DOI:10.11918/201908012

应用深度强化学习的压边力优化控制

张新艳, 郭 鹏, 余建波

(同济大学 机械与能源工程学院, 上海 201804)

摘 要: 为改善板料拉深制造的成品质量,采用深度强化学习的方法进行拉深过程的压边力优化控制. 提出一种基于深度强化学习与有限元仿真集成的压边力控制模型,结合深度神经网络的感知能力与强化学习的决策能力,进行压边力控制策略的学习优化. 基于深度强化学习的压边力优化算法,利用深度神经网络处理巨大的状态空间,避免了系统动力学的拟合,并且使用一种新的网络结构来构建策略网络,将压边力策略划分为全局与局部两部分,提高了压边力策略的控制效果. 将压边力的理论知识用于初始化回放经验池,提高了深度强化学习算法在压边力控制任务中的学习效率. 实验结果表明,与传统深度强化学习算法相比,所提出的压边力控制模型能够更有效地进行压边力控制策略优化,成品在内部应力、成品厚度以及材料利用率3个质量评价指标的综合表现优于传统深度强化学习算法. 将深度强化学习中的策略网络划分为线性部分与非线性部分,并结合理论压边力知识来初始化回放经验,能够提高深度强化学习在压边力优化控制中的控制效果,提高算法的学习效率.

关键词: 板材拉深成形;质量控制;深度强化学习;有限元仿真;优化控制

中图分类号: TG301 **文献标志码**: A **文章编号**: 0367-6234(2020)07-0020-09

Optimal control of blank holder force using deep reinforcement learning

ZHANG Xinyan, GUO Peng, YU Jianbo

(School of Mechanical Engineering, Tongji University, Shanghai 201804, China)

Abstract: To improve the quality of products in deep drawing process, the deep reinforcement learning method is used to optimize the blank holder force (BHF). A new BHF control model based on the integration of deep reinforcement learning and finite element simulation is proposed, and the BHF control strategy is optimized by combining the perception ability of deep neural network with the decision-making ability of reinforcement learning. The proposed control model uses the deep neural network to deal with huge state space and avoids the fitting of system dynamics. By utilizing a novel strategy network structure, the BHF control strategy is divided into global and local parts, and the control effect is improved. Meanwhile, the theoretical knowledge of BHF is used to initialize the replay experience, which improves the learning efficiency of deep reinforcement learning algorithm in BHF control tasks. Experiments show that the proposed BHF control model can optimize BHF control strategy more effectively than traditional deep reinforcement learning algorithm. The comprehensive performance of the proposed control model in three quality indicators (internal stress, thickness and material utilizing rate) is better than that of the traditional deep reinforcement learning algorithms.

Keywords: deep drawing; quality control; deep reinforcement learning; finite element analysis; optimal control

板材拉深成形作为一种基础零部件制造工艺,被广泛应用于汽车、机电、轻工和航空航天等诸多领域. 拉深成形通过压边力(blank holder force,BHF)来控制金属材料的流动,从而影响最终成品的成形质量. 在拉深过程中采用恒定的压边力容易导致起皱与破裂等质量缺陷,因此在拉深过程中合理地控制压边力参数就成为防止起皱、破裂和提高成品质量的重要手段之一.

收稿日期: 2019-08-02
基金项目: 国家自然科学基金(51375290)
作者简介: 张新艳(1974—),女,讲师;
余建波(1978—),男,教授,博士生导师
通信作者: 余建波,jbyu@tongji.edu.cn

在压边力控制领域,最优化理论与有限元模拟相结合是一类常用的方法. Ghouati 等[1]提出了网格法与单纯形法相结合的优化方法,其计算效率高,能够有效减少有限元仿真次数,但无法保证所求的解在可行域内. 包友霞等[2]在 Ghouati 提出的优化方法的基础上进行改进,使得优化过程中各变量点始终保持在可行域内,保证了解的可行性. 孙成智等[3]提出了一种集成了有限元模拟与自适应响应面法(adaptive response surface method,ARSM)的优化设计方法,并且应用信赖域模型管理来调节设计空间的变化,保证优化过程的收敛. Hillmann 等[4]将成形极限图上各点到成形极限和起皱极限距离的加权和作为目标函数,以压边力作为设计变量,在有

限元仿真环境下采用 BFGS 优化方法对压边力进行优化. Scott 等[5]以极限应变作为目标函数,以压边力作为设计变量,在 ABAQUS 仿真环境下利用灵敏度分析方法对盒形件进行优化. 以上方法准确性较高,但数值模拟速度无法满足优化迭代要求,限制了方法的使用,并且最佳压边力搜索方向也难以确定.

神经网络被广泛应用于处理压边力控制问题中的非线性关系. Senn 等[6]采用近似动态规划方法来进行压边力控制,利用神经网络来拟合系统动力学以及价值函数. 黄玉萍等[7]通过建立径向基网络,以应力、应变和减薄率作为输入,压边力曲线作为输出,构建了压边力优化模型. Qian 等[8]和 Manabe 等[9]利用神经网络进行材料参数和工艺参数的在线识别,并结合弹塑性理论预测压边力大小. 汪锐等[10]通过将模糊控制技术与神经网络相结合,构建模糊神经网络专家系统来进行压边力的智能控制.

传统的压边力控制方法往往需要对拉深过程进行建模或依赖一些先验知识. Dornheim 等[11]提出了一种无模型的压边力控制方法,避免了系统模型的拟合. 该方法将神经拟合 Q 迭代(neural fitted Q iteration, NFQ)算法与有限元仿真相结合,为每个控制步长建立一个 Q 值网络. 然而 NFQ 是一种基于价值的强化学习算法,只能用于离散动作空间的控制问题,无法用于连续动作空间的控制问题. 综合以上分析,目前压边力控制领域还存在难以获得精确动力学模型以及压边力控制效果无法达到最优化的问题.

本文提出了一种基于深度强化学习的压边力优化控制模型,提高了压边力的控制效果;引入一种新的策略网络结构,进一步提高了深度强化学习在压边力控制任务中的控制效果;将压边力理论知识引入网络训练中,用理论压边力公式进行回放经验池的初始化,提高了压边力策略的学习效率;以一个圆筒件的拉深成形过程为分析对象,通过有限元仿真验证了本文提出的压边力优化控制模型的有效性.

