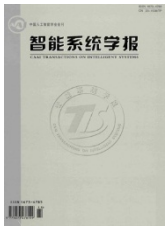

智能系统学报
CAAI Transactions on Intelligent Systems
ISSN 1673-4785,CN 23-1538/TP

《智能系统学报》网络首发论文

题目：　强化学习稀疏奖励算法研究——理论与实验
作者：　杨瑞，严江鹏，李秀
网络首发日期：　2020-09-21
引用格式：　杨瑞，严江鹏，李秀．强化学习稀疏奖励算法研究——理论与实验．智能系统学报. https://kns.cnki.net/kcms/detail/23.1538.TP.20200921.1556.004.html

智 能 系 统 学 报

CAAI Transactions on Intelligent Systems

DOI:10.11992/tis. 202003031

强化学习稀疏奖励算法研究——理论与实验

杨瑞[1]，严江鹏[1]，李秀[2]

（1. 清华大学 自动化系，北京 100084；2. 清华大学 深圳国际研究生院，广东 深圳 518055）

摘　要：近年来，强化学习在游戏、机器人控制等序列决策领域都获得了巨大的成功，但是大量实际问题中奖励信号十分稀疏，导致智能体难以从与环境的交互中学习到最优的策略，这一问题被称为稀疏奖励问题。稀疏奖励问题的研究能够促进强化学习实际应用与落地，在强化学习理论研究中具有重要意义。本文调研了稀疏奖励问题的研究现状，以外部引导信息为线索，分别介绍了奖励塑造、模仿学习、课程学习、事后经验回放、好奇心驱动、分层强化学习等方法。本文在稀疏奖励环境 Fetch Reach 上实现了以上 6 类方法的代表性算法进行实验验证和比较分析。使用外部引导信息的算法平均表现好于无外部引导信息的算法，但是后者对数据的依赖性更低，两类方法均具有重要的研究意义。最后，本文对稀疏奖励算法研究进行了总结与展望。

关键词：强化学习；深度强化学习；机器学习；稀疏奖励；神经网络；人工智能；深度学习；综述

中图分类号：TP181　　**文献标志码：**A

中文引用格式：杨瑞，严江鹏，李秀. 强化学习稀疏奖励算法研究——理论与实验[J]. 智能系统学报，DOI：10.11992/tis. 202003031.

英文引用格式：YANG Rui, YAN Jiangpeng, LI Xiu. Summary of Sparse Reward Algorithms in Reinforcement Learning -- Theory and Experiment [J]. CAAI transactions on intelligent systems，DOI：10.11992/tis. 202003031.

A Survey on Sparse Reward Algorithms in Reinforcement Learning -- Theory and Experiment

YANG Rui[1]，YAN Jiangpeng[1]，LI Xiu[2]

(1. Department of Automation, Tsinghua University, Beijing 100084, China; 2. Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China)

Abstract： In recent years, reinforcement learning has achieved great success in many sequential decision-making applications such as games and robot control. But the reward signals are very sparse in numerous real-world situations, which makes it difficult for agents to learn optimal strategy from interacting with the environment. Such problem is called sparse reward problem. The research on sparse reward can promote the actual applications of reinforcement learning and it is significant for reinforcement learning theory. We

基金项目：国家自然科学基金项目(41876098).

通讯作者：李秀. Email:li.xiu@sz.tsinghua.edu.cn.

investigated the current research status of sparse reward problem, and used the external information as the clue to introduce the following six classes of algorithms, reward shaping, imitation learning, curriculum learning, hindsight experience replay, curiosity-driven algorithms, and hierarchical reinforcement learning. We implemented typical algorithms of the six classes to perform experiments on the sparse reward environment Fetch Reach and did thorough comparison and analysis. Algorithms utilizing external information outperform algorithms without external information, but the latter is less dependent on data, and both methods are of great research significance. At the end of this paper, we summarized the existing sparse reward algorithms and future work.

Keywords: reinforcement learning;deep reinforcement learning; machine learning; sparse reward;neural networks; artificial intelligence;deep learning;summary

强化学习(reinforcement learning)是一类智能体在与环境的交互中不断试错来学习最优策略的机器学习方法[1][2]，主要用于解决序贯决策问题。在最近五年时间里，AlphaGO[3]、AlphaStar[4]、OpenAIFive[5]分别在围棋、星际争霸 II、Dota 2 击败了人类最高水平的玩家，强化学习一度成为了人工智能最热门的研究领域之一。AlphaGo 的主要作者 David Silver 认为[6]，强化学习与深度学习相结合，是实现通用人工智能（General Intelligence）的关键。

在强化学习中，奖励（Reward）起到了引导智能体学习方向的作用[7][8]，缺乏奖励信息将导致智能体学习缓慢甚至无法学习到最优策略，这就是稀疏奖励问题[9]（Sparse Reward Problem）。例如，在蒙特祖玛复仇游戏中，玩家需要依次执行上百个动作才能获得奖励，这使其成为了 Atari 游戏中最困难的任务之一。此外，在很多实际任务中，不存在现成的奖励值，人为设计的奖励函数又常常陷入局部最优[10]，这些问题限制了强化学习的实际应用。稀疏奖励问题的研究能够降低奖励函数的设计难度，提高学习算法的样本利用率，加速策略学习的速度，为强化学习的广泛应用与落地打下理论基础[11]。

本文总结了当前主流的稀疏奖励算法，围绕是否引入外部引导信息，将当下主流的稀疏奖励问题解决思路分为两类，分别介绍了奖励塑造[12]（Reward Shaping）、模仿学习[13]（Imitation Learning）、课程学习[14]（Curriculum Learning）和事后经验回放[10]（Hindsight Experience Replay）、好奇心驱动[15]（Curiosity-Driven Algorithms）、分层强化学习[16]（Hierarchical Reinforcement Learning）等 6 类算法，并在 Mujoco 的 Fetch Reach 环境[17]下进行了实验验证和分析，实验代码开源在以下地址：https://github.com/YangRui2015/Sparse-Reward-Algorithms。我们希望本文能够为稀疏奖励算法的进一步研究提供理论和实验基础。

1 强化学习与稀疏奖励问题数学模型

当强化学习问题满足马尔科夫性时，就能将其描述为由五元组(S, A, P, R, γ)定义的马尔科夫决策过程（MDP）[1][2]，其中S为状态空间，A为动作空间，P为状态转移概率矩阵，R为奖励值，$\gamma \in (0,1]$为折扣因子。智能体每个时刻观测到的状态$s_t \in S$，根据状态执行动作$a_t \in A$，环境接收到动作后转移到新的状态s_{t+1}并反馈一个数值的奖励r_t，如图 1 所示。

