智能集群系统的强化学习方法综述

李璐璐[1),2),3)] 朱睿杰[1),2),3)] 隋璐瑶[1),2),3)] 李亚飞[1),2),3)]
徐明亮[1),2),3)] 樊会涛[1),2)]

[1)](郑州大学计算机与人工智能学院 郑州 450001)
[2)](智能集群系统教育部工程研究中心 郑州 450001)
[3)](国家超级计算郑州中心 郑州 450001)

摘 要 智能集群系统是人工智能的重要分支,所涌现出的智能形态被称为集群智能,具有个体激发时的自组织性和群体汇聚时的强鲁棒性等特征.智能集群系统的协同决策过程是融合人-机-物,覆盖多元空间,囊括感知-决策-反馈-优化的复杂非线性问题,具有开放的决策模型和庞大的解空间.然而,传统的算法依赖大量的知识与经验,使其难以支持系统的持续演化.强化学习是一类兼具感知决策的端到端方法,其通过试错的方式不断迭代优化,具有强大的自主学习能力.近些年来,受生物群体和人工智能的启发,强化学习算法已由求解个体的决策问题,向优化集群的联合协同问题演进,为增强集群智能的汇聚和涌现注入了新动能.但是,强化学习在处理集群任务时面临感知环境时空敏感、群内个体高度自治、群间关系复杂多变、任务目标多维等挑战.本文立足于智能集群系统的协同决策过程与强化学习运行机理,从联合通信、协同决策、奖励反馈与策略优化四个方面梳理了强化学习算法应对挑战的方法,论述了面向智能集群系统的强化学习算法的典型应用,列举了相关开源平台及其适用算法.最后,从实际需求出发,讨论总结了今后的研究方向.

关键词 智能集群系统;集群智能;群体智能;强化学习;感知决策
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2023.02573

The Reinforcement Learning Approaches for Intelligent Collective System: A Survey

LI Lu-Lu[1),2),3)] ZHU Rui-Jie[1),2),3)] SUI Lu-Yao[1),2),3)] LI Ya-Fei[1),2),3)] XU Ming-Liang[1),2),3)]
FAN Hui-Tao[1),2)]

[1)](*School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou* 450001)
[2)](*Engineering Research Center of Intelligent Swarm Systems, Ministry of Education, Zhengzhou* 450001)
[3)](*National Supercomputing Center in Zhengzhou, Zhengzhou* 450001)

Abstract Intelligent Collective System (ICS) is an essential branch of artificial intelligence, encompassing various intelligent components that collectively give rise to an emergent phenomenon known as Collective Intelligence (CI). CI exhibits the characteristics of self-organization in individual excitation, strong robustness in swarm convergence, and other characteristics. Based on ICS, AI enables the emergence of CI, providing a powerful framework

收稿日期:2022-11-16;在线发布日期:2023-07-13. 本课题得到国家自然科学基金重点项目(62036010)、国家自然科学基金青年项目(62001422)、国家自然科学基金面上项目(61972362, 62372416)、国家重点研发计划课题(2021YFB3301504)资助. **李璐璐**,博士研究生,中国计算机学会(CCF)会员,主要研究领域为强化学习、智能集群系统等. E-mail:lll0626_zzu@gs.zzu.edu.cn. **朱睿杰**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为机器学习、智能集群系统等. **隋璐瑶**,硕士研究生,中国计算机学会(CCF)会员,主要研究领域为强化学习、智能集群系统等. **李亚飞**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为机器学习、城市计算、智能集群系统等. **徐明亮**(通信作者),博士,教授,国家杰出青年科学基金入选者,中国计算机学会(CCF)会员,主要研究领域为机器学习、集群智能系统等. E-mail:iexumingliang@zzu.edu.cn. **樊会涛**(通信作者),博士,研究员,中国工程院院士,主要研究领域为飞行器设计、智能集群系统等. E-mail:fanhuitao1962@163.com.

for harnessing the potential of intelligent systems. Specifically, the decision-making process of ICS is a multifaceted and intricate nonlinear problem that intricately integrates humans, machines, and objects. This process spans across diverse spaces and encompasses various stages, including perception, decision-making, feedback, and optimization, forming a dynamic loop of information flow. Within this intricate framework, there exist abundant decision models that enable the system to consider a wide range of possibilities and alternatives. The traditional algorithms mainly rely on a large amount of knowledge and experience, creating a significant challenge in supporting the development of the system. The reliance on vast amounts of explicit knowledge and predefined rules limits the system's ability to adapt and evolve in dynamic and complex environments. As the system encounters new situations or scenarios, its performance may suffer due to the lack of flexibility and adaptability inherent in these traditional algorithms. Reinforcement Learning (RL) is a powerful and comprehensive approach that seamlessly integrates perception and decision-making within an end-to-end framework. RL exhibits a remarkable autonomous learning capability, enabling systems to improve their performance through iterative optimization driven by trial and error. In RL, the system interacts with its environment, receiving feedback in the form of rewards or penalties based on its actions. Through this iterative process, the system learns to navigate complex decision spaces by exploring different actions and evaluating their consequences. By optimizing its decision-making policies over time, RL enables the system to acquire knowledge and adapt its behavior to maximize long-term rewards. Recently, there has been a remarkable evolution in RL algorithms, spurred by inspiration from both biological swarm behavior and artificial intelligence. These advancements have not only expanded the scope of RL from solving single-agent decision-making problems but have also paved the way for addressing joint collaboration problems involving multiple agents. As a result, RL has emerged as a new and promising avenue for the convergence and emergence of CI. However, the application of ICS faces significant challenges when dealing with various tasks. These challenges arise due to the unique characteristics of ICS, including the spatio-temporal sensitivity of the perceptual environment, the high autonomy of individuals within the swarm, the complex and variable relationships among agents, and the multi-dimensional nature of task goals. Based on the decision-making process of ICS and the operation mechanism of RL, this paper introduces RL algorithms that specifically target the challenges posed by ICS, focusing on four key aspects: joint communication, collaborative decision-making, reward feedback, and policy optimization. The paper further conducts an analysis of typical applications of RL algorithms in ICS, accompanied by a compilation of relevant open-source platforms and applicable algorithms. Finally, the paper addresses future research directions based on practical requirements.

Keywords intelligent collective system; collective intelligence; swarm intelligence; reinforcement learning; perception and decision-making

1 引 言

智能集群系统(Intelligent Collective System, ICS)是指在一定时空范围内聚集的多智能体,以集群协同形式完成特定任务的智能系统,如无人机集群[1]、地面机器人集群[2]等.相应的,本文将ICS中涌现出的智能形态称为集群智能(Collective Intelligence)[3-5].上述概念部分源于自然界普遍存在的细胞蠕动[6]、蜂群觅食[7]、狼群围捕[8]、人群流动[9]等生物群体行为,并与机器学习[10]、物联网[11]、虚拟现实[12]等领域密切相关.目前,智能集群系统已被广泛应用于交通管控[13]、物流调度[14]、工业制造[15]、农林保植[16]、医疗辅助[17]、国防军事[18]等领域,如图1所示.一些文

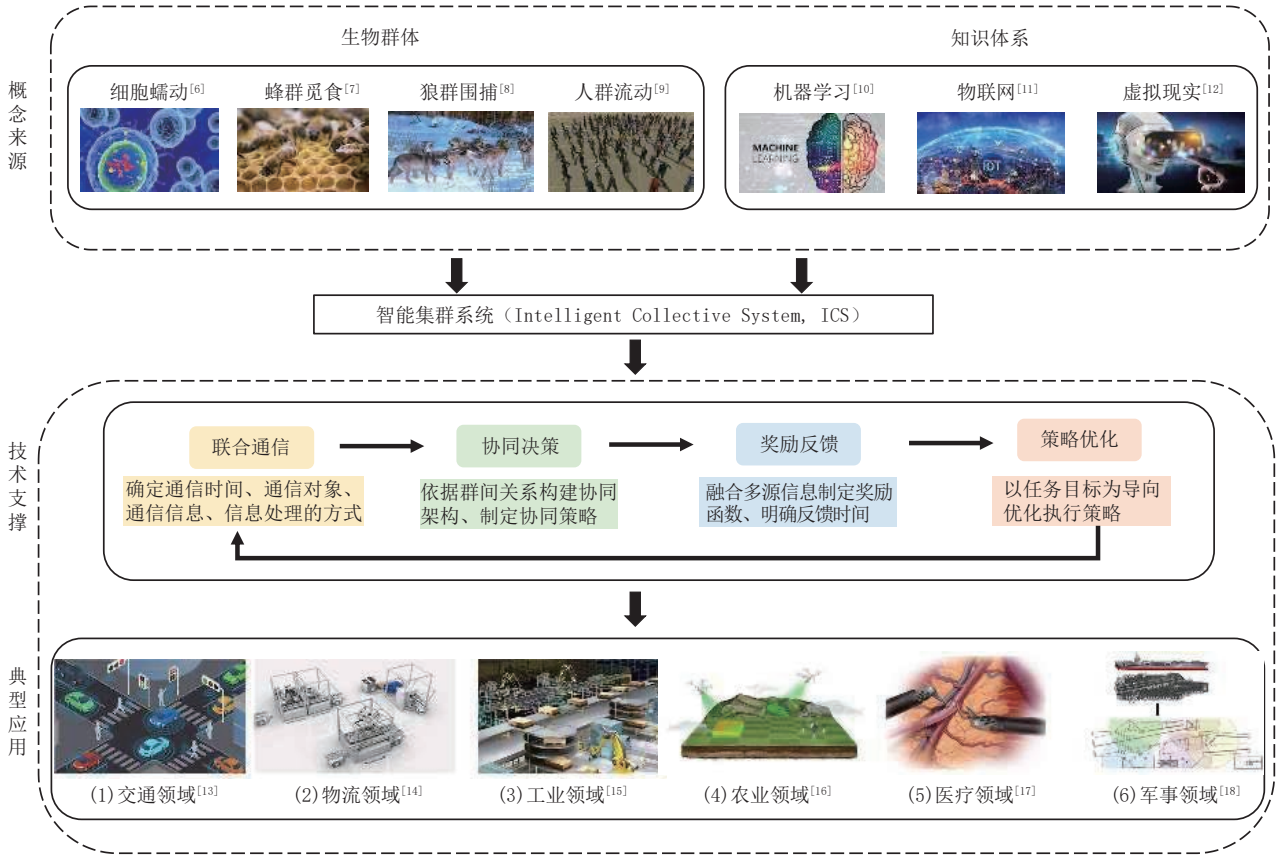


图1　智能集群系统(Intelligent Collective System, ICS)

献将具有此特征的系统表述为群体智能系统[19]，或将此智能形态定义为Swarm Intelligence[20]、Crowd Intelligence[21]等，本文在此将以上表述统一约定为智能集群系统(ICS)和集群智能(CI).

