DOI: 10.3901/JME.2020.20.047

基于深度强化学习与有限元仿真集成的 拉深成形控制^{*}

郭 鹏 张新艳 余建波

(同济大学机械与能源工程学院 上海 201804)

摘要: 金属板材拉深过程中的压边力是决定成品质量的关键参数,传统压边力控制方法往往需要对高度非线性的拉深过程进行建模,导致其控制结果与实际存在较大偏差。提出一种基于深度强化学习与有限元仿真集成的金属板材拉深过程控制模型,利用深度神经网络强大的预测能力来提取拉深加工过程中的状态信息并进行可靠预测,结合强化学习的决策能力来进行压边力控制策略的学习优化,避免了精确系统动力学模型的拟合以及先验知识的获取。同时,针对板材拉深加工中常见的拉裂质量缺陷与起皱质量缺陷,建立拉深成形性能评价函数,为深度强化学习提供回报信号来指导学习过程,并利用有限元仿真构成深度强化学习的环境模型。试验表明,深度强化学习模型能够有效地进行压边力控制策略优化,有效提高产品质量。所提出的压边力控制模型利用无模型的深度强化学习,能避免拉深过程的系统模型拟合,可提高压边力控制策略的控制效果,同时结合循环神经网络能解决板材拉深加工过程中的部分可观察性问题。

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: TG156

Control of Deep Drawing Process Based on Integration of Deep Reinforcement Learning and Finite Element Method

GUO Peng ZHANG Xinyan YU Jianbo

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

Abstract: The blank holder force in the deep drawing process is the key to determine the quality of finished products. The traditional blank holder force control methods often need to model the highly nonlinear deep drawing process, resulting in a large deviation of the control result from the real situation. The proposed control model based on integration of deep reinforcement learning and finite element method, uses the prediction ability of deep neural network to extract the state information of deep drawing process and predict system state, uses the decision ability of reinforcement learning to optimize control policy, and avoids the fitting of highly nonlinear process dynamic and the acquirement of prior knowledge. Besides, according to the common quality defects in deep drawing process, which are crack and wrinkle, the evaluation function of deep drawing process is established to provide the reward signal to guide the deep reinforcement learning, and the environment model of deep reinforcement learning is constructed by using finite element simulation. Experiments show that the deep reinforcement learning model can effectively optimize the blank holder force control policy and improve the product quality. The proposed blank holder force control model uses model-free deep reinforcement learning to avoid the fitting of deep drawing process, and improves the control effect of blank holder force control policy. The usage of recurrent neural network solves the problem of partial observability in the deep drawing process.

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

0 前言

板材拉深成形是指基于金属材料的可塑性,利

用外力及模具,将被加工板料制成零部件的一种加工方法,在汽车、造船、电器、航空等领域应用广泛^[1]。板料成形工艺涉及很多重要参数,其中压边力(Blank holder force, BHF)对于板材成形质量有着重要的影响。在板材拉深过程中,通过改变压边力大小来调节压边器作用于板料的正向压力,以改变板料与模具和压边器接触面间的摩擦力,从而增加

^{*} 国家自然科学基金资助项目(71777173)。20190818 收到初稿,20200418 收到修改稿

板料中的拉应力,减小切向应力,达到控制材料的流动的目的。在拉深加工过程中对压边力进行合理的调控是保证成品成形质量的重要手段。

传统的压边力控制方法通常将数值模拟技术与试验设计相结合,进行压边力优化控制。林晓娟等^[2]用有限元分析软件对压边力的恒定模式和变化模式分别进行数值模拟试验,分析其对圆锥型工件拉深成形过程以及最终成品质量的影响。YOUNIS^[3]对于四种不同的压边力加载模式进行有限元模拟试验,结果表明变化的非均匀压边力能够让最终产品的厚度分布最均匀。马瑞等^[4]通过数值模拟和试验,对成形过程中压边力随凸模行程的变化规律进行研究。谢延敏等^[5]利用有限元分析与灰色系统理论,对方盒件拉深成形中的压边力等可控因子进行试验设计,获得因子的最佳参数。刘罡^[6]将数值模拟和正交试验方法有机地结合起来寻找优化的工艺条件。虽然结合数值模拟与试验设计的方法可以提高压边力优化的效率,但是采用试验设计方法确定的最优压边力曲线只是实际最优压边力的一种近似,无法获得真正的最优压边力控制曲线。

国内外学者对板材拉深过程的变压边力控制技术进行了广泛研究。CAO 等^[7]分别采用了 PI 控制以及自回归滑动平均模型(Auto-regressive moving-average model, ARMA)^[8]方法对板材拉深过程的压边力进行优化。SHENG 等^[9]则采用了增量优化的方法对盒型件的最优压边力分布进行了研究。孙成智等^[10]采用响应面法,得到了盒形件在不同位置的最优压边力。HILLMANN 等^[11]将成形极限图上各点到成形极限和起皱极限的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。SCOOT 等^[12]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。

神经网络由于其表征非线性关系的强大能力,被广泛应用于处理板材拉深加工中的非线性动态。SENN 等^[13]使用近似动态规划方法来进行压边力控制策略的优化。该文献以板料 3 个位置的冯米塞斯应力作为系统状态,离散化的压边力值作为动作空间,并利用神经网络来拟合最优值函数与系统动力学模型。黄玉萍等^[14]针对复杂工件的深拉过程,建立了径向基神经网络来进行压边力控制,以最大等效应力、最大主应变、最大减薄率作为神经网络的输入参数,以冲头行程 5 个不同时刻对应的压边力为输出参数,构建了压边力优化神经网络模型。LI