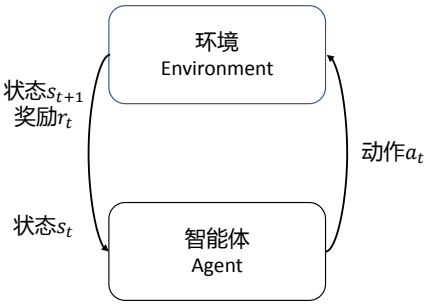

图 1 智能体与环境交互示意图

Fig.1 Schematic of interactions between agent and environment

强化学习的目标就是最大化累计折扣奖励值[1][2]：

$$R = \sum_{k=0}^{\infty} (\gamma^k r_{t+k}) . \tag{1}$$

为了完成最大化累计折扣奖励值的目标，需要引入策略（Policy）和值函数（Value Function）两

个重要概念。策略可以描述为当前状态s_t下选择动作a_t的概率：

$$\pi(a_t|s_t) = P(a_t|s_t)\,.$$

值函数是指从状态s或状态动作对(s,a)出发能获得累积奖励值的期望，用于评价状态、状态动作对的好坏。状态值函数$V(s)$，动作值函数$Q(s,a)$分别为：

$$V(s) = E[\sum_{k=0}^{\infty}(\gamma^k r_{t+k})\,|s_t = s]\,, \qquad (2)$$

$$Q(s,a) = E[\sum_{k=0}^{\infty}(\gamma^k r_{t+k})\,|s_t = s, a_t = a]\,. \qquad (3)$$

深度强化学习与传统强化学习的区别是使用了深度神经网络来拟合值函数、策略或环境动态模型[8]。神经网络的引入提高了强化学习解决大规模复杂问题的能力，在众多领域取得了令人瞩目的成绩[19]。目前的深度强化学习方法可以分为以下三类：基于值函数的方法、基于策略梯度（policy gradient）的方法以及 Actor-Critic 的方法[7]。三类方法的代表分别是 DQN[13][14]、REINFORCE[15]、Actor-Critic[16]，从 Actor-Critic 还衍生出 A3C[17]、PPO[18]、DDPG[19]等一系列当前主流的强化学习算法。

基于值函数的方法，以 DQN[20][21]为例，用w代表神经网络的参数，其损失函数为[20]：

$$loss = E\left[\left(r + \gamma \max_{a'} Q(s',a',w) - Q(s,a,w)\right)^2\right]. \quad (4)$$

根据式(1)、(3)、(4)，在奖励值几乎为零的稀疏奖励情况下，值函数的估计$Q(s,a)$接近零，值函数网络更新缓慢。

基于策略梯度的方法，以 REINFORCE[15]为例，θ代表神经网络的参数，其更新梯度为：

$$\nabla_\theta \pi(a|s;\theta) = \nabla_\theta log\pi(a|s;\theta)R_t\,. \qquad (5)$$

根据式(1)、(5)，在稀疏奖励的条件下，累积奖励值R_t接近零，因此策略网络更新缓慢。

基于 Actor-Critic[16]的方法同理，Critic 部分基于值函数更新，Actor 部分基于策略梯度更新，稀疏奖励的条件下两部分梯度更新均接近于零。

稀疏奖励问题除了奖励的稀疏性导致学习缓慢外，还可能存在稀疏性带来的估计不可靠的问题，由于奖励样本少，值函数估计的方差较大，这会导致模型训练难以收敛。研究者们为解决以上问题，进行了一系列的研究工作，我们将在第 2 节进行介绍。

2 稀疏奖励研究现状

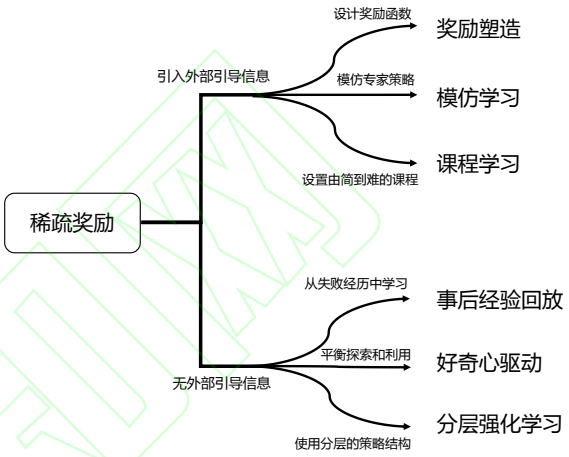

图 2 稀疏奖励主流算法

Fig.2 Mainstream sparse reward algorithms

目前解决稀疏奖励问题的算法主要有奖励塑造、模仿学习、课程学习、事后经验回放、好奇心驱动、分层强化学习等 6 类算法，我们可以根据是否引入外部引导信息将算法分为两大类，如图 2 所示。引入外部引导信息的算法通常针对特定问题，需要相应的领域知识或数据，泛化性较差，同时也因为先验知识的引入，降低了强化学习模型的学习难度，通常具有实现简单、学习速度快的特点，这一类型算法有奖励塑造、模仿学习、课程学习。无外部引导信息的算法通过挖掘模型、数据自身的潜能，泛化性能更好，但是模型通常更为复杂，这一类型算法包括好奇心驱动、事后经验回放、分层强化学习。接下来我们将对各个方法进行展开介绍。

2.1 奖励塑造

奖励塑造通常是利用先验知识人工设计附加奖励函数[12]来引导智能体完成期望任务的一类方法。合适的附加奖励函数能够有效克服稀疏奖励

问题中奖励的稀疏性，加快智能体学习速度。通常用$R(s,a,s')$表示原 MDP 的奖励函数，用$F(s,a,s')$ 表示附加奖励函数，使用奖励塑造后新 MDP 的奖励函数为：

$$R'(s,a,s') = R(s,a,s') + F(s,a,s').$$

但是在新的 MDP 问题中学习到的最优策略不一定是原 MDP 的最优策略，也就可能导致奖励塑造后学习到非理论最优的策略[27][28]。Ng 等[27]证明了当附加奖励函数可以表示为势能函数$\Phi(s)$（Potiental based Funciton）的差分形式时，能够保证最优策略不变。

$$F(s,a,s') = \gamma\Phi(s') - \Phi(s),$$

其中s'表示s的下一个状态，γ是原 MDP 中的折扣因子，势能函数$\Phi(s)$是状态到实数的映射。

Ng 等[27]使用距离、子目标来设计启发式的附加奖励函数，在表格问题中明显加速了学习过程；Jagodnik 等[29]使用距离信息计算和人为主观评价两种方式作为奖励函数来控制手臂仿真机器人，结果均好于优化的比例微分控制器（PD controller）；Ferreira 等[30]将奖励塑造的强化学习方法引入对话管理任务中，显著提高了前期训练阶段的表现。

为了克服奖励塑造泛化性较差的问题，研究者们提出了一些自动化地设计奖励函数的方法。Ng 等[31]提出逆强化学习（Inverse Reinforcement Learning）的概念，将专家示例看作为最优或者次优的策略，然后从专家示例的数据中学习出奖励函数用于指导智能体训练。Marthi 等[32]提出基于抽象函数（Abstrcat Function）的自动奖励塑造方法，通过求解抽象 MDP 问题的势能函数，再计算势能函数的差分就得到了附加奖励函数。

2.2 模仿学习

模仿学习是一类从示例数据中学习策略的方法[13]。稀疏奖励问题往往具有巨大的状态动作空间，难以直接进行探索和学习，使用示例数据进行监督式的学习能够使智能体快速掌握示例策略，极大减少了训练时间。

