

基于群体熵的机器人群体智能汇聚度量

冯埔¹, 吴文峻², 罗杰¹, 于鑫¹, 田雍恺¹

(1. 北京航空航天大学计算机学院, 北京 100191;

2. 北京航空航天大学人工智能研究院, 北京 100191)

摘要: 群体行为往往能产生远超个体行为的价值和复杂度。为了在个体智能的基础上更有效地衍生出群体智能, 需要基于群体熵来科学地衡量群体智能水平, 并以群体熵为引导目标, 推动群体智能的增强和演进。针对这个重要的科学问题, 以无人小车群体为研究对象, 提出基于参数共享和群体策略熵的多智能体 soft Q learning 算法, 通过共享智能体的观测信息, 并结合最大熵强化学习方法, 实现探索型任务中群体策略的持续学习更新。同时, 通过将群体熵定义为度量工具, 刻画群体学习中熵变化模式, 实现对群智汇聚过程的定量分析。

关键词: 群体熵; 群体智能; 深度强化学习

中图分类号: TP181

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202213

Emergence measurement of robot swarm intelligence based on swarm entropy

FENG Pu¹, WU Wenjun², LUO Jie¹, YU Xin¹, TIAN Yongkai¹

1. School of Computer Science and Engineering, Beihang University, Beijing 100191, China

2. Institute of Artificial Intelligence, Beihang University, Beijing 100191, China

Abstract: Swarm behavior can often produce value and complexity far beyond individual behavior. In order to more effectively derive swarm intelligence on the basis of individual intelligence, it is necessary to scientifically measure the level of swarm intelligence based on swarm entropy, and use swarm entropy as the guiding goal to promote the enhancement and evolution of swarm intelligence. Aiming at this important scientific problem, the unmanned car group as the research object was taken and a multi-agent soft Q learning method based on parameter sharing and group strategy entropy was proposed. Which by sharing the observation information of the agent, combined with the maximum entropy reinforcement learning method, to achieve continuous learning and updating of swarm strategies in exploratory tasks. At the same time, by defining swarm entropy as a measurement tool, characterizing the entropy change pattern in swarm learning, realizing the quantitative analysis of the gathering process of swarm intelligence.

Key words: swarm entropy, swarm intelligence, deep reinforcement learning

0 引言

个体间的交互等行为让群体现象变得更加复杂, 个体合作让群体发展出高于个体智能的更复杂的智能模式。人工智能的发展推动了人类社会的快

速发展, 在单体智能的基础上, 群体智能^[1]模式应运而生。例如无人机(车)集群协同对抗、群体软件开发等复杂群体行为, 都可以被视为群体智能的体现。不同时空节点上的个体通过通信、观测等方式, 以竞争、合作等自主协同模式共同完成复杂任

收稿日期: 2021-12-14; 修回日期: 2022-01-19

通信作者: 吴文峻, wwj09315@buaa.edu.cn

基金项目: 科技创新 2030—“新一代人工智能”重大项目(No.2018AAA0102300)

Foundation Item: Science and Technology Innovation 2030 — “New Generation Artificial Intelligence” Major Project (No.2018AAA0102300)

务,涌现出超越个体智能的群体智能。

对群体智能机理的研究主要从生物仿生和控制算法研究展开,一方面对生物体的群体行为进行仿生研究,主要包括模拟鸟群行为的 Boids 模型^[2]、自驱动粒子模型 (Vicsek 模型^[3])、进化算法^[4]、蚁群优化算法^[5]、粒子群优化算法等;另一方面不断对控制算法进行提升,主要包括机器人集群控制算法以及强化学习算法,如 Q-learning^[6]、PPO^[7]、多智能体强化学习^[8]等。

从上述分类可以看到,目前对群体智能的研究主要以模拟自然界和人类群体的群智行为来解决复杂问题,缺乏从机理层面对群智涌现汇聚过程的度量。基于群体中个体直接或间接交互信息的特性,也有很多研究者将熵引入群体智能的研究范畴中。

熵是热力学中计算描绘系统混乱程度的物理参数,熵增大表示系统无序程度升高,熵减小则表示系统无序程度下降,熵是一个可测的物理量,通过熵的变化可以观测整体系统无序程度的变化。Shannon C E^[9]将熵引入信息论,提出了信息熵的概念,利用信息熵来描述信息量的多少。李习彬^[10]从信息量的角度分析了系统的整体有序组织化程度和熵之间的关系。

熵在复杂系统和群智系统中也有广泛的应用。在系统熵的计算上, Kauffman S A^[11]提出了布尔网络模型,利用动力学方法对系统的演化和内部自组织现象进行刻画。在群体熵度量方面, Corneel C^[12]利用熵研究了群体运动行为中的涌现行为。Franceschini V 等人^[13]构建了复杂动力学系统熵的多项式求根方法。罗杰等人^[14]提出群体智能的动力学模型和群体熵度量,使用动力学模型刻画群智的激发汇聚过程。这些研究从复杂系统动力学的角度为群智的度量提供了丰富的理论基础。

相比于传统的控制与生物仿生方法,强化学习因其自身不断与环境交互更新策略的特点,而具备更强的适应性和动态性,被认为是通往强人工智能的重要途径。近年来,强化学习算法和逆强化学习^[15-16]不断迭代更新,包括以 DQN (deep Q learning)、PPO (proximal policy optimization)、DDPG (deep deterministic policy gradient) 为代表的单体强化学习算法与以 MADDPG^[17]、VDN^[18]、QMIX^[19]等为代表的多智能体强化学习算法。在群体任务中,强化学习的训练过程即群智汇聚的过

程,训练以最大化整体奖励为目标。也正因为强化学习以最大化奖励为学习目标,奖励函数设计尤为重要。引入群体熵度量,可以对强化学习的训练过程进行多维度的分析,打开深度网络训练的黑箱,同时也可以进一步度量群体智能程度。为了进一步提升强化学习的探索能力,有研究把熵作为正则项加入强化学习的奖励函数中,通过对不确定性策略的奖励让智能体在训练时对未知状态进行更大程度的探索。本文通过基于最大熵原理的 soft Q learning 算法^[20]完成雷区规避,并进行探索任务,在此过程中使用群体熵对训练过程进行度量,从群体熵的角度对群智汇聚过程进行度量。