等^[15]将深拉过程划分为 6 个阶段,为每个阶段确定一个压边力值。该文献利用有限元法仿真产生的训练样本训练神经网络来拟合目标函数,最后通过多目标遗传算法来求解最优压边力。神经网络结合弹塑性理论是另一种常用的压边力控制方法。QIAN 等^[16-17]利用神经网络来进行深拉过程中材料性能和操作条件参数的在线识别。根据识别模型得到的材料性能参数和摩擦因数,定量描述了法兰起皱和侧壁破裂的临界条件,并通过结合这些条件以及能源最小化的目标来预测压边力大小。

还有学者将模糊控制应用于压边力控制中。汪锐等^[18]结合模糊逻辑系统方法与人工神经网络方法,构建了模糊神经网络来进行压边力预测。MANABE 等^[19-20]提出一种基于模糊理念的新的控制方法,通过一个交替转换模块对冲压速度和压边力进行联合控制。基于模糊推理的压边力优化方法的缺点在于模糊控制规则的获取依赖于大量的试验与经验,并且控制精度有待提高。

传统压边力控制方法往往需要通过对拉深过程中复杂的非线性关系进行建模来进行压边力的预测与控制。由于精确的系统动力学模型难以获得,基于模型的方法得到的压边力控制方案往往可靠性不高。DORNHEIM 等^[21]提出了一种无模型的压边力控制方法,将神经拟合 Q 迭代(Neural fitted Q iteration, NFQ)算法与有限元仿真相结合来学习压边力控制策略。

本文提出了一种基于深度强化学习的板材拉深过程的压边力优化控制模型,本文的主要创新点如下。

(1) 通过有限元仿真来进行拉深加工过程的加工模拟,并结合基于策略的深度强化学习方法,提高了压边力的控制精度与控制策略的学习效率。

(2) 采用循环神经网络构建深度强化学习算法中的评论家网络与行动者网络,解决了板材拉深加工过程中的部分可观察性问题。

(3) 根据板材拉深加工过程的特殊性,建立了评估成品质量的回报函数,有效地量化了最终产品的成形性能。最后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深度强化学习方法在压边力优化控制中的有效性。

1 理论背景

1.1 部分可观察马尔科夫决策过程

部分可观察马尔科夫决策过程(Partially-observed markov decision process, POMDP)可以由以下部分来

描述: 状态集合 S , 动作集合 A , 初始状态分布 $p(s_1)$, 状态转移分布函数 $p(s_{t+1}|s_t, a_t)$ 以及回报函数 $r(s_t, a_t)$ 。当智能体(agent)无法直接观测到状态 s_t 时, 它以概率 $p(o_t|s_t)$ 接受来自观察集合 O 的观察 o_t 。

在 POMDP 中, 智能体只能通过观察历史 $h_t = (o_1, a_1, o_2, a_2, \dots, o_{t-1}, a_{t-1}, o_t)$ 来间接地观测系统状态。因此, 智能体的目标是学习到一个策略 $\pi(h_t)$ 来最大化未来预期收益。策略构建了从历史到一个动作分布 $P(A)$ 的映射。对于随机策略, 控制目标是最大化

$$J = E_{\tau} \left[\sum_{t=1}^T \gamma^{t-1} r(s_t, a_t) \right] \quad (1)$$

式中, 轨迹 $\tau = (s_1, o_1, a_1, s_2, o_2, a_2, \dots, s_T, o_T)$ 从轨迹分布中采样得到, 轨迹分布由策略 $\pi: p(s_1)p(o_1|s_1)\pi(a_1|h_1)p(s_2|s_1, a_1)p(o_2|s_2)\pi(a_2|h_2)\dots$ 推导得出, γ 为未来收益的折扣率。

1.2 强化学习

强化学习是一种从环境映射到动作的学习, 目的是使得智能体在与环境的交互过程中获得最大的累积回报^[22]。强化学习使用动作值函数 $Q^{\pi}(s_t, a_t)$ 来描述当遵循策略 π 时, 在状态 s_t 下采取动作 a_t 的未来期望回报

$$Q^{\pi}(s_t, a_t) = E_{\pi} \left[\sum_{i=t}^T \gamma^{i-t} r(s_i, a_i) | s_t, a_t \right] \quad (2)$$

在强化学习领域通常使用动作值函数的递推形式, 即贝尔曼方程

$$Q^{\pi}(s_t, a_t) = E_{s_{t+1}, a_t} \left[r(s_t, a_t) + \gamma E_{\pi} [Q^{\pi}(s_{t+1}, a_{t+1})] \right] \quad (3)$$

当目标策略为确定性策略 μ 时, 式(3)中的内层期望能够略去

$$Q^{\mu}(s_t, a_t) = E_{s_{t+1}} \left[r(s_t, a_t) + \gamma Q^{\mu}(s_{t+1}, \mu(s_{t+1})) \right] \quad (4)$$

由于外层的期望仅取决于环境的系统动态, 因此可以使用离策略(off-policy)的方法来学习 Q^{μ} , 即使用一个与所学策略不同的行为策略 μ' 来产生状态转移经验。

作为一种常用的离策略算法, Q 学习(Q-learning) $\mu(s) = \arg \max_a Q(s, a)$ 使用了贪婪策略。深度强化学习用参数化的函数近似器来表示 Q 函数, θ^Q 表示近似器的参数, 通过最小化误差项 $L(\theta^Q)$ 来优化近似器参数。误差项表达式为

$$L(\theta^Q) = E_{\mu'} \left[(Q(s_t, a_t | \theta^Q) - y_t)^2 \right] \quad (5)$$

式中, 目标值 y_t 表达式为

$$y_t = r(s_t, a_t) + \gamma Q(s_{t+1}, \mu(s_{t+1}) | \theta^Q) \quad (6)$$

在部分可观察的情况下, 由于无法直接得到状

态 S , 因此动作值函数应该被定义为针对于历史 h 的形式

$$Q^{\pi}(h_t, a_t) = E_{s_t|h_t} [r(s_t, a_t)] + E_{\tau_{>t}|h_t, a_t} \left[\sum_{i=1}^{T-t} \gamma^i r(s_{t+i}, a_{t+i}) \right] \quad (7)$$

