优势加权互信息最大化的最大熵分层强化学习

乌 兰[1] 刘 全[1),2)] 黄志刚[1] 朱 斐[1),2)] 张立华[1]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215006)

[2)](苏州大学江苏省计算机信息处理技术重点实验室 江苏 苏州 215006)

摘 要 近年来,深度强化学习在控制任务中取得了显著的效果.但受限于探索能力,难以快速且稳定地求解复杂任务.分层强化学习作为深度强化学习的重要分支,主要解决大规模问题.但是仍存在先验知识设定的不合理和无法有效平衡探索与利用等难题.针对以上问题,提出优势加权互信息最大化的最大熵分层强化学习(Maximum Entropy Hierarchical Reinforcement Learning with Advantage-weighted Mutual Information Maximization,HRL-AMIM)算法.该算法通过优势函数加权重要性采样与互信息最大化,解决由策略引起的样本聚类问题,增加内部奖励来强调Option的多样性.同时,将奖励引入最大熵强化学习目标,使策略具有了更强的探索性和更好的稳定性.此外,采用Option数量退火方法,不仅减少了先验知识对性能的影响,还平衡了算法的探索与利用,并获得了更高的样本效率和更快的学习速度.将HRL-AMIM算法应用于Mujoco任务中,实验表明,与传统深度强化学习算法和同类型的分层强化学习算法相比,HRL-AMIM算法在性能和稳定性方面均具有较大的优势.进一步通过消融实验和超参数敏感性实验,验证了算法的鲁棒性和有效性.

关键词 深度强化学习;分层强化学习;优势加权;互信息;最大熵

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2023.02066

Maximum Entropy Hierarchical Reinforcement Learning with Advantage-weighted Mutual Information Maximization

WU Lan[1] LIU Quan[1),2)] HUANG Zhi-Gang[1] ZHU Fei[1),2)] ZHANG Li-Hua[1]

[1)](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)

[2)](*Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, Jiangsu* 215006)

Abstract Reinforcement learning is a significant research area in machine learning. By interacting with the environment, agents can adapt to the dynamic environment. At the same time, this interactive learning approach allows an agent to progressively optimize its policy, which is promising for a wide range of applications. Deep reinforcement learning, a method that combines reinforcement learning with deep learning, plays a crucial role in artificial intelligence. This combination enables agents to learn and make autonomous decisions in complex and dynamic environments without complex supervised data. In recent years, deep reinforcement learning has achieved remarkable results in games and complex control tasks. For example, Deep Q Learning (DQN) algorithm uses a convolutional neural network to process the visual input from the game screen and continuously updates the policy through a Q-learning algorithm. In Atari 2600 games,

收稿日期:2022-10-27;在线发布日期:2023-07-10.本课题得到国家自然科学基金(62376179,61772355,61702055,61876217,62176175),新疆维吾尔自治区自然科学基金(2022D01A238)、江苏高校优势学科建设工程资助项目(PAPD)资助.**乌 兰**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为分层强化学习、离线强化学习. E-mail:20217927001@stu.suda.edu.cn. **刘 全**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为深度强化学习、自动推理. E-mail:quanliu@suda.edu.cn. **黄志刚**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为深度强化学习、分层强化学习. **朱 斐**,博士,副教授,中国计算机学会(CCF)高级会员,主要研究领域为强化学习、文本挖掘. **张立华**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为深度强化学习、逆向强化学习.

the DQN can learn advanced game strategies autonomously by looking at the game screen pixel information, even without human expert guidance. However, DQN is only applicable to discrete action space tasks. To solve this problem, Deep Deterministic Policy Gradient (DDPG) combines deterministic policy gradient algorithms with DQN algorithms to achieve policy optimization and learning in continuous action spaces. Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm uses a clipped double Q network to prevent the value function from being overestimated. Moreover, it introduces delayed policy updates and targeted policy smoothing to improve policy learning stability and exploratory power. Soft Actor-Critic (SAC) algorithm achieves efficient learning over a continuous action space by simultaneously learning a policy network and a value function network, combined with entropy regularization. The algorithm provides a useful learning framework for solving large-scale problems. However, deep reinforcement learning is difficult to solve complex tasks quickly and stably due to the limited exploration capability. Hierarchical reinforcement learning is an essential branch of deep reinforcement learning that focuses on solving large-scale problems. It is an effective solution to the problem of performance degradation of deep reinforcement learning when dealing with large-scale problems through time abstraction. However, there are still challenges such as the unreasonable setting of a priori knowledge and the inability to balance exploration and exploitation effectively. To address the above problems, Maximum Entropy Hierarchical Reinforcement Learning with Advantage-weighted Mutual Information Maximization (HRL-AMIM) algorithm is proposed. The method solves the sample clustering problem induced by the policy by weighted importance sampling of the advantage function and maximizing the average mutual information, adding internal rewards to emphasize the diversity of Options. Meanwhile, rewards are introduced into the maximum entropy reinforcement learning goal, which makes the policy more exploratory and better stable. In addition, the Option number annealing method not only reduces the impact of prior knowledge on performance but also balances the exploration and exploitation of the algorithm, achieving higher sample efficiency and faster learning speed. The HRL-AMIM algorithm is applied to the Mujoco task, and the experiments show that the algorithm is superior to the traditional deep reinforcement learning algorithms and similar hierarchical reinforcement learning algorithms in terms of performance and stability. Furthermore, the robustness and effectiveness of the algorithm are verified by ablation experiments and hyperparameter sensitivity experiments.

Keywords　deep reinforcement learning; hierarchical reinforcement learning; advantage-weighted; mutual information; maximum entropy

1　引　言

深度强化学习[1](Deep Reinforcement Learning，DRL)将深度学习[2](Deep Learning，DL)与强化学习[3](Reinforcement Learning，RL)相结合，兼备DL的信息感知能力和RL的决策控制能力，形成一种端到端的完整智能系统. Mnih等人[4]将Q-learning算法与深度学习算法结合，引入经验回放机制和目标网络，提出深度Q网络(Deep Q Network，DQN)，该算法在Atari 2600视频游戏上表现出超越人类专家的成果. 针对DQN算法只能用于离散动作空间任务的问题，Lillicrap等人[5]结合确定性策略梯度算法和DQN算法，提出适用于大规模连续动作空间的深度确定性策略梯度算法. 为了防止动作值函数被过高地估计，Fujimoto等人[6]提出了孪生延迟深度确定性策略梯度算法(Twin Delayed Deep Deterministic Policy Gradient，TD3)，该算法采用截断双Q学习、延迟策略更新和目标策略平滑3个关键技术. Haarnoja等人[7]提出了软性行动者-评论家

(Soft Actor-Critic,SAC)算法,将最大熵引入Actor-Critic(AC)算法[8],一方面鼓励智能体不断地探索,另一方面降低算法对模型与估计误差的敏感度.然而,当DRL任务的状态-动作空间非常庞大或在稀疏奖励[9]的环境下,常用的探索方案不仅对超参数的设置有较高要求[10],表现出脆弱的收敛性[11],而且无法引导智能体探索更广阔的动作空间.

分层强化学习[12](Hierarchical Reinforcement Learning,HRL)作为DRL的重要研究分支,是一种以半马尔可夫决策过程[13](Semi-Markov Decision Process,SMDP)为理论基础的降维学习.运用时序抽象技术将强化学习分解为不同层次的抽象间和抽象内部,有效地提高了智能体的探索能力和迁移能力.HRL可以将最终目标分解为多个子任务来学习层次化的策略,再通过组合多个子任务的策略形成有效的全局策略.从而解决DRL中难以解决的大规模问题,并提升样本的利用效率.为了有效地解决智能体在大规模任务中探索能力不足的问题,Zhou等人[14]提出基于Option的分层强化学习(Option也常被称为技能Skill[15]).在该学习中,Option选择策略和Option内部策略的训练过程是分离的.下层网络为Option内部策略,学习不同行为的控制方式,负责在接下来的N步内选择动作;上层策略为Option选择策略,学习如何选择策略,负责调用Option[16].Levy等人[17]提出增广分层策略(Augmented Hierarchical Policy,AHP)算法,利用基函数计算梯度,并构建一个同时学习Option间和Option内部的框架.在AHP的基础上,Bacon等人[18]提出Option-Critic(OC)算法,在AC框架上引入Option内部策略和中断函数来构建模型,自此OC成为HRL的经典算法之一.Zhang等人[19]在OC框架上提出双行动者-评论家算法,通过将一个SMDP划分为两个并行的扩展马尔可夫决策过程(Markov Decision Process,MDP),其中扩展的MDP采用AC算法,形成了双AC架构.这使得所有的策略优化算法都可以直接用于Option学习和策略学习.Kumar等人[20]提出一种新的基于互信息(Mutual Information,MI)的赋能学习算法,从信息论的角度评价智能体在某状态下采取某动作对环境影响的程度.多样性分层强化学习[21](Diversity Is All You Need,DIAYN)将赋能学习应用于HRL,充分利用信息论的观点来发现状态、动作和技能的分布关系.在DIAYN基础上,Baumli等人[22]使用两个逆向预测,分别从轨迹的初始状态和终止状态预测技能,从而学习技能差异性.Lin等人[23]运用上层网络控制Option策略,下层网络学习解决子任务.同时,对于每个子任务,采用表征学习与模仿学习来进行感知与探索.

然而,基于Option的HRL仍存在不足之处.通常只能在离散状态-动作空间上进行学习,导致算法稳定性较差.本文所强调的先验知识在于,传统HRL使用人为设定的固定数量Option.在经典OC算法中将数量取固定值,当取值过小或过大时,不仅会导致失去时序抽象意义,而且限制算法性能,同时也无法平衡探索与利用.探索和利用是HRL中的经典问题,目前可通过ε-贪婪策略[24]、添加噪音[25]、好奇心驱动[26]和变分信息最大化[27]等方法提高算法的探索性能.但是能否有效平衡探索与利用仍是HRL下的棘手难题.

针对以上问题,本文提出了一种基于优势加权互信息最大化的最大熵分层强化学习(Maximum Entropy Hierarchical Reinforcement Learning with Advantage-weighted Mutual Information Maximization,HRL-AMIM)算法.该算法运用加权重要性采样对优势函数模态进行转化,并利用互信息最大化来学习分层策略的潜在变量,结合这两种方法构建Option先验网络.特别地,通过内部奖励增加Option的探索性能.同时,采用Option数量退火方法,使数量动态地变化,减少先验知识对算法性能的影响,并平衡了探索与利用问题.此外,将该思想应用于最具竞争力的异策略连续控制SAC算法,提出基于Option框架的最大熵分层强化学习,提高智能体探索性能并稳定地更新网络.同时,实验结果表明,HRL-AMIM算法在不同的随机种子下具有稳定的性能.