模仿学习中如果只使用示例数据进行监督学习，难以泛化到陌生的环境中，且长时间步的决策将导致误差累积，逐渐偏离示例策略[33]。Ross 等[33]证明了误差与时间步的平方成正比，为此提出在交互中缓慢移动策略分布的 SMILe（Stochastic Mixing Iterative Learning）算法，并理论证明了其收敛性。Nair 等[34]在策略梯度算法中引入行为克隆损失（Behavior Clone Loss）来学习示例策略，能够完成复杂的机械臂控制任务，其行动损失函数为：

$$L_{BC} = \sum_{i=1}^{N} \|\pi(s_i|\theta_\pi) - a_i\|_2^2 ,$$

其中(s_i, a_i)是示例数据，π和θ_π分别是智能体的策略和策略的参数。Ho 等[35]将对抗生成网络的思想引入模仿学习提出生成对抗模仿学习（Generative Adversarial Imitation Learning，GAIL），使用生成模型产生行为数据，使用判别模型区分行为数据和专家策略数据，GAIL 在复杂高维环境的模仿任务中超过了现有方法，其优化目标函数为：

$$E_\pi\big[log\big(D(s,a)\big)\big] + E_{\pi_E}\big[log\big(1 - D(s,a)\big)\big] - \lambda H(\pi),$$

其中，π是生成模型的策略，π_E是专家策略，D是判别模型，H是熵函数，λ是系数。

通常模仿学习中的示例状态动作对(s_t, a_t)不容易获得，而示例状态序列$(s_0, s_1, \dots, s_t)$更容易获得，从示例状态序列进行模仿学习的任务被称为“从观测学习”（Learning from Observation）[36]。Torabi 等[37]提出 BCO（Behavioral Cloning from Observation），通过学习环境动态模型预测示例观测序列的动作，然后使用行为克隆进行模仿学习，在多个任务中的表现超过了 GAIL。

2.3 课程学习

课程学习是机器学习中逐步增加任务难度以加快学习速度的方法。在强化学习中课程学习实质上在逐步调整学习的任务分布，智能体在简单任务上更容易获得奖励，将相对简单的任务上学到的策略迁移到复杂任务中，降低了在复杂任务中探索的难度，因此课程学习能够用于解决稀疏奖励问题[14]。

Elman 等[38]最早提出在语法学习任务中使用逐步增加训练数据复杂度的方法来训练神经网络，能够解决直接使用全部数据难以训练的问题。Bengio 等[14]认为，课程学习的本质是逐步调整学习样本的分布，在简单的样本分布上更容易学习

到泛化性好的策略，并通过实验证明了课程学习能够提高训练速度和收敛到更优解。Bengio 等[14]给出了课程学习的数学定义，z表示训练样本，$P(z)$表示目标训练集分布，$\lambda \in [0,1]$表示课程学习的阶段，$W_\lambda(z)$是λ阶段样本z的采样权重，λ阶段训练分布$Q_\lambda(z)$为：

$$Q_\lambda(z) \propto W_\lambda(z)P(z), \forall z,$$

其中Q_λ满足：

$$\int Q_\lambda(z)\,dz = 1\,,$$

$$Q_1(z) = P(z), \forall z\ .$$

不同课程阶段对应的分布满足熵增和权重单调增加：

$$H\big(Q_\lambda(z)\big) < H\big(Q_{\lambda+\epsilon}(z)\big), \forall \epsilon > 0\,,$$

$$W_{\lambda+\epsilon}(z) < W_\lambda(z), \forall z, \forall \epsilon > 0\,.$$

课程学习的一个难点在于如何自动化地设计课程的分级，解决这个问题需要考虑训练模型时的反馈。Graves 等[39]提出一种自适应课程学习方法，通过预测正确率和网络复杂性的增长来自动调整课程的难度，在语言模型训练任务上显著提高了训练速度。Akkaya 等[40]提出 Automatic Domain Randomization（ADR），通过设定表现阈值，当正确率高于表现阈值时就扩大训练的分布提高课程难度，反之缩小训练分布，在复杂的魔方操作任务中取得了成功。

2.4 事后经验回放

事后经验回放(HER)[11]是一种从失败经历中学习的强化学习方法，通过修正失败经历的目标产生奖励信息，解决了奖励的稀疏性问题，同时对失败样本的利用极大提高了样本利用效率。G表示目标空间，$g \in G$是实验目标，$g' \in G$是实际实现的目标，HER 将经验数据中目标修改为g'，就产生了成功的回合数据用于策略训练。HER 的实现将在第 3 节的实验部分进行更具体的介绍。

目前对于事后经验回放算法的改进主要在于降低偏差、改进目标采样方式、适配在线策略算法等。Lanka 等[41]认为 HER 修改目标引入的新数据带来了偏差，提出通过调整真实奖励和 HER 的奖励的权重来降低偏差。Binyamin 等[42]指出，在目标物体未移动的情况下，采样的目标只与初始位置有关而与策略无关，这样的样本会给训练带来偏差，于是提出 Filtered-HER，通过滤去该类型目标来缓解该问题。Rauber 等[43]通过重要性采样将 HER 运用到策略梯度方法上，实验结果表明 HER 明显提高了策略梯度方法的样本利用效率。

2.5 好奇心驱动

好奇心驱动是一类使用内在奖励引导智能体探索环境的方法，高效的探索能够更快地获得外在奖励[15]，同时能够降低环境的奖励、状态转移的不确定性，平衡探索与利用[44]，因此好奇心驱动能解决稀疏奖励问题带来的稀疏性和不可靠性问题。目前的好奇心驱动算法根据内在奖励计算方式可以分为访问计数法和预测差法[45]。Bellemare 等[46]通过信息增益（Information Gain）来联系访问计数法和预测差法，证明了两者本质是相同的。

访问计数法使用访问次数定义状态的陌生程度，鼓励智能体探索更陌生的状态，以提高探索能力和降低对奖励估计的不确定性。Strehl 等[47]提出了一种基于模型的内在奖励方法，使用与状态动作对计数的平方根成反比的内部奖励，并理论证明了其最优性：

$$r_{in}(s,a) = \beta N(s,a)^{-\frac{1}{2}}\,,$$

其中β是常系数，$N(s,a)$是状态动作对(s,a)的计数值。为了将访问计数法推广到高维连续状态空间中，Tang 等[48]使用哈希函数将连续的状态空间离散化进行计数，该方法在多个连续动作控制问题中取得了成功。

预测差法通过学习环境的状态转移，使用预测误差作为内在奖励，能降低环境动态的不确定性。预测差法中使用状态s_t和动作a_t来预测新的状态s_{t+1}的方法被称作前向动态方法（Forward Dynamic）[49]。Stadie 等[50]提出一种根据编码后的状态$\phi(s_t)$和动作a_t来预测$\phi(s_{t+1})$的前向动态方法，使用归一化的预测误差计算内在奖励，预测误差为：

$$e(s_t, a_t, s_{t+1}) = \|\phi(s_{t+1}) - M(\phi(s_t), a_t)\|_2^2\,,$$

其中M表示预测网络，该方法能够有效解决大规模游戏环境的探索问题。Pathak 等[45]认为好奇心驱动存在电视噪声问题，于是提出 Intrinsic Curiosity Module（ICM），在前向动态模型的基础上增加

了使用$\phi(s_{t+1})$和$\phi(s_t)$来预测a_t的逆向模型，如图3所示。逆向模型的作用是提取对智能体选择动作有影响的特征[45]，能够缓解电视噪声问题。

