改进深度确定性策略梯度算法及其在控制中的应用

张浩昱 熊 凯
(北京控制工程研究所空间智能控制技术国家级重点实验室 北京 100190)

摘 要 深度强化学习往往存在采样效率过低的问题,优先级采样可以在一定程度上提高采样效率。将优先级采样用于深度确定性策略梯度算法,并针对普通优先级采样算法复杂度高的问题提出一种小样本排序的思路。仿真实验结果表明,这种改进的深度确定性策略梯度算法提高了采样效率,具有好的训练效果。将深度确定性策略梯度算法用于小车方向控制,相比于传统的 PID 控制,该算法避免了人工调整参数的问题,具有更广阔的应用前景。

关键词 深度强化学习,深度确定性策略梯度,优先级采样,方向控制
中图法分类号 TP183 **文献标识码** A

Improved Deep Deterministic Policy Gradient Algorithm and Its Application in Control

ZHANG Hao-yu XIONG Kai
(Science and Technology on Space Intelligent Control Laboratory, Beijing Institute of Control Engineering, Beijing 100190, China)

Abstract Deep reinforcement learning often has the problem of low sampling efficiency. Priority sampling can improve sampling efficiency to a certain extent. The prioritized experience replay was applied to the deep deterministic policy gradient algorithm, and a small sample sorting method was proposed for the high complexity of the general prioritized experience replay algorithm. Simulation results show that the improved deep deterministic policy gradient algorithm improves the sampling efficiency and has better training effect. The algorithm is applied in the direction control of a car, compared with traditional PID control, this algorithm can avoid the problem of manual adjustment of parameters and has a wider application prospect.

Keywords Deep reinforcement learning, Deep deterministic policy gradient, Prioritized experience replay, Direction control

1 引言

近年来,深度强化学习(DRL)在围棋[1-2]、视频游戏[3-4]等领域获得了巨大的成功。深度 Q 网络(DQN)通过引入经验回放技术[3]解决了使用神经网络去拟合值函数导致训练结果不收敛的问题,Lillicrap 等将经验回放技术与确定性策略梯度(DPG)[5]相结合,提出了深度确定性策略梯度(DDPG)[6],成功地将一般强化学习所能解决的低维离散的动作空间的问题扩展到高维连续动作空间。

经验回放技术通过采集大量训练样本,在训练时从样本集中随机挑选训练样本进行训练,从而打破训练数据之间的关联性,这种方法被证明是有效的。然而,这种方法的缺点在于采样效率很低,很难选取到有效的训练数据进行训练,因此算法的收敛速度很慢,甚至难以收敛。针对这一问题,Schual 等提出了经验优先级方法[7]并将其应用于深度 Q 网络,依据时序误差确定样本的重要程度,优先选择更重要的样本进行训练,从而加速训练过程。

PID 控制不依赖被控对象的数学模型,实用性很强,是目前应用最为广泛的控制策略,然而其控制参数往往需要设计者依靠自身经验去选择。相比于 PID 控制,深度确定性策略梯度算法同样不依赖被控对象的数学模型,同时可以避开人工调整参数,具有广泛的应用前景。

本文针对确定性策略梯度算法,研究将经验优先级方法应用于深度确定性策略梯度算法的可行性。由于经验优先级方法需要对所有样本进行优先级计算并排序,导致运算复杂度较高,本文提出了一种小样本排序思路,并结合倒立摆(CartPole)模型,验证了算法的有效性。将改进的深度确定性策略梯度算法用于四轮小车转向控制,并对比该算法的控制效果与 PID 控制算法的控制效果,仿真实验结果表明:本文提出的算法具有很好的控制效果,在实际应用中具有广阔的应用前景。

2 相关工作

2.1 深度确定性策略梯度算法

强化学习的描述通常是基于马尔科夫决策过程[8]的(MDP),其任务对应了一个四元数组 $E=\langle S,A,P,R\rangle$,其中,S 为状态集,A 为动作集,P 为状态转移概率,R 为奖赏函数。在强化学习中,策略记为 π,表示的是状态到动作