ICS是一类融合人-机-物，涵盖多元空间，囊括感知-决策-反馈-优化的复杂系统[22]. 在ICS中，物理空间内的群体自主交互，对外界环境的刺激保持响应，在信息空间中产生海量多源数据，再融合社会空间的认知与计算属性，进而激发群体产生协同决策能力[23-24]. 然而，目前群智激发机理尚不清晰，理论分析代价昂贵，群体系统难以建模[25]. 传统的监督学习算法、启发式算法等依赖大量的知识与经验，不能覆盖复杂非线性问题的庞大解空间，难以支持ICS的持续演化[5]. 强化学习(Reinforcement Learning, RL)[26]是一类依据感知信息，以目标为导向，通过迭代试错的方式优化策略的算法. 相较于其他算法，RL算法具有无需预先精准建模、能处理高维非线性数据、易融合多领域知识等优点，为智能集群系统中复杂任务的求解注入了新动能[26-29]. 但是，RL在处理集群任务时面临感知环境时空敏感、群内个体高度自治、群间关系复杂多变、任务目标多维等挑战[23, 30]. 据此，本文立足于智能集群系统的协同决策过程和强化学习的运行机理，对面向智能集群系统的强化学习方法进行了系统的梳理和综述，下述第2节介绍了相关背景知识；第3节先从联合通信、协同决策、奖励反馈和策略优化四个方面分析探究了RL应对挑战的方法，并总结梳理了相关典型应用和开源平台；第4节讨论了未来应关注的相关研究问题.

2　背景知识

强化学习算法由任务环境和智能体两个实体构成，并通过策略联接状态、动作、奖励三大要素，如图2所示[31]. 智能体将从环境感知到的状态信息传递到决策网络来获取动作，在任务环境中执行动作后，得到及时反馈与最新状态，并以目标为导向迭代优化策略[32]. 依据智能体的数量，强化学习分为了单智能体强化学习算法(Single-Agent Reinforcement Learning, SARL)和多智能体强化学习算法(Multi-Agent Reinforcement Learning, MARL)[30].

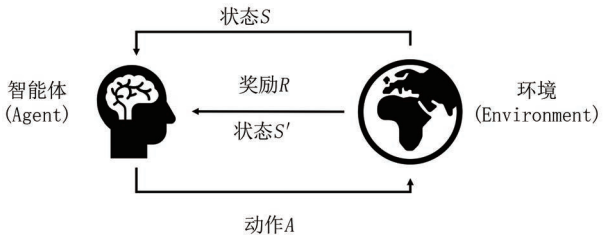


图2 强化学习的基本结构[31]

2.1 单智能体强化学习

SARL算法依据马尔可夫决策过程(Markov Decision Processes, MDP)[33]建模. MDP由5元组$<S,A,P,R,\gamma>$定义,其中S为智能体的状态集合,A为智能体的动作集合,$P:S\times A\to\triangle(S)$表示执行动作$a\in A$后,从状态$s\in S$转移到状态$s'\in S$的概率,$R$表示智能体的奖励集合,$\gamma$为奖励的折扣因子. MDP模型通过制定策略$\pi:S\to\triangle(A)$进而获取最优解. 为评价策略的优劣,RL算法定义了价值函数,其中状态-动作价值函数(State-Action Value Function)即Q函数,定义为:

$$Q_\pi(s,a)=\mathbb{E}\left[\sum_{t\geqslant 0}\gamma^t R(s,a,s')|a\sim\pi(\cdot|s),a_0=a,s_0=s\right]. \tag{1}$$

表示在初始状态s,执行动作a后,遵循策略π得到的期望回报. 状态价值函数(State-Value Function)即V函数,定义为:

$$V_\pi(s)=\mathbb{E}\left[\sum_{t\geqslant 0}\gamma^t R(s,a,s')|a\sim\pi(\cdot|s),s_0=s\right]. \tag{2}$$

表示在回合初始状态s,执行策略π,直到回合结束,累计得到的期望回报. Q函数与V函数的关系为[34]:

$$Q_\pi(s,a)=\mathbb{E}_{s'\sim p(\cdot|s,a)}\left[R+V_\pi(s')\right], \tag{3}$$

$$V_\pi(s)=\mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q_\pi(s,a)\right]. \tag{4}$$

单智能体强化学习算法根据求解目标为最优价值函数还是最优策略,分为基于值的算法和基于策略的算法.

Q学习(Q-Learning, QL)算法[35]是第一个基于值的强化学习算法,QL定义了状态动作评分表,使用贪心算法获得最大Q值,并通过时间差分(Temporal-Difference, TD)方法进行更新. 但是,QL仅能应对简单低维的离散动作问题. Mnih等人结合深度神经网络与QL算法,提出了深度Q网络(Deep Q-Network, DQN)[36]算法,将深度神经网络作为Q函数逼近器. 由于传统的DQN算法存在过估计、难收敛等问题,一系列优化算法被相继提出,如双DQN(Double DQN)[37]算法用于缓解过估计问题;对抗DQN(Dueling DQN)[38]算法加快了收敛速度;循环DQN(Recurrent DQN, DRQN)[39]算法缓解了存储有限的问题等.

基于策略的强化学习方法直接在策略空间内更新参数θ,旨在寻求最优策略$\pi^*\approx\pi_\theta(\cdot|s)$. 策略梯度(Policy Gradient, PG)[40]算法采用梯度下降的方法获取最优策略,进而决策出连续动作的概率值. 但是,PG依据随机机制进行筛选,存在强不确定性. 确定性策略梯度(Deterministic PG, DPG)[41]算法采用确定策略生成连续动作. 由于基于策略的算法极易受到参数的影响,信任区域策略优化(Trust Region Policy Optimization, TRPO)[42]算法通过限制改变范围来避免参数过大变化. 近端策略优化(Proximal Policy Optimization, PPO)[43]算法通过修剪的方式保证策略在置信域中进行优化.

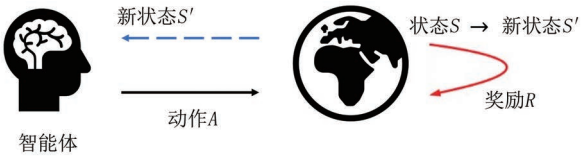
2.2 多智能体强化学习

MARL的决策过程被模型化为马尔可夫博弈(Markov Games, MGs)[44]过程. MGs由六元组$<\mathcal{N},\mathcal{S},\mathcal{A}_i,\mathcal{P},\mathcal{R}_i,\gamma>$构成,$\mathcal{N}$为智能体的数量,$\mathcal{S}$为所有智能体共享状态空间,$\mathcal{A}_i$为智能体$i$的动作空间,$i\in\{1,\ldots,\mathcal{N}\}$且$\mathcal{A}:=\mathcal{A}_1\times\cdots\times\mathcal{A}_\mathcal{N}$,$\mathcal{P}$为状态转移概率,$\mathcal{R}_i$为智能体$i$的奖励值. 每个智能体通过制定策略$\pi_i\in\prod_i(\mathcal{A}_i)$,获得使奖励值最大化的动作. MARL的价值函数定义为:

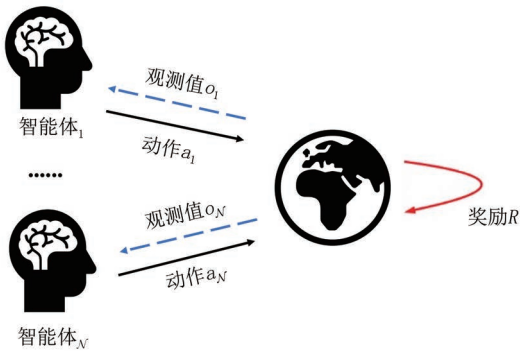
$$V^i_{\pi^i,\pi^{-i}}=\mathbb{E}\left[\sum_{t\geqslant 0}\gamma^t\mathcal{R}_i(s,a,s')|a_i\sim\pi(\cdot|s),s_0=s\right], \tag{5}$$

其中$-i$表示除了智能体i以外的其他智能体. 与公式(2)对比,智能体采用联合策略进行决策,最优表现受自身策略和其他智能体策略的影响.

在多智能体环境下,由于智能体很难获取到全部状态信息,部分观测马尔可夫决策过程(Partially Observable MDP, POMDP)额外关注了状态值的观测概率. 尽管此模型考虑了智能体的部分观测性质,但是其求解时间较长,很难应用于实际控制场景. 基于POMDP,离散马尔可夫决策过程(Decentralized-POMDP, Dec-POMDP)[45]被提出并广泛用于多智能体任务建模. Dec-POMDP由8元组$<\mathcal{N},\mathcal{S},\mathcal{A}_i,\mathcal{P},\mathcal{R}_i,\gamma,O_i,\mathcal{O}>$构成,$O_i$为智能体$i$的观测值集合,智能体的联合观测集合为$\mathbb{O}=O_1\times\cdots\times O_\mathcal{N}$,$\mathcal{O}$表示局部观测值的转移概率. MDP与Dec-POMDP的差异如图3所示[46].



(1) 马尔可夫决策过程(MDP)



(2) 离散马尔可夫决策过程(Dec-POMDP)

图3　MDP与Dec-POMDP[46]

相较于SARL算法，MARL算法涉及多个智能体同时与环境交互，致使MARL的任务环境存在不稳定、非静态的现象. 为缓解以上问题，一些算法采用经验重放方法稳定训练过程，如长短期记忆[13]、循环神经网络[47]等，但其计算量和存储开销较大. 分层强化学习算法[48]、元强化学习[49]、联邦强化学习[50]等算法通过融合多源知识、优化协同架构来缓解不稳定的问题，增强了系统的适应性，提升了学习效率. 此外，多个智能体在决策时需要考虑其他智能体的影响，导致状态空间提高了多个数量级，使MARL算法面临信息维度爆炸的挑战. MARL算法采用数据分析、特征提取、平均场理论等方式对状态信息降维[51-53]，但是这类方法依赖于数据的质量，且易造成数据丢失. 好奇心机制[54]、注意力机制[55]等方式通过探索有用范围、筛选必要状态信息等方式，提升了算法训练效率，被广泛用于MARL算法中.

3　面向智能集群系统的强化学习方法

智能集群系统的协同任务融合了人、机、物群体，覆盖多元空间，囊括感知-决策-反馈-优化环节[22]. 在执行集群协同任务时，稳定可靠的感知通信是获取集群行为的前提与保障[24]，集群凭借感知通信信息制定协同策略，进行联合决策，避免出现“群而不治”的现象[19]. 为了明确当前决策对环境的影响以及各智能体的贡献，智能体需得到及时的奖励反馈，以促进群智可预知、可持续的涌现[25]. 此外，为了实现集群系统的持续演化，需要针对不同的任务目标来优化协同策略，进而制定强稳健、高安全、易扩展的强化学习算法[30]. 鉴于此，我们对面向智能集群系统的强化学习方法进行了分析总结.

3.1　联合通信

在执行集群协同任务时，智能体的感知对象为环境和其他智能体，由于任务环境的演化和奖励的获取依赖于智能体的联合动作，致使强化学习算法面临着感知对象复杂多变、任务环境时空敏感等挑战. 稳健可靠的通信交互机制是集群内部高效协作的前提与保障. 因此，面向智能集群系统的强化学习算法在感知阶段更注重智能体之间的联合通信，以明确通信时间、交互对象、传输信息和处理方式.