本文第1节主要介绍相关背景,包括强化学习框架、任务介绍与最大熵原理。第2节主要介绍本文采用的基于参数共享的多智能体 soft Q learning 算法。第3节主要验证群体熵在群智汇聚过程中的度量作用。第4节给出本文的结论和未来可能的研究方向。

1 背景介绍

随着深度强化学习的快速发展,强化学习模型在很多游戏场景中已经取得超越人类玩家的成绩。强化学习因其可适应环境动态变化的特点被认为是通往强人工智能的重要途径。本节主要介绍深度强化学习的整体框架与相关方法以及群体熵计算过程中采用的最大熵原理。

1.1 马尔可夫决策

强化学习以马尔可夫决策过程为基础框架,智能体通过感知和执行动作与环境进行交互,基于当前观测和输入产生自身的动作行为 A , 以动作与环境交互实现状态转移,获得环境反馈的奖励。笔者以马尔可夫决策过程为整体决策框架。对于 N 个智能体,定义状态的集合 S , 将各个智能体的动作定义为 $A_1 \sim A_N$, 将各个智能体的观测定义为 $O_1 \sim O_N$ 。智能体根据自身的策略 π 选择当前状态下的动作,将基于交互的环境状态转移概率定义为 TP: $S \times A_1 \times \dots \times A_N$ 。每个智能体从环境中获得自身的奖励 R_i , 每个智能体以最大化自身奖励为优化目标,奖励如下:

$$R_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k} \quad (1)$$

其中, $\gamma \in [0,1]$ 表示折扣系数。Q 值为 S, A 对奖励

的映射, 表示为 $Q(s, a)$, 即某个状态下采取某个动作的奖励。策略 π 下最优的状态动作值函数为:

$$Q^*(s_t, a_t) = \max_{\pi} E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots] \quad (2)$$

根据贝尔曼方程可知:

$$Q^*(s_t, a_t) = E[r_t + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})] \quad (3)$$

强化学习整体上可以分为两大类: 基于价值函数的强化学习方法、基于策略梯度的强化学习方法, 分别以深度 Q 学习和策略梯度方法为代表。

基于价值函数的强化学习方法用深度神经网络来逼近价值函数, Mnih V 等人^[21]提出的 DQN 是深度强化学习领域开创性的工作, 在 Q 学习的基础上加入卷积神经网络 (convolutional neural network, CNN), 并利用经验回放机制和目标网络进行训练。在每个时间点存储智能体的经验信息 $e_t = (s_t, a_t, r_t, s_{t+1})$, 形成回放记忆序列 $D = \{e_1, \dots, e_N\}$, 每次从 D 中随机提取小批量的经验样本, 使用梯度下降方法更新网络参数。经验回放机制提升了数据使用效率, 降低了训练数据之间的关联性。在训练过程中深度 Q 学习将值函数参数化地表示为 $Q(s, a; \theta_i)$, Q 网络的优化目标为 $r + \gamma \max_{a'} Q(s', a'; \theta_i^-)$, 这个优化目标由另一个目标网络生成, 其中的 a' 为所有可能的动作选择, s' 为下一个时刻的状态, θ_i^- 为神经网络的参数。目标网络与估计网络都采用深度卷积网络, 结构相同, 目标网络的参数由估计网络每次通过 K 个迭代进行更新, 在此过程中目标网络参数保持不变。在学习过程中, 通过最小化以下损失函数进行更新:

$$L_i(\theta_i) = E_{(s, a, r, s')} [(y - Q(s, a; \theta_i))^2] \quad (4)$$

基于策略梯度的强化学习方法以深度神经网络计算策略梯度, 将 π 表示为 π_{θ} , 计算出动作的策略梯度, 并沿着策略梯度的方向不断调整动作, 逐渐得到最优策略。常见的策略梯度可以分为随机性策略和确定性策略。随机性策略 $\pi_{\theta}(a|s) = P[a|s; \theta]$, 表示当状态为 s 时, 动作 a 的选择由以 θ 为参数的概率分布决定, 因此即使在同一状态下也会有不同的动作选择。确定性策略则表示在某一个状态下只对唯一的动作。比较具有代表性的策略梯度方法有 PPO、DDPG, 分别对应随机性策略方法和确定性策略方法。

对于随机性策略, 梯度计算如下:

$$\begin{aligned} \nabla_{\theta} L(\pi_{\theta}) &= \int_{\mathcal{S}} p^{\pi}(s) \int_{\mathcal{A}} \nabla_{\theta} \pi_{\theta}(a|s) Q^{\pi}(s, a) da ds \\ &= E_{s \sim p^{\pi}, a \sim \pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(a|s) Q^{\pi}(s, a)] \end{aligned} \quad (5)$$

对于确定性策略, 梯度计算如下:

$$\begin{aligned} \nabla_{\theta} L(\mu_{\theta}) &= \int_{\mathcal{S}} \rho^{\mu}(s) \nabla_{\theta} \mu_{\theta}(s) \nabla_a Q^{\mu}(s, a) \Big|_{a=\mu_{\theta}(s)} ds \\ &= E_{s \sim \rho^{\mu}} [\nabla_{\theta} \mu_{\theta}(s) \nabla_a Q^{\mu}(s, a) \Big|_{a=\mu_{\theta}(s)}] \end{aligned} \quad (6)$$

1.2 最大熵强化学习

最大熵原理的基本思想是在满足某些约束的事件概率分布时, 信源的熵最大可以使研究人员依靠有限数据达到尽可能客观的效果。Ripley B D 等人^[22]提出了熵集中原理, 在满足给定约束的条件下, 概率分布绝大多数集中在使得熵最大的区域。熵较大的分布具有较高的多样性, 因此实现的方法也更多, 也越有可能被观察到。