本文受北京市自然科学基金(4162070),国家自然科学基金 (61573059)资助。
张浩昱(1994—),男,硕士生,主要研究方向为深度强化学习,E-mail:Haoy_Zhang@163.com;熊 凯(1976—),男,博士,研究员,主要研究方向为自适应滤波和航天器自主导航,E-mail:17600517255@163.com(通信作者)。

的映射，即$S\rightarrow A$。

通常使用累积折扣奖赏[9]来定义 t 时刻的状态回报，即

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

其中，γ 为折扣因子，表明越远处的奖赏对当前状态的评估影响越小，$r(s_i,a_i)$表示在状态 s_i 选择动作 a_i 所获得的奖赏值。设初始状态为 s_1，在某一策略 π 下，状态分布服从 ρ^{π}，则强化学习的任务为学习到一个策略 π，使得期望的初始状态回报达到最大。定义强化学习目标如下：

$$J=\mathbb{E}_{s\sim\rho,a\sim\pi}[R_1] \tag{2}$$

状态动作值函数 $Q^{\pi}(s,a)$表示从状态 S 出发，执行了动作 a 后，再执行策略 π 所带来的累积奖赏：

$$Q^{\pi}(s_t,a_t)=\mathbb{E}_{s\sim\rho,a\sim\pi}[R_t|s_t,a_t] \tag{3}$$

时序误差方法[10]应用了动态规划的思想，用下一时刻的状态动作值函数来估计当前的状态动作值函数：

$$Q^{\pi}(s_t,a_t)=\mathbb{E}_{r,s_{t+1}\sim E}[r(s_t,a_t)+\gamma\mathbb{E}_{a_{t+1}\sim\pi}[Q^{\pi}(s_{t+1},a_{t+1})]] \tag{4}$$

其中，r,s_{t+1}的分布服从于环境 E，a_{t+1}的分布服从于策略 π。

定义时序误差为：

$$\delta_t=r_t+\gamma Q^{\pi}(s_{t+1},a_{t+1})-Q^{\pi}(s_t,a_t) \tag{5}$$

在不需要指明特定策略的情况下，Q^{π}往往可以直接使用 Q 来代替。

使用神经网络去逼近状态动作值函数，设网络的参数为 $\boldsymbol{\theta}^Q$，可以定义网络的损失函数为

$$L(\boldsymbol{\theta}^Q)=\mathbb{E}_{s\sim\rho,a\sim\pi}[(y_t-Q(s_t,a_t|\boldsymbol{\theta}^Q))^2] \tag{6}$$

其中，

$$y_t=r_t+\gamma Q(s_{t+1},a_{t+1}|\boldsymbol{\theta}^Q) \tag{7}$$

在确定性策略梯度中，对于某一状态，策略不会以概率选取动作，而是输出确定的动作，相比于随机策略梯度不需要在动作空间积分，具有更高的效率。在确定性策略梯度中，策略以 μ 表示，其参数为 $\boldsymbol{\theta}$，强化学习的目标可以表示为策略 μ 的函数，可以使用梯度下降法去优化参数。可以证明目标函数 $J(\mu_\theta)$对参数 $\boldsymbol{\theta}$ 的梯度为[5]：

$$\nabla_\theta J(\mu_\theta)=\mathbb{E}_{s\sim\rho^{\mu}}[\nabla_\theta\mu_\theta(s)_a Q^{\mu}(s,a)|_{a=\mu_\theta(s)}] \tag{8}$$

借助于 Actor-Critic 框架[11]，我们可以使用 Critic 网络去评价动作，指导网络参数更新，其参数为 $\boldsymbol{\theta}^Q$；使用 Actor 网络去选择动作，其参数为 $\boldsymbol{\theta}^{\mu}$。根据选取的训练数据，可以得到时序差分(TD)误差为：

$$\delta_t=r_t+\gamma Q^{\boldsymbol{\theta}^Q}(s_{t+1},\mu_{\theta^{\mu}}(s_{t+1}))-Q^{\theta^Q}(s_t,a_t) \tag{9}$$