1　理论背景

1.1　马尔科夫决策过程

强化学习(reinforcement learning, RL)问题一般由马尔科夫决策过程(Markov decision process, MDP)进行建模[12]. 通常将 MDP 定义成一个四元组 (S,A,r,p), 其中:1) S 为所有系统状态集合, $s_t \in S$ 表示智能体(agent)在时刻 t 的系统状态;2) A 为动作集合, $a_t \in A$ 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. $p(s_{t+1} \mid s_t,a_t)$ 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.

在强化学习中,定义策略 $\pi:S \to A$ 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 $r(s_t,a_t)$ 并转移到下一状态 s_{t+1}. 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:

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

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

1.2　强化学习

强化学习的目标是寻找到一个最优策略 π_ϕ(参数为 ϕ)来最大化期望回报值 $J(\phi) = E_{s_i \sim p_\pi, a_i \sim \pi}[R_0]$ [13]. 在行动者-评论家(actor-critic)框架中,策略网络(actor)通过确定性策略梯度[14](deterministic policy gradient, DPG)进行网络更新:

$$\nabla_\phi J(\phi) = E_{s \sim p_\pi}[\nabla_a Q^\pi(s,a) \mid_{a=\pi(s)} \nabla_\phi \pi_\phi(s)],$$

式中: $Q^\pi(s,a) = E_{s_i \sim p_\pi, a_i \sim \pi}[R_t \mid s,a]$ 为动作值函数(critic),表示在遵循策略 π 情况下,在状态 s 采取动作 a 后的期望回报值.

Q 学习(Q-learning)使用时间差分算法进行动作值函数的学习,通过迭代贝尔曼方程求解 Q 函数:

$$Q^\pi(s_t,a_t) = r(s_t,a_t) + \gamma E_{s_{t+1},a_{t+1}}[Q^\pi(s_{t+1},a_{t+1})],$$
$$a_{t+1} \sim \pi(s_{t+1}).$$

对于巨大的状态空间,通常使用一个可微的函数近似器 $Q_\theta(s,a)$ 估计动作值,其参数为 θ. 深度 Q 学习[15](Deep Q-learning, DQN)算法采用了“目标网络”技术,在更新过程中使用另一个网络 $Q_{\theta'}(s,a)$ 计算目标值:

$$y_t = r(s_t,a_t) + \gamma Q_{\theta'}(s_{t+1},a_{t+1}), a_{t+1} \sim \pi_{\phi'}(s_{t+1}).$$

式中动作 a_{t+1} 根据目标策略网络 $\pi_{\phi'}$ 进行选择. 获得目标值后,DQN 通过最小化损失函数 $L(\theta)$ 进行动作值网络参数的更新:

$$L(\theta) = E_{s_t,a_t,r(s_t,a_t),s_{t+1}}[(y_t - Q_\theta(s_t,a_t))^2].$$

2　基于深度强化学习的压边力控制策略优化算法

通过将双延迟深度确定性策略梯度[16](twin delayed deep deterministic policy gradient, TD3)与结构化控制网络[17](structured control network, SCN)相结合,本文提出了 SCN-TD3 算法用于压边力控制策略的学习.

2.1 双延迟深度确定性策略梯度

TD3 是一种 actor-critic 框架的深度强化学习算法,在深度确定性策略梯度[18](deep deterministic policy gradient,DDPG)的基础上拓展而来. 为了解决 actor-critic 框架算法中的 Q 值过估计问题,TD3 采用 3 个关键技术提高算法的稳定性和性能.

1)actor-critic 框架下的剪裁双 Q 学习.受深度双 Q 学习[19](double deep Q-learning, DDQN)启发,TD3 使用当前 actor 网络选择最优动作,使用目标 critic 网络评估策略:

$$y_t = r(s_t, a_t) + \gamma Q_{\theta'}(s_{t+1}, \pi_{\phi}(s_{t+1}))\,.$$

在 actor-critic 框架中,目标 actor 与目标 critic 网络采用的"软更新"[18]方式使得当前网络与目标网络过于相似,无法有效分离动作选择与策略评估. 因此,算法保持了一对 actor 网络(π_{ϕ_1},π_{ϕ_2})和一对 critic 网络(Q_{θ_1},Q_{θ_2}). 其中,π_{ϕ_1} 根据 Q_{θ_1} 进行优化,π_{ϕ_2} 根据 Q_{θ_2} 进行优化:

$$\begin{cases} y_t^1 = r(s_t, a_t) + \gamma Q_{\theta_2'}(s_{t+1}, \pi_{\phi_1}(s_{t+1}))\,, \\ y_t^2 = r(s_t, a_t) + \gamma Q_{\theta_1'}(s_{t+1}, \pi_{\phi_2}(s_{t+1}))\,. \end{cases} \tag{1}$$

如果 critic 网络 Q_{θ_1} 与 Q_{θ_2} 相互独立,那么根据式(1)能有效避免由于策略更新所导致的偏差. 然而 Q_{θ_1} 与 Q_{θ_2} 在计算目标值时互相使用,并且基于相同的回放经验进行更新,因此两者并不互相独立. 为了进一步减小偏差,TD3 使用了剪裁双 Q 学习(Clipped Double Q-learning)算法计算目标值:

$$y_t^1 = r(s_t, a_t) + \gamma \min_{i=1,2} Q_{\theta_i'}(s_{t+1}, \pi_{\phi_1}(s_{t+1}))\,.$$

为了减少计算成本,TD3 使用了一个单独的 actor 网络以及两个 critic 网络. actor 网络 π_{ϕ} 根据 critic 网络 Q_{θ_1} 进行更新,critic 网络 Q_{θ_2} 的目标值 y_t^2 与 y_t^1 相等.

2)策略延迟更新.在深度强化学习算法中,目标网络被用于提供一个稳定学习目标. 通过多步更新,critic 网络能逐渐减小与目标 Q 值之间的误差;然而,在 critic 网络高误差情况下,进行 actor 网络的更新会导致策略的离散行为. 因此,actor 网络的更新频率应低于 critic 网络的更新频率,保证 actor 网络能在 Q 值误差较低的情况下进行更新,提高 actor 网络的更新效率. TD3 在 critic 网络每进行 d 次更新后,进行一次 actor 网络的更新.