式中, $\tau_{>t} = (s_{t+1}, o_{t+1}, a_{t+1}, \dots)$ 表示未来轨迹, 两个期望分别针对状态和轨迹分布的条件概率 $p(s_t|h_t)$ 与 $p(\tau_{>t}|h_t, a_t)$ 进行计算, 这两个概率与策略 π 相关。

1.3 长短期记忆网络

循环神经网络(Recurrent neural network, RNN)隐藏层的单元间存在有向连接, 适用于处理具有前后依赖关系的序列信息。常规 RNNs 由于梯度消失问题, 无法学习序列的长期依赖。长短期记忆网络(Long short-term memory network, LSTM)的提出解决了梯度消失与梯度爆炸问题^[23]。图 1 展示了一个标准 LSTM 单元的组成结构。

图 1 长短期记忆单元的基本结构

LSTM 单元结构引入了 3 个重要的门结构来有效地处理梯度。LSTM 单元的输出 h_t 依赖于当前输入序列 X_t 和 LSTM 单元前一时刻的状态 h_{t-1}

$$h_t = f(h_{t-1}, X_t) \quad (8)$$

3 个门结构与过程状态计算式如下

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \quad (9)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \quad (10)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \quad (11)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, X_t] + b_C) \quad (12)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (13)$$

$$h_t = o_t \odot \tanh(C_t) \quad (14)$$

式中, i_t 表示输入门的状态, f_t 表示遗忘门的状态, o_t 表示输出门的状态。 W 与 b 表示 LSTM 单元的权重与偏差。 C_t 表示当前状态的输出, $\tilde{C}$ 表示 C_t 的候补值。输入门、遗忘门与输出门的激活函数均为 sigmoid 函数, 将相关的输出限制在 0 到 1 之间。0 值表示门

结构过滤并完全阻隔了信息, 1 值表示所有信息都通过门结构。遗忘门 f_t 决定上一时刻单元状态信息被保留的程度。输入门 i_t 定义了输入序列中的哪些信息被用来更新单元状态。输出门 o_t 利用单元信息来决定输出信息。

2 基于深度强化学习的压边力控制策略优化算法

本文构建的压边力控制策略优化算法如图 2 所示。通过环境与智能体之间的交互, 不断更新行动者(Actor)网络与评论家(Critic)网络的参数, 其中网络参数的更新算法基于循环确定性策略梯度^[24](Recurrent deterministic policy gradient, RDPG)。

RDPG 是一种 actor-critic 结构的深度强化学习

图2 基于深度强化学习的压边力控制模型

在部分可观察性下, 最优策略以及相应的动作值函数都是关于整个过去的观察-动作历史 h_t 的函数。因此使用 $\mu(h)$ 与 $Q(h,a)$ 来替代 $\mu(s)$ 和 $Q(s,a)$, 得到以下策略更新方程

$$\frac{\partial J(\theta)}{\partial \theta} = E_{\tau} \left[\sum_t \gamma^{t-1} \frac{\partial Q^{\mu}(h_t, a)}{\partial a} \bigg|_{a=\mu^{\theta}(h_t)} \frac{\partial \mu^{\theta}(h_t)}{\partial \theta} \right] \quad (16)$$

式中, 期望针对整个轨迹 $\tau=(s_1, o_1, a_1, s_2, o_2, a_2, \dots, s_T, o_T)$ 进行计算, 轨迹从当前策略推导出的轨迹分布中采样得到。与完全可观测情况类似, 使用近似值 Q^{ω} 来替代 Q^{μ} , Q^{ω} 为参数为 ω 的循环神经网络。因此, 可以利用依时间反向传播(Back propagation through time, BPTT)来有效地训练循环神经网络, 从输入的观察历史中提取系统状态, 而不是直接依赖于整个观察历史。

本文将 RDPG 与有限元仿真相结合, 构成了基于深度强化学习的压边力控制策略优化算法, 如表 1 所示。算法输入为一个进行板材拉深过程仿真

算法, 在确定性策略梯度^[25](Deterministic policy gradient, DPG)上拓展而来, 用于解决连续动作空间的 POMDP 问题。DPG 算法的主要思想是对于一个以参数 θ 参数化的确定性策略 μ^{θ} , 在给定与当前策略相关的动作值函数 Q^{μ} 的情况下, 针对参数 θ 应用链式法则得到策略梯度, 并通过反向传播来更新策略网络。

$$\frac{\partial J(\theta)}{\partial \theta} = E_{s \sim \rho^{\mu}} \left[\frac{\partial Q^{\mu}(s, a)}{\partial a} \bigg|_{a=\mu^{\theta}(s)} \frac{\partial \mu^{\theta}(s)}{\partial \theta} \right] \quad (15)$$

式中, 期望是针对由当前策略 μ^{θ} 导出的状态访问分布 ρ^{μ} 来计算的。在实际中, 式中精确的动作值函数 Q^{μ} 由一个近似值 Q^{ω} 来替代, Q^{ω} 以参数 ω 进行参数化。近似值 Q^{ω} 对于动作 a 可微并且能够用 Q 学习来进行学习。