对于一般的强化学习来说, 学习目标为最大化累计奖励, 即:

$$\pi^* = \arg \max_{\pi} E_{(s_t, a_t) \sim \rho^{\pi}} \left[\sum_t R(s_t, a_t) \right] \quad (7)$$

而在最大熵强化学习中, 除了从外部环境获得奖励, 还将每一次动作选择的熵也加入优化目标。即最优策略为:

$$\pi^* = \arg \max_{\pi} E_{(s_t, a_t) \sim \rho^{\pi}} \left[\sum_t R(s_t, a_t) + \alpha H(\pi(\cdot|s_t)) \right] \quad (8)$$

相比于基础的强化学习方法, 最大熵强化学习有以下几点优势。

① 学习得到的策略可以作为更复杂任务的初始化策略网络, 通过最大熵, 策略网络可以学习到更具一般性的特征。

② 具有更强的探索能力, 相比于贪婪方法或者 ϵ 贪婪方法, 最大熵强化学习让每一个动作都有概率被选择, 提升了训练过程中的探索效率。

③ 更具鲁棒性, 面对动态变化的环境更容易做出相应调整。

2 基于参数共享的多智能体 soft Q learning

2.1 任务介绍

本文设置任务场景为多智能体穿越地雷区域到达目标位置, 智能体数量为 $N=4$, 地图大小为 10×10 方格, 实验仿真环境选择使用 Webots, 智能体使用 Webots 中的 E-puck 智能车。智能车的状态

用 S 表示, 观测信息用 O 表示, 动作 A 包括上、下、左、右 4 种, 环境的状态转移概率为 T 。智能体的任务目标为绕开地雷区域到达目标区域, 任务在智能体踩到地雷或者达到目标点时结束。奖励函数 R 设计为到达目标区域为正向奖励, 被炸毁为负向奖励。任务地图如图 1 所示。其中爆炸标识表示地雷区域, 红旗表示目标区域, 黑色小车为智能体 A~D。

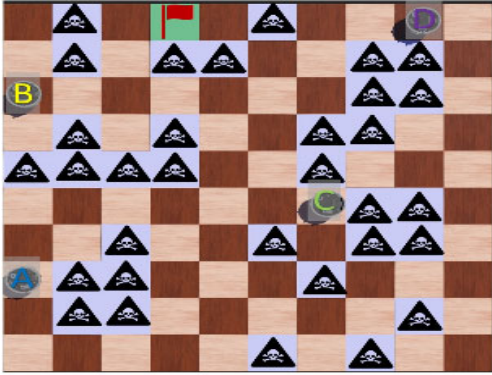


图1 任务地图

2.2 算法介绍

本文提出基于参数共享的多智能体 soft Q learning 算法, 主要包括两部分算法: soft Q learning 与参数共享。

传统的强化学习在面对多模 Q 值时往往只能收敛到其中的一个峰, 为了能让策略选择更加符合真实 Q 值的分布, 同时为了提升智能体在状态空间中的探索效率, 在深度 Q 学习的基础上引入最大熵原理, 形成 soft Q learning 算法^[20]。

在值函数中加入熵, 将其作为正则项, 使得在训练过程中尽可能地扩大探索范围, 熵越大, 奖励越高。由此可以得到价值函数 V_{soft} :

$$V_{\text{soft}}(s_t) = E_{a_t \sim \pi} [Q_{\text{soft}}(s_t, a_t) - \alpha \log \pi(a_t | s_t)] \quad (9)$$

其中, 策略函数 $\pi(a_t | s_t) \propto \exp(Q_{\text{soft}}(s_t, a_t))$, 策略动作选择由 Q 值生成的概率分布决定。

进一步展开:

$$\pi(s_t, a_t) = \exp\left(\frac{1}{\alpha} (Q_{\text{soft}}(s_t, a_t) - V_{\text{soft}}(s_t))\right) = \frac{\exp\left(\frac{1}{\alpha} Q_{\text{soft}}(s_t, a_t)\right)}{\exp\left(\frac{1}{\alpha} V_{\text{soft}}(s_t)\right)} \quad (10)$$

状态 s 下累计奖励为:

$$V_{\text{soft}}(s_t) = E_{a_t \sim \pi} [Q_{\text{soft}}(s_t, a_t) - \alpha \log \pi(a_t | s_t)] \quad (11)$$

$$Q_{\text{soft}}(s_t, a_t) = E[r_t + \gamma \text{soft max}_a Q(s_{t+1}, a)] \quad (12)$$

优化目标为最小化 Q 值估计误差:

$$J_Q(\theta) = E\left[\frac{1}{2} (\hat{Q}_{\text{soft}}(s_t, a_t) - Q_{\text{soft}}^\theta(s_t, a_t))^2\right] \quad (13)$$

$\hat{Q}_{\text{soft}}(s_t, a_t) = r_t + \gamma E_{s_{t+1} \sim p_s} [V_{\text{soft}}^\theta(s_{t+1})]$ 为目标 Q 网络的 Q 值, 根据 Q 值估计网络每隔一定步数进行参数更新。

利用 soft Q learning, 将 DQN 中的确定性策略变为随机性策略, 在优化目标中加入熵, 将其作为正则项, 扩大策略的策略空间。

多智能体之间的信息交互可以用参数共享的方式来表示, 智能体的观测用 O_t 表示, 传统多智能体任务中的强化学习方法多以单体模型或集中式训练分布式执行 (centralized training decentralized execution, CTDE) 模型为架构, 本文通过将智能体在与环境交互过程中的信息进行共享, 将个体探索的经验存入共同的经验池中 $D = \{e^1, e^2, e^3, \dots, e^N\}$, 其中 N 表示智能体数量, 训练时从经验池中进行采样。相比于 DQN 的经验池独立采样 $D = \{e_{t1}, e_{t2}, e_{t3}, \dots, e_{tk}\}$, 本文共享参数算法可以充分对智能体的观测信息进行整合, 并进行强化学习训练。智能体之间使用同一个策略网络参数, 完整的基于参数共享的多智能体 soft Q learning 算法如下, 多个智能体同时与环境进行交互, 共享观测信息。