使用梯度下降法更新网络参数，设学习率为：

$$\boldsymbol{\theta}^Q_{t+1}=\boldsymbol{\theta}^Q_t+\alpha_Q\delta_t\nabla_{\boldsymbol{\theta}^Q}Q^{\boldsymbol{\theta}^Q}(s_t,a_t) \tag{10}$$

$$\boldsymbol{\theta}^{\mu}_{t+1}=\boldsymbol{\theta}^{\mu}_t+\alpha_{\mu}\nabla_{\boldsymbol{\theta}^{\mu}}\mu_{\boldsymbol{\theta}^{\mu}}(s_t)\nabla_a Q^{\boldsymbol{\theta}^Q}(s_t,a_t)|_{a=\mu_\theta(s)} \tag{11}$$

深度确定性策略梯度采用了深度 Q 网络的思想，Actor 网络和 Critic 网络都设置了独立的目标网络，其参数更新速度更加缓慢，稳定性[12]得到了提升。

2.2 经验回放技术

在深度 Q 网络之前使用神经网络去拟合值函数往往会出现不收敛的问题，其原因就在于利用有标签的数据去训练神经网络，需要假设这些数据都是独立同分布的，但是通过强化学习采集到的数据之间存在着某些关联，因此直接用来训练神经网络会导致不稳定。深度 Q 网络的一个关键就是提出了经验回放的概念，该概念被用于打破训练数据之间的这种关联。这种方法建立了一个经验集 D，并将其用于存放历史经验 $e_t=(s_t,a_t,r_t,s_{t+1})$，迭代时从经验集中随机抽取一组历史经验用于估计 Q 算法中的期望值函数，再执行一次梯度下降。通过这种方法可以有效打破训练数据之间的联系，使得最终的训练可以收敛。

然而普通的经验回放技术从样本集中随机抽取数据进行训练，忽略了样本之间的差异，因此采样效率很低，往往需要很长时间才能收敛，甚至会出现不收敛的情况。Schaul 等针对深度 Q 网络的训练过程，将 TD 误差的大小作为衡量样本重要性的指标[7]，提出了优先级经验回放的方法，大大提高了样本的采样效率，提高了收敛速度。但是这种方法需要对样本集中的所有样本计算 TD 误差，并进行排序，算法的复杂程度被大幅提高。

3 改进深度确定性策略梯度算法

优先级经验回放技术在深度 Q 网络中取得了成功，本文针对深度确定性策略梯度训练缓慢的问题，将经验回放技术应用于深度确定性策略梯度算法，并针对优先级经验回放技术中算法复杂度较高的问题，提出了一种小样本优先级排序的方法。

深度确定性策略梯度采用 Actor-Critic 框架进行训练(见图 1)，Actor 使用确定性策略梯度方法选择动作；Critic 使用时序差分方法对值函数进行估计，用于评价 Actor 选择动作的优劣。在这个框架中，Critic 网络产生的 TD 误差信号同时用于指导 Critic 网络和 Actor 网络更新，因此使用 TD 误差来确定训练样本的优先级是十分合理的。

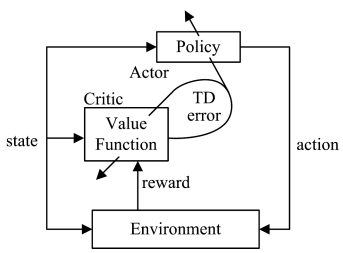

图 1 Actor-Critic 框架

对于任意时刻 t，每个样本的优先级表示为：

$$p_t=|\delta_t|=|r_t+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)| \tag{12}$$