3)目标策略平滑正则化.由于 TD3 中采用的是确定性策略,进行 critic 更新时目标值很容易受函数近似误差的影响,导致目标值不准确. 因此 TD3 引入了一个正则化方法来减少目标值的方差,通过自举相似状态动作对的估计值进行 Q 值估计平滑化:

$$y_t = r(s_t, a_t) + E_{\varepsilon}[\,Q_{\theta'}(s_{t+1}, \pi_{\phi'}(s_{t+1}) + \varepsilon)\,].$$

TD3 通过向目标策略添加一个随机噪声,并且在 mini-batches 上取平均的方法实现平滑正则化:

$$y_t = r(s_t, a_t) + \gamma \min_{i=1,2} Q_{\theta_i'}(s_{t+1}, \pi_{\phi'}(s_{t+1}) + \varepsilon)\,,$$
$$\varepsilon \sim \mathrm{clip}(N(0,\sigma)\,, -c, c)\,.$$

2.2 结构化控制网络

受传统非线性控制理论启发,文献[17]提出了结构化控制网络(structured control network,SCN),将 actor-critic 框架中的策略网络分为非线性部分与线性部分两个部分. 将上述两个部分的动作值相加得到最终动作:

$$\pi_{\phi}(s) = \pi^{\mathrm{n}}(s) + \pi^{\mathrm{l}}(s)\,.$$

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

2.3 理论压边力知识

在压边力控制领域,研究人员通过板材成形理论以及对拉深过程的简化假设,推导了圆筒件拉深过程的有效压边力区间. 通过预先确定有效压边力区间,能够得到相对合理的压边力曲线.

如图 1 所示,板材拉深过程的有效压边力范围由上限压边力 Q_{rup} 与下限压边力 Q_{fwr} 组成. Q_{rup} 表示在拉深过程中保证工件不产生破裂缺陷的最大压边力,Q_{fwr} 表示在拉深过程中保证工件法兰边不产生起皱缺陷的最小压边力. 其中,

$$Q_{\mathrm{rup}} = \frac{1}{\mu F(\alpha)}\left\{\frac{2R_B}{R_C + R_B}\sigma_{\mathrm{b}}\left(\frac{1+r}{\sqrt{1+2r}}\right)^{n+1} \cdot \right.$$
$$\left.[\,1 - \mu K_1(\alpha)\,] - 2\omega I(\alpha) - J(\alpha)\right\}.$$

式中:μ 为拉深过程中毛坯与模具间的摩擦因数,n 为材料的硬化指数,σ_{b} 为材料的抗拉强度,r 为材料的厚向异性系数. R_B、R_C、$F(\alpha)$、$K_1(\alpha)$、ω、$I(\alpha)$ 与 $J(\alpha)$ 是随着拉深过程变化的变量,具体物理意义与计算方式参见文献[20].

$$Q_{\mathrm{fwr}} = \frac{3}{8}\pi r_0^2 B \frac{\gamma_0}{t_0} F(n, m, \rho)\, F_m(\lambda_m)\,.$$

式中:t_0 为板材厚度,B 为材料的强度系数,γ_0 为单波的最大挠度,r_0 为法兰内半径,m 为拉深系数,$F(n,m,\rho)$ 与 $F_m(\lambda_m)$ 为随拉深过程变化的两个变量,具体物理意义与计算方式参见文献[21].

图 1 中位于上限压边力与下限压边力之间的 3 条压边力曲线是由 3 种深度强化学习算法优化学习得到的. 可以看出,它们在整个拉深过程中始终保持在 Q_{rup} 与 Q_{fwr} 之间,保证了最终成品不产生质量

缺陷.

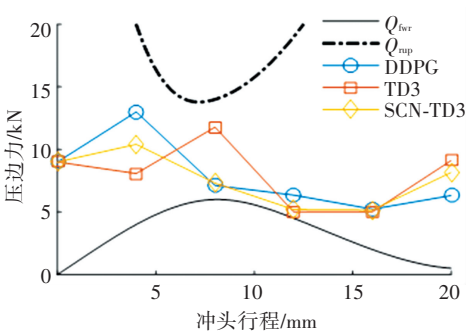

图 1　有效压边力

Fig.1　Effective blank holder force

2.4　算法描述

本文将 SCN-TD3 与有限元仿真相结合，构建了基于深度强化学习的压边力控制策略优化算法，算法描述如下.

输入：有限元模型

输出：actor 网络 π_ϕ

第 1 步：初始化 critic 网络 Q_{θ_1} 与 Q_{θ_2} 的参数 θ_1 与 θ_2，以及 actor 网络的参数 ϕ

第 2 步：初始化目标网络参数：$\theta_1' \leftarrow \theta_1, \theta_2' \leftarrow \theta_2, \phi' \leftarrow \phi$

第 3 步：初始化回放经验 B

第 4 步：For episode = 1, M do

第 5 步：初始化有限元模型状态 s_1

第 6 步：For t = 1, T do

第 7 步：选择动作 a_t，

$a_t \leftarrow \pi_\phi(s_t) + \varepsilon, \varepsilon \sim N(0, \sigma)$

第 8 步：在有限元模型中执行 a_t，输出 s_{t+1} 与 r_t

第 9 步：将转移经验存储到回放经验中，$(s_t, a_t, r(s_t, a_t), s_{t+1}) \rightarrow B$

第 10 步：从回放经验中采样出一个 mini-batch $(s_t^j, a_t^j, r(s_t^j, a_t^j), s_{t+1}^j), j = 1, \ldots, N$

第 11 步：利用目标网络得到动作

$a_{t+1}^j \leftarrow \pi_{\phi'}(s_{t+1}^j) + \varepsilon, \varepsilon \sim \mathrm{clip}(N(0, \sigma'), -c, c)$

第 12 步：为 mini-batch 中的每个转移经验计算目标值 $y_t^j \leftarrow r(s_t^j, a_t^j) + \gamma \min_{i=1,2} Q_{\theta_i'}(s_{t+1}^j, a_{t+1}^j)$.