算法 1 基于参数共享的多智能体 soft Q learning 算法

步骤 1 初始化网络分布参数: θ, ϕ

步骤 2 初始化目标网络参数: $\bar{\theta}, \bar{\phi}$

步骤 3 初始化经验池: D

步骤 4 对每轮训练循环, for i in range (episode_num)

步骤 5 对每个训练步循环

步骤 6 收集经验数据, 基于策略函数对 4 个智能体动作进行采样: $a_t \leftarrow f^\theta(\xi; s_t)$, 其中 $\xi \sim N(0, I)$

步骤 7 采样状态转移数据 $s_{t+1} \sim p_s(s_{t+1} | s_t, a_t)$, 并存入经验池 $D \leftarrow D \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$

步骤 8 当经验池中样本数量大于阈值时, 从经验池中随机选择小批量 (minibatch) 数据: $\{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)})\}_{i=0}^N \sim D$

步骤 9 输入 Q 值估计网络, 计算 Q 值

步骤 10 计算 soft Q 值

步骤 11 输入目标 Q 值网络, 计算 Q 值

步骤 12 计算目标网络的 soft Q 值

步骤 13 计算损失函数梯度 $\nabla J_Q(\theta)$

步骤 14 使用 Adam 算法优化更新 Q 值估计网络 θ

步骤 15 每隔 T 步用 Q 值估计网络 θ 更新 Q 值目标网络 $\bar{\theta}$

2.3 结果分析

在实验中, 本文设置了 4 个智能体共同进行探索, 地图大小为 10×10 方格, 地图中密集分布地雷区域, 当任何智能体进入地雷区域时, 视作被炸毁, 任务失败。任务开始时, 4 个智能体分布在图 1 所示的位置, 目标区域为红旗区域。奖励函数设计如下。

- ① 到达终点: 正向奖励+5。
- ② 炸毁: 负向奖励-0.5。
- ③ 遭遇边界: 负向奖励-0.025。
- ④ 未回到上一步位置: 正向奖励+0.1。
- ⑤ 回到上一步位置: 负向奖励-0.1。

Q 值估计网络和 Q 值目标网络在网络结构设计上一致, 由 2 层全连接层构成。多智能体探索任务流程如图 2 所示。首先初始化算法参数, 在每一个训练回合开始时, 复位 4 个智能体到初始位置, 并将网络参数初始化。设定目标区域, 完成地图和智能体的复位。对于每一次训练中的每一步, 由深度网络为 4 个智能体根据当前状态输出动作指令。由

智能体完成指令要求的动作, 在执行动作的过程中, 对状态转移进行判断, 首先判断是否达到边界, 如果超出边界, 则将智能体移动回该动作执行之前的位置, 并将奖励设置为-0.025, 对每一个智能体都执行上述判断。若智能体没有超出边界, 进行下一步是否遇到雷区的判断, 若遇到雷区, 则判定任务失败, 结束该回合。若未遇到雷区, 下一步判断是否到达目标区域, 若未到达, 则继续执行下一状态的动作; 若智能体均达到目标区域, 则判定任务完成, 回合结束, 将奖励设置为+5。如此循环, 当训练回合数等于预设的最大回合数时, 结束训练。

按照上述设计, 分别采用常用的独立 Q 学习 (independent Q learning, IQL) 算法和本文提出的基于参数共享的多智能体 soft Q learning 算法进行实验和分析。IQL 算法在多智能体任务场景中表现出色, 将该算法作为本文算法的对照实验。IQL 算法训练奖励如图 3 所示 (图 3、图 4、图 6、图 7 中蓝线为每 20 回合的平均奖励/熵, 黄线为每回合的奖励/熵), 训练过程中平均奖励从-40 逐渐收敛到 20 左右, 收敛所用回合数在 400 左右。基于参数共享的多智能体 soft Q learning 算法训练奖励如图 4 所示, 训练收敛速度更快, 在 200 回合左右即可使平均奖励收敛到 20 左右, 同时因为最大熵扩大了探索的因素, 使得在训练 400 回合之后, 平均奖励会大于 20, 即在训练较多轮次之后获得比 IQL 算法训练结果更优的策略。