确认通信时间的方式分为连续通信法和非连续通信法. 其中，连续通信法指在每一个执行时间步或每间隔一段时间进行通信. 如Foerster等[56]提出的强化智能体间学习(Reinforced Inter-Agent Learning，RIAL)和可微智能体间学习(Differentiable Inter-Agent Learning，DIAL)算法，规定了智能体在每个时间步传输联合通信信息；Sukhbaatar等[57]定义了通信网络(Communication Network，CommNet)，在每个时间步接受其他智能体传递的信息；目标多智能体通信(Target Multi-Agent communication，TarMac)[58]算法支持智能体在每个时间步进行多轮针对性交互；基于两阶段注意力网络的博弈抽象机制(Game Abstraction Mechanism based on Two-stage Attention Network，G2ANet)[59]在每个时间步判断智能体之间的关系；多表演者-注意力-评价者(Multi-Actor-Attention-Critic，MAAC)[60]算法在每一个时间步为每个智能体选取通信信息. 非连续通信方法规定智能体仅能在满足需求时进行通信，如独立控制持续通信(Individualized Controlled Continuous Communication，IC3)[61]算法仅在智能体认为有利时进行通信；多智能体图注意通信(Multi Agent Graph Attention Communication，MAGIC)[62]算法规定智能体仅在必要时进行交互；近似价值分解Q函数(Nearly Decomposable Q-functions，NDQ)[63]算法通过设置编码器来控制通信时间；注意力通信模型(Attention Communication Model，ATOC)[64]定义超参数判断是否通信等.

连续通信法能够持续收集系统的实时数据信息，避免了信息缺失，适用于通信要求较高的场景，如中小型智能机器人控制等. 但是，频繁的通信需要大量的通信资源，使其难以扩展应用到大型复杂

集群系统.非连续通信方式具有减少数据传输量、降低通信开销、提高通信有效性等优点,且具有可扩展性和鲁棒性,能够防止传递不相关或不利的交互信息,适用于通信受阻场景,如集群协同作战.然而通信的时机难确定,且减少通信频次的方式一定程度阻碍了群智的融合.

针对交互对象,鉴于智能体间的关系[65],我们将其分为了全通信、邻居通信和分组通信三种方式,如图4所示.全通信方式指当前智能体的通信对象为其他所有的智能体,典型的算法有DIAL[56]、TarMAC[58]、调度网络(Schedule Network, SchedNet)[66]、G2ANet[59]等.其中,G2ANet算法规定智能体在初始时全联接,再采用硬注意力机制(Hard Attention, HA)切断不相关的智能体之间的通信渠道.邻居通信方式依据欧氏距离、最短距离等确定通信范围,在通信范围内的智能体皆为其感知对象,代表性算法有深度图网络(Deep Graph Network, DGN)[67]、通信神经网络(Neural Communication, NeurComm)[68]、流通信(Flow Communication, FlowComm)[69]等.分组通信方法按照目标、智能体的等级、种类等分组,通过定义代理人(如图4的子图(3)中的斜纹圆形)进行组间通信后,将通信信息传递给组内智能体,组内智能体通常不发生交互.如CommNet[57]算法定义了星型通信结构;ATOC算法[64]设计了树形分组结构;学习结构化通信算法(Learning Structured Communication, LSC)[65]依据等级分组通信等.

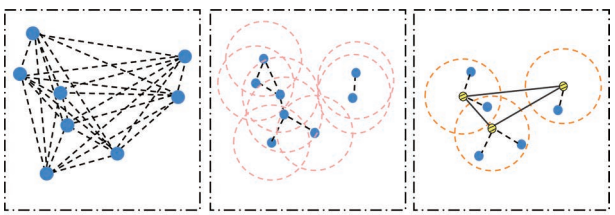


图4 强化学习中典型的通信拓扑[65]

全通信的方式能够使智能体感知到其他所有智能体的信息,提高了协同效率,适用于小型全局协同场景,如资源配置.然而,该方法可能抑制了智能体自身的探索能力,且其通信开销高,数据传输量较大.邻居通信的方式在一定程度上降低了通信开销,适用于局部信息需求较高场景,如集群交通控制等.但该方法存在通信冗余,且其确定感知范围的方式大多数依赖于手动配置.分组通信的方式能够避免传输不必要的信息,有效降低通信开销,适用于复杂通信场景,如医疗资源分配等.然而,此方法的效果依赖于分组质量,且分组也增加了计算的复杂性.

由于智能体在系统中具有局部观测性,需要依据通信对象传递的信息拓阔视野.强化学习的通信信息主要分为全局信息和局部信息.全局信息包括全局状态和动作等,如CommNet[57]算法的通信控制器接收总状态序列;双向协调网络(Bidirectionally Coordinated Network, BicNet)[70]算法共享全局状态空间;步进式互信息通信(Progressive Mutual Information Collaboration, PMIC)[71]算法根据全局状态和联合动作定义互信息等.局部信息指通信范围内的局部状态和动作,如G2ANet[59]算法对局部观测值进行编码;ATOC[64]算法使用局部观测值和动作来获取隐藏状态;多智能体激励式通信(Multi-Agent Incentive Communication, MAIC)[72]算法利用局部信息和通信对象的ID构建协同模型等.

广播全局信息的方式确保了智能体的全局视野,但其通信开销较大,仅适用于通信简单的小型集群协同场景.传输局部信息的方式降低了通信负担,避免了信息冗余.然而,局部视野易导致智能体陷入局部最优.

获取到通信信息后,强化学习算法对消息的处理方式分为了直接合并处理、价值平等处理和价值不平等处理三类[73].其中,直接合并的方式指以串联的方式拼接信息,扩大输入空间[56,66,74].例如,DIAL/RIAL[56]算法将通信动作拼接在局部观测值后;SchedNet[66]算法合并了局部观测值和联合编码信息等.平等处理方法即通过取均值或求和的形式来聚合所有智能体的通信信息[57,61,69],如CommNet[57]算法通过反向传播的方式处理累加后的通信信息;IC3Net[61]算法通过线性变换矩阵处理所有智能体的平均通信信息等.不平等处理方法为通过结合图神经网络、注意力机制、互信息等方法区别对待不同智能体的通信信息[58-60].例如,TarMAC[58]算法使用基于签名的软注意力机制确定通信对象的权重,再依据权重分配交互信息;G2ANet[59]算法则先通过HA确定智能体间的联通性,再通过软注意力机制获得交互信息的权重;MAIC[72]算法基于互信息理论先使用高斯编码对其他智能体的动作进行建模,再计算智能体的动作之间的互信息,提高通信的有效性等.

直接合并的方法操作简单且易实现,适用于小型同质集群控制场景,但其囿于集群的规模而难以扩展.价值平等处理的方式降低了信息的冗余度,

但是无差别的聚合会导致信息缺失，且协同效果易受单点影响．价值不平等处理的方式能够合理利用有效信息，提升了通信效率．但相较于其他方法，它需进行额外的训练，增加了算法复杂度．

表1梳理分析了相关算法及其特点与适用场景．目前越来越多的算法朝着简洁高效的感知交互方向发展，即在非连续通信时间内，以目标为导向确认通信对象，交互局部信息后，融合注意力机制、图神经网络等方式来提取重要特征，旨在避免通信冗余，提升联合通信效率．

表1　联合通信机制

要素	方法	特点	适用场景	文献
时间(When)	连续通信	优点：实时传输，加强协作 缺点：计算复杂度高，数据量大，难扩展	通信要求较高的场景(如中小型智能机器人控制等)	[56-60]
	非连续通信	优点：通信效率高，易扩展，鲁棒性强 缺点：难确定通信时机	通信受阻场景(如集群协同作战等)	[61-64]
对象(Who)	全部	优点：提高协作效率 缺点：通信开销高，数据传输量大	简单、对全局信息有需求的场景(如全局资源配置等)	[56, 58, 59, 66]
	邻居	优点：减少通信开销 缺点：通信冗余，依赖阈值设置	对局部信息有需求的场景(如交通灯控制，工业运输等)	[67-69]
	分组	优点：有效降低通信开销 缺点：依赖于分组质量，提升计算复杂度	安全需求高、通信资源有限的场景(如医疗资源分配等)	[57, 64, 65]
信息(What)	全局	优点：提供全局视野，增强协同性 缺点：通信开销大，通信冗余	通信简单的小型集群协同场景(如网络资源分配等)	[57, 70, 71]
	局部	优点：降低通信负担，高效易扩展 缺点：易陷入局部最优	复杂通信场景(如无人集群协同控制等)	[59, 64, 72]
处理(How)	直接合并	优点：简单，易实现 缺点：输入长度不固定，难扩展	小型同质集群控制场景(如无人农场等)	[56, 66, 74]
	价值平等	优点：易实现，降低冗余 缺点：信息缺失，易受影响	同质集群协同控制场景(如无人车协同控制等)	[57, 61, 69]
	价值不平等	优点：易扩展，利用率高 缺点：信息缺失，提高计算复杂度	复杂集群系统控制场景(如无人驾驶等)	[58-60]

3.2　协同决策

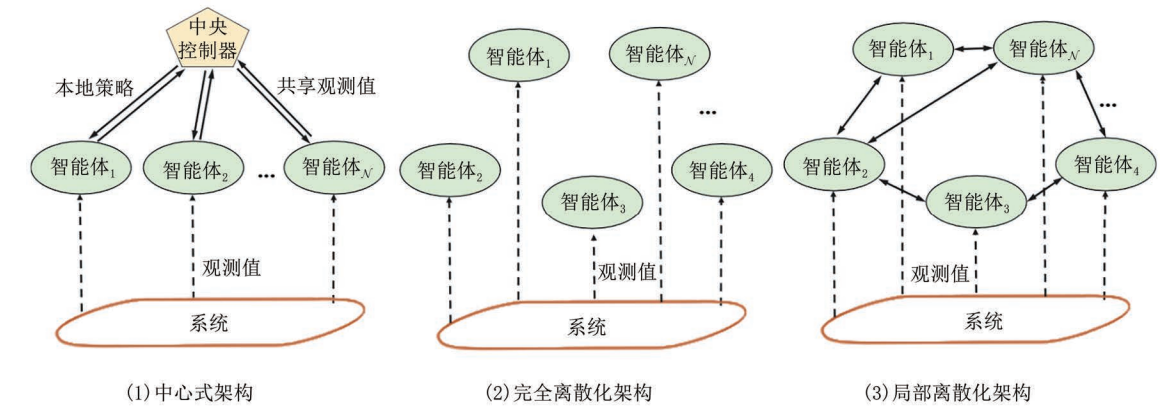
在获得集群联合信息后，集群内个体间关系复杂多变，基于强化学习的协同策略面临着“群而不智”的挑战．作为群智涌现的关键，集群进行协同决策时需依据群间关系构建协同架构并制定协同策略，以聚合个体激发过程中得到的知识和经验，进而一致化个体与全局的优化目标，获得奖励最大化策略．表2总结分析了协同决策方法的架构和策略．

表2　协同决策方法

协同决策	方法	特点		参考文献
架构	中心化	优点：简单易现，收敛快	缺点：计算开销大，易受单点故障影响	[76-80]
	完全离散化	优点：简单易现，隐私性强	缺点：不稳定，难收敛	[81-85]
	局部离散化	优点：增强交互，适应性强，易扩展	缺点：收敛较慢，存在通信开销，增加复杂度	[86-87]
策略	价值(value)	优点：明确群间关系，针对性训练，易收敛	缺点：探索能力弱，易积累误差，存在信息损失	[88-93]
	策略(policy)	优点：适应性强，能处理高维数据	缺点：收敛慢，不稳定	[77, 93-97]

强化学习算法的协同架构分为三类[75]，如图5所示．为了缓解智能体部分观测性，集中控制器多被用于处理联合状态、动作、奖励等信息[76]．多智能体深度确定策略梯度(Multi-Agent DDPG，MADDPG)[77]算法在此基础上设计了集中式训练分布式执行(Centralized Training with Decentralized Execution，CTDE)范式，使智能体在训练阶段考虑其他智能体影响，在决策阶段去中心化执行．CTDE架构降低了交互环境的不稳定性，被广泛应用于同质智能体的协同任务中[78-80]．中心化方法易于实现，收敛速度较快．但其对存储和计算资源需求较高，难以扩展，且易受单点故障的影响．相较于CTDE，完全离散化的架构使智能体根据局部观测值独立学习[81-83]．此类架构易于实现且隐私性强，适用于通信

成本高的实际应用场景[84-85]. 但是,完全分布式的架构使得智能体难以学习到整个系统的最优解,易造成环境不稳定、算法难收敛等问题. 为了缓解上述问题,近期一些算法采用建立通信渠道、共享参数等方式来构建局部离散化的协同架构. 以网络化去中心架构为例[86-87],其通过在智能体间定义通信渠道,建立网络拓扑,以传递通信信息. 该架构在增强智能体间交互的同时,降低了独立学习对环境造成的不稳定性. 但是,其收敛速度慢于中心化架构,且通信开销大于完全离散化架构.