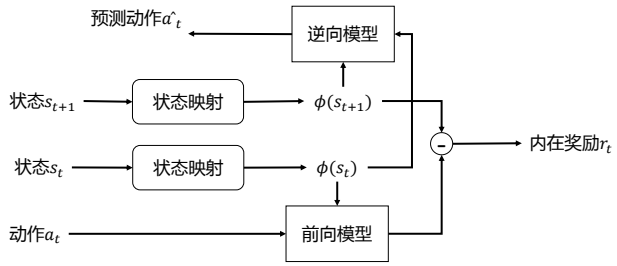

图3 ICM 原理图

Fig.3 Schematic of ICM

状态编码是高维连续状态空间下的好奇心驱动算法一个重要研究内容。Burda 等[15]对比了ICM、VAE[51]、Random Network、Pixels 四种编码方式在54个游戏中的实验结果，得出以下结论：ICM 的泛化性最好，Random Network 也足以在45%的游戏中超过 ICM，意味着很多游戏只需使用固定的随机网络就能够提取足够的特征用于策略学习，对后续的研究具有启示作用。

2.6 分层强化学习

分层强化学习（HRL）是一类使用分层策略结构的方法，分层的结构能够学习不同层次的策略，从而有效解决维度爆炸的问题[16][52] 。分层强化学习方法的上层策略往往能够处理更大时间尺度的决策，同时分层强化学习方法还能缩小各层策略的动作序列空间，进一步提高了解决稀疏奖励问题的能力。

目前用于稀疏奖励问题的分层强化学习算法主要有两类[52]，基于选项[53]（Option）的方法和基于子目标[54]（Subgoal）的方法。

基于选项的方法结构简单，上层策略在多个下层策略中进行选择，被选择的下层策略输出动作，如图4所示。Sutton 等[53]将基于选项的分层算法表述为半马尔可夫决策过程（SMDP），并推导出在动态规划、Q-Learning 中基于选项方法的公式。Bacon 等[55]将基于选项的分层方法和策略梯度法结合，提出了 Option-Critic 算法，并通过实验验证了该方法能够学习到具有实际意义的选项策略。Frans 等[56]结合元学习方法来训练基于选项的分层结构，在多个连续动作控制问题中显著提高了学习速度。

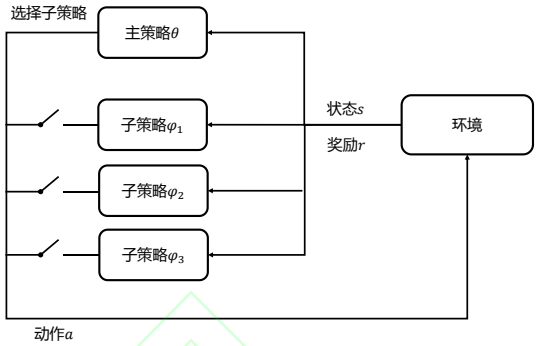

图4 基于选项的分层结构示意图

Fig.4 Schematic of option-based HRL methods

基于子目标的方法结构则是：上层策略生成子目标，下层策略输出动作去实现子目标，如图5所示。Vezhnevets 等[57]提出 FeUdal Networks，将子目标设定为隐状态空间中的方向，在蒙特祖玛复仇等多个 Atari 游戏中的表现均超过基线算法。Nachum 等[58]认为，采用离线策略训练分层结构将因为策略的改变产生偏差，可能导致训练不稳定，因此提出了使用离线数据校正的 HIRO 算法，在 Ant-Gather，Ant-Maze 等复杂的连续动作控制环境中表现均优于 FeUdal Networks。Levy 等[59]提出了 Hierarchical Actor Critic（HAC），该方法在基于子目标的分层算法基础上，结合了事后经验回放算法极大提升了学习速度且表现好于 HIRO。

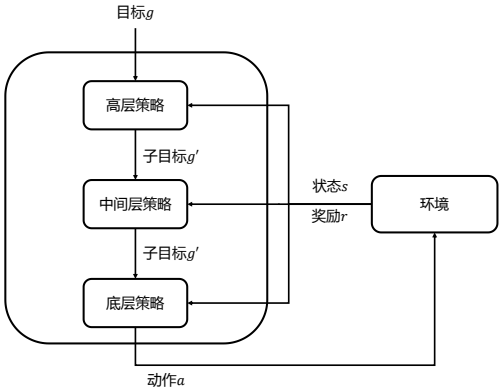

图5 基于子目标的分层结构示意图

Fig.5 Schematic of subgoal-based HRL methods

3 稀疏奖励算法实验

为了初步验证上述算法在稀疏奖励问题中作用，我们在 MuJoCo 的机器人环境 Fetch Reach[17] 下分别实现了 6 类算法的代表性算法（实验 2～7）：

（1） 奖励塑造：实验 2 实现了 Ng 等[27]提出的势能函数差分形式的奖励塑造方法，记为 Reward Shaping；

（2） 模仿学习：实验 3 实现了 Nair 等[34]提出的行为克隆算法，记为 Behavior Clone；

（3） 课程学习：实验 4 实现了符合 Bengio 等[14]定义的课程学习方法，记为 Curriculum Learning；

（4） 事后经验回放：实验 5 实现了事后经验回放算法[11]，记为 HER；

（5） 好奇心驱动：实验 6 实现了一种基于预测差的前向动态算法[50]，记为 Forward Dynamic；

（6） 分层强化学习：实验 7 实现了一种基于子目标的分层强化学习算法[54]，记为 HDDPG。

Fetch Reach 实验环境的任务是控制机械臂到达目标位置，目标范围在空间中用小红球表示，如图 6 所示，当机械夹顶端碰到小红球即为达到目标。对于机器人的每步操作，完成目标获得奖励 $+1$，否则获得奖励 -1。实验测试中执行随机动作 5×10^4步，获得正奖励的概率仅为 1.5%，因此本实验环境是典型的稀疏奖励环境。

各实验均进行 10 次随机试验，去除最大最小值绘制成功率均值标准差曲线（见 3.2），1 个 epoch 包括 100 局仿真，其余实验参数见附录。

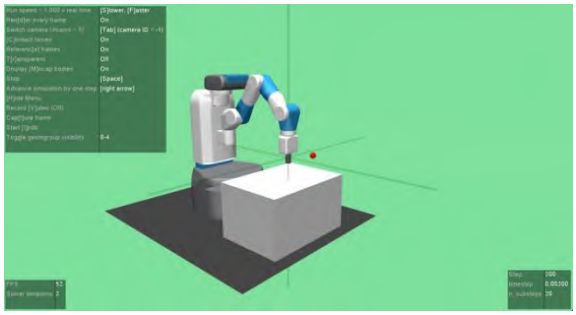

图 6 Fetch Reach 实验环境

Fig.6 Fetch Reach environment

表 1 DDPG 参数

Table 1 DDPG Parameters

名称	参数
优化器	Adam
学习率	0.001
折扣因子$\boldsymbol{\gamma}$	0.98
滑动平均比例$\boldsymbol{\tau}$	0.05
网络全连接层数	3
隐层节点数	64
激活函数	Relu
输出层激活函数	tanh
动作噪声	N(0,0.1)
最大经验池大小	10^6
单次更新次数	40
batch 大小	64

3.1 实验算法

3.1.1 实验 1 DDPG

本实验采用的基线算法是基于泛化值函数估计[60]的深度确定性策略梯度算法（以下简称 DDPG），DDPG 是代表性的基于 Actor-Critic 的连续动作控制方法，后续 6 个算法的实现均建立