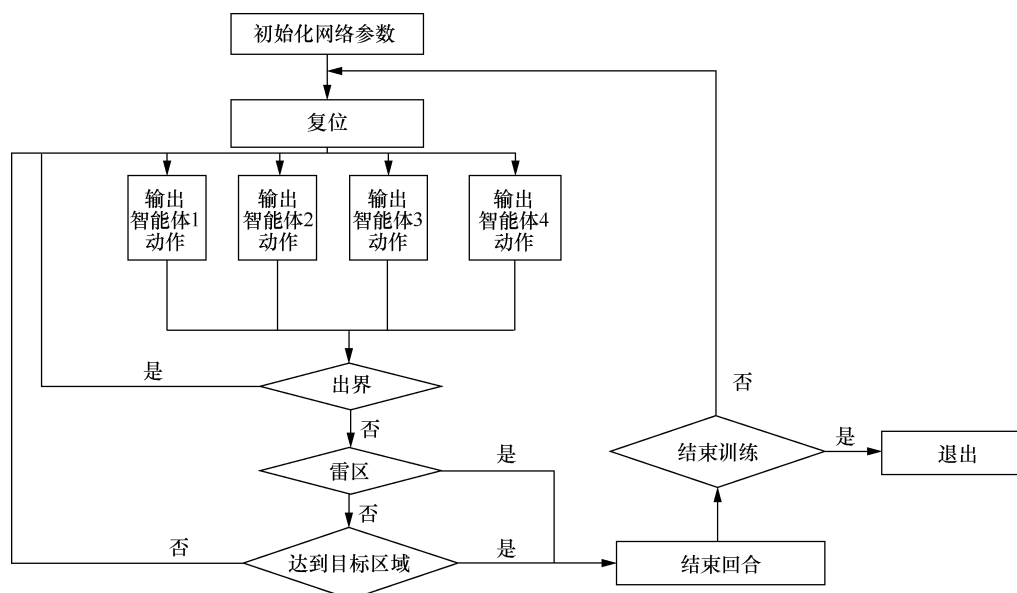


图 2 多智能体探索任务流程

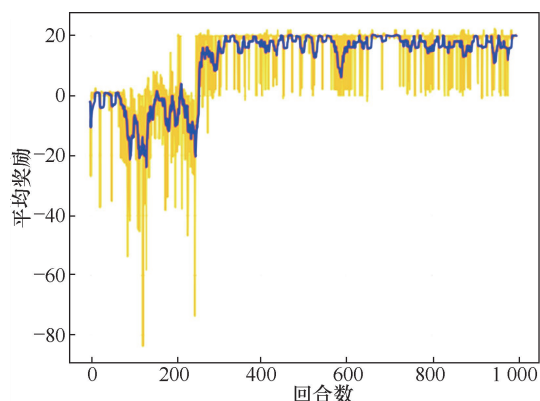


图3 IQL 算法训练奖励

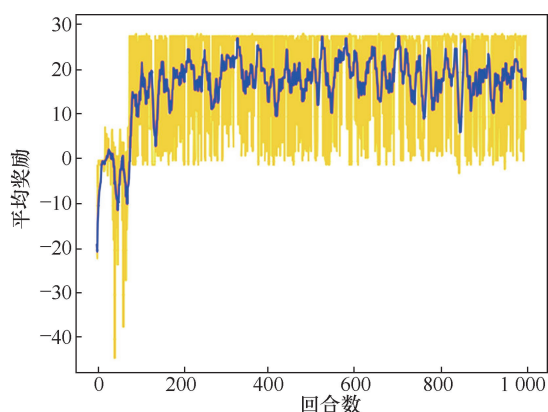


图4 基于参数共享的多智能体 soft Q learning 算法训练奖励

在探索任务场景中,本文提出的基于参数共享的多智能体 soft Q learning 算法基于参数共享的方法在智能体之间共享探索经验,使得探索效率更高,在采样同等数量样本情况下进行了更有效的策略网络更新,相比于原方法的 400 回合,本文方法收敛速度更快。且因为将熵作为正则项,使得对最优策略的探索更彻底,减少了在稀疏奖励情况下的探索过程中收敛到次优情况的可能性,最终收敛奖励大于 20,优于原方法。

群体任务中通过将熵作为正则项的方式,在原有 ϵ 探索的基础上,把 DQN 中的 Q 值转化为 soft Q 值,强化学习策略网络学习得到各个动作的选择概率,相比于原方法中的选择 Q 值最大的动作,soft Q learning 算法使得各个动作均有概率被选择,尤其在长周期稀疏奖励情况下,当部分采样并不能很准确地优化 Q 值时,通过将 Q 值转化为概率分布来扩大该情况下的探索范围,提升采样效率,减小训练收敛到次优的可能性。将 Q 值转化为选择概率,通过计算 soft Q 值的变化可以进一步度量训练策略的变化过程,这部分内容将在第 3 节中介绍。

在群体任务中,每个智能体利用参数共享的方

式进行训练,这种训练模式可以被视为自然个体间的信息共享。群体中个体之间通过简单的交互或者观测可以形成复杂的群体现象,在本文中,可以将群体智能的汇聚视作智能体之间通过参数共享实现了群体中的交互行为,并利用此类交互行为完成协作探索的群体任务。群体协作在探索效率上优于个体独立探索的效率。在 IQL 中最终收敛需要 400 回合左右,而通过参数共享的方式只需要 200 回合左右。在个体策略熵的基础上,本文分析了个体之间的交互对策略训练的影响,提出了群体熵(互信息)的定义,这部分内容将第 3 节中介绍。

3 基于群体熵的群智汇聚过程分析

本文算法在奖励函数中加入熵,并将其作为正则项。以此扩大了训练过程中的探索,取得相比标准 Q 学习更好的训练效果。在训练中,以负向奖励的形式加入熵,熵越小,奖励越低;熵越大,奖励越高。每个个体在每一步中的熵的计算式如下:

$$h_i(s) = -\sum P_i(s, a_i) \ln P_i(s, a_i) \quad (14)$$

其中, $P_i(s, a_i)$ 表示在状态 s 下选择动作 a 的概率分布, $h_i(s)$ 表示群体中个体 i 在状态 s 下的策略熵。在地图探索任务中,到达目标区域给予正向奖励,个体间共享观测信息,通过奖励设置可以看到群体策略的优化结果;但在训练过程中,并不能通过奖励函数来度量个体共享的观测信息的价值。本节将在 soft Q learning 实现探索任务的基础上,提出群体熵度量工具,在训练过程中对群体中个体间交互信息利用的有效程度进行度量。

3.1 群体熵工具介绍

个体在自身信息的基础上通过通信等方式与其他个体进行交互,形成了群体现象。在群体智能汇聚度量中需要考虑个体间因为交互导致的智能程度变化。群体熵由互信息 (mutual information, MI) 计算得到,本文将群体熵定义如下:

$$\text{Entropy}_{\text{swarm}} = \sum \alpha_{i,j} \text{MI}(\text{agent}_i, \text{agent}_j) \quad (15)$$

将互信息加入群体熵度量。互信息是衡量随机变量之间相互依赖程度的度量。群体熵可以描述群体中信息交互的价值。

对于随机变量 X 和随机变量 Y , 其自身的信息熵分别为 $H[X]$ 和 $H[Y]$:

$$H[X] = -\sum_{x \in X} p(x) \cdot \log_2 p(x) \quad (16)$$

$$H[Y] = -\sum_{y \in Y} p(y) \cdot \log_2 p(y) \quad (17)$$

X 和 Y 的联合熵为 $H[XY]$:

$$H[XY] = -\sum_{x \in X} \sum_{y \in Y} p(xy) \cdot \log_2 p(xy) \quad (18)$$

X 和 Y 的条件熵为 $H[X|Y]$:

$$H[X|Y] = -\sum_{x \in X} \sum_{y \in Y} p(xy) \cdot \log_2 p(x|y) \quad (19)$$

联合熵也可以表示如下:

$$H[XY] = H[X] + H[Y|X] = H[Y] + H[X|Y] \quad (20)$$

根据上述信息论中熵的计算, 推出个体间相互作用的互信息计算式, 具体如下:

$$MI[X;Y] = -\sum_{x \in X} \sum_{y \in Y} p(xy) \cdot \log_2 \frac{p(x) \cdot p(y)}{p(xy)} \quad (21)$$