本文主要贡献包括以下3个方面:

(1)本文提出一种新的基于Option框架的分层强化学习.在最大熵RL目标下,将内部奖励增广到环境奖励来鼓励Option策略的多样性;

(2)采用退火方法使Option数量随着迭代次数进行动态变化,减少先验知识对算法性能的影响,并解决了经典OC算法中无法平衡探索和利用的问题;

(3)在Mujoco实验环境中,将HRL-AMIM算法分别与DRL和HRL算法进行对比.同时,进行消融实验,并展开全面的超参数敏感性分析实验.此外,可视化了Option分布,从而验证了该方法的优越性.

2　背景知识

2.1　马尔可夫决策过程

用于描述强化学习的马尔可夫决策过程是一个五元组$M=(S,A,P,R,\gamma)$，其中S为有限的状态集，A为有限的动作集．$P=S\times A\times S\to[0,1]$为状态转移模型，$p(s'|s,a)$表示在状态$s$下执行$a$到达下一个状态$s'$的概率．$R=S\times A\to\mathbb{R}$为即时奖励函数，$\gamma\in[0,1]$为折扣因子．策略$\pi$表示状态到动作的映射函数$\pi:S\to A$，可分为确定性策略$a=\pi(s)$和随机性策略$\pi(a|s)$. 智能体从初始状态分布中产生一个初始状态$S_0$，根据策略$\pi$执行动作$A_0$，与环境进行交互获得奖赏$R_1$，并且转移到下一个状态$S_1$，不断重复得到一个状态-动作序列$\rho_\pi=S_0,A_0,R_1,S_1,A_1,R_2,\cdots$. 带折扣回报定义为从$t$时刻开始的累计奖励$G_t=R_{t+1}+\gamma R_{t+2}+\gamma^2R_{t+3},\cdots$. 强化学习的目标是最大化期望回报，定义状态值函数为在状态s处遵循策略π所得到的期望回报$V(s)=_\pi(G_t|S_t=s)$，动作值函数为在状态s处执行动作a，遵循策略π所得到的期望回报$Q(s,a)=_\pi(G_t|S_t=s,A_t=a)$. 为了提高策略的学习率并减小方差，将策略的优势函数[28]定义为$A_\pi(s,a)=Q(s,a)-V(s)$.

2.2　互信息

互信息最大化可用来学习可解释表征[29]. 熵表示为不确定性的描述，提供的信息量多少，而互信息表示为不确定性的消除程度，获得的信息量多少．互信息公式为

$$\begin{aligned}I(X,Y)&=\mathcal{H}(X)-\mathcal{H}(X|Y)\\&=\mathcal{H}(Y)-\mathcal{H}(Y|X)\end{aligned}\tag{1}$$

其中，$\mathcal{H}(X)$为信息熵，用来描述X的不确定性．$\mathcal{H}(X|Y)$为条件熵，表示收到Y后，关于X尚存的平均不确定性．$I(X,Y)$为互信息，表示从Y获得的关于X的平均信息量．

Gomes等人[30]提出基于正则化信息最大化(Regularized Information Maximization，RIM)来学习条件概率模型．尽管RIM属于无监督聚类问题，但同样适用于解决学习隐性离散表征的各种问题．从而将学习Option策略的潜在变量o的问题转化为学习状态-动作空间的潜在表征问题．

2.3　最大熵强化学习

最大熵强化学习使用熵正则化来优化目标函数[31]. 在最大化期望奖励的同时引入最大熵，促进智能体提高探索能力，同时在面对扰动时策略可以表现得更加稳定．因此，构建一个最大熵目标函数为

$$\begin{aligned}J(\pi)=&\sum_{(s,a,o)\sim\pi,\pi_o}p(s,a)\cdot\\&[Q(s,a)+\alpha\mathcal{H}(\pi(\cdot|s))]\end{aligned}\tag{2}$$

其中，$p(s,a)=d_\pi(s)\pi(a|s)$表示为由策略π诱导的状态-动作对的概率，$d_\pi(s)=\sum\gamma^t p(s_t=s)$为由策略$\pi$所诱导的具有折扣的访问频率，$\pi(a|s)=\sum\pi(o|s)\pi_o(a|s,o)$为分层策略，等同于每个Option下Option选择策略和Option内部策略之积．并且利用动作价值函数$Q(s,a)$来近似预期回报．此外，α为温度系数，表示着熵项$\mathcal{H}(\pi(\cdot|s))$的重要程度．当系数越高，探索率越高，意味着越能得到随机性较强的策略．

2.4　基于Option的分层强化学习

优势函数作为基于策略学习的重要方法之一，可显著提高算法学习效率[32]. Li等人[33]提出了基于优势函数与辅助奖励的分层强化学习(HRL with Advantage-based Auxiliary Rewards，HAAR)算法，利用增广状态的上层优势函数来构建内部奖励，并同时更新上下层策略，但是采用同策略的更新方法，降低了样本的利用率．将信息论与DRL相结合是HRL的一个研究热点[34]. Osa等人[35]提出了具有优势加权重要性的信息最大化(HRL Via Advantage-weighted Information Maximization，adInfoHRL)算法，通过信息最大化准则，以发现Option对应模态为解决方案，保证了Option的可区分度．Hou等人[36]在adInfoHRL算法基础上结合SAC算法，提出了具有优势加权重要性的软性行动者-评论家(Soft Actor-Critic with Advantage Weighted Mixture Policy，SAC-AWMP)算法，大幅提升了adInfoHRL算法的性能，但是没有体现分层学习带来的收益．Zhang等人[37]提出HIDIO算法，该算法引入了判别器与子轨迹特征提取器，并运用自监督学习对潜在Option进行表达．为了在不同Option之间进行自动选择，Gehring等人[38]提出一个3层的分层策略，来减轻对先验知识的需求．Rao等人[39]提出分层混合潜变量模型，该方法从离线数据集中学习不同抽象级别的Option，使学习的Option可以灵活地转移到不同的模态．

本文提出的HRL-AMIM算法与以上算法的区别在于，在最大熵分层强化学习框架下，利用互信息和TV距离设计内部奖励，并增广到环境奖励中，从

而构造一个采用异策略更新方式的伪奖励函数. 特别地,采用退火方法使Option数量随着迭代次数改变而非固定不变. 此外,通过消融实验和超参数敏感性分析,在参数的选择上给出实用性建议,构建完整的实验体系.

3 HRL-AMIM算法

3.1 整体框架

本章介绍了HRL-AMIM算法,由Option数量的设定和Option先验网络两个方面组成. 为减少先验知识对性能的影响,运用退火方法使Option数量的学习过程是动态的,并对Option数量进行限制,防止分层抽象失去意义. 利用adInfoHRL算法中的思想,构建了Option先验网络,实现了潜在变量的转化过程. 特别地,我们构造一个新的伪奖励函数,丰富了Option的多样性选择,可解决DRL中探索能力不足的问题.

HRL-AMIM算法的整体框架,如图1所示.

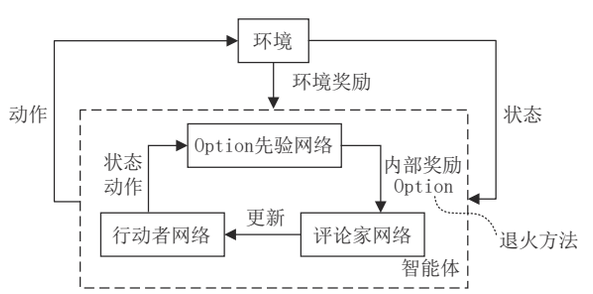


图1 HRL-AMIM框架图

该算法由Option先验网络、行动者网络和评论家网络构成. 智能体根据当前状态与环境交互得到动作和环境奖励. 对于Option先验网络的训练,其中该网络输入为状态-动作对,输出为内部奖励和Option. 评论家网络根据奖励与Option进行梯度更新,行动者网络根据评论家的更新调整策略参数,利用更新后的策略在下一状态选择动作.

3.2 Option数量的设定

以往基于Option分层强化学习工作中[18-19],Option数量$\mathcal{O}^{num}$通常是固定的. 当数量太少时,Option选择策略的范围很广,使得探索能力缓慢降低. 同时,数量较少会导致时序抽象作用大大减弱. 但当数量太多时,Option选择策略变得灵活性较差且很难找到最优策略.

针对以上问题,提出一种Option数量退火方法. 采用迭代次数将Option数量$\mathcal{O}^{num}$退火,公式为

$$\mathcal{O}_k^{num}=f(\mathcal{O}_k^{num};\xi)=\mathcal{O}_i^{num}e^{-\xi k}, \tag{3}$$

其中,k为迭代次数,$\mathcal{O}_i^{num}$为初始Option数量,ξ为退火温度. 并定义了一个最少Option数量$\mathcal{O}_s^{num}$,当$\mathcal{O}_k^{num}<\mathcal{O}_s^{num}$时,令$\mathcal{O}_k^{num}=\mathcal{O}_s^{num}$,来防止Option崩溃为单个动作.

当Option数量增多时,可以提高Option的选择多样性,但随之会产生部分低利用率的Option,导致算法性能效果不佳. 因此,在训练网络时,我们采用了丢弃机制. 丢弃机制的思想是对于一次迭代中的最后一层神经网络,选中利用率低下的神经元将其丢弃,再进行网络的训练和优化,直至训练结束. 本文对于低利用率的解释是指在最近的5次采样中出现次数最少的,这不会影响训练模型本身. 我们对于丢弃后的网络,在进行训练时,可看成一种数据增强(Data Augmentation)的过程.

3.3 Option先验网络

3.3.1 优势函数的模态转化

优势函数通常具有多种模态,理想情况下,每个Option策略都对应优势函数的单个模态,找到这些模态就可以调用不同的Option策略. 但在实践中准确找到优势函数的模态并非易事. 因此,提出将寻找优势函数模态的问题转化为寻找状态-动作对的概率密度模态问题.