在其基础上。基于泛化值函数估计的方法认为智能体的策略和值函数与目标有关，G表示目标空间，$g \in G$表示目标，a为动作，s为状态，d^{π}是策略π的采样分布，则 DDPG with UVFA 的损失函数分别为：

$$Q_{target} = r_{t+1} + \gamma Q(s_{t+1}, \pi(s_{t+1}, g), g),$$

$$L_{critic} = E_{s_t \sim d^{\pi}}\left(Q_{target} - Q(s_t, a_t, g)\right)^2,$$

$$L_{actor} = -E_{s_t \sim d^{\pi}}[Q(s_t, \pi(s_t, g), g)].$$

实验 1 比较了 DDPG 与随机策略（Random）的表现，实验结果如图 7 所示。DDPG 经过 370 个 epoch 的学习平均表现才超过随机策略，在 500 个 epoch 内最高平均成功率 21%。通过实验 1，我们可以验证常规的深度强化学习算法 DDPG 在稀疏奖励任务中学习缓慢。

3.1.2 实验 2 Reward Shaping

实验 2 根据机械夹顶端到目标位置的距离设计奖励函数。g表示目标位置，ϕ表示状态到位置的映射，$\phi(s_t)$表示状态s_t的位置，我们设计势能函数为距离目标位置的距离：

$$\Phi(s_t, g) = -\|g - \phi(s_t)\|_2^2.$$

根据 Ng 等[27]的理论，我们设计奖励函数r为势能函数的差分形式：

$$\begin{aligned} r(s_t, s_{t+1}, g) &= \lambda(\Phi(s_{t+1}, g) - \Phi(s_t, g)) \\ &= \lambda(\|g - \phi(s_t)\|_2^2 - \|g - \phi(s_{t+1})\|_2^2), \end{aligned}$$

其中λ是大于 0 的奖励系数。在本实验中，由于单步位移较小（10^{-3}量级），因此设置了$\lambda = 10^3$。

实验 2 比较了奖励塑造和 DDPG 的表现，实验结果如图 8 所示。奖励塑造明显加快了学习速度，40 个 epoch 就收敛到 100%成功率，验证了按照势能函数差分形式设计奖励函数的有效性。

3.1.3 实验 3 Behavior Clone

实验 3 实现的是 Nair 等[34]使用的行为克隆（Behavior Clone）算法，利用 HER 训练的成功率 100%的智能体随机产生 100 局的交互数据作为示例数据，同时最大化累计奖励和最小化与示例策略的误差。由于示例数据有限，不一定能获得示例策略在当前状态下的决策，为此我们使用从示例数据中的采样(s_e, g_e, a_e)来计算策略误差：

$$\begin{aligned} L_{actor} = &-E_{s_t \sim d^{\pi}}[Q(s_t, \pi(s_t, g), g)] + \\ &\alpha E_{s_e \sim d^{\pi_e}}\|a_e - \pi(s_e, g_e)\|_2^2, \end{aligned}$$

其中π_e、π分别表示示例策略和实际策略，s_t、g分别表示实际状态和目标，α是常系数，d^{π}、d^{π_e}分别代表智能体采样数据分布和示例数据分布。

实验 3 比较了不同α下行为克隆的表现，其中$\alpha = 0$即代表 DDPG，实验结果如图 9 所示。结果显示，随着示例策略损失的系数α增加，学习速度加快，当$\alpha = 1$时，8 个 epoch 就收敛到 100%成功率。实验结果验证了引入示例策略损失的行为克隆方法能够显著提高学习速度，$\alpha = 1$时 Behavior Clone 在本实验实现的算法中最快收敛到 100%成功率。

3.1.4 实验 4 Curriculum Learning

实验 4 通过设计目标分布范围逐渐增大的任务序列来实现课程学习，能够证明我们的设计方式符合 Bengio 等[14]对课程学习的定义。实验任务中目标位置在各维度变化范围是$(-0.15, 0.15)$，记做$range = 0.15$，我们设计的课程目标分布序列为：

$$range = 0.05 + i \times \frac{0.1}{c-1}, i \in [0, c-1],$$

其中c为课程数。我们的课程序列保证了最后阶段$i = c - 1$时目标分布范围与其余实验相同。

实验 4 比较了不同课程数c对课程学习的影响，实验结果如图 10 所示。当课程数$c = 1$时，课程难度较大，难以进行策略学习；当课程数$c = 2$时，在 50 个 epoch 处存在任务切换导致的断层；当课程数继续增加，课程跨度变小，策略迁移更快，学习速度进一步提高，但当课程数大于 10 后，提高课程数几乎不带来性能的提升。实验中$c = 10$时学习速度最快，70 个 epoch 收敛到 100%成功率。

3.1.5 实验 5 HER

实验 5 实现的 HER[11]算法具体描述为：在每局仿真结束后，对该局的数据$(s_t, a_t, s_{t+1}, r_t, g)$采样多个状态$s \in S$，利用状态到目标的映射$\phi$得到该局完成的目标$g' = \phi(s)$。对每条数据$(s_t, a_t, s_{t+1}, r_t, g)$用$g'$计算新的奖励值$r_t' = -(\|\phi(s_{t+1}) - g'\|_2^2 > threshold)$，完成目标奖励为 0，否则奖励为$-1$。最后将生成的新数据

$(s_t, a_t, s_{t+1}, r_t', g')$和原数据一起存入经验池，使用 DDPG 进行策略更新。

本实验中，状态s是包括位置、速度信息的 15 维向量，前三维是机械夹顶端的位置，因此状态到位置的映射$\phi = s[0:3]$。本实验采用的每局采样目标数为 4。

实验 5 比较了事后经验回放与 DDPG 的表现，实验结果如图 11 所示。HER 在 15 个 epoch 左右就收敛到 100%成功率，且稳定性高，在无外部引导信息算法中表现最好。

3.1.6 实验 6 Forward Dynamic

实验 6 实现的好奇心驱动算法是基于预测差的 Forward Dynamic[50]，使用前向模型M学习环境动态，同时使用标准化的预测误差$loss_{norm}$作为内在奖励r_{in}：

$$loss_{forward} = E_{s_t \sim d^{\pi}} \| s_{t+1} - M(s_t, a_t; \theta) \|_2^2 ,$$

$$loss_{forward} \xrightarrow{update} \mu, \sigma ,$$

$$loss_{norm} = \frac{loss_{forward} - \mu}{\sigma},$$

$$r_{in} = loss_{norm} ,$$

其中μ, σ分别表示预测误差的均值和标准差。实验中使用 3 层 64 节点的神经网络来作为前向模型M。

实验 6 比较了 Forward Dynamic 与 DDPG 的表现，实验结果如图 12 所示。Forward Dynamic 在 80 个 epoch 左右收敛到 100%成功率，验证了 Forward Dynamic 方法对稀疏奖励问题的有效性。