的有限元模型。在进行环境与智能体的交互之前, 首先初始化回放缓冲 R 、评论家网络 Q^{ω} 与行动者网络 μ^{θ} 以及对应的目标网络的参数。在每个 episode 开始时, 先初始化一个空白历史 h_0 以及有限元模型的初始状态 s_1 。在 episode 开始后的每个控制步长 t , 将接收到的观察 o_t 与上一步长的控制动作 a_{t-1} 加入到历史 h_t 中, 然后利用行动者网络得到动作 $\mu^{\theta}(h_t)$ 并加上一个探索噪声 ε (本文使用 Ornstein-Uhlenbeck 过程来产生探索噪声 ε) 形成动作 a_t , 最后在有限元模型中执行动作 a_t 并输出观察 o_{t+1} 与回报 r_t 。当一次完整的拉深加工过程仿真完成后, 形成一个完整的情节序列 $(o_1, a_1, r_1, \dots, o_T, a_T, r_T)$, 并将其存入回放缓冲 R 中。在网络训练阶段, 从回放缓冲中采样出 N 个情节序列 $(o_1^i, a_1^i, r_1^i, \dots, o_T^i, a_T^i, r_T^i)_{i=1,2,\dots,N}$, 并为每一条序列的每一个时间步长构建历史 $h_t^i = (o_1^i, a_1^i, \dots, a_{t-1}^i, o_t^i)$ 用于网络训练, 然后利用目标网络为每一个采样的历史计算目标值 $(y_1^i, \dots, y_T^i)$, 最后进行评论家网络与行动者网络参数 ω 与 θ 以及对应的目标网络参数 ω' 与 θ' 的更新。

表 1 基于深度强化学习的压边力控制策略优化算法

算法 1 基于深度强化学习的压边力控制策略优化算法

输入：有限元模型

- 1: 初始化 critic 网络 $Q^{\omega}(h_t, a_t)$ 以及 actor 网络 $\mu^{\theta}(h_t)$ 的网络参数 ω 和 θ
- 2: 初始化目标网络 $Q^{\omega'}$ 和 $\mu^{\theta'}$ 的网络参数, $\omega' \leftarrow \omega$, $\theta' \leftarrow \theta$
- 3: 初始化回放缓冲 R
- 4: **For** episode = 1, M **do**
- 5: 初始化空白历史 h_0
- 6: 初始化有限元模型
- 7: **For** $t = 1, T$ **do**
- 8: 接受观察 o_t
- 9: 将接受到的观察与上一步长的控制动作加入到历史中, $h_t \leftarrow h_{t-1}, a_{t-1}, o_t$
- 10: 得到此步长的控制动作 $a_t = \mu^{\theta}(h_t) + \varepsilon$
- 11: 在有限元模型中执行 a_t 并输出 o_{t+1} 和 r_t
- 12: **End for**
- 13: 将一个完整的情节序列 $(o_1, a_1, r_1, \dots, o_T, a_T, r_T)$ 存储到回放缓冲 R 中
- 14: 从回放缓冲中采样出一个 minibatch $(o_1^i, a_1^i, r_1^i, \dots, o_T^i, a_T^i, r_T^i)_{i=1,2,\dots,N}$
- 15: 构建历史 $h_t^i = (o_1^i, a_1^i, \dots, o_{t-1}^i, a_{t-1}^i, o_t^i)$ 用于训练
- 16: 计算目标中 $y_t^i = r_t^i + \gamma Q^{\omega'}(h_{t+1}^i, \mu^{\theta'}(h_{t+1}^i))$
- 17: 根据梯度 $\Delta\omega = \frac{1}{NT} \sum_i \sum_t (y_t^i - Q^{\omega}(h_t^i, a_t^i)) \frac{\partial Q^{\omega}(h_t^i, a_t^i)}{\partial \omega}$ 更新 ω
- 18: 根据梯度 $\Delta\theta = \frac{1}{NT} \sum_i \sum_t \frac{\partial Q^{\omega}(h_t^i, \mu^{\theta}(h_t^i))}{\partial a} \frac{\partial \mu^{\theta}(h_t^i)}{\partial \theta}$ 更新 θ
- 19: 更新目标网络参数, $\omega' \leftarrow \tau\omega + (1-\tau)\omega'$, $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$
- 20: **End for**

输出：actor 网络 μ^{θ}

RDPG 主要用于解决 POMDP 问题, 需要处理时间序列信息, 从观测历史中提取系统的真实信息, 因此 RDPG 使用了 LSTM 网络来表示值函数与策略函数。本文使用了 4 层的 LSTM 网络来分别表示值函数与策略函数, 隐藏层节点为 80。在深度强化学习中, 使用深度神经网络来表示值函数和策略函数会导致训练过程不稳定。为了解决此问题, 算法采用了两种方法: 经验回放与目标网络^[26]。在学习过程中, 算法将经历过的轨迹历史储存在一个称为回放缓冲的数据库中, 通过从数据库中随机采样的轨迹数据来进行参数更新。对于有监督学习的深度神经网络模型, 其训练过程的假设前提通常是训练数据独立同分布, 然而深度强化学习过程中得到的样本数据前后是有关系的。经验回放技术能够打破数据间的关联性, 克服训练数据的相关性和非平稳分布问题, 有效地提高了数据效率和训练稳定性。目标网络是另外一种能够有效提高训练稳定性的技巧。算法在训练过程中针对评论家网络 Q^{ω} 与行动者网络 μ^{θ} 分别建立了用于计算目标值 y_t^i 的目标网络 $Q^{\omega'}$ 与 $\mu^{\theta'}$, 其中参数 ω' 与 θ' 以

一定时间延迟跟随 ω 与 θ 的值变化。算法采用“软更新”^[27]方式来更新目标网络参数 ω' 与 θ' 。由于使用被更新的评论家网络与行动者网络来计算目标值会导致 Q 值更新过程发散, 因此通过目标网络技术能够打破目标值与当前评论家网络和行动者网络之间的相关性, 进一步提升训练过程的稳定性。

