASIRI S. Injection molding manufacturing process: review of case-based reasoning applications[J]. Journal of Intelligent Manufacturing, 2020, 31: 847-864.

[5] OZCELIK B, ERZURUMLU T. Determination of effecting dimensional parameters on warpage of thin shell plastic parts using integrated response surface method and genetic algorithm[J]. International Communications in Heat and Mass Transfer, 2005, 32(8): 1085-1094.

[6] OKTEM H, ERZURUMLU T, UZMAN I. Application of Taguchi optimization technique in determining plastic injection molding process parameters for a thin-shell part[J]. Materials & Design, 2007, 28(4): 1271-1278.

[7] SHEN C, WANG L, LI Q. Optimization of injection molding process parameters using combination of artificial neural network and genetic algorithm method[J]. Journal of Materials Processing Technology, 2007, 183(2-3): 412-418.

[8] ALTAN M. Reducing shrinkage in injection moldings via

the Taguchi, ANOVA and neural network methods[J]. Materials & Design, 2010, 31(1): 599-604.

[9] WANG G, ZHAO G, LI H, *et al*. Research on optimization design of the heating/cooling channels for rapid heat cycle molding based on response surface methodology and constrained particle swarm optimization[J]. Expert Systems with Applications, 2011, 38(6): 6705-6719.

[10] XU Y, ZHANG Q W, ZHANG W, *et al*. Optimization of injection molding process parameters to improve the mechanical performance of polymer product against impact[J]. The International Journal of Advanced Manufacturing Technology, 2015, 76: 2199-2208.

[11] TSAI K M, LUO H J. An inverse model for injection molding of optical lens using artificial neural network coupled with genetic algorithm[J]. Journal of Intelligent Manufacturing, 2017, 28: 473-487.

[12] MOAYYEDIAN M, ABHARY K, MARIAN R. Optimization of injection molding process based on fuzzy quality evaluation and Taguchi experimental design[J]. CIRP Journal of Manufacturing Science and Technology, 2018, 21: 150-160.

[13] ALAM S T, AMIN M A. Determining optimum design parameters of foldable product using response surface methodology and genetic algorithm[J]. Engineering, 2020, 12(11): 839-850.

[14] 胡雷, 高煌, 李阳, 等. 一种基于分类模型的注塑工艺参数优化方法[J]. 塑料工业, 2017, 45(3): 74-78.
HU L, GAO H, LI Y, *et al*. An optimization method of injection molding process parameters based on classification model[J]. Plastics Industry, 2017, 45(3): 74-78.

[15] RIBEIRO B. Support vector machines for quality monitoring in a plastic injection molding process[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) , 2005, 35(3): 401-410.

[16] KAISER L, BABAEIZADEH M, MILOS P, *et al*. Model-based reinforcement learning for Atari [EB/OL]. arXiv: 1903.00374 (2019-03-01) [2024-03-11]. https://arxiv.org/abs/1903.00374.

[17] YE D, CHEN G, ZHANG W, *et al*. Towards playing full moba games with deep reinforcement learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 621-632.

[18] KOBER J, BAGNELL J A, PETERS J. Reinforcement learning in robotics: a survey[J]. The International Journal of Robotics Research, 2013, 32(11): 1238-1274.

[19] WANG N, GAO Y, ZHAO H, *et al*. Reinforcement learning-based optimal tracking control of an unknown unmanned surface vehicle[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(7): 3034-3045.

[20] ZHAO Z, HE W, MU C, *et al*. Reinforcement learning control for a 2-DOF helicopter with state constraints: theory and experiments[J]. IEEE Transactions on Automation Science and Engineering, 2022, 21(1): 157-167.

[21] 董豪, 杨静, 李少波, 等. 基于深度强化学习的机器人运动控制研究进展[J]. 控制与决策, 2022, 37(2): 278-292.
DONG H, YANG J, LI S B, *et al*. Research progress of robot motion control based on deep reinforcement learning[J]. Control and Decision, 2022, 37(2): 278-292.

[22] AFSAR M M, CRUMP T, FAR B. Reinforcement learning based recommender systems: a survey[J]. ACM Computing Surveys, 2022, 55(7): 1-38.

[23] CHEN X, YAO L, MCAULEY J, *et al*. Deep reinforcement learning in recommender systems: a survey and new perspectives[J]. Knowledge-based Systems, 2023, 264: 110335.

[24] LIU C, DING J, SUN J. Reinforcement learning based decision making of operational indices in process industry under changing environment[J]. IEEE Transactions on Industrial Informatics, 2020, 17(4): 2727-2736.

[25] OLIFF H, LIU Y, KUMAR M, *et al*. Reinforcement learning for facilitating human-robot-interaction in manufacturing[J]. Journal of Manufacturing Systems, 2020, 56: 326-340.

[26] LENG J, RUAN G, SONG Y, *et al*. A loosely-coupled deep reinforcement learning approach for order acceptance decision of mass-individualized printed circuit board manufacturing in industry 4.0[J]. Journal of Cleaner Production, 2021, 280: 124405.

[27] 唐振韬, 邵坤, 赵冬斌, 等. 深度强化学习进展: 从 AlphaGo 到 AlphaGo Zero[J]. 控制理论与应用, 2017, 34(12): 1529-1546.
TANG Z T, SHAO K, ZHAO D B, *et al*. Recent progress of deep reinforcement learning: from AlphaGo to AlphaGo Zero[J]. Control Theory and Technology, 2017, 34(12): 1529-1546.

[28] 李明磊, 章阳, 康嘉文, 等. 基于多智能体强化学习的区块链赋能车联网中的安全数据共享[J]. 广东工业大学学报, 2021, 38(6): 62-69.
LI M L, ZHANG Y, KANG J W, *et al*. Secure data sharing in blockchain-enabled vehicular networks based on multi-agent reinforcement learning[J]. Journal of Guangdong University of Technology, 2021, 38(6): 62-69.

[29] 郭心德, 丁宏强. 离散制造智能工厂场景的 AGV 路径规划方法[J]. 广东工业大学学报, 2021, 38(6): 70-76.
GUO X D, DING H Q. An AGV path planning method for discrete manufacturing smart factory[J]. Journal of Guangdong University of Technology, 2021, 38(6): 70-76.

[30] DAYAN P, WATKINS C. Q-learning[J]. Machine Learning, 1992, 8(3): 279-292.

[31] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 2018.

[32] SCHULMAN J, WOLSKI F, DHARIWAL P, *et al*. Proximal policy optimization algorithms [EB/OL]. arXiv: 1707.06347 (2017-07-20) [2024-03-11]. https://arxiv.org/abs/1707.06347.

[33] MNIH V, KAVUKCUOGLU K, SILVER D, *et al*. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[34] ZHANG K, YANG Z, BAŞAR T. Multi-agent reinforcement learning: a selective overview of theories and algorithms[M/OL]. Handbook of Reinforcement Learning and Control, (2021-06-24)[2024-11-20]. 2021: 321-384.

（责任编辑：张玮欣　英文审核：熊荣斌）