通过个体的信息熵和联合熵可以计算互信息:

$$MI[X;Y] = H[X] + H[Y] - H[XY] \quad (22)$$

由此给出群体熵中互信息的定义:

$$\sum \alpha_{i,j} MI(\text{agent}_i, \text{agent}_j) = \sum (H(\text{agent}_i) + H(\text{agent}_j) - H(\text{agent}_i, \text{agent}_j)) = \sum (H(\text{agent}_i) - H(\text{agent}_i | \text{agent}_j)) \quad (23)$$

其中, 个体熵 $H(\text{agent}_i)$ 可以通过个体策略的不确定性获得:

$$H(\text{agent}_i) = \text{mean} \left(\sum_{j \in T} \sum_{i \in N} h_i(s) \right) \quad (24)$$

其中, T 表示每个训练回合中的步数, N 表示智能体总数, s 表示智能体状态。在本文的算法中, 个体间采用参数共享的方法进行经验共享。在训练过程中, 通过对群体熵的计算可以观测群体中信息交互对智能水平提升的影响。

当智能体到达地雷区域时, 收集到较大负向奖励经验, 智能体自身会避开对应区域, 该位置中相应的动作选择概率也会降低。而通过参数共享加入经验池, 其他智能体在进行动作选择时也会考虑引入此经验的影响, 从而避开导致进入地雷区域的动作。

由于任务奖励的稀疏性, 在未到达目标区域时不能获得正向奖励。故在训练过程中, 智能体采样

得到的经验并不一定是有效的, 采样返回的 Q 值并不一定能为策略优化提供有益的梯度。可以通过互信息的变化来度量个体之间共享参数的有效性。这正是群体智能汇聚过程的体现。

互信息较大表明群体中交互对群体决策的提升较大。而当群体中相互间的经验数据对提升策略帮助不大时, 互信息相应变小。

为了验证群体熵工具对群体智能汇聚的度量作用, 计算个体之间的互信息, 并将其作为探索过程中因个体间相互作用对群智汇聚过程产生的影响的度量。当个体间无相互影响时, 互信息为 0; 当两者间存在相互影响时, 互信息为正。

通过群体熵的定义, 对比训练回合之间熵值的变化可以对训练过程进行除奖励函数变化外的进一步描述。

3.2 群智汇聚度量

随机动作奖励如图 5 所示。图中黄线表示每回合的奖励, 蓝线表示每 30 回合的平均奖励。由于每个智能体只进行随机动作, 并不存在策略优化的过程, 随机动作导致的奖励也是随机的, 群体熵保持不变。在动作选择完全随机的情况下, 奖励函数收敛在 0 左右, 并没有对环境进行探索进而得到较优策略, 个体间的通信并未对群体行为决策提供帮助, 个体的行为纯随机, 故并没有智能的提升, 群体熵保持在 0 左右。

在采用独立 Q 学习时, 个体间完全没有交互, 个体的策略学习完全依靠自身的探索, 并没有考虑其他智能体经验的影响, 基于互信息计算的群体熵在此时保持为 0。训练过程中只有个体自身对环境学习带来的智能, 并没有出现群体智能, IQL 个体熵变化如图 6 所示, 随着训练进行, 个体学习策略不断优化, 个体熵趋于收敛。

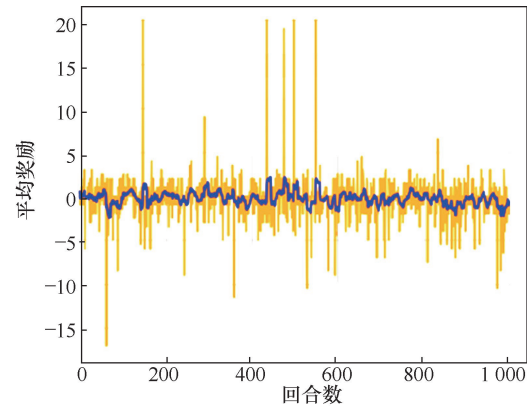


图 5 随机动作奖励

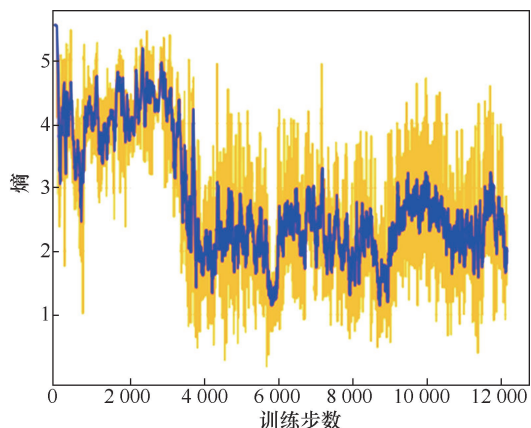


图6 IQL 个体熵变化

在多智能体强化学习算法训练中,一方面可以通过奖励函数来评估训练过程与效果,同时也可以利用本节介绍的群体熵来度量群体智能的汇聚。在本实验开始阶段,初始化所有网络参数,个体层面和群体层面的策略选择都是随机的,此时动作获得的奖励很低。根据 soft Q learning 计算得到在状态 s 下动作选择的概率比较平均,群体中各个个体的个体熵和在 5.54 左右,与随机选择动作策略时一致。