3 基于深度强化学习的拉深控制模型

3.1 问题描述

本文的研究对象为板材拉深过程, 目的是得到拉深过程中压边力的最优控制策略。如图 3 所示, 拉深装置主要由 3 个部分组成: 冲头、压边器和凹模。拉深加工开始时, 毛坯被放置在压边器与凹模之间, 通过压边器作用于毛坯表面的力将毛坯夹紧。拉深加工开始后, 冲头以一定的速度向下冲压, 将毛坯压入凹模形成工件。由于在拉深加工过程中, 我们仅能观测到与过程相关的物理量(如力与位移), 无法进行实际系统状态的实时测量, 因此将拉深加工过程建模成离散时间的 POMDP。

图3 板材拉深加工过程示意图

3.2 压边力控制模型

板材拉深加工过程压边力控制模型如图2所示,主要由环境与智能体两部分组成。环境由有限元模型(Finite element model, FEM)、回报函数和感知模型组成;智能体主要由评论家网络 $Q^o(h, a_t)$ 和行动者网络 $\mu^o(h_t)$ 组成。其中评论家网络与行动者网络均为由LSTM构成的循环神经网络。环境接受到动作 a ,根据前一时间环境状态得到当前的回报 r 与观察值 o 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压边力控制策略优化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。

3.2.1 有限元模型

本文建立有限元仿真模型来模拟板材拉深过程,有限元模型如图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。为了实现拉深加工过程的时间离散控制,在有限元模型中建立了5个时间步长。有限元模型根据上一时间步长的仿真结果以及控制智能体给出的当前压边力值,计算出下一时间步长的状态。

图4 有限元模型示意图

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(Punch)、凹模(Die)和压边器(Holder)。冲头半径 R_p 为25.0 mm,冲头圆角半径 r_p 为4.0 mm,凹模内径 R_d 为26.2 mm,凹模圆角半径 r_d 为6.0 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,材料型号为08F低碳钢,相关材料属性参数参见文献[28]。冲头以恒定速度4 mm/s向下冲压,将毛坯压入凹模中,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,压边力从当前值线性变化到步长开始时给定的值。压边力大小的变化范围为3~13 kN。

本文使用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观察来设置动作。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ABAQUS脚本与分析重启技术。

3.2.2 部分可观测性

拉深过程中,智能体无法直接得到系统状态 s ,仅能在每个步长开始时得到观察值 o 。本文中设计观察值为:①实际冲击力 F_y ;②毛坯径向进给量 pos_x ;③压边器的轴向实际位移 pos_y 。

3.2.3 回报函数

回报函数由局部成本和全局成本组成。局部成本与每一个步长的状态和动作有关,全局成本则根据终止状态表征最终成品质量的优劣。本文假设局部成本均为0,优化过程仅依赖于终止状态产生的回报值。在板材拉深中,产品的主要失效形式为拉裂和起皱。本文的控制目标是最终产品不产生失效且成形性能最优,采用基于成形极限图(Forming limit diagram, FLD)的成形性能评价函数^[28]作为回报函数。

(1) 成形极限图。在金属成形领域,FLD图被用来判定材料失效情况,表示了板料成形过程中在不同应力状态下的变形极限。FLD图以平面应变状态中的工程面内最小主应变为横坐标,工程面内最大主应变为纵坐标,包含了一条由所有成形极限点构成的成形极限曲线(Forming limit curve, FLC)。如果板上某点的平面主应变位于FLC的上方,则该点所处区域材料出现拉裂;如果该点处于FLC的下方,则该点变形状态安全,且距离FLC越远,安全裕度越大。对于低碳钢来说,成形极限曲线左半部分的形状几乎是一样的,曲线沿主应变轴的高低(FLD_0)只是钢材性能的函数。根据文献[28]可得FLC左半部分的经验公式

$$\varepsilon_{\max} + \varepsilon_{\min} = \ln(1 + FLD_0) \quad (17)$$

$$FLD_0 = (23.3 + 14.13d)n / 21 \quad (18)$$

式中, d 为板料厚度, n 为应变硬化指数, $\varepsilon_{\max}$ 、 $\varepsilon_{\min}$ 为真实的最大与最小主应变。

对于金属板料拉深成形问题来说, 板料上各点应变状态均处于 FLD 左侧, 因此只需要确定 FLC 的左半部分就可以用来制定该金属板料拉深成形问题的成形极限图。

(2) 拉裂评价函数。ABAQUS 拉裂判定示意图如图 5 所示, A 点为板料上某点的面内主应变状态, B 点位于 FLC 上, 最小主应变与 A 点相同, 定义

$$\omega_{FLD} = \frac{\varepsilon_{\max}^A}{\varepsilon_{\max}^B} \quad (19)$$

当 $\omega_{FLD}=1$ 时, A 点出现拉裂; 当 $\omega_{FLD}<1$ 时, A 点是安全的, 且 ω_{FLD} 越小, 该点的安全裕度越大。

图 5 ABAQUS 中拉裂判定示意图

在图 6 中, 实线表示 $\omega_{FLD}=1$ 的曲线, 即成形极限曲线, 用 $\psi(\varepsilon_{\min})$ 表示; 虚线表示 $\omega_{FLD}=0.6$ 的曲线, 即安全裕度为 40% 的曲线, 用 $\phi(\varepsilon_{\min})$ 来表示。当某点在 $\phi(\varepsilon_{\min})$ 下方, 即该点安全裕度大于 40% 时, 假设其对拉裂可能性不造成影响; 当某点在 $\phi(\varepsilon_{\min})$ 上方时, 该点离 $\psi(\varepsilon_{\min})$ 越近, 拉裂可能性越大。基于以上分析, 定义 r_f^i

$$r_f^i = \begin{cases} [\varepsilon_{\max}^i - \phi(\varepsilon_{\min}^i)]^2, & \varepsilon_{\max}^i \geq \phi(\varepsilon_{\min}^i) \\ 0, & \varepsilon_{\max}^i < \phi(\varepsilon_{\min}^i) \end{cases} \quad (20)$$