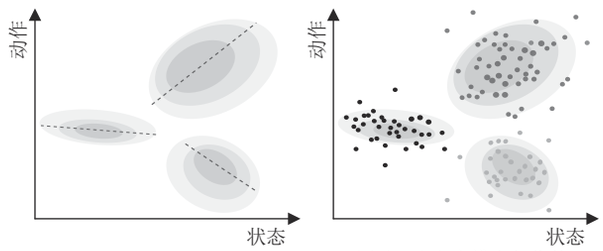
首先提出一个基于优势函数的策略为

$$\pi_A(a|s)=\frac{f(A_\pi(s,a))}{M}=\frac{\exp(A_\pi(s,a))}{M} \tag{4}$$

其中,$A_\pi(s,a)=Q(s,a)-V(s)$为优势函数,M为配分函数,f为一个单调递增的指数函数且满足$f>0$. 当遵循此策略时,能发现具有更大优势的动作.

基于此,优势函数的模态相当于由$\pi_A(a|s)$诱导的密度模态,则优势函数模态的问题可简化为由$\pi_A(a|s)$引起的样本聚类问题. 但直接进行聚类,会浪费计算资源,故将聚类样本问题再次转化为最大化$I(o,(s,a);\eta)$的问题. $I(o,(s,a);\eta)$划分状态-动作空间,不同于采用$\mathcal{H}(o|s)$和$I(s,o)$划分状态空间,具体如图2所示.

在图2中,(a)可视化了状态-动作空间中多模态优势函数,蓝色阴影为不同模态下优势函数的轮廓,红色虚线为预期的Option策略. (b)可视化了在优势加权重要性估计下,状态-动作对的密度模态与优势函数的模态是一一对应的,散点是由策略诱导的密度模态.



(a) 优势函数多种模态图　(b) (s,a) 密度模态与优势函数模态对应图

图2　优势函数模态的转化

因此，通过互信息最大化可有效地解决由策略 $\pi_A(a|s)$ 引起的样本聚类问题.

3.3.2　内部奖励的目标函数

HRL-AMIM算法利用内在动机丰富Option，不仅强调了Option多样化的重要性，而且提高了智能体向更广阔空间搜索的能力. 结合2.2节提到的基于正则化信息最大化(RIM)理论，引入互信息 $I(o,(s,a);\eta)$ 和正则化项 $\ell(\eta)$ 作为内部奖励 $R_{in}=\delta_2\ell(\eta)-\delta_1 I(o,(s,a);\eta)$. 一是互信息作为内部奖励可以增强每个内部Option策略的可判别性；二是通过增加基于TV距离的内部奖励，使选择Option策略在分配Option时，状态-动作对中的扰动项并不会导致选择Option策略有根本上的变化，充分地考虑了状态-动作空间的连续性.

因此，HRL-AMIM算法提出用 η 参数化一个神经网络 $p(o|s,a;\eta)$，即Option先验网络. 该网络为了最大化互信息的状态-动作对，通过最小化Option先验网络的目标函数来更新参数 η.

目标函数表示为

$$J_o(\eta)=\delta_2\ell(\eta)-\delta_1 I(o,(s,a);\eta) \tag{5}$$

其中，公式(6)为互信息，公式(7)为TV距离正则化项. δ_1 为噪声权重系数，δ_2 为互信息权重系数.

互信息公式表示为

$$I(o,(s,a);\eta)=\mathcal{H}(o;\eta)-\mathcal{H}(o|s,a;\eta) \tag{6}$$

其中，$\mathcal{H}(o;\eta)$ 是由潜在变量 o 和参数 η 组成的熵项，$\mathcal{H}(o|s,a;\eta)$ 表示为条件熵.

TV距离是基于虚拟对抗训练(Virtual Adversarial Training，VAT)的正则化项，具有鼓励探索的含义，公式为

$$\ell(\eta)=D_{\mathrm{TV}}(p(o|\tilde{s},\tilde{a};\eta)\,\|\,p(o|s,a;\eta)) \tag{7}$$

其中，$\tilde{s}=s+\varepsilon_s$，$\tilde{a}=a+\varepsilon_a$，且 ε_s 和 ε_a 为高斯噪声. $p(o|s,a;\eta)$ 为Option先验网络. 正则化项 $\ell(\eta)$ 惩罚原始状态-动作对和扰动之间的不相似性，且这种正则化提高了潜在表征学习的能力，算法性能也在许多实验任务中得到验证.

为了强调内部奖励的重要性，将内部奖励 R_{in} 作用在环境奖励 R 上，得到增广奖励函数 R_{aug}，设定一个新的伪奖励函数为

$$R_{aug}(s,a)=(1-\upsilon)R(s,a)+\upsilon R_{in} \tag{8}$$

其中，υ 是超参数，用于控制多样性项对奖励的相对重要性，υ 的取值范围为 $0<\upsilon<1$. 当 υ 无限趋近于0时，增广奖励函数退化成环境奖励函数，即 $R_{aug}(s,a)=R(s,a)$.

在实践中需要近似优势函数，并估计当前优势函数模态对应的潜在变量 o. 在由策略 $\pi_A(a|s)$ 诱导的状态-动作对的概率密度函数 $p_{\pi_A}(s,a)$ 下，潜在变量 o 的概率密度为

$$p(o;\eta)=\int p_{\pi_A}(s,a)\,p(o|s,a;\eta)\,dads. \tag{9}$$

熵项 $\mathcal{H}(o;\eta)$ 表示为

$$\mathcal{H}(o;\eta)=-\int p(o;\eta)\log p(o;\eta)\,do. \tag{10}$$

相似地，条件熵 $\mathcal{H}(o|s,a;\eta)$ 表示为

$$\mathcal{H}(o|s,a;\eta)=-\int p_{\pi_A}(s,a)\,p(o|s,a;\eta)\log p(o|s,a;\eta)\,dads. \tag{11}$$

因此，在由策略 $\pi_A(a|s)$ 引起的状态-动作对概率密度 $p_{\pi_A}(s,a)$ 下，计算互信息项扮演着重要角色.

3.3.3　优势函数加权重要性采样

虽然在学习过程中由策略 $\pi_A(a|s)$ 引起的样本是可用的，但从这些样本中得到的离散表征并不对应优势函数的模态. 为了解决该问题，引入优势加权重要性采样方法.

通过此方法，我们将计算得到的权重代替条件熵中的概率密度函数. $d_{\pi_A}(s)=\sum_{t=0}^{T}\gamma^t p(s_t=s)$ 和 $d_b(s)=\sum_{t=0}^{T}\gamma^t p(s_t=s)$ 分别为由策略 π_A 和 b 所诱导的具有折扣的访问频率，其中 b 表示用于生成经验的行为策略. 假设策略更新引起的状态分布变化足够小，可近似分布 $d_{\pi_A}(s)\approx d_b(s)$，则重要性采样权重表示为

$$\begin{aligned}\omega_A(s,a)&=\frac{p_{\pi_A}(s,a)}{p_b(s,a)}\\&=\frac{d_{\pi_A}(s)\pi_A(a|s)}{d_b(s)b(a|s)}\approx\frac{\pi_A(a|s)}{b(a|s)}\\&=\frac{f(A_\pi(s,a))}{Mb(a|s)}=\frac{\exp(A_\pi(s,a))}{Mb(a|s)}.\end{aligned} \tag{12}$$

为了避免训练时间过长，同时提高训练的稳定性，该算法对重要性采样权重进行归一化：

$$\bar{\omega}_A(s,a)=\frac{\omega_A(s,a)}{\sum_{i=1}^{N}\omega_A(s_i,a_i)}=\frac{\dfrac{f(A_\pi(s,a))}{Mb(a|s)}}{\sum_{i=1}^{N}\dfrac{f(A_\pi(s_i,a_i))}{Mb(a_i|s_i)}}=\frac{\dfrac{\exp(A_\pi(s,a))}{b(a|s)}}{\sum_{i=1}^{N}\dfrac{\exp(A_\pi(s_i,a_i))}{b(a_i|s_i)}} \tag{13}$$

$\bar{\omega}_A$称为优势加权重要性采样权重,用于估计潜在变量的目标函数.其中,配分函数M可以被消去,所以在实际运算过程中不需要计算M,这大大节省了计算时间成本.N是经验池中样本大小.

通过优势加权重要性采样权重,潜在变量o的概率密度公式可被估计为

$$\hat{p}(o;\eta)=\int\bar{\omega}_A(s,a)p(o|s,a;\eta)dads. \tag{14}$$

则熵$\hat{\mathcal{H}}(o;\eta)$的经验估计为

$$\hat{\mathcal{H}}(o;\eta)=-\int\hat{p}(o;\eta)\log\hat{p}(o;\eta)do. \tag{15}$$

相似地,条件熵$\hat{\mathcal{H}}(o|s,a;\eta)$的经验估计为

$$\hat{\mathcal{H}}(o|s,a;\eta)=-\int\bar{\omega}_A(s,a)p(o|s,a;\eta)\log p(o|s,a;\eta)dads. \tag{16}$$

为了训练Option先验网络,行为策略最新收集的样本存入Option先验网络经验池$\mathcal{D}$.本文使用此经验池进行潜在变量的表征学习,可以有效提高Option先验网络的性能.虽然从行为策略中进行采样Option,理论上仍是异策略的,但在真正实现时,使用了最新的行为策略收集到的样本,在某种程度上是"半"异策略的.此外,我们采用了延迟更新的技术,使Option先验网络的更新频率得以推迟,不仅可以获得低方差的函数估计,还提高了Option先验网络的稳定性.

3.4 HRL-AMIM算法

针对经典OC算法中将数量取固定值,从而无法平衡探索和利用的问题,本文提出了基于优势加权互信息最大化的最大熵分层强化学习(HRL-AMIM)算法.在由φ参数化的Option策略网络$\pi(a|s;\varphi)$下,原始最大熵强化学习模型引入两个内部奖励$I(o,(s,a);\eta)$和$\ell(\eta)$,重新定义HRL-AMIM算法的最大熵目标函数为

$$J(\pi)=\sum_{(s,a,o)\sim\pi_o(a|s,o),\pi(o|s)}d^\pi(s)\pi(a|s;\varphi)[Q(s,a)+\alpha_\pi\mathcal{H}(\pi(\cdot|s;\varphi))+\delta_2\ell(\eta)-\delta_1I(o,(s,a);\eta)] \tag{17}$$

其中,$Q(s,a)$表示软动作值函数.

根据2.3节得到$\pi(a|s;\varphi)=\sum\pi(o|s)\pi_o(a|s,o;\varphi)$,$\pi(o|s)$为Option选择策略;$\pi_o(a|s,o;\varphi)$为由$\varphi$参数化的Option内部策略,是一个随机的高斯混合策略.采取随机策略的原因是该策略能在大型复杂连续的空间更好地搜索,并且可以将探索和利用集成到同一个策略中.α_π为控制HRL-AMIM随机性的温度系数.