3.1.7 实验 7 HDDPG

实验 7 实现的分层强化学习算法 HDDPG（Hierarchical DDPG）结构上和 HAC[59]相近，区别在于 HDDPG 不使用事后经验回放。HDDPG 属于基于子目标的分层强化学习方法，假设层数为H，最高层策略输入状态s_t、系统目标g，输出子目标：

$$subgoal_H = \pi_H(s_t, g) .$$

中间层策略输入状态、子目标，输出子目标：

$$subgoal_{h-1} = \pi_h(s_t, subgoal_h), h \in (1, H).$$

最下层策略输入状态、子目标，输出动作：

$$a_t = \pi_1(s_t, subgoal_1) .$$

各层每一步如果完成目标获得奖励$+1$，否则获得奖励-1，如果超过最大步数未完成目标还将获得负奖励作为惩罚。

实验 7 比较了不同层数 HDDPG 的表现，其中层数为 1 即为 DDPG，实验结果如图 13 所示。本实验验证了 HDDPG 分层的结构提高了解决稀疏奖励问题的能力，同时能够得到层数对 HDDPG 的影响，当层数小于等于 3 的时候，HDDPG 表现随层数增加而提高，层数为 3 时最高平均成功率达到 66%。此外，我们也进行了层数大于 3 的实验，结果接近 DDPG，我们认为这是因为分层的离线策略算法存在的偏差[58]随着层数增加累积，当层数过大时会导致训练不稳定，甚至难以进行策略学习。

3.2 实验结果

3.2.1 实验 1 DDPG

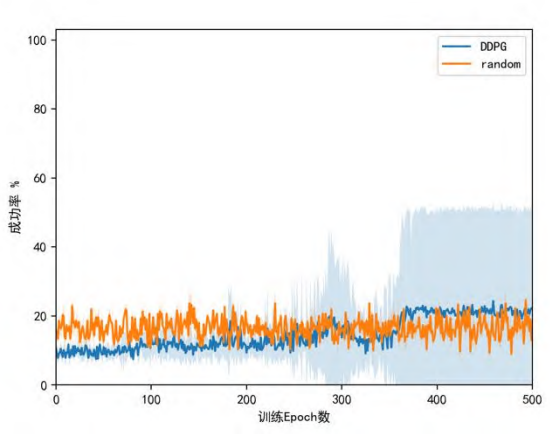

图 7 DDPG 与随机动作成功率学习曲线

Fig.7 Performance curve of DDPG and Random

3.2.2 实验 2 Reward Shaping

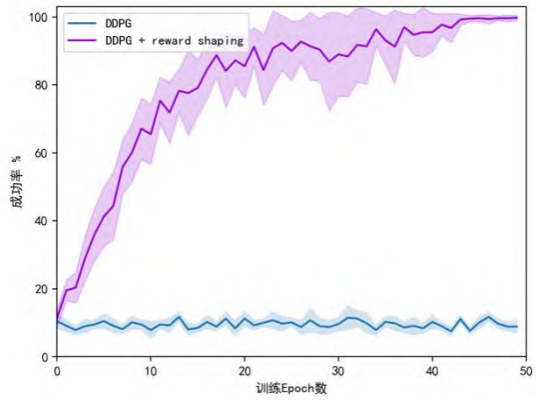

图 8 奖励塑造与 DDPG 成功率学习曲线

Fig.8 Performance curve of Reward Shaping and DDPG

3.2.3 实验 3 Behavior Clone

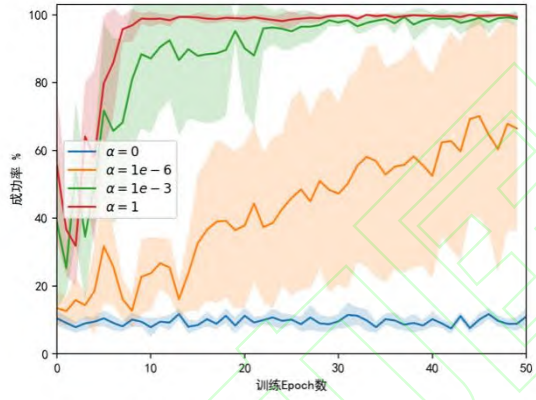

图 9 Behavior Clone 成功率学习曲线

Fig.9 Performance curve of Behavior Clone

3.2.4 实验 4 Curriculum Learning

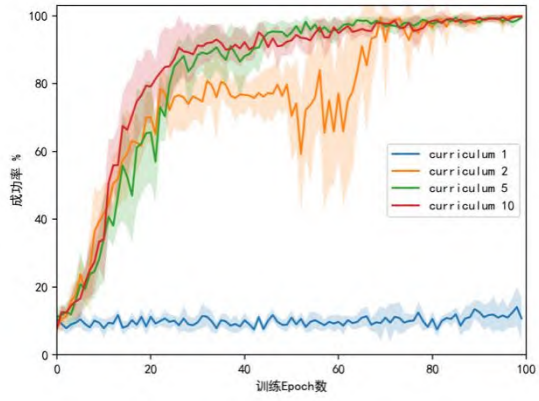

图 10 课程学习成功率学习曲线

Fig.10 Performance curve of Curriculum Learning

3.2.5 实验 5 HER

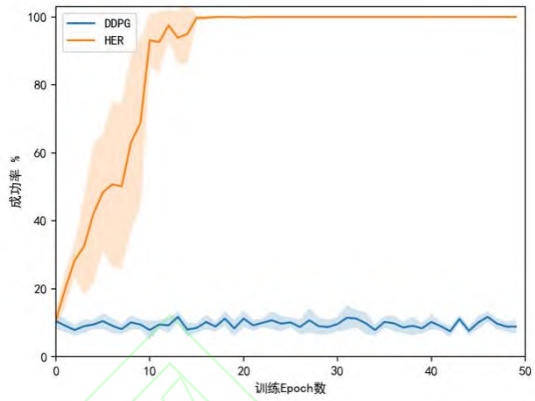

图 11 HER 与 DDPG 学习曲线

Fig.11 Performance curve of HER and DDPG

3.2.6 实验 6 Forward Dynamic

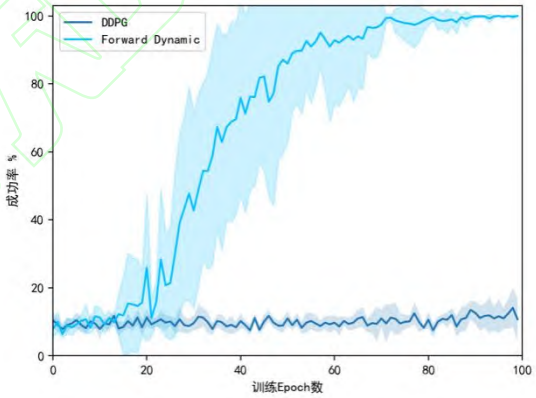

图 12 Forward Dynamic 与 DDPG 成功率学习曲线

Fig.12 Performance curve of Forward Dynamic and DDPG

3.2.7 实验 7 HDDPG

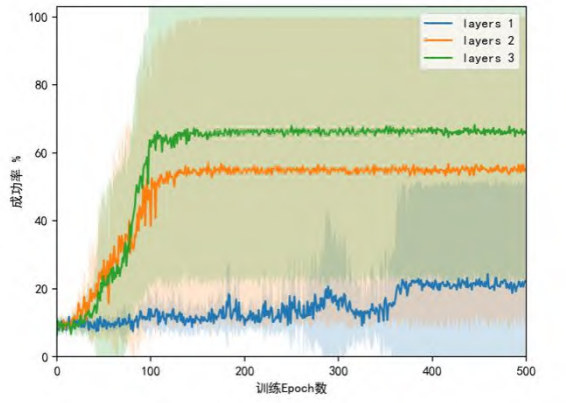

图 13 不同层数 HDDPG 成功率学习曲线

Fig.13 Performance curve of different layers of HDDPG

3.3 实验总结

表 2 实验结果

Table 2 Experiment results

算法	有无外部引导信息	最高平均成功率（%）	收敛 epoch 数
Random	～	17	～
DDPG	～	21	370
Reward Shaping	有	100	43
Behavior Clone(α=1)	有	100	8
Curriculum 10	有	100	70
HER	无	100	15
Forward Dynamic	无	100	80
HDDPG layers3	无	66	100