并定义拉裂评价函数

$$r_f = \sum_{i=1}^n r_f^i \quad (21)$$

图 6 成形极限图中曲线 $\psi(\varepsilon_{\min})$ 与 $\phi(\varepsilon_{\min})$ 示意图

(3) 起皱评价函数。在成形极限图中添加一条虚线表示的曲线 $\varphi(\varepsilon_{\min}) = -\varepsilon_{\min}$, 如图 7 所示。当点位于曲线 $\varphi(\varepsilon_{\min})$ 下方时, 根据体积不变定义, 该点的厚向应变大于 0, 有起皱趋势。由此定义 r_w^i

$$r_w^i = \begin{cases} [\varphi(\varepsilon_{\min}^i) - \varepsilon_{\max}^i]^2, & \varepsilon_{\max}^i \leq \varphi(\varepsilon_{\min}^i) \\ 0, & \varepsilon_{\max}^i > \varphi(\varepsilon_{\min}^i) \end{cases} \quad (22)$$

并定义起皱评价函数

$$r_w = \sum_{i=1}^n r_w^i \quad (23)$$

图 7 成形极限图中曲线 $\varphi(\varepsilon_{\min})$ 示意图

(4) 整体性能评价函数。减小拉裂可能性与减小起皱可能性是一对相互矛盾的目标, 减小压边力可以有效避免拉裂, 但同时过小的压边力又会引起起皱。为了对这两个目标进行协调与折中, 采用加权系数法来构建整体性能的评价函数

$$r = -(\omega_1 \times r_f' + \omega_2 \times r_w') \quad (24)$$

式中, r_f' 、 r_w' 定义如下

$$r_f' = \frac{r_f - \min r_f}{\max r_f - \min r_f} \quad (25)$$

$$r_w' = \frac{r_w - \min r_w}{\max r_w - \min r_w} \quad (26)$$

式中, ω_1 、 ω_2 为权重系数, 且有

$$\omega_1 + \omega_2 = 1 \quad (27)$$

由于在实际应用中, 考虑抗拉裂性比抗起皱性重要, 取 $\omega_1=0.6$, $\omega_2=0.4$ 。

4 试验与结果分析

本文利用圆筒件拉深成形的有限元模型进行试验来验证压边力控制模型的有效性, 并将本文的控制模型与基于 NFQ 算法的压边力控制模型^[21]以及几种典型压边力加载模式的控制结果进行对比。

4.1 RDPG 建模过程分析

在 RDPG 中, 行动者网络与评论家网络的结构

均为4层LSTM循环神经网络,隐藏层节点为80,学习率分别为0.001与0.002;在NFQ中,网络结构与文献[21]中一致,学习率为0.001,折扣率为0.9,动作空间为离散空间{3 000, 4 000, 5 000, 6 000, 7 000, 8 000, 9 000, 10 000, 11 000, 12 000, 13 000}。在前250个episode,利用初始化网络与探索噪声来产生初始回放经验。开始训练后,每5个episode输出一次回报值。

图8展示了在不同算法下,回报值随训练局数的变化情况。由图可知,在RDPG控制下,回报值大约在第800个episode收敛;在NFQ控制下,回报值大约在第1 000个episode收敛。这表明RDPG算法的策略学习效率比NFQ算法更高,体现了基于梯度的强化学习算法相对于基于价值的强化学习算法具有收敛更加迅速的优势。

图8 RDPG与NFQ训练过程回报值变化

RDPG控制下的最优回报值为-0.29, NFQ控制下的最优回报值为-0.32。两者最优回报值对应的压边力控制轨迹如表2所示。由表可知,这两种控制策略的趋势大致相同,但压边力的控制精度不同。RDPG的动作空间为连续动作空间,而NFQ的动作空间为离散动作空间,因此RDPG能够学习到更加精确的控制策略从而得到更优的回报值。

表2 最优控制策略

	BHF1	BHF2	BHF3	BHF4	BHF5
RDPG	12 890	12 364	7 444	3 401	3 430
NFQ	13 000	13 000	8 000	4 000	3 000

4.2 成品质量分析

利用RDPG与NFQ所学习到的最优压边力控制策略,在ABAQUS中分别进行板材拉深仿真。根据仿真结果输出U2等值线分布云图、Mises应力分布云图、厚度分布云图、FLDCRT分布云图和成形极限图,对于这两种算法控制下的成品的成形质量进行分析。

U2等值线分布云图表示成品的每个有限元单元在轴向上的位移值。由图9和图10可知,RDPG控制下成品的法兰边位移为0.101 8 mm, NFQ算法控制下成品法兰边位移为0.064 1 mm。这表示在

RDPG控制下成品的起皱情况比在NFQ算法控制下严重一点,但是两者的起皱情况均在合理范围内,且从最终的成品三维图中没有发现明显起皱现象。

图9 RDPG控制下成品的U2等值线分布云图

图10 NFQ控制下成品的U2等值线分布云图

Mises应力分布云图展示了成品的各有限元单元Mises应力的大小分布情况。从图11和图12可以看出,在RDPG算法控制下,最大与最小应力分别为558.8 MPa和116.1 MPa;在NFQ算法控制下,最大与最小应力分别为559 MPa和218.3 MPa。这表明在NFQ控制下的成品的整体内部应力更大。

图11 RDPG控制下成品的Mises应力分布云图

图12 NFQ控制下成品的Mises应力分布云图

厚度分布云图体现了成品各处厚度的分布情况。由图 13 和图 14 可知, RDPG 控制下成品最小厚度为 0.873 2 mm, NFQ 控制下的成品最小厚度为 0.868 1 mm。从总体上看, RDPG 算法控制下的成品厚度分布更加均匀, 减薄率更低。