基于参数共享的多智能体 soft Q learning 个体熵变化如图 7 所示,奖励如图 4 所示。对比奖励函数的变化,随着训练回合数增多,个体熵从一开始的稳定在高点到快速下降,从一开始的盲目探索逐渐收敛到有方向性的运动。2 000 步(100 个回合)之后,个体熵 $\Sigma(\text{agent}_i)$ 快速下降,在 3 000 步(200 个回合)之后收敛,表明在积累了一定新状态探索经验之后,个体的行为更加收敛,参照奖励函数变化,也可以看到个体的奖励逐渐增大,智能程度增强。图 6 中 IQL 的个体熵在 2 500 步之后快速下降,在 4 500 步左右收敛。相比于 IQL,共享参数之后个体的策略学习更快收敛。

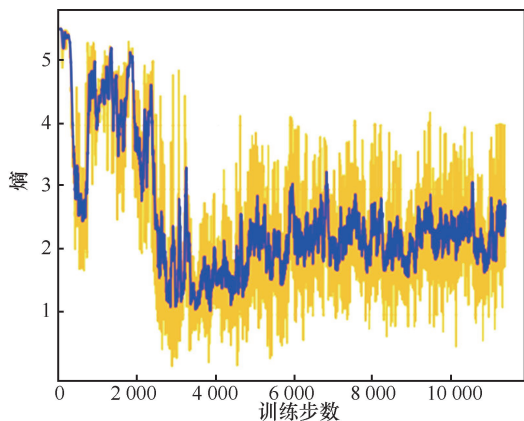


图7 基于参数共享的多智能体 soft Q learning 个体熵变化

在训练初期,因为个体的未探索区域较大,智能体之间信息收集更分散,训练经验差异较大,相互间的互信息也更大。随着训练的开展,个体一方面通过自身的探索,另一方面通过个体间的参数共享,各个位置上不同动作神经网络估计的 soft Q 值进一步接近真实值,动作选择更加明确,个体熵也相应下降,在 200 回合左右收敛到 1~2。

个体可以通过个体熵衡量群体中个体智能水平随着训练过程的变化,并不能真实地刻画群体内个体间的相互作用。本文提出的群体熵可以通过对比智能体间相互作用对个体熵的影响来描述群体因交互导致的智能水平变化。

基于参数共享的多智能体 soft Q learning 群体熵变化如图 8 所示。训练初期群体熵的值较大,群体探索程度低,智能体之间参数共享可以很大程度拓展其中个体的探索,由 soft Q 值产生的个体熵在此时变化较大,这部分的变化主要由共享的其他智能体的参数产生。智能体与其他个体的交互导致了自身熵的变化,其他智能体的经验数据让智能体自身的智能水平产生了变化,群体熵可以很好地刻画交互导致的群体智能提升。

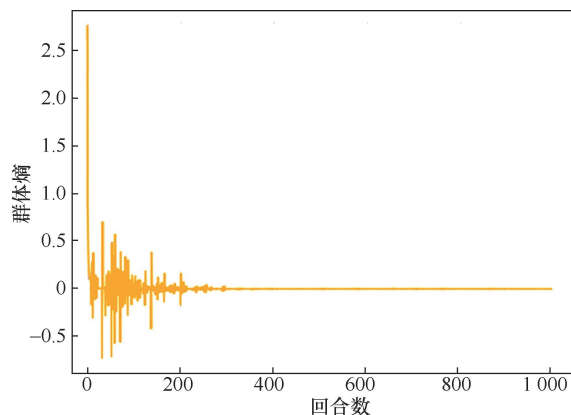


图8 基于参数共享的多智能体 soft Q learning 群体熵变化

随着训练的进行,群体熵下降,在 200 回合之前,群体熵保持在 -0.5 至 0.5 之间波动,此时处于训练效果快速提升阶段,对应图 4 中奖励为 0~20 的阶段。在此阶段,因为信息共享,经验快速积累,给予了决策网络大量的有效样本,群体的策略处于优化过程中,群体状态处于群智汇聚的中间态。

在训练后期,奖励函数逐渐收敛,在 200 回合之后收敛在 20 左右。相对应的群体熵也收敛到 0 左右。群体策略的不确定程度基本为 0,此时群体的策略稳定,已经探索得到一个较优的策略,达到

群智汇聚的完成态。

对于群体而言，整体奖励更大，任务完成度更高。当群体熵值不断减小时，个体间的交互有效性降低，群体整体的系统向低能量状态的有序结构演化。在群智汇聚的过程中，个体间通过信息共享提升了智能水平，在利用个体熵衡量个体智能的基础上，群体熵对交互的影响进行了量化，从相互作用的角度对群智汇聚进行度量。

基于参数共享（每隔3步）的多智能体 soft Q learning 个体熵变化如图9所示，可以看到在初始阶段群体熵较大，在3次独立探索后共享参数对 soft Q 值的更新较大。在收敛速度上，相比于每步共享参数方法，基于参数共享（每隔3步）的多智能体 soft Q learning 算法收敛更慢，群体熵在300~400回合之后趋于0。