根据每个样本的优先级，确定第 i 个样本选取的概率分布：

$$P(i)=\frac{p_i}{\sum_k p_k} \tag{13}$$

依据上式确定的样本选取概率分布会优先选取具有较大优先级的样本进行训练，但是在训练初期，TD 误差并不准确，因此在训练初期，需要适当减少优先级在采样中的权重，同时为了使得每个样本都有机会被选中，引入偏置量。则样本选取的概率分布可以改为：

$$P(i)=a\times\frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}}+b \tag{14}$$

其中，α 为优先级权重，当 $\alpha=0$ 时，每个样本的优先级相同，完全不使用优先级采样。b 为一个小量，使得优先级几乎为 0 的样本也有可能被选中，α 为归一化系数，使得最终概率和仍为 1。

一般的优先级经验回放技术需要对所有样本计算 TD 误差并排序，增加了算法的计算成本。本文设计了一种小样本排序方法，即在抽取批量样本数据时，首先随机抽取数倍的批量样本数据，在随机抽取样本的基础上进行优先级排序，按照

优先级选取批量的训练样本。通过这种小样本排序方法，可以大大减少算法的复杂度，同时在一定程度上提高了低优先级样本被选中的概率，增加了训练样本的多样性，防止出现对值函数的过估计[13]。

改进深度确定性策略梯度算法如算法 1 所示。

算法 1　改进深度确定性策略梯度

随机初始化 Actor 网络 μ 和 Critic 网络 Q，参数 $\boldsymbol{\theta}^{\mu}$ 和 $\boldsymbol{\theta}^{Q}$

初始化目标网络 μ' 和 Q'，参数 $\boldsymbol{\theta}^{\mu'}\leftarrow\boldsymbol{\theta}^{\mu}$ 和 $\boldsymbol{\theta}^{Q'}\leftarrow\boldsymbol{\theta}^{Q}$

初始化经验回放集 R

设定参数 a，b，α，τ

在每一个训练回合中

　初始化动作选择噪声 N

　初始化起始状态 S_1

　在每一步中

　　根据当前策略和探索噪声选择动作 $a_t=\mu(s_t\mid\boldsymbol{\theta}^{\mu})+N$

　　执行动作 a_t 得到回报 r_t 和下一个状态 s_{t+1}

　　将 (s_t,a_t,r_t,s_{t+1}) 存入 R

　　从 R 中随机抽取数倍的批量样本数据

　　计算每个样本的优先级 $P(i)=a\times\frac{p_i^{\alpha}}{\sum_k p_k^{\alpha}}+b$

　　根据优先级选取 N 个样本数据

　　设 $y_i=r_i+\gamma Q'(s_{i+1},\mu'(s_{i+1}\mid\boldsymbol{\theta}^{\mu'})\mid\boldsymbol{\theta}^{Q'})$

　　通过最小化损失 $\frac{1}{N}\sum_i(y_i-Q(s_i,a_i\mid\boldsymbol{\theta}^{Q}))^2$ 更新 Critic 网络

　　通过策略梯度更新 Actor 网络

$$\nabla_{\theta^{\mu}}J\approx\frac{1}{N}\sum_i\nabla_{\theta^{\mu}}\mu(s\mid\theta^{\mu})\mid_{s_i}\nabla_a Q(s,a\mid\boldsymbol{\theta}^{Q})\mid_{a=\mu_\theta(s)}\mid_{s=s_i,a=\mu(s_i)}$$

　　更新目标网络

$$\boldsymbol{\theta}^{Q'}\leftarrow\tau\boldsymbol{\theta}^{Q}+(1-\tau)\boldsymbol{\theta}^{Q'}$$

$$\boldsymbol{\theta}^{\mu'}\leftarrow\tau\boldsymbol{\theta}^{\mu}+(1-\tau)\boldsymbol{\theta}^{\mu'}$$

　结束

结束

4　实验结果

本文将改进的深度确定性策略梯度用于经典的倒立摆控制问题以验证算法的有效性，并与传统的深度确定性策略梯度算法进行对比分析。在改进的深度确定性策略梯度算法的基础上，将该算法用于小车方向控制，并将其与传统的 PID 控制算法进行对比分析。