在每个环境步下采样Option,Option通过与环境交互来获得下一时刻的状态.在给定状态s下,通过softmax策略计算Option选择策略$\pi(o|s)$为

$$\pi(o|s)=\frac{\exp(U(s,o))}{\sum_{o\in\mathcal{O}}\exp(U(s,o))} \tag{18}$$

其中,$U(s,o)$称软Option值函数.

当Option策略为随机策略时,不仅需要近似软Option值函数$U(s,o)$,还需要近似软动作值函数$Q(s,a)$.但是在没有探索的情况下学习策略时,则用$o=\arg\max U(s,o)$代替Option选择策略公式(18).在此实验中,Option选择策略每隔d个时间步骤确定一次Option,即$t=0,d,2d\cdots$.

在Option内部策略$\pi_o(a|s,o;\varphi)$下,软Option值函数$U(s,o)$为

$$U(s,o)=_{a\sim\pi_o(a|s,o)}[R|s,o]=\int\pi_o(a|s,o;\varphi)(Q(s,a)-\alpha_\pi\log\pi_o(a|s,o;\varphi))da. \tag{19}$$

在Option选择策略$\pi(o|s)$下,软状态值函数$V(s)$为

$$V(s)=_{o\sim\pi(o|s)}[U(s,o)-\alpha_o\log\pi(o|s)]=\int\pi(o|s)(U(s,o)-\alpha_o\log\pi(o|s))do \tag{20}$$

其中,α_o表示Option内部策略的温度系数.

软动作值函数$Q(s,a)$为

$$Q(s,a)=\gamma_{p(s'|s,a)}V(s')+R(s,a)+\delta_2\ell(\eta)-\delta_1I(o,(s,a);\eta)-\log p(o|s,a;\eta) \tag{21}$$

其中,Option先验网络中的优势函数为$A_\pi(s,a)=Q(s,a)-V(s)$,它需要用函数逼近器来进行估计.

在大规模连续动作空间下,需要对软策略迭代进行近似.考虑使用函数逼近估计状态值函数、动作值函数和策略.并不对收敛性进行评估与改进,而是交替使用随机梯度下降(SGD)算法优化网络.因此,评论家网络由ψ参数化的软Option值网络$U(s,o;\psi)$和由θ参数化的软动作值网络$Q(s,a;\theta)$组成.行动者网络由φ参数化的Option策略网络$\pi(a|s;\varphi)$组成.

软Option值网络$U(s,o;\psi)$的参数可以被训练

为最小化残差平方：

$$J_U(\psi_i)=_{(s,o)\sim\mathcal{R}}[\frac{1}{2}(U(s,o;\psi_i)-\hat{U}(s,o))^2] \quad (22)$$

其中，状态s从回放经验池$\mathcal{R}$中采样得到.

为了防止Q值被过高估计，减小误差偏置，我们采用双Q值学习$y=r+\gamma\min_{i=1,2}Q(s',a',\theta_i)$. 使用两个独立的$Q$值函数并选择两者中较小的$Q$值来计算目标值：

$$\begin{aligned}\hat{U}(s,o)=&_{a\sim\pi_o(a|s,o)}[\min_{i=1,2}Q(s,a;\theta_i)\\&-\alpha_\pi\log\pi_o(a|s,o;\varphi)]\end{aligned} \quad (23)$$

其中，动作a根据当前策略π_o采样. 潜在变量o根据softmax策略进行采样.

软动作值网络$Q(s,a;\theta)$的参数可以被训练为最小化软贝尔曼残差：

$$J_Q(\theta_i)=_{(s,a)\sim\mathcal{R}}[\frac{1}{2}(Q(s,a;\theta_i)-\hat{Q}(s,a))^2] \quad (24)$$

其中，状态-动作对(s,a)从回放经验池$\mathcal{R}$中采样得到. 类似地，采用双U值计算目标值$\hat{Q}(s,a)$，最后得到新函数：

$$\begin{aligned}J_Q(\theta_i)=&_{(s,a)\sim\mathcal{R}}[\frac{1}{2}[Q(s,a;\theta_i)-\\&(R(s,a)+\delta_2\ell(\eta)-\delta_1 I(o,(s,a);\eta)\\&-\log p(o|s,a;\eta))+\gamma_{o'\sim\pi(o|s)}\\&(\min_{i=1,2}U(s',o',\psi_i)-\alpha_o\log\pi(o'|s'))]^2].\end{aligned} \quad (25)$$

用似然比梯度估计来学习策略梯度，使算法不需要通过策略和软动作值网络进行反向传播梯度. 可以代替公式(17)，通过最小化期望KL散度，确定策略网络新的目标函数为

$$J_\pi(\varphi,\theta_i)=_{s\sim R}\left[D_{KL}\left(\pi(\bullet|s;\varphi)\left\|\frac{\exp(Q(s,\bullet;\theta_i,\varphi))}{M_\theta(s)}\right.\right)\right] \quad (26)$$

其中，$M_\theta(s)=\int\pi(a|s;\varphi)\exp((Q(s,a;\theta_i))$为配分函数，对分布进行了归一化，通常难以计算. 动作a根据当前策略π采样. 潜在变量o根据Option先验网络$p(o|s,a;\eta)$获得.

软动作值网络由神经网络表示，并且可以进行微分，为了降低方差，采用重参数化技术，使用函数f重参数化策略. 特别地，HRL-AMIM算法引入了Option先验网络中的潜在变量o.

$$a=f(\varepsilon;s,\varphi)=\int o_i f_i(\varepsilon;s,\varphi) \quad (27)$$

其中，动作a_i是第i个Option下的动作，o_i是潜在变量的第i个元素. ε是从固定各向同性的高斯分布即球形高斯分布采样的输入噪声. 最终目标函数为

$$\begin{aligned}J_\pi(\varphi,\theta_i)=&_{s\sim\mathcal{R},\varepsilon\sim\mathcal{N}}[\log\pi(f(\varepsilon;s,\varphi)|s)-\\&\min_{i=1,2}Q(s,f(\varepsilon;s,\varphi);\theta_i)+\log M_\theta(s)]\end{aligned} \quad (28)$$

其中，噪声ε从高斯分布$\mathcal{N}(0,I)$中采样得到.

配分函数$M_\theta(s)$与φ无关，所以对策略梯度没有贡献，可忽略其计算. 则目标函数的梯度近似为

$$\begin{aligned}\hat{\nabla}_\theta J_\pi(\varphi,\theta_i)=&\nabla_\theta\log\pi(a|s;\varphi)+\\&(\nabla_a\log\pi(a|s;\varphi)-\nabla_a Q(s,a;\theta_i)\nabla_\varphi f(\varepsilon;s,\varphi))\end{aligned} \quad (29)$$

其中，a由f估计.

函数进行多次梯度更新应该仍然保证其收敛性，因此需要提供一个稳定的目标. 如果缺少固定的目标，就会导致每次更新引入新的残余误差，并开始累积此误差. 目标网络可以用于减少多步更新的误差，并且策略在高误差状态更新会导致发散，那么应该每隔k轮更新一次策略网络和两个目标网络. 因此，我们使用具有平滑系数τ的指数移动加权平均来同时软更新目标Q网络和目标V网络：

$$\begin{aligned}\bar{\psi}_i&\leftarrow\tau\psi_i+(1-\tau)\bar{\psi}_i\\\bar{\theta}_i&\leftarrow\tau\theta_i+(1-\tau)\bar{\theta}_i.\end{aligned} \quad (30)$$

τ太大会导致训练不稳定，而太小会让训练速度大大降低.

算法1介绍了HRL-AMIM的整个过程，也是对图1结构框架的详细描述.

算法1.　HRL-AMIM

输入：学习率l，折扣因子γ

1. 初始化：全局经验池$\mathcal{R}$，Option先验网络经验池$\mathcal{D}$，小批量样本$\mathcal{B}_{s-off}$和$\mathcal{B}_{off}$
2. 初始化：初始Option数量$\mathcal{O}_i^{num}1$，最少Option数量$\mathcal{O}_s^{num}$
3. 初始化：Option先验网络$p(\bullet|\bullet,\bullet;\eta)$，Option策略网络$\pi(\bullet|\bullet,\bullet;\varphi)$，评论家网络$U(\bullet|\bullet,\bullet;\psi)$和$Q(\bullet|\bullet,\bullet;\theta)$，目标评论家网络$U(\bullet|\bullet,\bullet;\bar{\psi})$和$Q(\bullet|\bullet,\bullet;\bar{\theta})$
4. FOR 每轮迭代k DO
5. 　FOR 每个环境步 DO
6. 　　$o\sim\pi(o|s)$
7. 　　$a\sim noise+\pi(a|s;\varphi)$
　　$=noise+\sum\pi(o|s)\pi_o(a|s,o;\varphi)$
8. 　　$R,s'\sim p(s'|s,a)$
9. 　　$R_{aug}=(1-\upsilon)R+\upsilon R_{in}$
10. 　　$\mathcal{R}\leftarrow\mathcal{R}\cup(s,a,R_{aug},s')$
11. 　END FOR
12. 　IF Option先验网络经验池$\mathcal{D}$已满 THEN
13. 　　从$\mathcal{D}$中取出“半”异策略小批量样本$\mathcal{B}_{s-off}$
14. 　　$\eta\leftarrow\eta-\alpha_\eta\hat{\nabla}_\eta J_o(\eta)$
15. 　　清除Option先验网络经验池$\mathcal{D}$

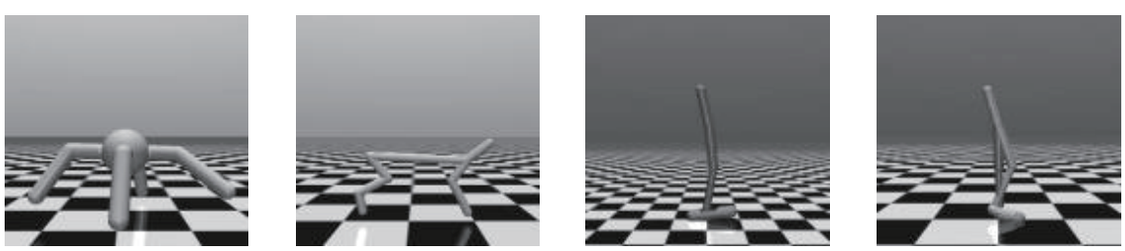
END IF
FOR 每个更新步 DO
从经验池$\mathcal{R}$中取出异策略小批量样本$\mathcal{B}_{off}$
IF t走完d步 THEN
$o_i=p(o_i|s_i,a_i;\eta)$
$\psi_i\leftarrow\psi_i-\alpha_{\psi_i}\hat{\nabla}_{\psi_i}J_U(\psi_i)$, for $i\in\{1,2\}$
$\theta_i\leftarrow\theta_i-\alpha_{\theta_i}\hat{\nabla}_{\theta_i}J_Q(\theta_i)$, for $i\in\{1,2\}$
$\varphi\leftarrow\varphi-\alpha_{\varphi}\hat{\nabla}_{\varphi}J_\pi(\varphi)$
END IF
END FOR
$\bar{\psi}_i\leftarrow\tau\psi_i+(1-\tau)\bar{\psi}_i$, for $i\in\{1,2\}$
$\bar{\theta}_i\leftarrow\tau\theta_i+(1-\tau)\bar{\theta}_i$, for $i\in\{1,2\}$
$\mathcal{O}_{k+1}^{num}=\max(f(\mathcal{O}_k^{num}),\mathcal{O}_s^{num})$
END FOR

4 实 验

将HRL-AMIM算法应用于一系列连续环境的Mujoco实验. 不仅与经典深度强化学习中的SAC算法和TD3算法进行对比,还与同类型分层强化学习中的SAC-AWMP算法和adInfoHRL算法相比较. 此外,进行消融实验和超参数敏感性实验来验证HRL-AMIM算法的有效性和鲁棒性. 最后通过可视化Option分布来验证Option选择策略的合理性.