第 13 步：根据梯度

$\nabla_{\theta_i} L(\theta_i) = N^{-1} \sum_j (y_t^j - Q_{\theta_i}(s_t^j, a_t^j)) \nabla_{\theta_i} Q_{\theta_i}(s_t^j, a_t^j)$ 更新 θ_i

第 14 步：If t mod d then

第 15 步：根据梯度

$\nabla J(\phi) = N^{-1} \sum_j (\nabla_{a_t^j} Q_{\theta_1}(s_t^j, a_t^j) \mid_{a_t^j = \pi_\phi(s_t^j)} \nabla_\phi \pi_\phi(s_t^j)$，更新 ϕ

第 16 步：更新目标网络

$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i', \phi' \leftarrow \tau\phi + (1-\tau)\phi'$

第 17 步：End if

第 18 步：End for

第 19 步：End for

3　基于深度强化学习的拉深控制模型

3.1　问题描述

本文针对板材拉深过程进行压边力控制优化，得到成形质量合格的成品件. 如图 2 所示，板材拉深装置主要由毛坯、冲头、压边圈和凹模 4 部分组成. 毛坯被放置在压边圈与凹模法兰之间，由压边圈夹紧. 整个加工过程被分为 5 个控制步长，每个控制步长内压边力的大小相等，冲头以恒定速度向下冲压，将毛坯压入凹模腔体. 本文将板材拉深控制过程建模成离散时间的马尔科夫决策过程，以板材内部的 Mises 应力分布作为系统状态 s，每个控制步长内的压边力大小作为系统动作 a. 由于本文所建立的有限元模型被划分为 527 个单元，使用全体单元的 Mises 应力分布作为系统状态会使得状态空间过于庞大，不利于问题的有效求解. 因此，本文采用图 2 标记的部分有限元的 Mises 应力作为系统状态，在反应系统状态特征的同时将系统状态缩小为 27 维.

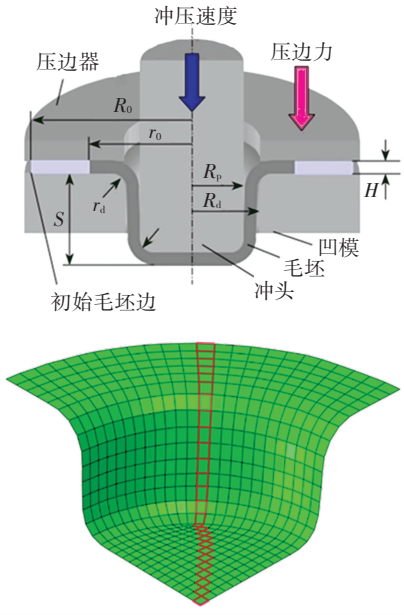

图 2　拉深模型

Fig.2　Deep drawing finite element model

3.2　压边力控制模型

压边力控制模型如图 3 所示，主要由环境与智能体两部分组成. 其中，环境由有限元模型与成本函数组成；智能体由两个价值网络以及一个策略网络组成. 环境接受到动作 a，根据前一时刻环境状态得到当前的回报 r 与观察值 s 并将其输入智能体，智

能体输出下一步长动作,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压边力控制策略.

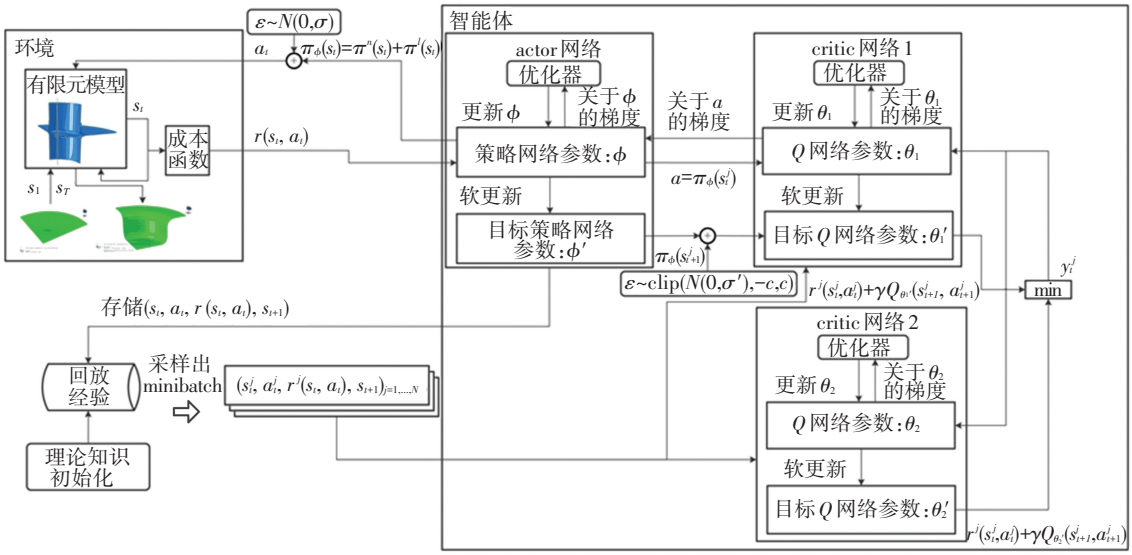

图 3 压边力控制模型

Fig.3 Blank holder force control model

3.2.1 有限元模型

本文建立的板材拉深仿真模型如图 4 所示. 通过假设模型对称性与材料各向同性,建立了 1/4 的板材拉深三维模型,并将拉深过程划分为 6 个离散的时间步长. 其中前 5 个为控制步长,完成向下拉深过程,最后 1 个步长为卸载步长,冲头恢复到原始位置同时将压边力卸载. 模型根据上一步长的状态与输入的压边力值,计算出下一步长的状态.

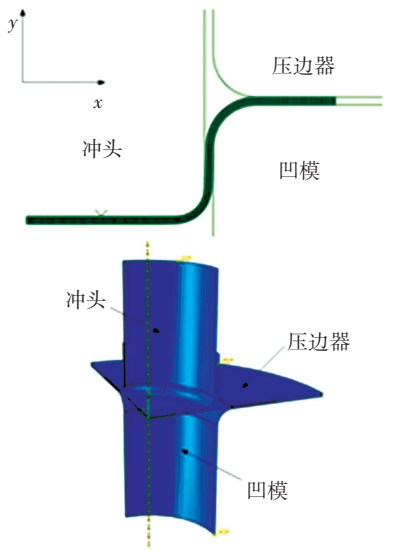

图 4 有限元模型

Fig.4 Finite element model

有限元模型由 3 个刚体部件与 1 个可变形毛坯组成. 刚体部件分别为冲头、凹模和压边圈. 冲头半径 R_p 为 25 mm,冲头圆角半径 r_p 为 4 mm,凹模内径 R_d 为 26.2 mm,凹模圆角半径 r_d 为 6 mm. 圆形毛坯厚度 H 为 1 mm,半径 R_0 为 50 mm. 毛坯材料属性为弹塑性材料,材料为 08F 低碳钢[22]. 材料的弹性模型为线弹性模型,塑性模型为符合 Mises 屈服准则的各向同性模型. 毛坯的有限元单元类型为可变形的 4 节点 4 边形壳单元(S4R).