我们通过表 2 总结了实验实现的 6 类算法的代表性算法在稀疏奖励问题中的表现，相比于 DDPG 均有显著提升。其中，使用示例策略辅助学习的 Behavior Clone 表现最佳，8 个 epoch 就收敛到 100%成功率，HER 没有使用外部引导信息但是表现接近 Behavior Clone，15 个 epoch 收敛到 100%成功率。

从表 2 我们可以看到使用外部引导信息的算法平均表现好于无外部引导信息的算法。实际上我们使用的 Fetch Reach 实验环境是相对简单的，容易构建可靠的外部引导信息，在更多的稀疏奖励任务中，外部引导信息是难于构建的，这就需要无外部引导信息算法的研究。在我们的实验中，HER、Forward Dynamic 等算法的表现已经接近使用外部引导信息的算法，具有重要的研究价值和意义。

4 展望

稀疏奖励算法未来的研究方向可以是根据各算法存在的问题进行改进，或对多种稀疏奖励算法进行结合。自适应课程学习是重要的研究方向，结合生成对抗思想[61]或元学习[62]的自动课程学习具有广阔的研究前景。好奇心驱动存在被称作“电视噪声”的问题[45]：当环境中出现难以预测的随机噪声时，会导致智能体驻留而不去完成任务。为解决电视噪声问题，未来可以研究好奇心驱动的状态编码方式[15]或结合注意力机制[63]等方法，让智能体更关注与任务有关的状态和奖励信息。事后经验回放算法基于“存在状态到目标的映射”的假设，但是大量实际问题中并不容易找到这样的映射关系，比如视频游戏、机器人导航[64]等问题，研究通用的目标映射方法有助于将 HER 这种高效的稀疏奖励算法应用到更多问题中。分层强化学习还需要进一步研究分层的数据矫正[58]以及更高效的分层结构。基于目标的分层强化学习中，自动学习目标空间的表示[65]也是重要的研究内容。最后，多种稀疏奖励算法的结合算法能够融合不同算法的优势，为稀疏奖励问题提供更加强有力的解决方案，例如 Levy 等[59]提出的分层与 HER 结合的 HAC、Lanier 等[66]提出的好奇心驱动与 HER 结合的算法、Nair 等[34]提出的模仿学习与 HER 结合的算法。

5 结束语

本文将目前主流的稀疏奖励算法按是否引入外部引导信息分为两类，分别介绍了奖励塑造、模仿学习、课程学习和事后经验回放、好奇心驱

动、分层强化学习等 6 类算法的发展和应用。本文还通过实验验证了 6 种算法在稀疏奖励问题中的有效性，相比于 DDPG 均有显著提升，部分无外部引导信息的算法表现已经接近使用外部引导信息的算法。本文的实验比较了种算法解决稀疏奖励问题的能力，为进一步的研究提供了实验基础。

稀疏奖励问题的研究对强化学习理论的拓展以及算法的实际落地具有重要意义，我们希望在未来看到更高效的算法用于解决复杂的稀疏奖励问题，同时也希望看到更多强化学习算法的实际落地与应用，为社会创造更多价值。

参考文献:

[1] SUTTON R S, BARTO A G. Reinforcement Learning: An Introduction[M]. MIT Press, US, 1998.

[2] SUTTON R S, BARTO A G. Reinforcement learning: An introduction(2nd Edition) [M]. MIT Press,US, 2018.

[3] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484.

[4] VINYALS O, BABUSCHKIN I, CZARNECKI WM, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575:350–354.

[5] BERNER C, BROCKMAN G, CHAN B, et al. Dota 2 with Large Scale Deep Reinforcement Learning[J]. ArXiv Preprint ArXiv:1912.06680, 2019.

[6] SILVER D. Tutorial: Deep reinforcement learning[C]//Proc. of the 33rd Int. Conf. on Machine Learning (ICML 2016). 2016.

[7] LI Y. Deep reinforcement learning: An overview[J]. ArXiv Preprint ArXiv:1701.07274, 2017.

[8] LI Y. Deep reinforcement learning[J]. ArXiv Preprint ArXiv:1810.06339, 2018.

[9] Riedmiller M, Hafner R, Lampe T, et al. Learning by playing-solving sparse reward tasks from scratch[J]. ArXiv Preprint ArXiv:1802.10567, 2018.

[10] HOSU I A, REBEDEA T. Playing atari games with deep reinforcement learning and human checkpoint replay[J]. EGPAI 2016. ArXiv Preprint ArXiv:1607.05077, 2016.

[11] ANDRYCHOWICZ M, WOLSKI F, RAY A, et al. Hindsight experience replay[C]//Advances in Neural Information Processing Systems. 2017: 5048-5058.

[12] GULLAPALLI V, BARTO A G. Shaping as a method for accelerating reinforcement learning[C]//Proceedings of the 1992 IEEE International Symposium on Intelligent Control. IEEE, 1992: 554-559.

[13] HUSSEIN A, GABER M M, ELYAN E, et al. Imitation learning: A survey of learning methods[J]. ACM Computing Surveys (CSUR), 2017, 50(2): 1-35.

[14] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]//Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 41-48.

[15] BURDA Y, EDWARDS H, PATHAK D, et al. Large-scale study of curiosity-driven learning[J]. ArXiv Preprint ArXiv:1808.04355, 2018.

[16] 周文吉，俞扬. 分层强化学习综述[J]. 智能系统学报，2017, 12 (5) : 590-594.

ZHOU WENJI，YU YANG. Summarize of hierarchical reinforcement learning[J]. CAAI Transactions on Intelligent Systems，2017，12(5): 590-594.

[17] Plappert M, Andrychowicz M, Ray A, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research[J]. ArXiv Preprint ArXiv:1802.09464, 2018.

[18] 杨惟轶，白辰甲，蔡超，等. 深度强化学习中稀疏奖励问题研究综述[J/OL]. 计算机科学:1-13[2020-03-10]. http://kns.cnki.net/kcms/detail/50.1075.TP.20191122.1628.023.html.

YANG WEIYI, BAI CHENJIA, CAI CHAO, et al. Survey on Sparse Reward in Deep Reinforcement Learning[J/OL]. Computer Science:1-13[2020-03-10].

[19] 万里鹏，兰旭光，张翰博，郑南宁. 深度强化学习理论及其应用综述[J].模式识别与人工智能, 2019, 32(01):67-81.

WAN LIPENG，LAN XUGUANG，ZHANG HANBO, et al. A Review of Deep Reinforcement Learning Theory and Application[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(01):67-81.

[20] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. ArXiv Preprint ArXiv:1312.5602, 2013.

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

[22] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning, 1992, 8(3-4): 229-256.

[23] KONDA V R, TSITSIKLIS J N. Actor-critic algorithms[C]//Advances in Neural Information Processing Systems. 2000: 1008-1014.

[24] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//International Conference on Machine Learning. 2016: 1928-1937.