4.1 实验环境

Gym[40]作为OpenAI的仿真平台,是深度强化学习中一个重要的开源工具包. 其提供了丰富的实验环境,包含Atari游戏、Mujoco、经典控制和Box2D等. 其中,Mujoco[41]是一个免费的开源物理引擎,不仅用于实现基于模型的计算,还可以用作传统的模拟器,包括游戏和交互式虚拟环境. 同时,Mujoco作为高维度环境可解决复杂的大规模问题. 所以,本文采用在Gym中开发的一系列大规模连续控制任务作为实验环境,如图3所示.



(a) Ant-v2 (b) HalfCheetah-v2 (c) Hopper-v2 (d) Walker2d-v2

图3 Mujoco物理引擎环境

针对一系列连续控制目标,我们选取了Ant-v2、HalfCheetah-v2、Hopper-v2和Walker2d-v2这4个任务进行分层强化学习算法的实现. 这些任务可以有效评估HRL-AMIM算法在大规模连续状态-动作空间环境下的性能,其详细介绍如表1所示.

表1 Mujoco实验任务介绍

环境名称	动作维度	状态维度	实验任务
Ant-v2	8	111	训练一个四足行走智能体
HalfCheetah-v2	6	17	训练一个双足奔跑智能体
Hopper-v2	3	11	训练一个单足跳跃智能体
Walker2d-v2	6	17	训练一个双足行走智能体

4.2 实验设置

所有任务的每个算法,均以5个不同的随机种子独立运行,随机种子的取值范围是[0,3000]的整数. 每次实验为100万步,期望回报通过每1000步评估10次来进行估计.

在实验对比图中,实线表示经过当前时间步评测后,该算法5次独立运行训练得到的平均性能结果. 阴影部分为5次独立训练的平均性能的误差,阴影部分越大,意味着该算法的训练稳定性越差.

Option策略网络采用有两个卷积层的一维卷积神经网络. Option先验网络和评论家网络均采用有两个隐藏层的线性神经网络,其中第一层有400个神经元,第二层有300个神经元. 同时,所有网络均采用Relu函数作为激活函数,并使用Adam优化器以梯度下降的方式更新神经网络参数. HRL-AMIM其他的超参数设置如表2所示.

为了科学地对比不同算法间的性能差异,涉及的其他算法均使用与HRL-AMIM算法相似的行动者-评论家网络结构,其超参数设置也尽可能地与该算法保持一致.

4.3 实验结果与分析

4.3.1 与DRL的对比实验

为了研究HRL-AMIM算法与经典深度强化学习算法的对比性能. 在大规模连续型任务Ant-

表2　HRL-AMIM算法超参数

超参数	取值	参数描述	超参数	取值	参数描述
$\mathcal{D}$	5000	Option先验网络经验池	α_π	0.2	对于π的熵值
$\mathcal{R}$	1e6	全局经验池	α_o	0.001	对于o的熵值
υ	0.8	奖励相对重要性	ξ	5e-3	退火温度
$\mathcal{O}_s^{num}$	4	最少Option数量	δ_1	1	互信息权重系数
$\mathcal{O}_i^{num}1$	50	初始Option数量	δ_2	5	噪声权重系数
$\mathcal{B}_{s-off}$	50	Option先验网络训练批量数	l	3e-4	学习率
$\mathcal{B}_{off}$	100	评论家网络训练批量数	γ	0.99	折扣因子
	$100\mathcal{O}^{lnum}$	Option策略网络训练批量数	τ	5e-3	目标平滑系数

v2、HalfCheetah-v2、Hopper-v2和Walker2d-v2中，我们将提出的HRL-AMIM算法与SAC算法和TD3算法进行性能对比，绘制学习曲线如图4所示.

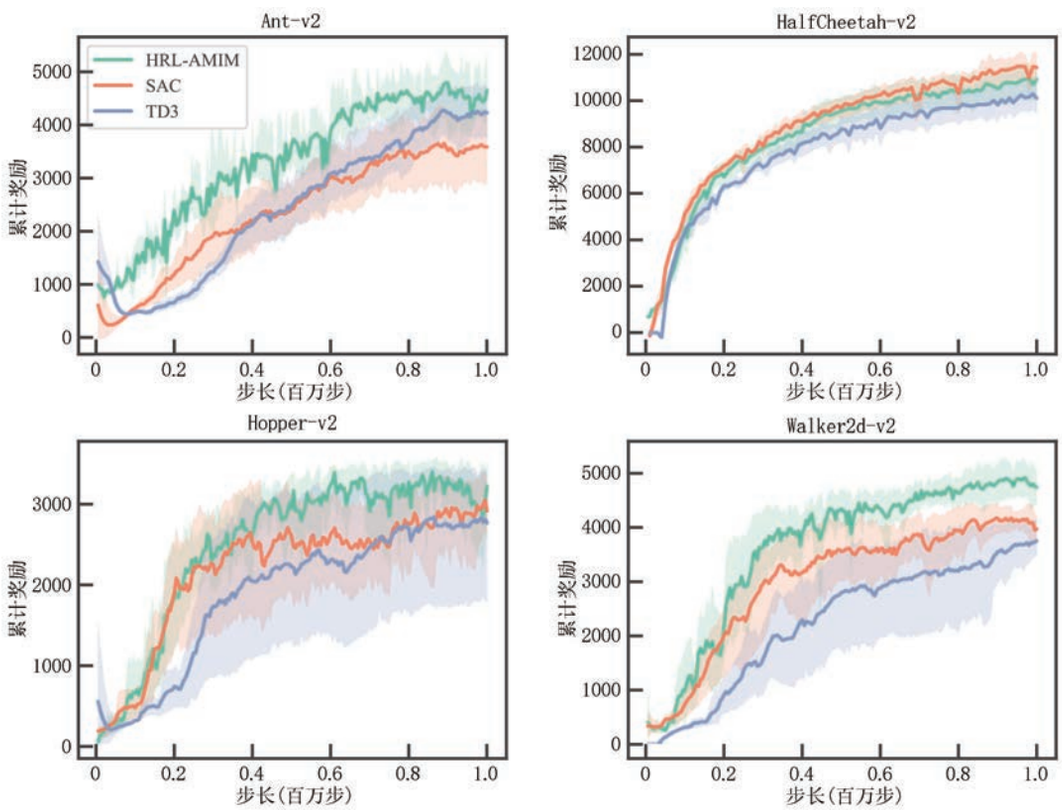


图4　HRL－AMIM与经典深度强化学习算法学习曲线

从图4可以看出，在Ant-v2、Hopper-v2和Walker2d-v2中，HRL-AMIM算法的表现均优于SAC算法和TD3算法. 然而，在HalfCheetah-v2中，虽然3个算法都以较快的速度达到收敛，但HRL-AMIM算法略逊于SAC算法. 究其原因是分层强化学习在稀疏奖励环境下更能表现出优于其他深度强化学习的探索能力. SAC算法不仅采用策略熵方法，还能自动调节温度系数，减轻了超参数调节的难度，同时更加强调探索的能力. 相比较而言，HRL-AMIM算法采用分层技术，Option策略的多样性提高了Option评论家的探索与性能. 但是不像原始动作每个时间步都可用，Option在满足终止条件之前执行可变时间步，在此期间其他Option保持休眠状态. 故对于这个密集奖励环境存在不确定性，有时目标对于Option还不如原始动作般有效.

表3给出了3个算法在5次独立训练终止时，策略在对应任务上所获得的累计奖励的平均值、标准差和中位数. 此外，以平均值作为最终评判标准，类似地，以下表格数据均按此标准进行分析表述.

从表3可以看出，HRL-AMIM算法在Ant-v2、

表3 HRL-AMIM与经典深度强化学习算法最终性能

算法	指标	Ant-v2	HalfCheetah-v2	Hopper-v2	Walker2d-v2
HRL-AMIM	平均值	**4656**	10938	**3223**	**4745**
	标准差	**504**	706	**436**	**318**
	中位数	**4493**	11 351	**3380**	**4835**
SAC	平均值	3591	**11 438**	2918	3977
	标准差	802	**700**	349	266
	中位数	3860	**11 450**	2938	3921
TD3	平均值	4234	10 112	2770	3754
	标准差	584	764	970	392
	中位数	4229	10 601	3149	3648

Hopper-v2和Walkerd-v2上表现最为优异. 相对于SAC算法,HRL-AMIM算法的性能分别提高了30%、2%和19%. 同时,相对于TD3算法,HRL-AMIM算法的性能分别提高了10%、16%和21%. 在HalfCheetah-v2上,其最终性能优于TD3算法且仅次于SAC算法,而且SAC算法仅比该算法的最终累计奖励高出5%. 同时,对比这3个算法,HRL-AMIM算法的标准差相对较低,说明训练过程较为稳定.

4.3.2 与HRL的对比实验

根据以上实验结果,可验证HRL-AMIM算法在经典Mujoco任务上大体优于传统的深度强化学习. 为了进一步研究该算法与其他分层强化学习算法的对比性能,将该算法与其他同类型分层算法SAC-AWMP算法和adInfoHRL算法进行比较. 特别地,这两个算法同样采用了优势加权重要性采样的思想,绘制学习曲线如图5所示.