强化学习中往往通过环境回报的奖赏值作为训练效果的评估指标，本文使用平均奖赏来评估算法，并比较算法的收敛速度。

为使实验结果更具说服力，本文记录了不同算法仿真 10 次的数据，并取平均值绘制结果。由于强化学习的训练效果受初始状态影响，为了排除随机干扰，在所有实验中使用了相同的初始随机种子。图 2 展示了两种算法在倒立摆环境下的平均奖赏。

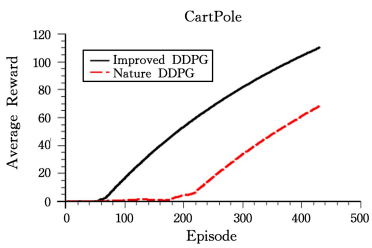

图 2　倒立摆平均奖赏

可以看出：改进的深度确定性策略梯度算法从第 60 个回合开始，平均奖赏就不断提高，而原始的深度确定性策略梯度算法需要从第 170 个回合开始。图 3 展示了每回合的奖赏变化，可以更直观地反映训练过程。

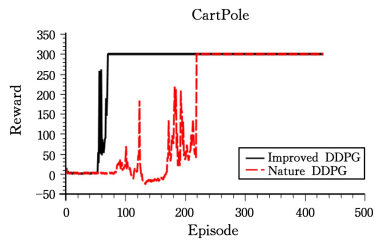

图 3　倒立摆每回合奖赏

本文将改进的深度确定性策略梯度算法用于四轮小车的方向控制。小车采用后轮驱动方式，转向通过后轮差动进行控制。图 4 展示了在方向控制中每回合获得的奖赏值，从图中可以看出：随着训练过程的进行，获得的奖赏逐渐增大，最终奖赏值收敛于 0，即误差为 0。

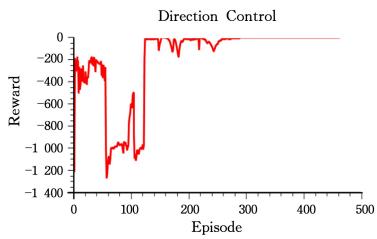

图 4　小车方向控制每回合奖赏曲线

图 5 展示了训练完成的网络与 PID 控制误差的对比曲线，其中 PID 控制器选择了两组较好的控制参数进行仿真实验。

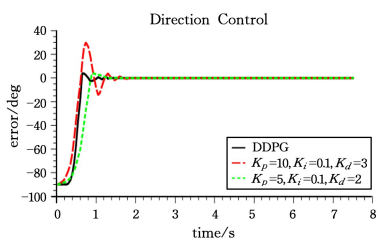

图 5　小车方向控制对比曲线

从仿真结果可以看出，由于 PID 控制的参数为定值，因此响应迅速的控制器会有较大的超调量，而超调量较小的控制器响应时间较长。深度确定性策略梯度算法结合了不同的 PID 控制参数的优点，可以在误差较大时快速响应，减小误差；在误差减小时相应减少控制量，从而减少超调量，具有更好的控制效果。

结束语　本文针对深度确定性策略梯度算法，研究使用优先级经验回放加速深度确定性策略梯度训练过程的可行性，同时针对优先级经验回放算法复杂度高的问题，提出一种基于样本优先级的小样本排序方法。本文通过经典倒立摆控制问题对原始深度确定性策略梯度算法与改进深度确定性策略梯度算法进行了对比研究，结果表明：本文提出的改进深度确定性策略梯度方法都够显著提高采样效率，具有更好的训练效果。在此基础上进行小车的方向控制，并与传统的 PID 控制进行对比，实验结果表明深度确定性策略梯度具有更好的控制效果。从实验结果来看，将深度强化学习方法用于控

（下转第 570 页）