(1)中心式架构 (2)完全离散化架构 (3)局部离散化架构

图5 强化学习算法中经典的集群协同架构[75]

针对协同策略,基于价值分解的协同策略将局部最优作为全局最优的一部分,使个体的优化目标与全局优化目标保持一致,进而进行更有针对性的训练. 价值分解网络(Value-Decomposition Networks, VDN)[88]算法累加各个智能体的最大动作价值之和来获得全局总价值,总价值再以反向误差的方式传递给每个智能体,使智能体能够明确自身对全局的影响,并以全局最优为目的进行更新. 但是,简单累加的形式难以表示复杂的状态特征. 为了更好地表示个体和群体间的复杂关系,Q混合(QMIX)[89]算法构建混合网络来获取全局价值,再通过约束全局价值与个体价值的关系,确保两者的单调一致性,使得个体的局部最优动作为全局最优动作的一部分. 但是,无论采用串联相加还是单调限制都存在一定的局限性,且近似的全局价值可能与真实值有很大的差距,这样的更新方式收效甚微. QTRAN[90]算法旨在释放累加性和单调性的限制,对任何可分解的因式进行分解. QTRAN构建了联合动作价值网络、独立动作价值网络和状态价值网络,并为每个网络定义了用于训练的损失函数,提升了算法的稳定性、加快了收敛速度. 此外,在满足个人-全局-最大(Individual-Gloal-MAX, IGM)的基础上,双对抗多智能体Q学习(Duplex dueling multi-agent Q-learning, QPLEX)[91]算法使用对决Q网络将协同问题转换为优势函数约束问题. 分队协同Q学习(Q-learning with Sub-team Coordination, QSCAN)[92]算法依据分队协同模式细分优势值,将全局价值先分解再传递. 由于基于IGM的算法存在观测值不充分、价值分解有损等问题,导致误差不断累积. IGM-匕首(IGM-Dagger, IGM-DA)[93]算法结合模仿学习算法来防止误差的累积.

除了基于价值分解的方法,一系列基于策略的协同算法被提出. MADDPG[77]方法结合CTDE架构和DDPG算法,将智能个体表示为独立执行动作的表演者,再通过全局评价者来指导训练. 但是,MADDPG的策略网络输出的动作具有单一确定性,导致其训练效率低,易收敛于次优策略. 为了缓解以上问题,共享经验AC(Share Experience AC, SEAC)[94]算法在策略更新时引入共享经验指导智能体协作,多智能体软演员评论家(Multi-Agent Soft Actor Critic, MASAC)[95]为每个智能体的独立评价网络增加动作熵,鼓励智能体进行探索,使算法更加稳定. 除了丰富策略的输入外,多智能体变化探索(Multi-Agent Variational Exploration, MAVEN)[96]算法引入分层控制的潜在空间,将基于值和基于策略的方法混合,以获取交互信息. 此外,MAVEN算法还关注了因分布式执行算法导致策略探索不充分的问题,通过最大化交互信息,加强智能体协同能力. 由于基于CTDE的算法易导致局部最优策略和全局最优策略不匹配,分解策略梯度(Decomposed Policy gradient, DOP)[97]算法将价值分解引入集中评价,以优化局部与全局的联合策略.

基于价值分解的策略能够明确智能体间的关系，使算法进行针对性训练，具有良好的收敛性．但是，在价值分解的过程中易存在探索不足、误差积累、信息损失等问题．与基于值的算法相比，基于策略的方式提升了算法的适应性，更适合于高维状态空间任务．但是，其收敛速度慢于基于值的协同策略，且通过策略直接选择执行动作的方式也缺乏稳定性．目前，有许多算法将两者结合，同时关注了算法的稳定性与高效性．

3.3　奖励反馈

在执行集群决策后，由于系统内个体高度自治且多样，致使智能呈现的程度具有强不确定性[25]．强化学习算法通过制定奖励机制，及时反馈决策的优劣，增进集群协作，推动集群智能可度量、可预知、可持续地涌现[98]．

智能集群系统具有个体奖励和群体奖励，强化学习算法依据任务类型定义个体奖励，评估各个智能体执行动作的优劣．例如，在交通信号协同控制任务中，奖励根据路口拥堵情况定义[99]；无人集群协同对抗任务中，奖励依据无人机与目标的位置、能耗、碰撞等元素定义[100]．此外，奖励反馈的时间可分为仅在关键时刻反馈、任务结束时反馈、每个时间步反馈．前两种方式提供了较稀疏的个体奖励，能够增强智能体的探索能力，但是训练效率较低．第三种方式的个体奖励较稠密，智能体能够得到及时的反馈指导，被广泛应用于各个算法中，但是易造成个体奖励稠密、群体奖励稀疏的现象．强化学习将群体奖励定义为个体奖励的累加或加权和，呈现了整个智能集群系统的效能，鼓励群内个体为了集体利益而行动．但是，个体奖励基于利己主义定义，智能体以最大化个体奖励为目的进行探索．在个体与集体利益出现冲突时，个体往往倾向于牺牲群体利益．如图6所示，群体奖励（红线）能够指导策略达到群体最优，但其稀疏且优化的速度慢．个体奖励（黄线）稠密且学习速度快，但因其利己的特性，容易陷入次优策略[101]．为了对系统进行有效的反馈，增强工作效率，构成荣辱与共的集群，需要制定折中的奖励机制（蓝线），在保证个体奖励的同时最大化群体奖励．

奖励塑造（Reward Shaping，RS）[102]是强化学习算法缓解个体奖励稠密、群体奖励稀疏问题的方法之一．RS使用额外的个人奖励编码先验知识，以帮助集群取得最大回报[103]．由于传统的RS算法需要依据任务类型预先设定势函数，以表征奖励的优劣，Bhaskara Marthi等[104]提出了一种自动学习势函数的方法，通过定义抽象的MDP模型，将价值函数近似为势函数．基于势函数的方法保证了群体与个体的学习目标一致，且能够独立优化个体奖励与群体奖励．但是，其需要根据任务类型定制，不具有泛化性，难以指导算法获取最优策略．一些算法在此基础上关注了如何有效利用个体奖励来指导集体奖励．Majumdar等[105]基于进化学习方法，使用梯度优化器最大化个体奖励，再利用个体学习到的技能最大化群体奖励．但是，进化过程需要大量的计算资源和高内存成本，且该算法只关注个体奖励向群体的单向转移，没有关注群体对个体的指导．个体奖励辅助团队奖励策略学习（Individual Reward Assisted Team Policy Learning，IRAT）[100]算法通过约束群体与个体的差异，鼓励个体奖励向群体奖励靠拢，群体则依据个体探索到的数据进行优化．

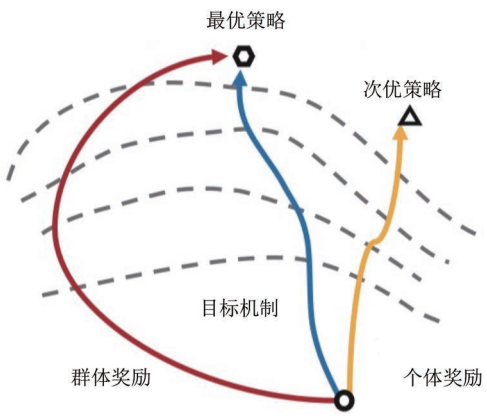


图6　群体奖励与个体奖励[101]

此外，面向智能集群系统的强化学习方法能够注入人类经验，构造以人为本的奖励机制，使之能够朝着人类喜好的方向进行训练[106]．根据引入人类评价的频次，我们将此奖励反馈机制分为了交互反馈和监督反馈．交互反馈指人类与智能体在发生沟通、协作等交互行为时构建的奖励机制，如人类观察者会塑造奖励函数以训练智能体执行任务[107-108]．监督反馈仅用于评价智能体的行为，其通过提供不同形式的反馈信号，引导智能体训练．具体来说，一些算法根据智能体是否执行正确动作来赋予离散信号，如积极奖励、消极奖励、积极惩罚和消极惩罚[109-110]，或将人类反馈作为标签，使用标签来反馈当前动作的优劣[111-112]．

交互反馈能够实时修正智能体的行为，具有较高的灵活性，适用于安全精密的场景．但是，频繁地

引入人类指导降低了智能体的探索能力，且在不易评估的场景下易出现人类误导的现象．监督反馈简单易实现，能够降低智能体的探索难度．但是，离散的奖励信号无法给予明确的反馈．综合而言，人在回路的奖励方式能够提供多维指导，辅助智能体获得最优决策．为了得到有效的指导信号，在制定奖励反馈机制时需要融合多元知识来丰富人为指导信息，且需明确引入指导的时机和频次．

3.4 策略优化

智能集群系统涵盖民生、经济、军事等领域，强化学习在执行集群任务时面临约束复杂、优化目标多等挑战．作为群智涌现的关键，强化学习算法需以强鲁棒、高安全、可泛化等目标为导向进行策略优化．鉴于此，本节分析总结了强化学习优化策略的方法[25,113]．

强化学习算法通过约束和干扰等方式提升策略的鲁棒性．其中，约束是指通过添加正则项等元素来限制策略间的差异．保守的Q学习（Conservative Q-Learning，CQL）[114]算法将保守思想注入值函数中来确定Q函数估计的下限值，进而削弱了因数据集与策略分布偏差造成的干扰．但是，这类算法的复杂度较高．Scott Fujimoto[115]等使用简单的方法对学习的策略与数据进行约束，在双延迟深度确定策略梯度（Twin Delayed Deep Deterministic Policy Gradient，TD3）[116]算法中添加行为克隆（Behavior Cloning，BC）正则项，规范了智能体的行为．但是，BC方法的性能过度依赖于数据的质量．为避免此现象，表演者修正的离线MARL（Offline MARL with Actor Rectification，OMAR）[117]算法为非数据集中的动作添加矫正项，辅助算法进行更新．干扰是指通过添加噪声、对抗等方式，使智能体在最坏情况下进行策略训练．具有噪声优势函数（Noise-Advantage，NA）和噪声价值函数（Noise-Value，NV）[118]的算法使用权重混合函数值与噪声值，鼓励个体进行多样性探索，提升了算法的鲁棒性．当智能体优化策略时，极大极小MADDPG（MiniMax MADDPG，M3DDPG）[119]算法假定其他智能体采用最小Q值所对应的动作，使智能体能够在最差的条件下训练策略．鲁棒多智能体强化学习（Robust MARL，RMARL）[120]算法基于M3DDPG，充分考虑了模型的不确定性，将该求解问题定义为鲁棒的马尔可夫决策过程，在保证收敛性的前提下，寻求所有智能体都不偏离的均衡策略．