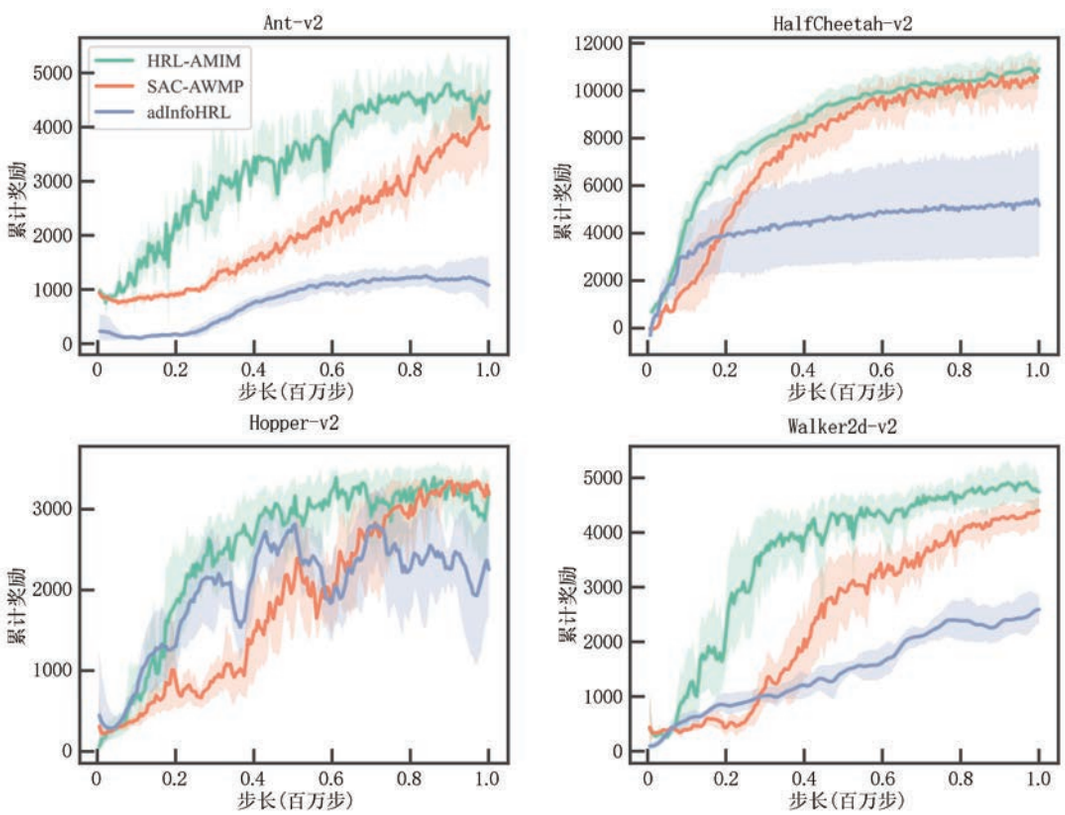


图5 HRL-AMIM与同类型分层强化学习算法学习曲线

从图5可以看出,HRL-AMIM算法在4个任务中表现均较为突出. 该算法可看成在adInfoHRL算法上的扩展,不同的是,采用了退火方法处理Option数量,大大加快了前期策略的学习速度. 并把内部奖励扩充到环境奖励,使智能体向更好的方向进行探索,验证了所提算法的优越性. 但是,与同类型HRL算法相比,HRL-AMIM算法在HalfCheetah-v2中的标准差最大. 究其原因是HRL-AMIM算法通过加强Option的多样性来提升智能体的探索能力,而强探索力会对算法产生不稳定因素,进而导致标准差无法降到足够水平.

表4给出了3个算法在5次独立训练终止时,策略在对应任务上所获得的累计奖励的平均值、标准差和中位数.

从表4可以看出,在所有任务中HRL-AMIM算法的训练效果最佳. 在Ant-v2、HalfCheetah-v2、

表4　HRL-AMIM与同类型分层强化学习算法最终性能

算法	指标	Ant-v2	HalfCheetah-v2	Hopper-v2	Walker2d-v2
HRL-AMIM	平均值	**4656**	**10 938**	**3223**	**4745**
	标准差	**504**	**706**	**436**	**318**
	中位数	**4493**	**11 351**	**3380**	**4835**
SAC-AWMP	平均值	4022	10 558	3185	4399
	标准差	718	587	312	331
	中位数	3895	10 770	3350	4494
adInfoHRL	平均值	1087	7244	2257	2594
	标准差	574	394	582	304
	中位数	885	7590	2290	2504

Hopper-v2和Walkerd-v2中，相比较SAC-AWMP算法，HRL-AMIM算法的性能分别提高了16%、4%、2%和8%. 同时，相对于adInfoHRL算法，HRL-AMIM算法的性能分别提高了328%、51%、43%和83%. 故该算法在同类型的分层强化学习中表现出巨大优势.

4.3.3　消融实验

根据以上实验结果，可验证HRL-AMIM算法在经典Mujoco任务上不仅优于传统的深度强化学习还优于同类型的分层强化学习算法. 为了进一步研究该算法的哪部分模型构建对实验性能有着重要影响，进行了消融实验. 首先，以Hopper-v2为例，进行增广环境奖励(Augmented)和不增广环境奖励(No augmented)的性能对比实验，并绘制了学习曲线图，如图6所示. 根据对比图6，可以看出增广环境奖励方法的算法性能更优异.

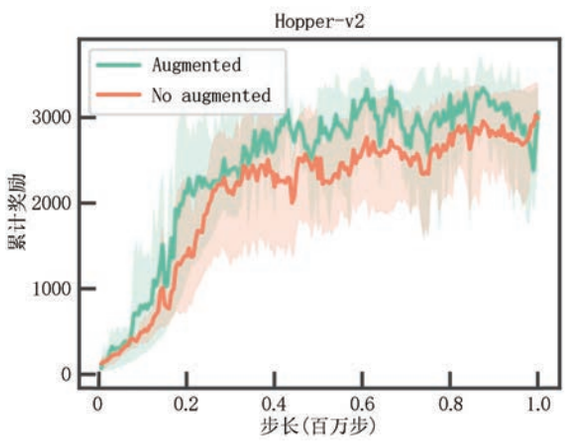


图6　增广环境奖励的消融实验性能曲线图

表5给了HRL-AMIM算法在5次独立训练终止时，增广与不增广环境奖励的策略在对应任务上所获得的累计奖励的平均值、标准差和中位数.

表5　增广环境奖励的消融实验最终性能

指标	平均值	标准差	中位数
增广环境奖励	**3223**	**436**	**3380**
不增广环境奖励	2995	479	3220

从表5可以看出，相比较不将内部奖励增广到环境奖励的方法而言，增广环境奖励方法具有更高的性能，最终累计奖励值提高了8%，标准差降低了10%.

此外，本文对退火方法进行消融实验. 使用退火方法的优点是Option数量随着迭代次数变化，使得搜索进程加快，缺点是需要设置一个合适的最少Option数量. 在adInfoHRL算法中，Option数量是固定不变的. 所以，本文以Hopper-v2为例，进行Option数量为退火(Annealing)和固定(Fixing)$\mathcal{O}^{num}=4$的性能对比实验，并绘制了学习曲线图，如图7所示. 根据对比图7，可以看出退火方法大大加快了策略的训练速度.

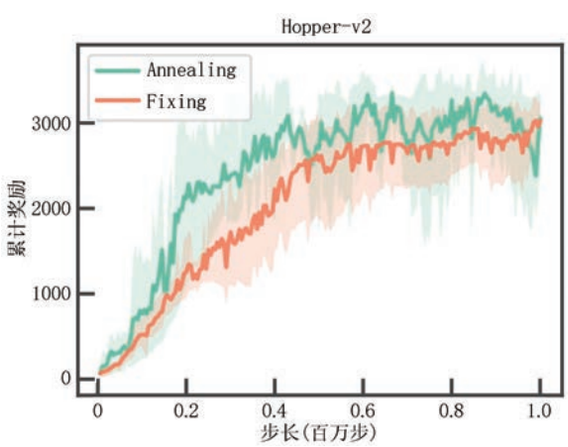


图7　退火方法的消融实验性能曲线图

表6给出了HRL-AMIM算法在5次独立训练终止时，退火与固定数量的策略在对应任务上所获得的累计奖励的平均值、标准差和中位数.

表6　退火方法的消融实验最终性能

指标	平均值	标准差	中位数
退火方法	**3223**	**436**	**3380**
$\mathcal{O}^{num}=4$	3030	409	3148

从表6可以看出，以Hopper-v2为例，退火方法不仅学习速度更快，而且相比较固定数量的Option而言，最终累计奖励值提高了6%，但标准差增加了7%.

对比两个主要模块的消融实验，增广环境奖励

不仅提升了算法性能，还提高了算法的稳定性．而退火方法虽然学习速度更快，但是在提升算法性能上不如增广环境奖励模块，且该方法导致标准差增加．总体来看，增广环境奖励对提升算法性能更加重要．

4.3.4 超参数敏感性实验

根据消融实验结果，证明了使用退火算法控制Option数量的变化和增广环境奖励是有意义的．下面进一步研究该算法对重要超参数的敏感性．目标网络是强化学习中常用的技巧，通过使用具有平滑系数的指数移动平均值来更新目标网络权重，从而有效提高算法的稳定性．绘制学习曲线如图8所示．

从图8可知，HRL-AMIM算法在$\tau=0.005$时表现最为优异．对于τ较小的情况，取$\tau=0.0005$时，智能体在初始阶段的学习速度很慢，经过50万时间步后才明显提高学习效率．对于τ较大的情况，取$\tau=0.5$时，HRL-AMIM算法在经过1百万时间步的训练后，在Hopper-v2和Walker2d-v2下的累计奖励仍然没有任何变化始终为0．这说明在该平滑系数下算法没有任何效果．在已被证实的经验中表明目标平滑系数用于稳定性训练，τ太大即快速移动目标会导致性能不稳定，而τ太小即缓慢移动目标则会使训练变慢．因此，加入$\tau=0.1$与$\tau=0.05$的情况，发现曲线的误差区间显著变宽，有效验证上述经验．故τ的合适值范围相对较宽．