[25] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. ArXiv Preprint ArXiv:1707.06347, 2017.

[26] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. ArXiv Preprint ArXiv:1509.02971, 2015.

[27] NG A Y, HARADA D, RUSSELL S. Policy invariance under reward transformations: Theory and application to reward shaping[C]//ICML. 1999, 99: 278-287.

[28] RANDLØV J, ALSTRØM P. Learning to Drive a Bicycle Using Reinforcement Learning and Shaping[C]//ICML. 1998, 98: 463-471.

[29] JAGODNIK K M, THOMAS P S, VAN DEN BOGERT A J, et al. Training an actor-critic reinforcement learning controller for arm movement using human-generated rewards[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017, 25(10): 1892-1905.

[30] FERREIRA E, LEFÈVRE F. Expert-based reward shaping and exploration scheme for boosting policy learning of dialogue management[C]//2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, 2013: 108-113.

[31] NG A Y, RUSSELL S J. Algorithms for inverse reinforcement learning[C]//Icml. 2000, 1: 663-670.

[32] MARTHI B. Automatic shaping and decomposition of reward functions[C]//Proceedings of the 24th International Conference on Machine learning. 2007: 601-608.

[33] ROSS S, BAGNELL D. Efficient reductions for imitation learning[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010: 661-668.

[34] NAIR A, MCGREW B, ANDRYCHOWICZ M, et al. Overcoming exploration in reinforcement learning with demonstrations[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 6292-6299.

[35] HO J, ERMON S. Generative adversarial imitation learning[C]//Advances in Neural Information Processing Systems. 2016: 4565-4573.

[36] LIU Y X, GUPTA A, ABBEEL P, et al. Imitation from observation: Learning to imitate behaviors from raw video via context translation[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 1118-1125.

[37] TORABI F, WARNELL G, STONE P. Behavioral cloning from observation[J]. ArXiv Preprint ArXiv:1805.01954, 2018.

[38] ELMAN J L. Learning and development in neural networks: The importance of starting small[J]. Cognition, 1993, 48(1): 71-99.

[39] GRAVES A, BELLEMARE M G, MENICK J, et al. Automated curriculum learning for neural networks[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1311-1320.

[40] AKKAYA I, ANDRYCHOWICZ M, CHOCIEJ M, et al. Solving Rubik's Cube with a Robot Hand[J]. ArXiv Preprint ArXiv:1910.07113, 2019.

[41] LANKA S, WU T. ARCHER: Aggressive Rewards to Counter bias in Hindsight Experience Replay[J]. ArXiv Preprint ArXiv:1809.02070, 2018.

[42] MANELA B, BIESS A. Bias-Reduced Hindsight Experience Replay with Virtual Goal Prioritization[J]. ArXiv Preprint ArXiv:1905.05498, 2019.

[43] RAUBER P, UMMADISINGU A, MUTZ F, et al. Hindsight policy gradients[J]. ArXiv Preprint ArXiv:1711.06006, 2017.

[44] SCHMIDHUBER J. A possibility for implementing curiosity and boredom in model-building neural controllers[C]//Proc. of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats. 1991: 222-227.

[45] PATHAK D, AGRAWAL P, EFROS A A, et al. Curiosity-driven exploration by self-supervised prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 16-17.

[46] BELLEMARE M, SRINIVASAN S, OSTROVSKI G, et al. Unifying count-based exploration and intrinsic motivation[C]//Advances in Neural Information Processing Systems. 2016: 1471-1479.

[47] STREHL A L, LITTMAN M L. An analysis of model-based interval estimation for Markov decision processes[J]. Journal of Computer and System Sciences, 2008, 74(8): 1309-1331.

[48] TANG H, HOUTHOOFT R, FOOTE D, et al. # exploration: A study of count-based exploration for deep reinforcement learning[C]//Advances in Neural Information Processing Systems. 2017: 2753-2762.

[49] BURDA Y, EDWARDS H, STORKEY A, et al. Exploration by random network distillation[J]. ArXiv Preprint ArXiv:1810.12894, 2018.

[50] STADIE B C, LEVINE S, ABBEEL P. Incentivizing exploration in reinforcement learning with deep predictive models[J]. ArXiv Preprint ArXiv:1507.00814, 2015.

[51] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. ArXiv Preprint ArXiv:1312.6114, 2013.

[52] RAFATI J, NOELLE D C. Learning representations in model-free hierarchical reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 10009-10010.

[53] SUTTON R S, PRECUP D, SINGH S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning[J]. Artificial Intelligence, 1999, 112(1-2): 181-211.

[54] KULKARNI T D, NARASIMHAN K, SAEEDI A, et al. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation[C]//Advances in Neural Information Processing Systems. 2016: 3675-3683.

[55] BACON P L, HARB J, PRECUP D. The option-critic architecture[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[56] FRANS K, HO J, CHEN X, et al. Meta learning shared hierarchies[J]. ArXiv Preprint ArXiv:1710.09767, 2017.

[57] VEZHNEVETS A S, OSINDERO S, SCHAUL T, et al. Feudal networks for hierarchical reinforcement learning[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 3540-3549.

[58] NACHUM O, GU S S, LEE H, et al. Data-efficient hierarchical reinforcement learning[C]//Advances in Neural Information Processing Systems. 2018: 3303-3313.

[59] LEVY A, KONIDARIS G, PLATT R, et al. Learning multi-level hierarchies with hindsight[J]. ArXiv Preprint ArXiv:1712.00948, 2017.

[60] SCHAUL T, HORGAN D, GREGOR K, et al. Universal value function approximators[C]//International Conference on Machine Learning. 2015: 1312-1320.

[61] SUKHBAATAR S, LIN Z, KOSTRIKOV I, et al. Intrinsic motivation and automatic curricula via asymmetric self-play[J].ArXiv Preprint ArXiv:1703.05407, 2017.

[62] JABRI A, HSU K, GUPTA A, et al. Unsupervised Curricula for Visual Meta-Reinforcement Learning[C]//Advances in Neural Information Processing Systems. 2019: 10519-10530.

[63] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[64] SAHNI H, BUCKLEY T, ABBEEL P, et al. Visual hindsight experience replay[J]. ArXiv Preprint ArXiv:1901.11529, 2019.

[65] SUKHBAATAR S, DENTON E, SZLAM A, et al. Learning Goal Embeddings via Self-Play for Hierarchical Reinforcement Learning[J]. ArXiv Preprint ArXiv:1811.09083, 2018.

[66] LANIER J B, MCALEER S, BALDI P. Curiosity-Driven Multi-Criteria Hindsight Experience Replay[J]. ArXiv Preprint ArXiv:1906.03710, 2019.

作者简介：

杨瑞，硕士研究生，主要研究方向为机器学习与强化学习。

严江鹏，博士研究生，主要研究方向为人工智能与计算机视觉。

李秀，教授，博士生导师，主要研究方向为智能系统、数据挖掘与模式识别。作为项目负责人，完成国家自然科学基金项目 3 项，深圳市基础研究项目 2 项，深圳市技术开发项目 1 项；作为子课题负责人，完成国家 863 项目 4 项；目前在研 863 重大项目 1 项，国家自然科学基金 1 项。获得国家发明专利授权 7 项，国家软件著作权 5 项。在国内外重要学术期刊或会议上发表学术论文 100 余篇。