约束的方法能够确保智能体的决策符合期望目标，可预防智能体做出对系统有害动作，进而增强系统的鲁棒性．但是，复杂动态的任务场景难以设定约束条件，且过度的约束会抑制智能体的探索能力，致使产生次优解．干扰的方法能够增强智能体的自适应能力和探索能力，防止过拟合．但是此方法增加了求解难度，且不恰当的干扰易导致系统不稳定．

强化学习通过改变学习目标、修改学习过程或引入人为干预的方式来提升算法的安全性．改变学习目标的方式即在奖励函数或目标函数中添加安全因素，如将风险因素添加到奖励函数或目标函数中，并通过奖励塑造、正则化等方式惩罚高风险的动作[121-122]．正则化SoftMax（Regularized SoftMax，RES）[123]算法通过在价值函数中添加正则项来惩罚偏离基础值的策略，再使用SoftMax计算目标价值，以避免出现过高估计．修改学习过程方法即在策略部署时考虑安全因素，如约束策略优化（Constrained Policy Optimization，CPO）[124]、最坏情况SAC（Worst-Case SAC，WCSAC）[125]等算法采用带约束的拉格朗日乘子来计算变量，将约束问题转换为无约束问题，使智能体在信任区域内进行探索．为了使此类算法满足高稳定性、最优性、高采样效率三个指标，约束变分策略优化（Constrained Variational Policy Optimization，CVPO）[126]算法结合凸优化、监督学习对策略进行安全约束．人工干预法则采用人为干预、专家指导等方式纠正智能体的执行动作，构建可信的策略模型．William Saunders等[127]提出了人类干预强化学习（Human Intervention RL，HIRL）机制，引入人类监督器和模拟器，以防止智能体出现不安全的动作．相较于HIRL，基于模型的人类干预（Model-Based Human Intervention，MBHI）[128]架构添加了动态模型，感知和规避未来潜在的灾难．

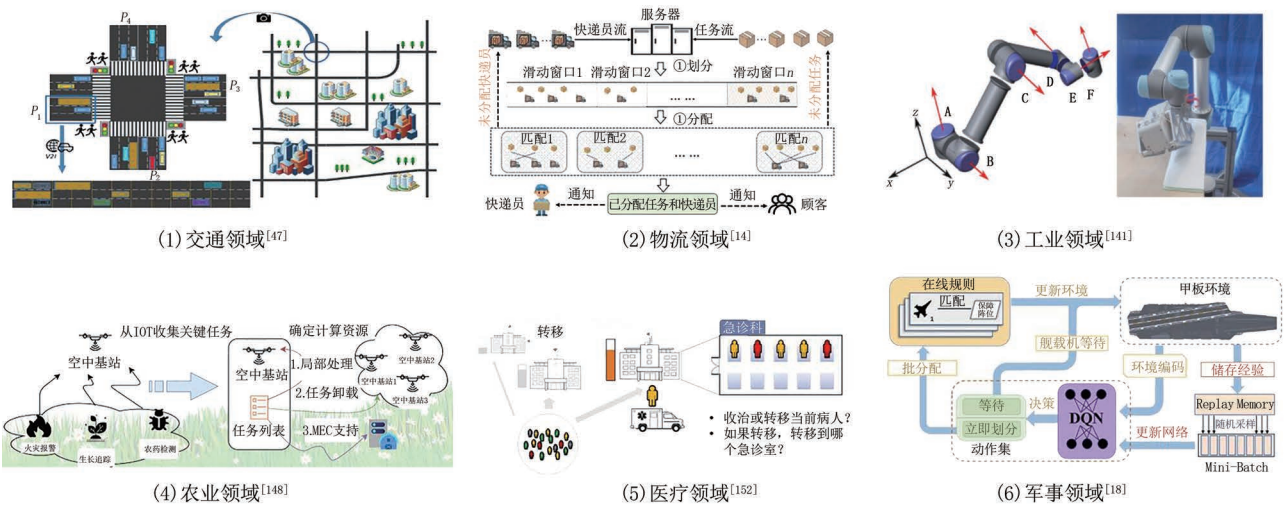
改变学习目标的方法能够引导智能体朝着安全的方向学习，但是其需要提前对任务环境进行有效评估，且对参数的选取十分敏感．修改学习过程的方法能够提高算法的自适应能力，但是其收敛速度较慢．人为干预的方式能够简化学习过程，加快训练速度，但其受限于应用场景．目前，大多数算法通过约束策略、限制优化区间等方法修改学习过程，使智能体在信任域探索，在满足安全约束的同时提升探索效率．

系统内的个体会随时加入或退出，使得任务空间处于动态变化中．为了提升策略的泛化性，需要制定不受输入大小限制和数据次序变化影响的机制．均值嵌入、DeepSets、注意力机制、图神经网络、循

环神经网络等皆被用于塑造置换不变性．均值嵌入方法将变长的状态维度编码为定长的维度[129]．DeepSets定义了置换不变的函数，提升了算法的可扩展性[130]．基于DeepSets，进化种群课程（Evolutionary Population Curriculum，EPC）[131]定义了总数恒定的训练架构，对输入进行约束，采用交叉、变异、选择等方式选出优异的群组模型用于更新，减小了群内个体的性能受群组规模大小的影响．图神经网络是基于图结构的DeepSets，其依据节点向量和传输信息来聚合学习特征．Iou Jen等基于图神经网络，提出置换不变评论家（Permutation Invariant Critic，PIC）[132]算法，构建不受输入顺序影响的聚合机制．循环神经网络则序列化输入状态，并结合LSTM等机制整合输入序列．注意力机制能够捕获输入的关键特征，从而泛化到参数固定的任意智能体中．结合注意力机制，动态多智能体课程学习（Dynamic Multi-Agent Curriculum Learning，DyMA-CL）[133]算法从小规模场景开始训练，通过逐步增加智能体的数目来求解大规模动态问题．均值嵌入方法简单易实施，但是其难以处理复杂场景的非线性数据．DeepSets能够处理高维数据，但是其依赖大量的训练．图神经网络能够处理复杂关系的集群协同问题，相较于其他算法，它需要较大的计算资源和通信开销．循环神经网络能够利用先验知识，捕获智能体的隐藏状态．但是，它需要反复迭代且存储负担较大．注意力机制虽然需要额外的训练，但是其具有自适应性，能够依据信息的重要程度进行聚合处理，被广泛应用于目前的算法．

3.5　典型应用

随着机器学习、物联网、虚拟现实等技术的深入运用，强化学习为智能集群系统的发展注入新动能，推动了人-机-物三元的高效协作与融合发展，并广泛应用于交通、物流、工业、农业、医学和军事等领域，如图7所示．



(1) 交通领域[47]　(2) 物流领域[14]　(3) 工业领域[141]

(4) 农业领域[148]　(5) 医疗领域[152]　(6) 军事领域[18]

图7　典型应用

（1）面向交通领域的集群任务类型包括交通流量预测[134]、车辆自动驾驶[135]、车辆路径规划[136]、交通信号灯控制[47]等．在此类任务环境下，交通环境中的信号灯、车辆、行人等群体被模型化为智能体，构成了人车物融合的复杂系统．智能体通过行车传感器、道路摄像头等途径获取状态信息，且这些信息能够被描述为序列化特征．此外，智能体当前时刻的决策影响下一时刻的特征，所以此过程能够被定义为Dec-POMDP，并可以采用强化学习算法优化协同决策．以交通信号控制为例，多智能体循环DDPG（Multi-Agent Recurrent DDPG，MARDDPG）[47]算法将每个路口模型化为智能体，智能体的状态值为路口信号灯的相位、车辆速度、车辆排队长度和等待通过行人的数量等特征，智能体的动作决定着交通灯的相位变化，奖励由队列长度、车辆和行人等待时间等多种因素决定．由于各个路口的观测视野有限，MARDDPG算法采用LSTM降低了因部分观测性导致的环境不稳定性，并采用参数共享来加速训练过程．由于各交叉口的拥堵状况存在差异，多个路口通过交换路况信息来缓解拥堵．然而，海量的交互数据会带来计算和存储压力．基于对抗生成网络（Generative Adversarial Network，GAN）的交通信号控制算法[99]调用GAN模型来自主生成交互信息，减少信息传输，增强训练效率．此外，为了考

虑人车混合的真实场景，基于上下文感知的多智能体宽度强化学习（Context-Aware Multi-Agent Broad Reinforcement Learning，CAMABRL）[13]算法基于真实数据构建了人车数据集，并采用宽度强化学习架构[137]加快了训练速度．相关网站①梳理了开源的交通数据集和模拟器，以及基于强化学习的交通信号控制算法．目前，交通智能集群系统的理论研究已取得一定的成果，但其大多数依赖于模拟仿真器来实现，未来将朝着Sim2Real的方向发展，并实现真正的落地应用．

（2）针对物流领域，因仓储吞吐量的快速增长，要求物资转运系统具有高自动化水平和高运营效率[14]．智能物流集群系统由无人车、无人机等众多运输物资的自动化设备构成，这些设备与工作人员相互协作，构成了高效运作的系统，多用于解决任务部署、路径规划等具有大量状态和动作的多周期序列决策问题．随着实际问题规模的扩大，经典的组合优化方法的性能受限于求解时间和计算能力，难以获得可行的最优方案．强化学习算法具有泛化能力强、求解速度快等优势，能够将各个实体模型化为智能体，再使用深度神经网络捕捉动态的状态特征，进而获得最优协同策略．在码头运输场景中，自动引导车是衔接各个人、机、物任务区域的关键因素．相较于传统的Dijkstra方法，MADDPG算法有效地缩短了自动引导车的等待时间[138]．它将每个自动引导车模型化为一个智能体，状态信息为当前位置，动作决定移动速度和方向，奖励为车辆与任务点的距离，再通过采用Gumbel-SoftMax方法离散化节点网络场景，预防路径冲突和缩短车辆等待时间．无人机集群能够在危险灾难场景补给物资，进行物流配送．Daiki等[139]将无人机集群的路径规划任务定义为多目标、多行程、多任务的NP-Hard问题．鉴于启发式算法难以解决动态环境中的大规模复杂问题，他们采用QL算法构建具有动态学习能力的智能体，并从供给速度、紧急程度、配送数量三方面评测了算法策略的有效性．因为无人机的运行时间和作业范围有限，Sangwoo等[140]提出融合多表演者-注意力评价者（Fusion MultiActor-Attention-Critic，F-MAAC）算法，定义了传感器融合层使智能体获取到有效的传感器信息，设计了权重层补偿模型在注意力层丢失的信息．在自建的仿真平台上，对比了DDPG、MADDPG、MAAC等算法，验证了F-MAAC算法在物流配送任务的有效性．