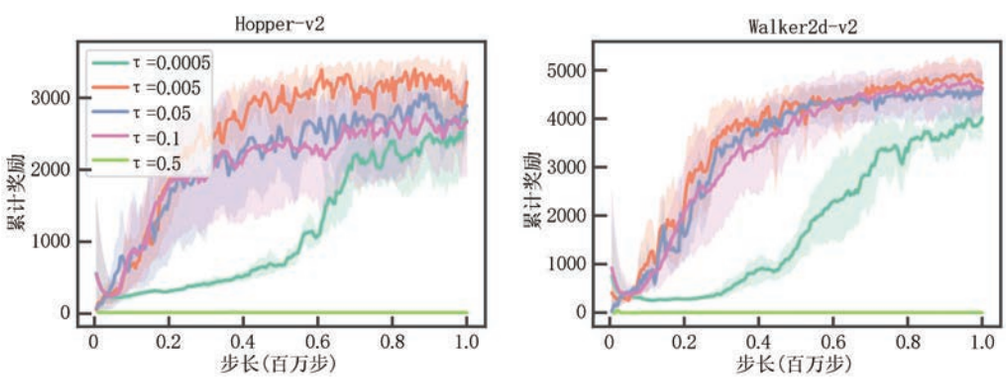


图8 不同目标平滑系数的HRL-AMIM算法学习曲线

表7给出了4个不同目标平滑系数的HRL-AMIM算法在5次独立训练终止时，策略在对应任务上所获得的累计奖励的平均值、标准差和中位数．

表7 不同目标平滑系数的HRL-AMIM算法最终性能

平滑系数	指标	Hopper-v2	Walker2d-v2
$\tau=0.0005$	平均值	2680	4025
	标准差	461	444
	中位数	2422	4074
$\tau=0.005$	平均值	**3223**	**4745**
	标准差	**436**	**318**
	中位数	**3380**	**4835**
$\tau=0.05$	平均值	2890	4634
	标准差	549	505
	中位数	2900	4731
$\tau=0.1$	平均值	2668	4626
	标准差	793	547
	中位数	3109	4524
$\tau=0.5$	平均值	17	3
	标准差	18	18
	中位数	6	2

从表7可以看出，HRL-AMIM算法的鲁棒性较好．选择目标平滑系数$\tau=0.005$时实验效果最佳，但当$\tau=0.5$时实验没有效果．对应Hopper-v2和Walker2d-v2，与$\tau=0.005$对比，当$\tau=0.0005$时算法的性能分别降低了20%和18%，且标准差分别增加了6%和40%；当$\tau=0.05$时算法的性能分别降低了12%和2%，且标准差分别增加了26%和59%；当$\tau=0.1$时算法的性能分别降低了21%和3%，且标准差分别增加了82%和72%．

为了进一步研究超参数对实验的影响，下面分别对互信息与噪声权重系数、奖励相对重要性、初始Option数量、最少Option数量与退火温度进行超参数敏感性实验．

表8给出了两个不同互信息与噪声权重系数的HRL-AMIM算法在5次独立训练终止时，策略在Hopper任务上所获得的累计奖励的平均值、标准差和中位数．

从表8可以看出，当选择互信息与噪声权重系数分别为$\delta_1=1$、$\delta_2=5$时实验效果最佳．相比较

表8　不同互信息与噪声权重的HRL-AMIM算法最终性能

互信息权重系数 噪声权重系数	指标	Hopper-v2
$\delta_1=1$ $\delta_2=5$	平均值	**3223**
	标准差	**436**
	中位数	**3380**
$\delta_1=0.1$ $\delta_2=0.5$	平均值	3188
	标准差	536
	中位数	3254
$\delta_1=0.05$ $\delta_2=0.1$	平均值	2955
	标准差	579
	中位数	3159

$\delta_1=0.1$、$\delta_2=0.5$而言，当$\delta_1=1$、$\delta_2=5$时HRL-AMIM算法性能提高了1%，且标准差降低了23%. 此外，相比较$\delta_1=0.05$、$\delta_2=0.1$而言，当$\delta_1=1$、$\delta_2=5$时HRL-AMIM算法性能提高了9%，且标准差降低了33%.

表9给出了3个不同奖励相对重要性的HRL-AMIM算法，在Hopper-v2任务上所获得的累计奖励的平均值、标准差和中位数.

表9　不同奖励相对重要性的HRL-AMIM算法最终性能

奖励相对重要性	指标	Hopper-v2
$\upsilon=0.8$	平均值	**3223**
	标准差	**436**
	中位数	**3380**
$\upsilon=0.5$	平均值	2747
	标准差	452
	中位数	2889
$\upsilon=0.2$	平均值	2662
	标准差	563
	中位数	2681

从表9可以看出，HRL-AMIM算法下不同奖励相对重要性系数的敏感性. 选择奖励相对重要性系数$\upsilon=0.8$时实验效果最佳. 相比较$\upsilon=0.5$而言，当$\upsilon=0.8$时算法的性能提高了17%，并且标准差降低了4%. 此外，相比较$\upsilon=0.2$而言，当$\upsilon=0.8$时算法的性能提高了21%，并且标准差降低了29%.

表10给出了两个不同初始Option数量的HRL-AMIM算法，在Hopper-v2任务上所获得的累计奖励的平均值、标准差和中位数.

由表10可以看出，对于不同初始Option数量而言，HRL-AMIM算法累计奖励的各项指标相差不大. 说明$\mathcal{O}_i^{num}$的合适值范围相对较宽. 以平均值为衡量标准，选择不同初始Option数量为$\mathcal{O}_i^{num}=50$时实验效果最佳. 相比较$\mathcal{O}_i^{num}=150$而言，当$\mathcal{O}_i^{num}=50$时算法的性能提高了2%，但是标准差也增加了2%.

表10　不同初始Option数量的HRL-AMIM算法最终性能

初始Option数量	指标	Hopper-v2
$\mathcal{O}_i^{num}=50$	平均值	**3223**
	标准差	**436**
	中位数	**3380**
$\mathcal{O}_i^{num}=150$	平均值	3157
	标准差	428
	中位数	3189

表11给出了3个不同最少Option数量的HRL-AMIM算法，在Hopper-v2任务上所获得的累计奖励的平均值、标准差和中位数.

表11　不同最少Option数量的HRL-AMIM算法最终性能

最少Option数量	指标	Hopper-v2
$\mathcal{O}_s^{num}=4$	平均值	**3223**
	标准差	**436**
	中位数	**3380**
$\mathcal{O}_s^{num}=2$	平均值	2983
	标准差	390
	中位数	3070
$\mathcal{O}_s^{num}=8$	平均值	3159
	标准差	493
	中位数	3365

从表11可以看出，HRL-AMIM算法下最少Option数量的敏感性. 选择最少Option数量$\mathcal{O}_s^{num}=4$时实验效果最佳. 相比较$\mathcal{O}_s^{num}=2$而言，当$\mathcal{O}_s^{num}=4$时算法的性能提高了8%，但是标准差也增加了12%. 此外，相比较$\mathcal{O}_s^{num}=8$而言，当$\mathcal{O}_s^{num}=4$时算法的性能提高了2%，并且标准差降低了13%.

表12给出了两个不同退火温度的HRL-AMIM算法，在Hopper-v2任务上所获得的累计奖励的平均值、标准差和中位数.

从表12可以看出，对于不同退火温度而言，

表12　不同退火温度的HRL-AMIM算法最终性能

退火温度	指标	Hopper-v2
$\xi=0.005$	平均值	**3223**
	标准差	**436**
	中位数	**3380**
$\xi=0.01$	平均值	2959
	标准差	581
	中位数	3018

HRL-AMIM算法累计奖励的各项指标相差不大.当不同退火温度为$\xi=0.005$时实验效果最佳.相比较$\xi=0.01$而言,当$\xi=0.005$时算法的性能提高了9%,且标准差降低了33%.

为了直观地看出各个超参数的指标值,下面绘制了误差柱状图,如图9所示.柱状图数值表示累计奖励在不同超参数下的平均值,黑棒为误差线,本文采用标准差来衡量该误差.

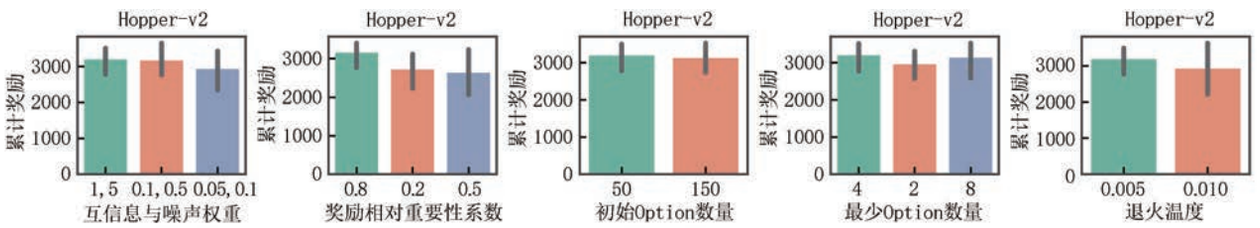


图9 HRL-AMIM算法的超参数敏感性实验柱状图

从图9可以看出,超参数敏感性实验的总体情况.对于初始Option数量和退火温度而言,超参数不敏感,性能指标基本稳定.对于互信息与噪声权重系数、奖励相对重要性和最少Option数量而言,超参数较为敏感.

4.3.5 Option分布

为了验证Option选择策略是合理的,利用t-SNE方法[42]对状态-动作空间进行降维,从而直观地说明Option的分布范围.为了便于观察选定Option数量递减到4时的情况.以Hopper-v2和Walker2d-v2为例,可视化不同Option对应的状态-动作空间,如图10所示.

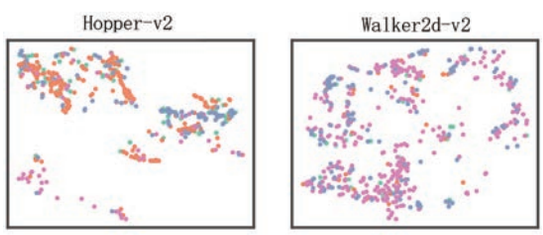


图10 运用t-SNE可视化Option分布

从图10可以看出,在两个任务中不同Option都对应着不同聚类.说明该Option选择策略训练有素,可以针对不同情况分配合适的Option.

5 总 结

本文提出了一种结合优势加权重要性采样和互信息最大化的异策略最大熵分层强化学习HRL-AMIM算法.将两个内部奖励增广到环境奖励,并构造了伪奖励函数,该方法增强了Option的多样性选择.同时,对Option数量进行动态的退火,目的是减少先验知识设定对算法性能的影响,并有效地解决了探索与利用不平衡的问题.此外,在基于SAC最大熵模型的结构上进行改进,使得HRL-AMIM算法具备强有力的探索性和良好的稳定性.