图 13 RDPG 控制下成品的厚度分布云图

图 14 NFQ 控制下成品的厚度分布云图

FLDCRT 分布云图表征了成品内各有限元单元的拉裂倾向分布情况, 数值越大则该处的拉裂倾向越严重。由图 15 和图 16 可知, RDPG 控制下成品最小安全裕度为 38%; NFQ 控制下成品最小安全裕度为 36%。两种算法控制下的拉裂倾向最严重的区域均出现在杯壁上中部分, NFQ 控制下成品的拉裂倾向更加严重。

图 15 RDPG 控制下成品的 FLDCRT 分布云图

最后输出 RDPG 与 NFQ 控制下成品的成形极限图进行成形质量分析(图 17)。由于 NFQ 控制下的整体压边力较 RDPG 大, 因此与 RDPG 相比, NFQ 控制下成品的成形极限图中各点整体位置靠上, 产品整体的拉裂倾向更加严重。总体上, 图中各点均处于较为合理的区域范围, 没有出现破裂的点。这

表明在 RDPG 与 NFQ 控制下成品的成形性能均良好, 然而 RDPG 控制下成品的成形极限图上各点的分布更集中, 成形质量更好。

图 16 NFQ 控制下成品的 FLDCRT 分布云图

图 17 RDPG 与 NFQ 控制下成品的成形极限图

在 RDPG 控制下, 破裂倾向指标 r_f 为 0.059, 起皱倾向指标 r_w 为 0.671; 在 NFQ 控制下, 破裂倾向指标 r_f 为 0.075, 起皱倾向指标 r_w 为 0.678。在 RDPG 控制下, 成品的破裂倾向指标与起皱倾向指标均小于 NFQ 控制下的情况。

4.3 RDPG 与传统压边力加载模式控制结果比较

本节将拉深加工中典型的 10 种压边力控制模式的控制结果与 RDPG 的控制结果作比较。传统压边力控制模式如图 18 所示。

图 18 传统压边力加载模式图

这些压边力加载模式涵盖了拉深加工领域大部分的压边力加载方式。RDPG 与传统压边力加载模式的控制结果回报值对比如表 3 所示。

表 3 不同加载模式回报值对比

压边力控制模式	回报值
恒定加载模式	-0.61
渐增加加载模式	-0.46
递减加载模式	-0.36
成形最后减小加载模式	-0.43
峰型加载模式	-0.38
谷型加载模式	-0.45
成形最后渐增加加载模式	-0.39
成形最后突增加加载模式	-0.40
折线形加载模式	-0.35
V 型加载模式	-0.33
RDPG 学习到的加载模式	-0.29

由表 3 可知, RDPG 学习到的压边力控制策略的结果优于所有的传统压边力加载模式的控制结果。

4.4 LSTM 网络重要性及输出结果分析

为了探究 LSTM 循环神经网络对于处理 POMDP 问题的重要性, 将 RDPG 与另一种采用普通神经网络的深度强化学习算法, 深度确定性策略梯度^[29](Deep deterministic policy gradient, DDPG)的控制结果进行比较。如图 19 所示, 在 DDPG 控制下回报值并未有明显收敛, 表示 DDPG 无法对压边力的控制策略进行优化。通过对比可知, 强化学习与深度 LSTM 网络的结合能够很好地解决 POMDP 问题, 从观测历史中获取隐藏的系统状态

用于策略学习。

以 RDPG 学习到的最优压边力控制策略下产生的拉深过程轨迹历史 h_r 作为输入, 考察 LSTM 网络在各控制步长输出的压边力变化情况, 如图 20 所示。

图 19 RDPG 与 DDPG 训练过程回报值变化

图 20 LSTM 网络输出动作值变化

初始经验回放中存在相当一部分回报值较小的转移经验, 对应于成品拉裂趋势较大或产生拉裂缺陷的拉深过程。在训练初期, 由于这部分经验参与更新梯度的计算, LSTM 网络在各控制步长的输出均趋向于较小的压边力值。随着训练进行, 转移经验中拉裂趋势较大的经验所占比重逐渐减小, 起皱趋势较大的经验所占比重逐渐上升, 各步长的压边力值逐渐增大并收敛于最优压边力值。因为本文仅在最后一个步长的压边力执行后得到一个回报值, 所以后期步长所接收到的回报信号更强, 这导致了越靠后的步长的压边力收敛越迅速。

目前本文所建立的基于深度强化学习的压边力控制模型仅仅在圆筒件拉深过程中进行了试验验

证, 未来该控制模型还能够延伸到更加复杂的非对称拉深过程的压边力控制问题中, 比如在压边器的不同位置施加不同的压边力。此外, 由于本文提出的基于深度强化学习的压边力控制模型往往需要进行大量的离线训练, 未来可考虑结合迁移学习来提升控制模型在不同压边力控制任务中转换的效率。

5 结论

(1) 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, 相较于传统的压边力控制方法, 本模型控制下成品的成形质量更好。