（3）针对工业领域，智能制造朝着新兴的Self-X（即自优化、自组织、自调节等）方向发展，旨在达到更高水平的自动化，实现人机物的高度融合[15]．智能机器人集群是工业领域提升任务效率的核心组件．传统的任务规划方法通过定义关键位置和动作来指导机器人执行任务．然而，针对多变的装配任务，此类方法需要耗费较长的时间调整参数，且缺乏精准性、适应性和灵活性．强化学习能够模仿人类学习方式，提供了一类利用传感器的反馈信息进行迭代优化的高精度方法，补偿了传统方法的低精度问题[141]．其中，为了缓解机器人集群在执行任务时因拓扑变化导致的计算复杂度高、难扩展的问题，分布式协调学习（Distributed Coordinated Learning，DCL）[142]算法在MARL模型上采用迁移学习方法传递学习信息，自适应拓扑结构变化．Yudha等[141]基于AC算法提出了参考补偿法，使用矫正信号补偿因未建模造成的偏差，并在6自由度的工业机械臂上进行了评估．此外，工业通信网络的维护与部署是循序渐进的序列化决策问题，能够采用RL算法来提高网络通信率、降低网络时延[143]．具体地说，为了满足复杂的流量需求，基于协作MADRL的带宽分配（Cooperative MADRL-based Bandwidth Allocation，CMD-BA）[144]算法将每一个波束表征为一个智能体，智能体的状态由波束、带宽块、活跃用户的数量、分配状态和其他智能体的带宽状态构成，动作则决定了如何进行资源块的分配．为了提高通信网络的可扩展性，Mee等[145]基于强化学习提出集群规模调节信任模块，协助集群增长到合适的规模，实现网络的可伸缩性．

（4）农业领域对环境的变化十分敏感，随着智慧农业的发展，一些机构利用物联网设备监测农作物的生长，并采用深度神经网络检测害虫、火灾和生长情况，进而辅助农技人员工作，提高生产质量[146]．然而，为了平衡任务的时效性与设备能耗间的关系，需要在感知的神经网络中添加决策过程．强化学习能够根据实况数据及时做出决定，可用于农业集群设备的部署与控制．智慧农场使用机器人监测环境，实现自动化控制．针对多机器人在农场的区域覆盖问题，Ahmad Din等[147]提出了基于集中卷积神经网络的DDQN算法，处理集群设备产生的大量数据，并在拥有不同地图和不同数目智能体的动态环境中，验证了算法的有效性．由于需要同时控制多个动态变化的传感器，智能体的动作空间具有复杂

① https://traffic-signal-control.github.io/

时变等特征. 为了增强时变复合动作空间的鲁棒性和可转移性，李文昊等[16]提出了基于CTDE的结构化协作强化学习算法（Structured Cooperative RL algorithm based on the CTDE，SCORE）. 在进行温室种植番茄的模拟实验中，SCORE算法根据天气和土壤的特征定义智能体的状态，通过动作控制温度、二氧化碳浓度、光照强度，并以温室内番茄增加的重量为奖励值. 实验对比证明，SCORE算法下的番茄的重量明显高于其他对比算法. 此外，一些农场使用空中基站（Aerial Base Stations，ABSs）来监测环境的异样，为了最大限度降低ABSs执行任务的能耗，Turgay等[148]将此问题定义为有约束的MDP，通过定义阈值来约束风险任务的边界，进而自适应地改变学习策略，最大化任务距离. 在大型农业工作场景，Viktor等[149]使用强化学习的感知决策技术，控制大型农业用车在崎岖的地形进行农作任务.

（5）针对医学领域，异构多源的医疗数据对病情的诊断至关重要，深度学习系统能处理大型数据集且支持多维输入. 强化学习结合深度学习算法，对数据感知后进行分析决策，并能够与医务人员协作诊断，可适用于医疗保健. 以机器人辅助手术（Robotic-Assisted Surgery，RAS）[17]为例，RL算法能够增强RAS的鲁棒性和自适应能力. 手术姿势识别是RAS的关键，鉴于现有研究未能重视未来信息对姿态识别的影响，RL算法结合树搜索方法[150]，提出了关节手术姿势分割与分类框架，将手术视频分割与分类任务定义为MDP，视觉特征定义为智能体的状态值，进而训练有关手术姿势识别的策略. 此外，RL算法能够提升RAS中重复性子任务的准确性，但传统方法缺乏风险评估和安全约束. 鉴于此，Ameya等[151]提出安全深度强化学习框架，将问题规约到安全工作空间内，用于组织收缩手术的自动化子任务. 除了RAS外，灾害环境下的救灾人员需要在资源高度限制的环境下做出决策并对救治人群进行医疗物资调配. HyunRok等[152]基于MAAC算法，将有限病房的入住问题模型化为Dec-POMDP，使用行为克隆的方法预训练神经网络，减少计算时间，以提高任务效率.

（6）军事领域的任务类型多为武器分配、协同作战、敌情勘探等，其本质为连续决策问题，属于RL的适用范围. 航母是衡量国家军事力量的主要武器，具有作业空间狭窄、任务密集多样、安全攸关等特点，相关任务被定义为复杂动态序列决策问题[153-154]. 由于启发式的算法复杂度较高、收敛速度慢、准确性低，李亚飞等[18]将舰载机保障任务建模为POMDP，提出战略DQN（Strategy-Deep Q-network，S-DQN）算法，提高了决策效率. 武器目标分配是军事指挥与控制的关键任务，无论是在航母还是陆地基地都被证明为NP-完全问题，其最优解源于整个解空间[155]. 为了减轻搜索解空间的计算压力，策略优化深度强化学习（Policy Optimization with DRL，PODRL）[156]算法将导弹的数量、消耗、类型定义为智能体的状态，针对打击目标分配不同的导弹，设计以人为本的复合奖励函数进行公平采样，提升了任务的完成率. 除了任务部署问题，RL算法还被用于解决军事模拟场景中的集群对抗问题. 为了验证RL算法在通信拒止环境的适用性，汪亮等[157]结合真实环境，搭建集群对抗模拟场景，依据通用RL框架，选择典型RL算法进行测试，如DQN[36]、线性Q学习（Linear QL，L-QL）[158]、异步优势AC（Asynchronous Advantage AC，A3C）[159]、近端分布策略优化（Distributed Proximal Policy Optimization，DPPO）[160]等算法，证明了基于RL的集群模型捕获目标数量最多，具有较高的任务完成率. 但是，此研究仅在网格状的二维环境进行了模拟仿真，未能考虑到大规模真实场景的复杂性. 针对大规模集群智能体协同交互时信息量多、易干扰的问题，蒲志强等[161]基于注意力机制提出了专注网络（Concentration Network，ConcNet），通过评估、排序、剪枝等过程汇聚集群信息，提升对抗的胜利率. 此外，基于平均场的强化学习算法[100]也被用于解决无人集群协同对抗问题.

3.6 开源平台

智能集群系统的任务具有复杂动态、高风险、强约束等特点，适配度高的RL模型需要进行多次迭代优化. 国内外相关机构研发了多种智能集群系统仿真平台，旨在加速开发周期、降低实验成本，为Sim2Real奠定基础，如表3所示. 本节以实际应用为出发点，阐述了智能集群系统开源平台及其可部署的RL算法.

图8展示了无人集群系统开源平台，此类平台能够模拟多个自主设备（如多无人机、无人车、机器人等）执行集群任务，可应用于物流、工业、农业、医学等不同任务环境. 其中，为了能够在真实环境中为自主设备部署RL算法，微软公司研发了基于虚幻引擎的AirSim模拟器[162]，旨在减小模拟与仿真的差距. AirSim能够渲染出复杂多样的真实环境，搭建具备精确动力学的集群自主移动设备模型，为

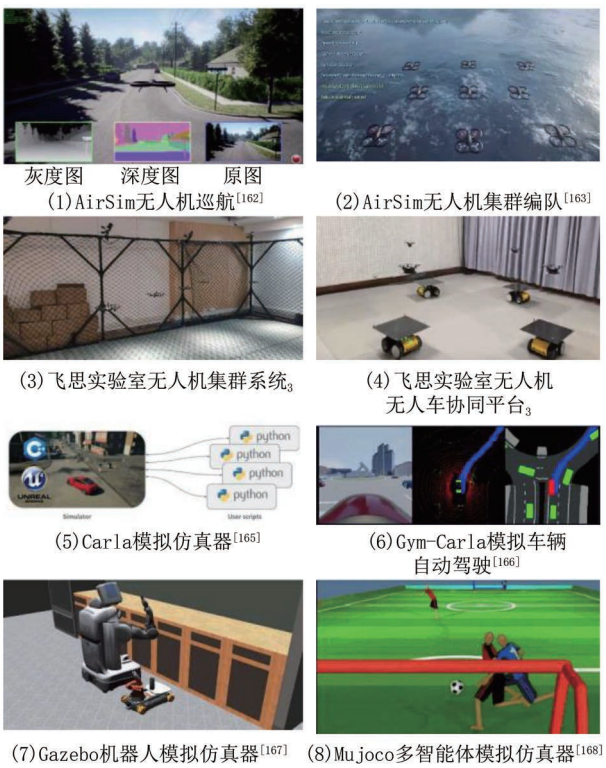
表3 开源平台及其支持算法

应用场景	平台名称	平台功能	适用系统	编程语言	可视化	动力建模	支持基础算法
无人系统	Airsim②	多无人车、无人机等自主移动设备高仿真模拟	Windows Linux	C C# C++ Python	3D	✓	QL, DQN, DRQN, DDQN, PPO, A3C, DDPG, TD3, …
	FeiSiLab③	多类型无人机、无人车的控制和测试	Windows	C++ Python	3D	✓	QL, DQN, DDPG…
	Gazebo④	搭载传感器的3D动态机器人模拟器	Linux	C++ Python	3D	×	QL, DQN, DDPG, PPO, DDQN, TD3, AC, …
	CARLA⑤	丰富环境的无人车模拟器	Linux	C++ Python	3D	✓	QL, DQN, PG, PPO, DDPG, Double DQN, Dueling DQN, A3C, PG, TRPO, TD3, …
	MuJoCo⑥	机器人动力学模拟仿真器	Windows Linux	Python	3D	✓	QL, DQN, PG, AC, TD3, A2C, PPO, DDPG, A3C, TRPO, …
军事推演	Mozi⑦	可在陆海空等全域联合作战的兵棋推演平台	Windows	Python	2D	×	QL, DQN, AC, PPO, DDPG, …
	JSBSim⑧	飞行动力学软件库，能够定义飞机、火箭等控制器	Windows Linux	C++ Python	无	✓	QL, DQN, AC, PPO, DDPG, …
交通管控	CityFlow⑨	大规模城市交通仿真平台	Windows Linux	Python	2D	×	DQN, DDQN, Dueling DQN, PPO, DDPG, TD3, SAC, A3C, PPO, QMIX, …
	SUMO⑩	能够刻画路网、人、车的交通仿真平台	Windows Linux	Python	2D	×	QL, DQN, SARSA, PPO, A3C, …
算法测试	StarCraft⑪	多智能体对抗游戏平台	Windows Linux	Python	2D	×	MADDPG, QMIX, QTRAN, VDN…
	MPE⑫	多粒子群环境	Windows Linux	Python	2D	×	MADDPG, DDPG, PPO, OMAR, TarMAC, …
	MAgent⑬	支持数百万简单个体的平台	Windows Linux	Python	2D	×	DQN, DRQN, A2C, MADDPG, …