冲头以恒定速度 4 mm/s 向下冲压,将毛坯压入凹模腔中,拉深深度 S 为 20 mm. 压边力在每个控制步长开始时给出,在每个控制步长内压边力保持不变,均匀施加在压边圈上. 压边力变化范围为 5 000~13 000 N.

本文用 ABAQUS 进行有限元模型的搭建. 在线优化控制环境中,智能体基于当前得到的系统状态来设置动作. 为了符合在线优化控制环境的要求,保证有限元模型的有效性与可重用性,使用了 ABAQUS 脚本与分析重启动技术.

3.2.2 回报函数

回报函数仅由终止状态产生的回报值组成. 控制目标为生产出的工件内部应力低,材料厚度充足,并且材料利用率低. 针对以上 3 个目标,分别建立评价指标函数并通过三者的加权和得出总的质量评价函数[11].

有限元模型由 527 个单元组成. 根据有限元仿真输出的 Mises 应力分布云图、厚度分布云图与 U1 位移分布云图,可以得到最终成品每个单元的 Mises 应力值 $m_i(i=1,2,\ldots,527)$、单元厚度 $h_i(i=1,2,\ldots,527)$ 以及毛坯边在 x 轴方向上的位移 d. 根据以上的数据,建立成本函数:

$$R_i(s_T)=10*\left(1-\frac{C_i(s_T)-C_i^{\min}}{C_i^{\max}-C_i^{\min}}\right),i\in\{a,b,c\},$$

$$C_a(s_T)=\sum_{i=1}^{527}m_i,$$

$$C_b(s_T)=-\min h_i,$$

$$C_c(s_T)=-d.$$

式中：$C_a(s_T)$、$C_b(s_T)$ 与 $C_c(s_T)$ 分别为内部应力项，最小厚度项与材料消耗项，$C_i^{\max}$ 与 $C_i^{\min}$ 为 100 次随机压边力策略控制下仿真产生的数据中的最大与最小成本项值.

最后，本文以加权调和平均的形式给出总的质量评价函数：

$$R(s_T)=\begin{cases}H(s_T,W),\text{if }\forall i\in a,b,c:\ R_i(s_T)\geqslant 0;\\ 0,\qquad\quad \text{otherwise.}\end{cases},$$

$$H(s_T,W)=\sum_i w_i/\left(\sum_i\frac{w_i}{R_i(s_T)}\right).$$

式中权重值 w_i 用于控制各个成本项的重要性，本文中权重 w_i 均为 1.

4　实验与结果分析

受硬件因素影响，实际实验验证十分困难. 本文参考文献[11]的压边力控制仿真实验设计，利用圆筒件拉深成形的有限元仿真进行实验.

4.1　训练过程分析

在 SCN-TD3 中，策略网络与价值网络的结构均为 4 层神经网络，隐藏层节点为 300. DDPG 与 TD3 的网络结构与 SCN-TD3 一致. SCN-TD3 算法中，学习率为 0.000 1，探索率 σ 为 0.1，目标动作噪声方差 σ' 为 0.2，目标动作截断值 c 为 0.2，策略网络更新间隔 d 为 2. DDPG 和 TD3 的参数与 SCN-TD3 一致. 训练过程中，算法在每个训练步长进行 10 次网络更新.

图 5 为不同算法回报值随训练步长数的变化情况. 本文将各控制步长下压边力的相邻训练步长的差作为算法收敛的判断依据. 当连续 100 个训练步长下，各压边力相邻训练步长的差均<1 000 N时，认为算法收敛. 在 SCN-TD3 控制下，回报值大约在第 1 500个步长收敛，而在 DDPG 与 TD3 控制下，回报值大约在第1 800与第1 700个步长收敛. 从回报值的整体变化趋势上看，SCN-TD3 控制下的回报值收敛最快，并且最终收敛到的回报值水平高于 TD3. TD3 控制下的回报值收敛略快于 DDPG，并且最终收敛到的回报值水平高于 DDPG. 这主要是由于 1) TD3 算法中采用的剪裁双 Q 学习、延迟策略更新和目标策略平滑正则化这 3 种技术，有效地缓解了价值网络的过估计问题以及过估计问题给策略网络更新所带来的影响；2) 策略网络的非线性结构与线性结构能够同时结合全局控制与局部控制的优点. 各算法的优势对比如表 1 所示.

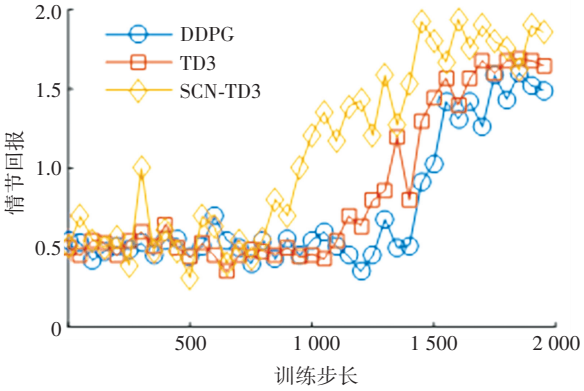

图 5　回报值随训练步长变化

Fig.5　Variation of episode reward with step

表 1　算法优势对比

Tab.1　Comparison of different algorithms

算法	收敛速度	控制效果	过估计问题	策略网络结构改进
DDPG	一般	一般	严重	无
TD3	较快	较好	较轻	无
SCN-TD3	快	好	较轻	线性与非线性策略相结合

在训练过程中，每 5 个训练 episode 结束后，利用当前的策略网络进行 10 次拉深仿真控制，取平均值作为验证回报值. SCN-TD3 得到的最优验证回报值为 1.928 8，而 DDPG 与 TD3 控制下的最优验证回报值分别为 1.602 9 与 1.690 8. 各算法最优验证回报值所对应的压边力控制策略如表 2 所示.

表 2　最优控制策略

Tab.2　Optimal control policy　　N

算法	压边力 1	压边力 2	压边力 3	压边力 4	压边力 5
DDPG	13 000	7 123	6 354	5 227	6 313
TD3	8 078	11 778	5 001	5 010	9 117
SCN-TD3	10 407	7 316	5 200	5 155	8 162

4.2　训练过程压边力变化分析

为了探究压边力在训练过程中的变化情况，本文给出了 DDPG、TD3 与 SCN-TD3 控制下各控制步长的压边力随训练步长的变化情况(见图 6). 由图 6 可知，在训练的早期，步长 2 到步长 5 的压边力聚集在最小值5 000附近，3 种算法均陷入局部最优. 随着训练的进行，算法逐渐跳出局部最优点，最终收敛于一定的压边力水平. SCN-TD3 控制下各步长压边力收敛速度均快于其他两者，体现了 SCN-TD3 在性能上的优势.