实验部分选取Mujoco环境中4个经典的连续控制任务验证了算法的性能.对比经典深度强化学习中的SAC算法和TD3算法,HRL-AMIM算法的最终总体性能分别提高了12%和14%.对比同类型分层强化学习中的SAC-AWMP算法和adInfoHRL算法,HRL-AMIM算法的最终总体性能分别提高了8%和116%.同时,分别对Option数量和增广环境奖励进行消融实验,验证了退火方法优于固定数量且增广环境奖励提高算法性能.并通过对重要超参数进行敏感性实验,说明了该算法具有很好的鲁棒性.此外,根据可视化Option分布可知,该Option选择策略可以针对不同情况分配合适的Option.

在未来的研究中,分层强化学习将更多地面向稀疏奖励问题.该学习利用抽象技术,不仅能快速捕捉外部奖励,还能收集内部奖励,有效地克服了稀疏奖励的问题.所以,我们下一步研究重点是将该方法应用于稀疏奖励环境中,验证其有效性.

参 考 文 献

[1] Liu Quan, Zhai Jian-Wei, Zhang Zong-Chang, et al. A survey on deep reinforcement learning. Chinese Journal of Computers, 2018, 41(1):1-27 (in Chinese)
(刘全,翟建伟,章宗长等.深度强化学习综述.计算机学报,2018,41(1):1-27)

[2] Goodfellow I, Bengio Y, Courville A, et al. Deep learning. Cambridge, USA: MIT press, 2016

[3] Sutton R S, Barto A G. Reinforcement learning: An introduction. Cambridge, USA: MIT press, 2018

[4] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control

through deep reinforcement learning. Nature, 2015, 518 (7540): 529-533

[5] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning//Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2016: 1-14

[6] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods//Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 2018: 1587-1596

[7] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor//Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 2018: 1861-1870

[8] Konda V R, Tsitsiklis J N. Actor-critic algorithms//Proceedings of the Advances in Neural Information Processing Systems, Cambridge, USA, 1999: 1008-1014

[9] Liu Cheng-Hao, Zhu Fei, Liu Quan. Option-critic algorithm based on sub-goal quantity optimization. Chinese Journal of Computers, 2021, 44(9): 1922-1933 (in Chinese)
(刘成浩，朱斐，刘全. 基于优化子目标数的 option-critic 算法. 计算机学报, 2021, 44(9): 1922-1933)

[10] Henderson P, Islam R, Bachman P, et al. Deep reinforcement learning that matters//Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 2374-3468

[11] Haarnoja T, Tang H, Abbeel P, et al. Reinforcement learning with deep energy-based policies//Proceedings of the International Conference on Machine Learning, Sydney, Australia, 2017: 1352-1361

[12] Sutton R S, Precup D, Singh S. Intra-option learning about temporally abstract actions//Proceedings of the International Conference on Machine Learning, Madison, USA, 1998: 556-564

[13] Sutton R S, Precup D, Singh S. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence, 1999, 112(1-2): 181-211

[14] Zhou Wen-Ji, Yu Yang. Summarize of hierarchical reinforcement learning. CAAI Transactions on Intelligent Systems, 2017, 12 (5): 590-594 (in Chinese)
(周文吉，俞扬. 分层强化学习综述. 智能系统学报, 2017, 12 (5): 590-594)

[15] Thrun S, Schwartz A. Finding structure in reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems, Denver, USA, 1994: 385-392

[16] Daniel C, Neumann G, Kroemer O, et al. Hierarchical relative entropy policy search. Journal of Machine Learning Research, 2016, 17(1): 1-50

[17] Levy K Y, Shimkin N. Unified inter and intra options learning using policy gradient methods//Proceedings of the European Workshop on Reinforcement Learning, Athens, Greece, 2011: 153-164

[18] Bacon P-L, Harb J, Precup D. The option-critic architecture // Proceedings of the AAAI Conference on Artificial Intelligence, California, USA, 2017: 1726-1734

[19] Zhang S, Whiteson S. Dac: The double actor-critic architecture for learning options//Proceedings of the Advances in Neural Information Processing Systems, 2019: 2010-2020

[20] Kumar N M. Empowerment-driven exploration using mutual information estimation. arXiv preprint arXiv:1810.05533, 2018

[21] Eysenbach B, Gupta A, Ibarz J, et al. Diversity is all you need: Learning skills without a reward function//Proceedings of the International Conference on Learning Representations, New Orleans, USA, 2019: 1-22

[22] Baumli K, Warde-Farley D, Hansen S, et al. Relative variational intrinsic control//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 6732-6740

[23] Lin Z, Li J, Shi J, et al. Juewu-mc: Playing minecraft with sample-efficient hierarchical reinforcement learning//Proceedings of the International Joint Conference on Artificial Intelligence, Vienna, Austria, 2022: 3257-3263

[24] Watkins C J, Dayan P. Q-learning. Machine learning, 1992, 8 (3): 279-292

[25] Fortunato M, Azar M G, Piot B, et al. Noisy networks for exploration//Proceedings of the International Conference on Learning Representations, Vancouver, Canada, 2018: 1-21

[26] Pathak D, Agrawal P, Efros A A, et al. Curiosity-driven exploration by self-supervised prediction//Proceedings of the International Conference on Machine Learning, Sydney, Australia, 2017: 2778-2787

[27] Houthooft R, Chen X, Duan Y, et al. Vime: Variational information maximizing exploration//Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 2016: 1109-1117

[28] Schulman J, Moritz P, Levine S, et al. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015

[29] Chen X, Duan Y, Houthooft R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 2016: 2172-2180

[30] Gomes R, Krause A, Perona P. Discriminative clustering by regularized information maximization. //Proceedings of the Advances in Neural Information Processing Systems, Vancouver, Canada, 2010: 775-783

[31] Ziebart B D. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. USA: Carnegie Mellon University, 2010

[32] Zhu A, Chen F, Xu H, et al. Empowering the diversity and individuality of option: Residual soft option critic framework. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(12): 1-10

[33] Li S, Wang R, Tang M, et al. Hierarchical reinforcement

learning with advantage-based auxiliary rewards//Proceedings of the Advances in Neural Information Processing Systems, Vancouver, Canada, 2019: 1407-1417

[34] Wang R, Yu R, An B, et al. I^{2}hrl: Interactive influence-based hierarchical reinforcement learning//Proceedings of the International Joint Conferences on Artificial Intelligence, Yokohama, Japan, 2020: 3131-3138

[35] Osa T, Tangkaratt V, Sugiyama M. Hierarchical reinforcement learning via advantage-weighted information maximization // Proceedings of the International Conference on Learning Representations, New Orleans, USA, 2019: 1-16

[36] Hou Z, Zhang K, Wan Y, et al. Off-policy maximum entropy reinforcement learning: Soft actor-critic with advantage weighted mixture policy (sac-awmp). arXiv preprint arXiv: 2002.02829, 2020

[37] Zhang J, Yu H, Xu W. Hierarchical reinforcement learning by discovering intrinsic options//Proceedings of the International Conference on Learning Representations, Vienna, Austria, 2021: 1-18

[38] Gehring J, Synnaeve G, Krause A, et al. Hierarchical skills for efficient exploration//Proceedings of the Advances in Neural Information Processing Systems, Los Angeles, USA, 2021: 11553-11564

[39] Rao D, Sadeghi F, Hasenclever L, et al. Learning transferable motor skills with hierarchical latent mixture policies//Proceedings of the International Conference on Learning Representations, 2022: 1-22

[40] Brockman G, Cheung V, Pettersson L, et al. Openai gym. arXiv preprint arXiv:1606.01540, 2016

[41] Todorov E, Erez T, Tassa Y. Mujoco: A physics engine for model-based control//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 2012: 5026-5033

[42] Van der Maaten, Laurens, Hinton G. Visualizing data using t-sne. Journal of Machine Learning Research, 2008, 9(11): 2579-2605

WU Lan, Ph. D. candidate. Her main research interests include hierarchical reinforcement learning and offline reinforcement learning.

LIU Quan, Ph. D., professor, Ph. D. supervisor. His research interests include deep reinforcement learning and automated reasoning.

HUANG Zhi-Gang, Ph. D. candidate. His research interests include deep reinforcement learning and hierarchical reinforcement learning.

ZHU Fei, Ph. D., associate professor. His research interests include reinforcement learning and text mining.

ZHANG Li-Hua, Ph. D. candidate. His research interests include deep reinforcement learning and inverse reinforcement learning.

Background

Deep reinforcement learning faces many difficulties when dealing with complex environments, such as high-dimensional state spaces, difficulty in convergence, and training instability. To overcome these problems, academics are exploring novel research methods. Among them, hierarchical reinforcement learning has emerged as a promising direction. Option-based hierarchical reinforcement learning is a prominent subfield within this area of research. Hierarchical reinforcement learning offers a framework for learning and executing a variety of behaviors by decomposing tasks into options, enabling more efficient exploration in complex environments.

In order to address the issue of balancing exploration and exploitation on the continuous large-scale state-action space, this paper proposes the Maximum Entropy Hierarchical Reinforcement Learning with Advantage-weighted Mutual Information Maximization (HRL-AMIM) algorithm. The algorithm takes an innovative approach to solving the policy-induced sample clustering problem and adds internal rewards to enhance the diversity of the Option. By using weighted importance sampling of the advantage function, and mutual information maximization, the algorithm is able to better explore the environment and improve the stability of training. At the same time, by introducing a reward mechanism to the maximum entropy reinforcement learning objective, the policy becomes more exploratory, allowing the algorithm to better adapt to various environments. In addition, Option number annealing reduces prior knowledge, effectively balances

exploration and exploitation, and improves sampling efficiency and learning speed. Hierarchical reinforcement learning uses time abstraction techniques to quickly capture both external and internal rewards, effectively overcoming the intractable problems of deep reinforcement learning in large-scale spaces. Therefore, hierarchical reinforcement learning has a role in artificial intelligence that cannot be ignored.

This paper is supported by the National Natural Science Foundation of China (62376179,61772355,61702055,61876217, 62176175), the National Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01A238), and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). These research projects aim to drive the further development of reinforcement learning theory and design efficient algorithms to enhance the power and applicability of reinforcement learning in a number of fields. Through these efforts, we can gain a deeper understanding of the core principles and methods of reinforcement learning and contribute to continuous progress and innovation in the field of artificial intelligence.