部署RL算法提供了高保真的数据信息. 此外，AirSim还能够注入人工指令进行真实的指挥控制. 依据AirSim提供的状态、环境信息，RL基于任务目标构建决策模型，能够测试DQN、PPO等强化学习算法在无人集群决策任务中的表现[163]. Rflysim[164]是北航可靠飞行控制组研制的无人集群系统，集成了视觉、算法与控制等技术，实现了软 件和硬件在环仿真. 飞思实验室(FeiSiLab)以Rflysim为平台，研发了多个无人集群仿真开发系统，并部署在国内多家实验室. 这些系统使用光学运动捕获个体的位置信息，采用视觉传感器采集环境数据，根据系统提供的应用程序接口部署RL算法，实现天地集群协同任务，如航迹规划、通信组网等. Carla[165]开源仿真平台构建了丰富的无人车模拟器，支持城市级别自动驾驶集群系统的开发. 该平台采用了服务器-客户端架构，其中服务器用于操控模型器、渲染环境等，客户端用于接受传感器数据，实现智能体与服务器之间的交互. 智能体从感知模块获得状态值，根据环境提供的奖励信号训练深度神经网络，决策出执行的动作，进而评估基于强化学习的自动驾驶方法的优劣[166]. 此外，Carla结合Gym开发了轻便的应用程序接口Gym-carla，能够支持多种强化学习算法. Gazebo[167]结合机器人操作系统(Robot Operating System, ROS)提供了3D动态机器人模拟器，设计

② https://github.com/microsoft/AirSim
③ http://www.feisilab.com/
④ https://classic.gazebosim.org/tutorials
⑤ http://carla.org/
⑥ https://github.com/deepmind/mujoco
⑦ https://gitee.com/hs-defense/moziai/
⑧ https://github.com/JSBSim-Team/jsbsim
⑨ https://github.com/cityflow-project/CityFlow
⑩ https://www.eclipse.org/sumo/
⑪ https://starcraft.com/zh-cn/
⑫ https://github.com/openai/multiagent-particle-envs
⑬ https://github.com/geek-ai/MAgent

了高保真的物理模型，可达到真实工业机器人控制水准．Gazebo能提供毫米范围内的良好精度，可部署基于值或基于策略的强化学习算法，指导多个机器人完成集群任务．MuJoCo[168]为跨平台的机器人动力学模拟仿真器，关注了机器人自身协调性与多机器人间的协作，如多机器人的多关节拆分等．为了支持大规模集群任务，MuJoCo通过部署强化学习算法，并行估计不同状态和动作的动态性，以获取集群最优控制策略．目前MuJoCo已被诸多RL算法作为对比实验平台．



(1) AirSim无人机巡航[162]　(2) AirSim无人机集群编队[163]

(3) 飞思实验室无人机集群系统$_3$　(4) 飞思实验室无人机无人车协同平台$_3$

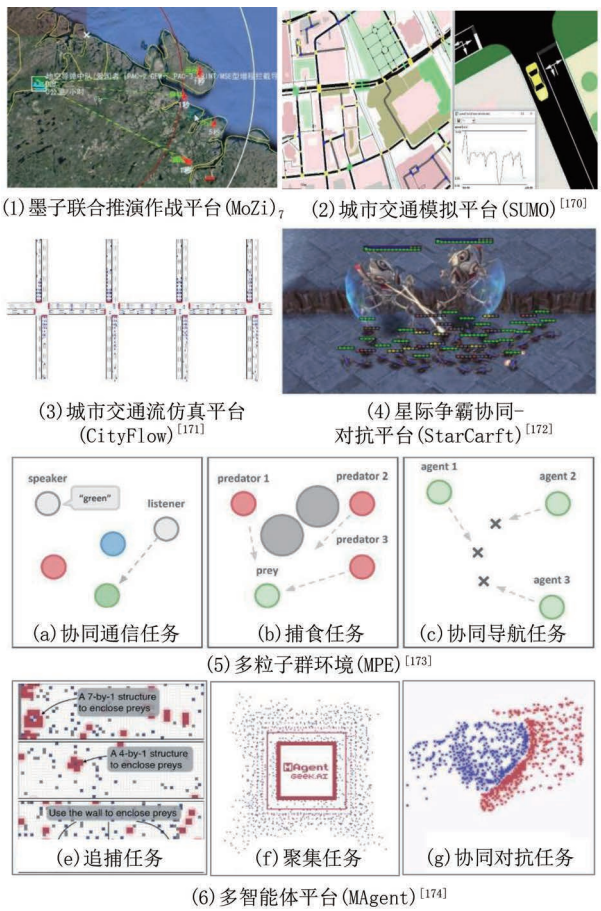
(5) Carla模拟仿真器[165]　(6) Gym-Carla模拟车辆自动驾驶[166]

(7) Gazebo机器人模拟仿真器[167]　(8) Mujoco多智能体模拟仿真器[168]

图 8　无人集群模拟仿真平台

指挥：现代海空行动(Command：Modern Air/Naval Operations，CMANO)是美国典型军事兵棋推演平台．北京华成公司受CMANO的启发，研发了墨子(Mozi)联合作战推演系统，其中墨子AI平台⑦可用于智能军事集群系统的开发与研究．墨子AI平台能够制定任务想定，明确作战目标，了解兵力集群编成，设定状态、动作、奖励，实现海陆空全域联合集群作战推演．强化学习算法依据系统特性，训练智能体的决策部署能力，进而提升反应速度、优化作战效能．JSBSim[169]是一个易融合于多平台的通用飞行动力学模型，能够部署到批处理的模拟飞行器上，或集成于虚幻等模拟环境中．JSBSim能够构造飞机、火箭等在自然作用力下的物理数学模型，为RL算法提供准确的武器状态数据．JSBSim为测试与训练真实空军战斗机提供了有效平台，美国国防部高级研究计划局在模拟空战比赛中，采用JSBSim模拟智能飞行器，并在多轮空战比赛中取得胜利．

模拟智能交通集群任务时，首先需从公开地图(OpenStreetMap，OSM)⑭获得路网信息，并从物联网设备获取车辆交通数据，再将数据信息加载于各个模拟平台．如图9的子图(2)所示，城市交通模拟



(1) 墨子联合推演作战平台(MoZi)$_7$　(2) 城市交通模拟平台(SUMO)[170]

(3) 城市交通流仿真平台(CityFlow)[171]　(4) 星际争霸协同-对抗平台(StarCarft)[172]

(a) 协同通信任务　(b) 捕食任务　(c) 协同导航任务

(5) 多粒子群环境(MPE)[173]

(e) 追捕任务　(f) 聚集任务　(g) 协同对抗任务

(6) 多智能体平台(MAgent)[174]

图 9　军事推演、交通管控与算法测试平台

(Simulation of Urban Mobility，SUMO)[170]平台是一个微观、多模态、空间连续、时间离散的交通仿真平台，能够读取不同规格的交通数据，依据需求生成道路信息，构建交通集群模型．交通控制接口(Traffic Control Interface，TraCI)为部署RL算法提供了在线控制接口和Python接口．由于SUMO无法扩展到大型路网和部署大量交通数据，相关研究

⑭ http://www.openstreetmap.org/

人员开发了城市交通流(CityFlow,CF)[171]平台,旨在支持灵活的路网和交通流.CF提供了用户友好的RL界面,且运行速度比SUMO快20倍.CF和SUMO已被公认为强化学习用于交通管控的基础平台,诸多算法在此平台上进行训练测试,如DQN、MARDDPG、CAMABRL等.

为了研发不受任务限制的通用算法,若干算法测试平台被相继开发,为在智能集群系统中部署RL算法奠定了理论基础.星际争霸(StarCraft)[172]是一款多智能体协同-对抗游戏,为MARL提供了多种任务环境,并发布了PyMARL⑮(Python MARL framework)开源项目.PyMARL囊括了最前沿的RL算法,如COMA[79]、VDN[88]、QMIX[89]、QTRAN[90]等,为训练与评估RL算法提供了标准对比平台.多智能体粒子群环境(Multi-agent Particle Environment, MPE)[173]将每个智能体模拟为粒子,粒子间通过协同或对抗完成多种任务类型,如通信、捕食、导航等,为应用于实际领域提供了理论基础.诸多RL算法使用MPE平台测试算法性能,如TarMAC[58]、PPO[43]、MADDPG[77]、OMAR[117]等.StarCraft和MPE平台重点关注了数十级别的智能体间的协同任务,为训练大规模智能集群的决策能力,MAgent[174]平台专注于支持数亿个智能体的集群任务.如图9的子图(4)至子图(6)所示,StarCraft、MPE、Magent为测试强化学习算法的集群决策性能提供了开源平台.但相对于面向应用的平台,面向算法测试的平台不具备精准物理动力学模型和逼真的实验环境.

4 研究展望

本文将智能集群系统和强化学习作为研究对象,讨论了面向智能集群系统的强化学习方法在联合通信、协同决策、奖励反馈、策略优化的相关工作,并介绍了典型应用领域和开源平台.最后,提出未来值得探讨的研究点,也是我们致力的研究方向:

(1)简洁有效的通信机制

在智能集群系统中,由于智能体具有部分可观性,需要确定通信时间、交互对象、传输信息以及信息的处理方式.目前,非连续的通信方式使得集群仅在必要时进行通信,分组通信综合了全通信与邻居通信模式来确认通信对象,交互局部通信信息的方式减轻了计算和存储的压力,基于互信息理论或注意力机制的处理方式能够提高通信的有效性.结合上述4个方面,以智能集群任务为核心,制定简明有效的感知通信机制,是实现集群协同交互的前提与保障.

(2)明确群间关系的局部离散协同架构

目前大多数的算法采用CTDE的协同架构,能够明确各个智能体的贡献,提升了算法性能.然而,其很难应用于参与个体多、环境动态复杂的真实任务场景.完全去中心化的协同架构虽然能够用于真实环境,但独立学习的方式会延迟智能涌现的时间.融合CTDE和完全去中心化的协同架构,设计明确贡献的局部离散协同架构,是集群智能涌现的关键.

(3)可衡量的奖惩反馈机制

在个体激发和群智汇聚的过程中,疏密不均的奖励机制会导致智能集群系统涌现出现象难解释、指标难量化、趋势难掌控的局面.在制定奖励反馈机制时,需均衡个体奖励与集群奖励,明确两者间的关系,注入以人为本的指导经验,设计可度量的奖惩反馈机制,在确保个体利益的前提,最大化群体奖励.

(4)稳健可信的策略优化方式

智能集群系统多面向民事应急、军事保障等复杂艰巨的任务场景.目前,强化学习算法通过干扰、约束等方式,采用人工干预、好奇心等机制,使智能体在最坏情况进行训练,进而训练出鲁棒、安全的策略,提升算法的泛化性和有效性.但是,这些优化方式存在"黑盒"现象,会出现参数难追溯、结构难分析的问题.面向智能集群系统的强化学习方法优化策略时,需要综合考虑任务需求,制定稳健可信的策略优化方式.

(5)大规模异构复杂模型

目前强化学习方法解决的智能集群任务规模多为中小型且只考虑同质个体,还未涉及到高复杂异构任务场景的应用.一个完整的集群任务不仅仅由同质群组构成,还可能由多种结构差异、目的多样、任务不同的异构群体构成.在同一环境执行任务时,异构群体会在多维空间中时空交汇,进行任务融合.针对以上问题,亟须构建大规模异构复杂模型,均衡异构群体间的利益,促进复杂任务的完成,提高作业的规模,保障任务的多样性.