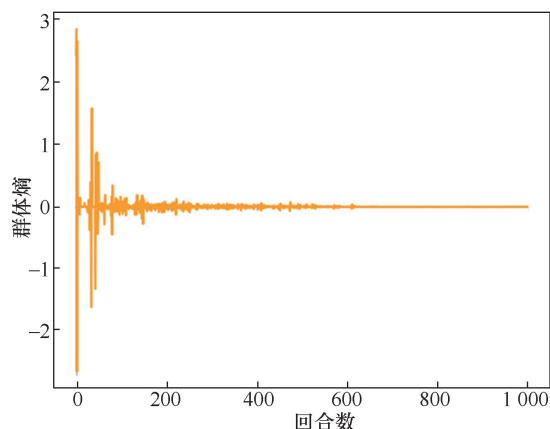


图9 基于参数共享（每隔3步）的多智能体 soft Q learning 个体熵变化

基于参数共享的多智能体 soft Q learning 在第300回合时的智能体轨迹如图10所示，可以看到4个智能体在训练之后已经学习得到了较优的策略。

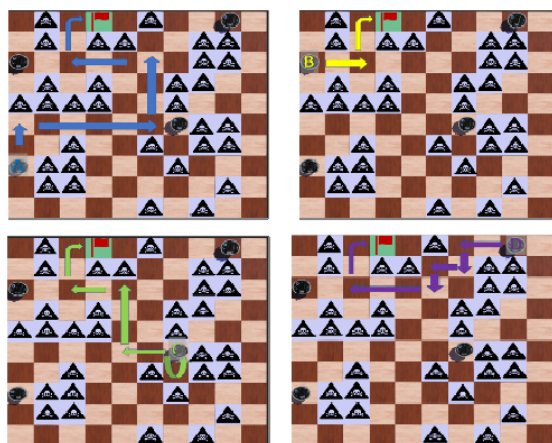


图10 基于参数共享的多智能体 soft Q learning 在第300回合时的智能体轨迹

4 结束语

本文首先利用基于深度强化学习方法实现了在带障碍物地图中的多智能体协作搜索任务，利用 Webots 仿真对提出的基于参数共享的多智能体 soft Q learning 算法进行了验证。并提出了群体熵的计算式，将群体熵作为多智能体强化学习训练效果的度量工具对搜索任务中的群智汇聚过程进行了分析，验证了群体熵可以作为对群体汇聚现象进行描述的有效工具。

本文基于同质性假设，将互信息作为群体熵计算方法，在考虑不同个体间互信息的权重时未进行差异化处理。未来将引入更复杂的任务场景，对抗任务与异构合作任务进行分析，加入个体差异性与结构特征，进一步验证将群体熵作为群智汇聚度量工具的有效性，同时在实际 E-puck 智能车场景中进行测试。

参考文献：

- [1] ZHU Y F, TANG X M. Overview of swarm intelligence[C]//Proceedings of 2010 International Conference on Computer Application and System Modeling (ICCSM 2010). Piscataway: IEEE Press, 2010.
- [2] REYNOLDS C W. Flocks, herds and schools: a distributed behavioral model[C]//Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM Press, 1987: 25-34.
- [3] VICSEK T, CZIRÓK A, BEN-JACOB E, et al. Novel type of phase transition in a system of self-driven particles[J]. Physical Review Letters, 1995, 75(6): 1226-1229.
- [4] DEB K, PRATAP A, AGARWAL S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182-197.
- [5] DORIGO M, BLUM C. Ant colony optimization theory: a survey[J]. Theoretical Computer Science, 2005, 344(2/3): 243-278.
- [6] KRÖSE B J A. Learning from delayed rewards[J]. Robotics and Autonomous Systems, 1995, 15(4): 233-235.
- [7] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv preprint, 2017, arXiv:1707.06347.
- [8] FOERSTER J N, ASSAEL Y M, DE FREITAS N, et al. Learning to communicate with deep multi-agent reinforcement learning[J]. Advances in Neural Information Processing Systems, 2016, 29.
- [9] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379-423.
- [10] 李习彬. 熵—信息理论与系统工程方法论的有效性分析[J]. 系统工程理论与实践, 1994, 14(2): 37-42.
- LI X B. Entropy-information theory and an analysis of the effectiveness of systems engineering's methodology[J]. Systems Engineering

Theory and Practice, 1994, 14(2): 37-42.

- [11] KAUFFMAN S A. The origins of order: self-organization and selection in evolution[M]. New York: Oxford University Press, 1993.
- [12] CORNEEL C. Emergent collective motion from local interactions[D]. Ghent: Ghent University, 2017.
- [13] FRANCESCHINI V, GIBERTI C, ZHENG Z M. Characterization of the Lorentz attractor by unstable periodic orbits[J]. Nonlinearity, 1993, 6(2): 251-258.
- [14] 罗杰, 姜鑫, 郭炳晖, 等. 群体智能系统的动力学模型与群体熵度量[J]. 中国科学: 信息科学, 2022, 52(1): 99-110.
LUO J, JIANG X, GUO B H, et al. Dynamic model and crowd entropy measurement of crowd intelligence system[J]. Scientia Sinica (Informationis), 2022, 52(1): 99-110.
- [15] 于鑫, 吴文峻, 罗杰, 等. 面向群体共识机制的逆强化学习辨识方法[J]. 中国科学: 技术科学, 2021.
YU X, WU W J, LUO J, et al. Inverse reinforcement learning identification method for group consensus mechanism [J]. Scientia Sinica (Technologica), 2021.
- [16] YU X, WU W J, FENG P, et al. Swarm inverse reinforcement learning for biological systems[C]//Proceedings of 2021 IEEE International Conference on Bioinformatics and Biomedicine. Piscataway: IEEE Press, 2021: 274-279.
- [17] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. arXiv preprint, 2017, arXiv:1706.02275.
- [18] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-decomposition networks for cooperative multi-agent learning[J]. arXiv preprint, 2017, arXiv:1706.05296.
- [19] RASHID T, SAMVELYAN M, DE WITT C S, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning[J]. arXiv preprint, 2018, arXiv:1803.11485.
- [20] HAARNOJA T, TANG H, ABBEEL P, et al. Reinforcement learning with deep energy-based policies[C]//Proceedings of the International Conference on Machine Learning. [S.l.:s.n.], 2017: 1352-1361.
- [21] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning [J]. Computer Science, 2013.
- [22] RIPLEY B D, ROSENKRANTZ R D E T. Jaynes: papers on probability, statistics and statistical physics[J]. The Statistician, 1990, 39(1): 90.

[作者简介]



冯埔(1995-), 男, 北京航空航天大学计算机学院博士生, 主要研究方向为多智能体强化学习。



吴文峻(1973-), 男, 博士, 北京航空航天大学人工智能研究院教授、博士生导师, 主要研究方向为群体智能、智能微服务、认知建模等。



罗杰(1981-), 男, 博士, 北京航空航天大学计算机学院副教授, 主要研究方向为知识图谱构、群体智能等。



于鑫(1994-), 男, 北京航空航天大学计算机学院博士生, 主要研究方向为群体智能、强化学习。



田雍恺(1997-), 男, 北京航空航天大学计算机学院硕士生, 主要研究方向为多智能体强化学习。