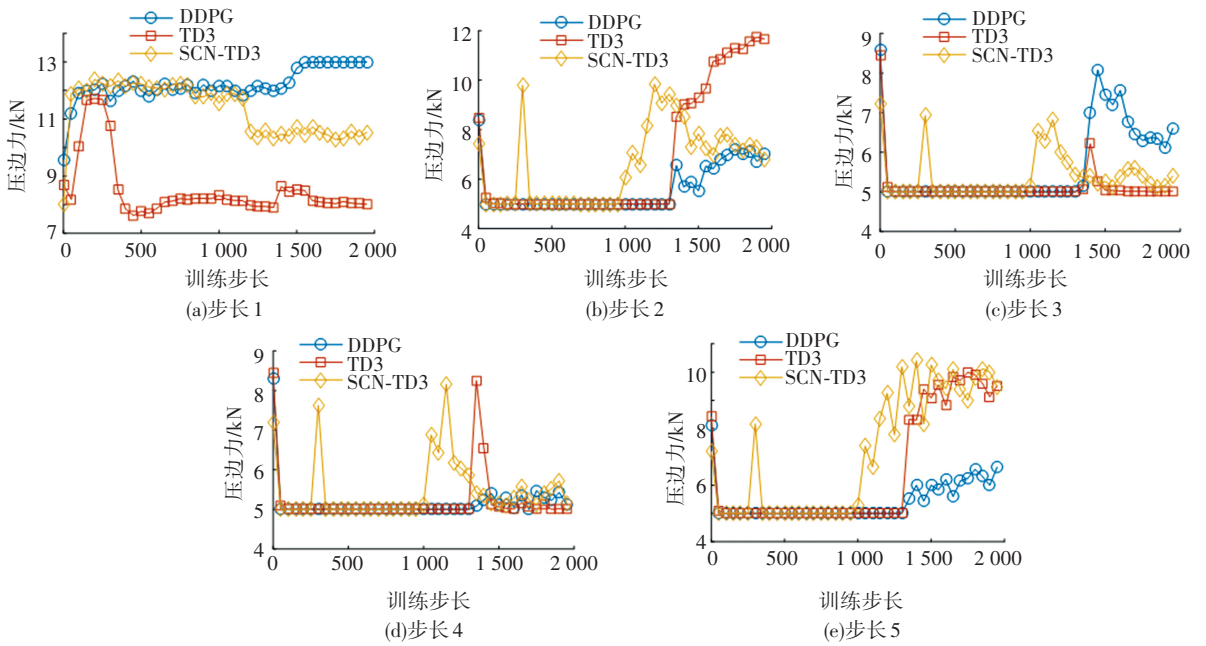

图 6　各控制步长压边力变化

Fig.6　Variation of blank holder force with each control step

4.3　成品质量分析

根据 DDPG、TD3 以及 SCN-TD3 学习到的最优压边力控制策略，在 ABAQUS 中分别进行板材仿真拉深. 根据仿真结果输出的 Mises 应力分布云图、厚度分布云图与 U1 位移分布云图，进行成品质量分析. Mises 应力分布云图展示了成品各有限元单元的 Mises 应力分布情况. 根据图 7～9，可以得到三者的内部应力项指标分别为4 221、4 475以及3 708. 从总体分布上看，TD3 控制下的成品的内部应力和最小，DDPG 控制下的成品的内部应力和最大.

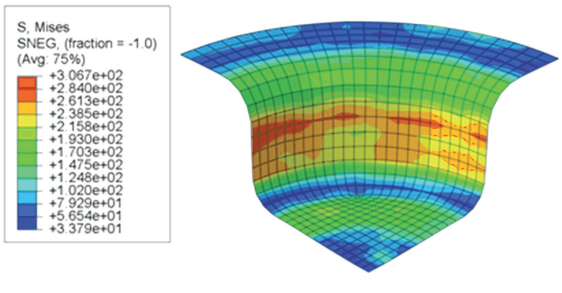

图 7　SCN-TD3 控制下的 Mises 应力分布云图

Fig.7　Mises stress distribution under SCN-TD3

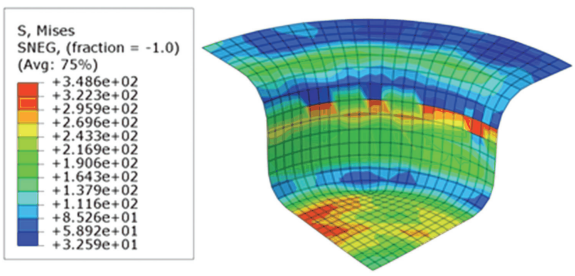

图 8　DDPG 控制下的 Mises 应力分布云图

Fig.8　Mises stress distribution under DDPG

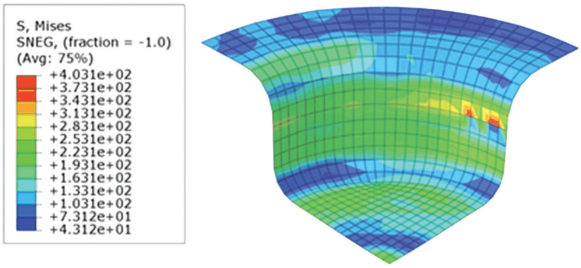

图 9　TD3 控制下的 Mises 应力分布云图

Fig.9　Mises stress distribution under TD3

厚度分布云图体现了成品各处厚度的分布情况. 根据图 10～12，SCN-TD3 控制下成品的最小厚度为0.858 4 mm，DDPG 控制下成品的最小厚度为 0.853 3 mm，TD3 控制下成品的最小厚度为 0.850 3 mm. SCN-TD3 控制下的成品厚度最为充足，TD3 控制下的成品厚度最薄.

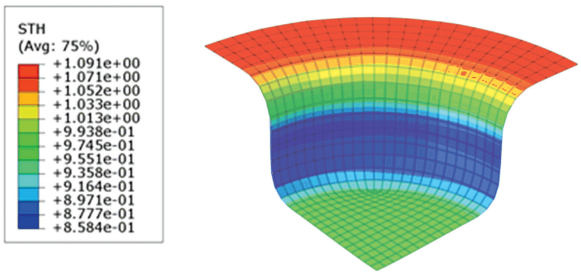

图 10　SCN-TD3 控制下的厚度分布云图

Fig.10　Thickness distribution under SCN-TD3

U1 位移分布云图表示成品的每个有限元单元在 x 轴向上的位移. 根据图 13～15 可知，SCN-TD3 控制下成品的法兰边位移为7.781 7 mm，DDPG 控制下成品的法兰边位移为 8.0837 mm，TD3 控制下成品的法兰边位移为7.944 6 mm. 表明 SCN-TD3 控制下的材料消耗比 DDPG 与 TD3 都要小.