(2) 相较于基于价值的强化学习算法 NFQ, 采用基于策略梯度的深度强化学习算法 RDPG, 使得策略的学习效率和控制精度更高, 成品质量更好。

(3) LSTM 网络能够从观察历史中有效地提取拉深过程的状态信息。

参 考 文 献

- [1] 陈一哲, 刘伟, 苑世剑. 薄板液压成形起皱预测及控制研究进展[J]. 机械工程学报, 2016, 52(4): 20-28.
CHEN Yizhe, LIU Wei, YUAN Shijian. Research development on wrinkling prediction and suppression for sheet hydroforming of thin-walled deep drawing parts[J]. Journal of Mechanical Engineering, 2016, 52(4): 20-28.
- [2] 林晓娟, 于宽, 潘悦飞, 等. 变压边力拉深工艺的研究[J]. 锻压装备与制造技术, 2009, 44(1): 59-62.
LIN Xiaojuan, YU Kuan, PAN Yuefei, et al. Application of response surface method in optimization of multi-stage deep drawing parameters[J]. China Metal Forming Equipment and Manufacturing Technology, 2009, 44(1): 59-62.
- [3] YOUNIS A D. The effect of using non-uniform blank holder force in deep drawing process on the thickness distribution along the cup[J]. Al Rafdain Engineering Journal, 2013, 21(2): 25-31.
- [4] 马瑞, 赵军, 屈晓阳. 盒形件智能化拉深变压边力控制规律及其预测[J]. 机械工程学报, 2010, 46(8): 32-36.
MA Rui, ZHAO Jun, QU Xiaoyang. Control law of variable BHF and its prediction in intelligent deep drawing for rectangular box[J]. Journal of Mechanical Engineering, 2010, 46(8): 32-36.
- [5] 谢延敏, 于沪平, 陈军, 等. 基于灰色系统理论的方盒件拉深稳健设计[J]. 机械工程学报, 2007, 43(3): 54-59.
XIE Yanmin, YU Huping, CHEN Jun, et al. Rectangular box deep drawing robust design based on gray system theory[J]. Journal of Mechanical Engineering, 2007, 43(3): 54-59.
- [6] 刘罡. 基于回弹控制的提高轿车冲压件成形精度方法研究[D]. 上海: 上海交通大学, 2001.
LIU Gang. Study on the method to improve forming accuracy of autobody panels based on springback control[D]. Shanghai: Shanghai Jiao Tong University, 2001.
- [7] CAO Jun, BOYCE M C. A predictive tool for delaying wrinkling and tearing failures in sheet metal forming[J]. Journal of Engineering Materials and Technology, 1997, 119(4): 354-365.
- [8] KRISHNAN N, CAO Jun. Estimation of optimal blank holder force trajectories in segmented binders using an ARMA model[J]. Journal of Manufacturing Science and Engineering, 2003, 125(4): 763-770.
- [9] SHENG Ziqiang, JIRATHEARANAT S, ALTAN T. Adaptive FEM simulation for prediction of variable blank holder force in conical cup drawing[J]. International Journal of Machine Tools and Manufacture, 2003, 44(5): 487-494.
- [10] 孙成智, 陈关龙, 林忠钦, 等. 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): 1086-1090.
SUN Chengzhi, CHEN Guanlong, LIN Zhongqin, et al. The optimization of variable blank-holder forces based on numerical simulation[J]. Journal of Shanghai Jiao Tong University, 2004, 38(7): 1086-1090.
- [11] HILLMANN M, KUBLI W. Optimization of sheet metal forming processes using simulation programs[C]//Proceedings of Numisheet'99, France: Besancon, 1999: 287-292.
- [12] 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, France: Besancon, 1999: 305-310.
- [13] 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): 133-143.
- [14] 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): 9-12.
HUANG Yuping, RUAN Feng, CAI Zhixing. Using neural network to optimize blank holder force[J]. Die and Mould Industry, 2008, 34(7): 9-12.
- [15] LI Qihan, LI Mengzhe, TIAN Ye. Optimization technology

- of sheet metal deep drawing with variable blank holder force[C]//Proceedings of 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, IEEE, 2010, 2: 495-497.
- [16] QIAN Zhiping, MA Rui, ZHAN Jun, et al. Intelligent control technology for deep drawing of sheet metal[J]. Journal of Central South University of Technology, 2010, 15(S2): 273-277.
- [17] 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 (1): 421-426.
- [18] 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): 411-415.
- 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 Jiao Tong University, 2001, 35(3): 411-415.
- [19] MANABE K, KOYAMA H, YOSHIHARA S, et al. Development of a combination punch speed and blank-holder fuzzy control system for the deep-drawing process[J]. Journal of Materials Processing Technology, 2002, 125: 440-445.
- [20] YAGAMI T, MANABE K. FE analysis on deformation mechanism of strain-rate-sensitive materials in cylindrical deep-drawing with combination punch speed and blank holder control[J]. Journal of Solid Mechanics and Materials Engineering, 2007, 1(12): 1385-1396.
- [21] 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>.
- [22] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. Cambridge: MIT Press, 2018.
- [23] HOCHREITER S. The vanishing gradient problem during learning recurrent neural nets and problem solutions[J]. International Journal of Uncertainty, Fuzziness And Knowledge-Based, Systems, 1998, 6(2): 107-116.
- [24] HEES N, HUNT J J, LILLICRAP T P, et al. Memory-based control with recurrent neural networks [EB/OL]. (2015-12-14) [2019-08-05]. <https://arxiv.org/abs/1512.04455>.
- [25] SILVER D, LEVER G, HEES N, et al. Deterministic policy gradient algorithms[C]//Proceedings of the International Conference on Machine Learning. Beijing: ACM, 2014: 387.
- [26] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529.
- [27] 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.
- [28] 毛狄评. 方盒件拉深成形问题的有限元分析[D]. 哈尔滨: 哈尔滨工业大学, 2010.
- MAO Diping. Investigation of the square cup deep drawing by finite element method[D]. Harbin: Harbin Institute of Technology, 2010.
- [29] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. Computer Science, 2015, 8(6): A187.

作者简介: 郭鹏, 男, 1994 年出生。主要研究方向为生产系统控制。

E-mail: guopeng19940821@163.com

张新艳, 女, 1974 年出生, 讲师, 硕士研究生导师。主要研究方向为物流系统规划与分析。

E-mail: alicexy@tongji.edu.cn

余建波(通信作者), 男, 1978 年出生, 博士, 教授, 博士研究生导师。主要研究方向为设备智能预诊维护与可靠性、复杂制造过程质量控制、机器学习、生产系统设计优化。

E-mail: jbyu@tongji.edu.cn