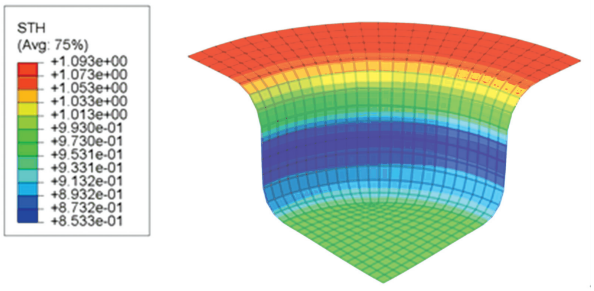

图 11　DDPG 控制下的厚度分布云图

Fig.11　Thickness distribution under DDPG

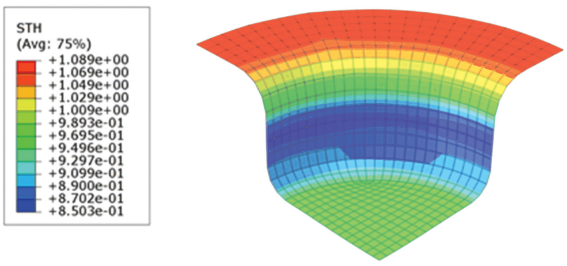

图 12　TD3 控制下的厚度分布云图

Fig.12　Thickness distribution under TD3

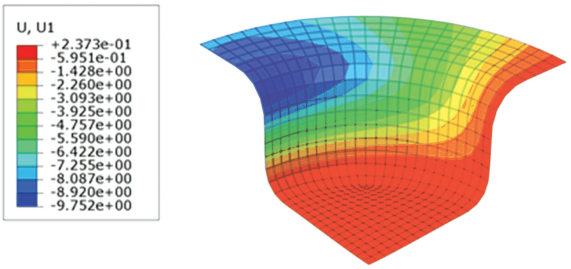

图 13　SCN-TD3 控制下的 U1 位移分布云图

Fig.13　U1 displacement distribution under SCN-TD3

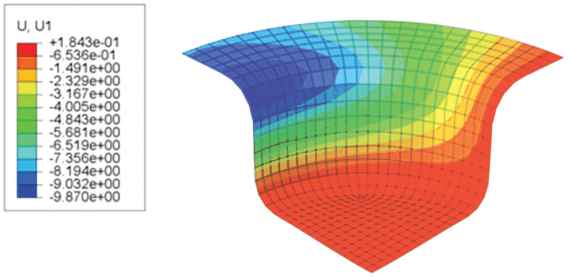

图 14　DDPG 控制下的 U1 位移分布云图

Fig.14　U1 displacement distribution under DDPG

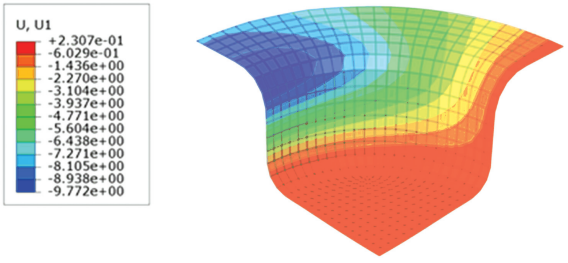

图 15　TD3 控制下的 U1 位移分布云图

Fig.15　U1 displacement distribution under TD3

由于成本函数的组成为内部应力项、最小厚度项与材料消耗项的调和平均，因此尽管 SCN-TD3 在内部应力和指标上的表现不如 TD3，但是其在 3 个成本项中的综合表现最优. 综合以上分析可知，相较于 DDPG 与 TD3，SCN-TD3 控制下成品的内部应力和较小，材料最小厚度充足，材料消耗程度最低，总体质量最优.

4.4　理论知识对于训练过程的影响

根据图 1 的理论有效压边力区间产生多条可行的压边力轨迹以及转移经验. 将有效压边力产生的转移经验加入初始经验回放池，达到将理论知识引入压边力策略优化过程的目的.

通过对各控制步长所对应拉深行程下的有效压边力区间进行随机采样，得到了1 000条有效压边力轨迹及5 000个有效转移经验. 为了探究理论知识对于训练过程的影响，控制初始转移经验中有效压边力转移经验所占比例分别为 0%、25%、50%、75%和 100%，输出所对应的回报值随训练步长的变化情况，如图 16 所示.

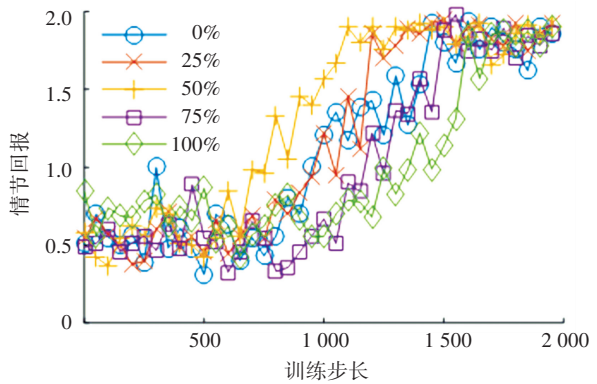

图 16　不同比例有效转移经验下的回报值变化

Fig.16　Variation of episode reward with different percentages of efficient transition experience

由图 16 可知，随着有效转移经验所占比例的增加，训练过程中回报值的收敛越来越迅速，在 50%达到最快收敛速度，随后收敛速度随有效转移经验所占比例的增加开始下降. 这表明，在初始经验回放中添加适量的有效转移经验能够为网络的训练提供一个良好的初始训练数据，让策略网络的参数更快地往回报值高的参数空间进行梯度下降. 然而当初始回放经验中的有效转移经验过多时，由于缺少低回报值的转移经验，策略网络的更新无法有效地远离低回报值的参数空间，反而使得回报值收敛速度下降. 根据以上分析可知，在初始经验回放池中保持经验样本的多样性有助于策略网络的训练.

5　结　论

1）本文将深度强化学习与有限元仿真进行集

成,建立了板材拉深过程压边力控制模型,避免了系统动态的拟合.

2)对策略网络的结构进行改进,并将压边力理论知识引入网络训练中,建立了一个更加有效的深度强化学习算法,提高了成品的成形质量.

3)有限元仿真实验验证了本文所提出的 SCN-TD3 算法的有效性,并与 DDPG 与 TD3 算法进行了压边力控制效果比较. 实验表明,SCN-TD3 控制下成品的内部应力和较小,材料最小厚度充足,材料消耗程度最低,总体质量最优.

参考文献

[1] GHOUATI O, LENOIR H, GELIN J C, et al. Design and control of forming processes using optimization techniques[C]//Proceedings of the ASME Design Engineering Technical Conference. Las Vegas: Nevada Press, 1995

[2] 包友霞, 徐伟力, 刘罡, 等. 薄板成形中变压边力优化设计方法[J]. 机械工程学报, 2001, 37(2): 105
BAO Youxia, XU Weili, LIU Gang, et al. Optimization design variable blank holder force in sheet metal forming[J]. Journal of Mechanical Engineer, 2001, 37(2): 105. DOI: 10.3321/j.issn:0577-6686.2001.02.024

[3] 孙成智, 陈关龙, 林忠钦. 基于数值模拟的变压边力优化设计[J]. 上海交通大学学报, 2004, 38(7): 1086
SUN Chengzhi, CHEN Guanlong, LIN Zhongqin. The optimization of variable blank-holder forces based on numerical simulation[J]. Journal of Shanghai JiaotongUniversity, 2004, 38(7): 1086. DOI: 10.3321/j.issn:1006-2467.2004.07.012

[4] HILLMANN M, KUBLI W. Optimization of sheet metal forming processes using simulation programs[J]. Numisheet, 1999, 99(1): 287

[5] SCOOT M A, CARDEW H M, HODGSON P, et al. Novel criteria and tools for the FE optimization of sheet metal forming process[C]// Proceedings of Numisheet'99. Besancon: [s.n.], 1999: 305

[6] SENN M, LINK N, POLLAK J, et al. Reducing the computational effort of optimal process controllers for continuous state spaces by using incremental learning and post-decision state formulations[J]. Journal of Process Control, 2014, 24(3): 133. DOI: 10.1016/j.jprocont.2014.01.002

[7] 黄玉萍, 阮锋, 蔡志兴. 应用神经网络优化压边力[J]. 模具工业, 2008, 34(7): 9
HUANG Yuping, RUAN Feng, CAI Zhixing. Using neural network to optimize blank holder force[J]. Die & Mould Industry, 2008, 34(7): 9. DOI: 10.3969/j.issn.1001-2168.2008.07.004

[8] QIAN Zhiping, MA Rui, ZHAO Jun, et al. Intelligent control technology for deep drawing of sheet metal[J]. Journal of Central South University of Technology, 2008, 15(2): 273. DOI: 10.1007/s11771-008-0470-4

[9] MANABE K, YANG M, YOSHIHARA S. Artificial intelligence identification of process parameters and adaptive control system for deep-drawing process[J]. Journal of Materials Processing Technology, 1998, 80: 421. DOI: 10.1016/s0924-0136(98)00121-6

[10] 汪锐, 罗亚军, 何丹农, 等. 基于模糊神经网络的压边力优化控制专家系统[J]. 上海交通大学学报, 2001, 35(3): 411
WANG Rui, LUO Yajun, HE Dannong, et al. Expert system based on fuzzy neural network for the optimal control of blank holder force[J]. Journal of Shanghai Jiaotong University, 2001, 35(3): 411. DOI: 10.3321/j.issn:1006-2467.2001.03.021

[11] DORNHEIM J, LINK N, GUMBSCH P. Model-free adaptive optimal control of sequential manufacturing processes using reinforcement learning[EB/OL]. (2018-09-18)[2019-08-05]. https://arxiv.org/abs/1809.06646

[12] 杜海文, 崔明朗, 韩统, 等. 基于多目标优化与强化学习的空战机动决策[J]. 北京航空航天大学学报, 2018, 44(11): 2247
DU Haiwen, CUI Minglang, HAN Tong, et al. Maneuvering decision in air combat based on multi-objective optimization and reinforcement learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(11): 2247. DOI: 10.13700/j.bh.1001-5965.2018.0132

[13] 张彦铎, 闵锋. 基于人工神经网络的强化学习在机器人足球中的应用[J]. 哈尔滨工业大学学报, 2004, 36(7): 859
ZHANG Yanduo, MIN Feng. Application of reinforcement learning based on artificial neural network to robot soccer[J]. Journal of Harbin Institute of Technology, 2004, 36(7): 859. DOI: 10.3321/j.issn:0367-6234.2004.07.004

[14] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]//Proceedings of the International Conference on Machine Learning. Beijing: ACM, 2014: 387

[15] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529. DOI: 10.1038/nature-14236

[16] FUJIMOTO S, VAN H H, MEGER D. Addressing function approximation error in actor-critic methods[EB/OL]. (2018-09-28)[2019-08-05]. https://arxiv.org/abs/1802.09477

[17] SROUJI M, ZHANG J, SALAKHUTDINOV R. Structured control nets for deep reinforcement learning[EB/OL]. (2018-09-22)[2019-08-05]. https://arxiv.org/abs/1802.08311

[18] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. (2015-09-09)[2019-08-05]. https://arxiv.org/abs/1509.02971

[19] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double *Q*-learning[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix: AAAI, 2016: 1813

[20] 赵军, 郑祖伟, 潘文武. 拉深过程智能化控制中的破裂失稳临界条件[J]. 燕山大学学报, 2000(4): 335
ZHAO Jun, ZHENG Zuwei, PAN Wenwu. The critical condition for the side-wall rupture in the intelligent control of deep drawing[J]. Journal of YanshanUniversity, 2000(4): 335. DOI: 10.3969/j.issn.1007-791X.2000.04.009

[21] 赵军, 张双杰, 曹宏强, 等. 拉深过程智能化控制中的法兰起皱临界条件[J]. 燕山大学学报, 1998(3): 197
ZHAO Jun, ZHANG Shuangjie, CAO Hongqiang, et al. Critical flange wrinkle condition in intelligent control of deep drawing process[J]. Journal of Yanshan University, 1998(3): 197

[22] 范泽. 冲压速度对板料成形性能的影响研究[D]. 合肥: 合肥工业大学, 2015
FAN Ze. Study of the effect of punch speed on sheet metal forming[D]. Hefei: Hefei University of Technology, 2015

(编辑 杨 波)