⑮ https://github.com/oxwhirl/pymarl

(6) 人机物共融的协同机制

虽然在计算、执行等方面，机器的能力凌驾于人类之上. 但是，在探索任务时，人类的主观感知能力；处理状态时，人类的态势理解能力；决策制定时，人类的判断指导能力；奖励激励时，人类的反馈评价能力等，都对系统的智能化起到了决定性的作用. 鉴于此，在执行协同任务，需将人群的主观意图和能动意识与机器群体相融合，制定人机物共融的协同机制.

(7) 创建高保真公开集群数据集

高保真、强可信、数量足的数据集是策略制定的先决条件. 集群智能领域涵盖了数学、物理、心理、生物等多门学科，且应用场景多为保密级，难以获取源数据. 现有研究多依据自建的模拟平台，构建私有的仿真数据集合. 针对上述问题，面向不同任务场景和异构集群，获取其时空数据信息，通过清洗、归类、整合后开源布公，为构建数据驱动的协同模型提供有力支持.

(8) 融合多领域知识的虚拟-真实学习环境

目前，计算机图形学、具身学习、Sim2Real、数字孪生、平行系统等多领域为智能集群系统在虚拟环境下学习训练提供了可能. 以智能集群与任务场景的交互为切入点，构建能够同时在虚拟仿真空间和真实物理空间学习的环境，增强智能集群自主探索和交互能力，打通模拟环境与真实环境之间的壁垒，将是未来的一个重要研究方向.

(9) 搭造一体化多应用可落地平台

集群智能已渗透到各个领域，但是目前算法测试和落地应用的平台功能单一，且多关注同质集群的协同任务. 为实现智能集群系统的完全自主控制，需要衔接模型测试和落地应用，搭建各平台间的通信通道，权衡平台间的利益，统一技术体系和评估标准，进而使算法策略快速得到真实的反馈响应.

致 谢 本文受国家自然科学基金重点项目(批准号:62036010)，国家自然科学基金青年项目(批准号:62001422)，国家自然科学基金面上项目(批准号:61972362, 62372416)，国家重点研发计划课题(课题编号:2021YFB3301504)资助；感谢王华、李书攀等老师提出了许多富有启发的意见；感谢各位审稿人的专业严谨评审！

参 考 文 献

[1] Zhang Ting-Ting, Song Ai-Guo, Lan Yu-Shi. Adaptive structure modeling and prediction for swarm unmanned system. Scientia Sinica Informationis, 2020, 50(3): 347-362 (in Chinese)
(张婷婷，宋爱国，蓝羽石. 集群无人系统自适应结构建模与预测. 中国科学：信息科学，2020，50(3)：347-362)

[2] Wang Wei-Jia, Zheng Ya-Ting, Lin Guo-Zheng, et al. Swarm robotics: A review. Robot, 2020, 42(2): 232-256 (in Chinese)
(王伟嘉，郑雅婷，林国政等. 集群机器人研究综述. 机器人，2020，42(2)：232-256)

[3] Chen Wei-Neng, Lu Tun, Jiang Wei-Chuan, Tang yong, Wang Hua, Li Chao-Chao, Xu Ming-Liang. Advances and tendency of research on collective intelligence modeling and evolutionary computation. China Computer Federation Proceedings. Beijing, China: China Machine Press, 2021 (in Chinese)
(陈伟能，卢暾，蒋巍川，汤庸，王华，李超超，徐明亮. 群智建模与演化计算研究进展与趋势. 中国计算机科学技术发展报告. 北京，中国：机械工业出版社，2021)

[4] Malone T W, Bernstein M S. Handbook of collective intelligence. Cambridge, USA: MIT Press, 2022

[5] Pu Zhi-Qiang, Yi Jian-Qiang, Liu Zhen, et al. Knowledge-based and data-driven integrating methodologies for collective intelligence decision making: A survey. ACTA Automatica Sinica, 2022, 48(3): 1-17 (in Chinese)
(蒲志强，易建强，刘振等. 知识和数据协同驱动的群体智能决策方法研究综述. 自动化学报，2022，48(3)：1-17)

[6] Passino K M. Biomimicry of bacterial foraging for distributed optimization and control. IEEE Control Systems Magazine, 2002, 22(3): 52-67

[7] Karaboga D. An idea based on honey bee swarm for numerical optimization. Anatolia, Turkey: Erciyes university, Technical report: TR06, 2005

[8] Duan Hai-Bin, Zhang Dai-Feng, Fan Yan-Ming, et al. From wolf pack intelligence to UAV swarm cooperative decision-making. Scientia Sinica Informationis, 2019, 49(1): 112 - 118 (in Chinese)
(段海滨，张岱峰，范彦铭等. 从狼群智能到无人机集群协同决策. 中国科学：信息科学，2019，49(1)：112-118)

[9] Xu Ming-Liang, Jiang Hao, Jin Xiao-Gang, et al. Crowd simulation and its applications: Recent advances. Journal of Computer Science and Technology, 2014, 29(5): 799-811

[10] The State Council of People's Republic of China. New Generation Artificial Intelligence Development Plan, State Development No. 35, 2017 (in Chinese)
(中华人民共和国国务院. 新一代人工智能发展规划. 国发35号，2017)

[11] Chen W, Qiu X, Cai T, et al. Deep reinforcement learning for internet of things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1659-1692

[12] Zhao Qin-Ping. Ten scientific and technical problems in virtual reality. Scientia Sinica Informationis, 2017, 47(6): 800-803 (in Chinese)
(赵沁平. 虚拟现实中的10个科学技术问题. 中国科学: 信息科学, 2017, 47(6): 800-803)

[13] Zhu R, Wu S, Li L, et al. Context-aware multi-agent broad reinforcement learning for mixed pedestrian-vehicle adaptive traffic light control. IEEE Internet of Things Journal, 2022, 9(20): 19694-19705

[14] Li Y, Wu Q, Huang X, et al. Efficient adaptive matching for real-time city express delivery. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(6): 5767-5779

[15] Zheng P, Xia L, Li C, et al. Towards Self-X cognitive manufacturing network: An industrial knowledge graph-based multi-agent reinforcement learning approach. Journal of Manufacturing Systems, 2021, 61: 16-26

[16] Li W, Wang X, Jin B, et al. Structured cooperative reinforcement learning with time-varying composite action space. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 8618-8634

[17] Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nature medicine, 2019, 25(1): 24-29

[18] Li Ya-Fei, Wu Qing-Shun, Xu Ming-Liang, et al. Real-time scheduling for carrier-borne aircraft support operations: a reinforcement learning approach. Scientia Sinica Informationis, 2021, 51(2): 247-262 (in Chinese)
(李亚飞, 吴庆顺, 徐明亮等. 基于强化学习的舰载机保障作业实时调度方法. 中国科学: 信息科学, 2021, 51(2): 247-262)

[19] Luo Jie, Jiang Xin, Guo Bing-Hui, et al. Dynamic model and crowd entropy measurement of crowd intelligence system. Scientia Sinica Informationis, 2022, 52(1): 99-110 (in Chinese)
(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. 中国科学: 信息科学, 2022, 52(1): 99-110)

[20] Bonabeau E, Dorigo M, Theraulaz G. Inspiration for optimization from social insect behaviour. Nature, 2000, 406(6791): 39-42

[21] Li W, Wu W, Wang H, et al. Crowd intelligence in AI 2.0 era. Frontiers of Information Technology & Electronic Engineering, 2017, 18(1): 15-43

[22] Chen Jie, Fan Bang-Kui, Deng Fang, et al. Evolution and coordination of intelligent group systems. Academic Summary of the 252th Shuangqing Forum. China Science Foundation, 2021 (in Chinese)
(陈杰, 樊邦奎, 邓方等. 智能群系统的衍化与协同. 第252期双清论坛学术综述. 中国科学基金, 2021)

[23] Ma Hua-Dong, Zhao Dong, Wang Xin-Bing, et al. A novel crowdsensing system architecture model and its implementation methods. Scientia Sinica Informationis, 2023 (in Chinese)
(马华东, 赵东, 王新兵, 王甲海, 华蓓, 童剑军. 一种新型群智感知系统架构模型和实现方法. 中国科学: 信息科学, 2023)

[24] Sun Jia-Chen, Wang Jin-Long, Chen Jin, et al. Cooperative communication based on swarm intelligence: vision, model, and key technology. Scientia Sinica Informationis, 2020, 50(3): 307-317 (in Chinese)
(孙佳琛, 王金龙, 陈瑾等. 群体智能协同通信: 愿景、模型和关键技术. 中国科学: 信息科学, 2020, 50(3): 307-317)

[25] China Artificial Intelligence 2.0 Development Strategy Research Project Organization. Research on the Development Strategy of Artificial Intelligence 2.0 in China. Hangzhou, China: Zhejiang University Press, 2018 (in Chinese)
(中国人工智能2.0发展战略研究项目组. 中国人工智能2.0发展战略研究. 杭州, 中国: 浙江大学出版社, 2018)

[26] Minsky M L. Theory of neural-analog reinforcement systems and its application to the brain-model problem [Ph.D. Thesis]. Princeton University, USA, 1954

[27] Ren T, Niu J, Shu L, et al. Enabling efficient model-free control of large-scale canals by exploiting domain knowledge. IEEE Transactions on Industrial Electronics, 2020, 68(9): 8730-8742

[28] Sun Chang-Yin, Mu Chao-Xu. Important scientific problems of multi-agent deep reinforcement learning. Acta Automatica Sinica, 2020, 46(7): 1301-1312 (in Chinese)
(孙长银, 穆朝絜. 多智能体深度强化学习的若干关键科学问题. 自动化学报, 2020, 46(7): 1301-1312)

[29] Lv Shuai, Gong Xiao-Yu, Zhang Zheng-Hao, Han Shuai, Zhang Jun-Wei. Survey of deep reinforcement learning methods with evolutionary algorithms. Chinese Journal of Computers, 2022, 45(7): 1478-1499 (in Chinese)
(吕帅, 龚晓宇, 张正昊, 韩帅, 张峻伟. 结合进化算法的深度强化学习方法研究综述. 计算机学报, 2022, 45(7): 1478-1499)

[30] Li T, Zhu K, Luong N C, et al. Applications of multi-agent reinforcement learning in future internet: A comprehensive survey. IEEE Communications Surveys & Tutorials, 2022, 24(2): 1240-1279

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

[32] Sun Zheng-Lun, Qiao Peng, Dou Yong, Li Qing-Qing, Li Rong-Chun. PALA: Parallel actor-learner architecture for distributed deep reinforcement learning. Chinese Journal of Computers, 2023, 46(2): 229-243 (in Chinese)
(孙正伦, 乔鹏, 窦勇, 李青青, 李荣春. 面向执行-学习者的在线强化学习并行训练方法. 计算机学报, 2023, 46(2): 229-243)

[33] Puterman, Martin L. Markov decision processes: discrete stochastic dynamic programming. New York, USA: John Wiley & Sons, 2014

[34] Yang Y, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv: 2011.00583, 2020

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

[36] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. arXiv preprint arXiv: 1312.5602, 2013.

[37] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 2094-2100

[38] Wang Z, Schaul T, Hessel M, et al. Dueling network

architectures for deep reinforcement learning//Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York, USA, 2016: 1995-2003

[39] Hausknecht M, Stone P. Deep recurrent q-learning for partially observable mdps//Proceedings of the AAAI fall symposium series. Arlington, USA, 2015: 29-37

[40] Sutton R S, McAllester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation//Proceedings of the Conference on Neural Information Processing Systems, Denver, USA, 1999, 12: 1057-1063

[41] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms//Proceedings of the 31st International Conference on International Conference on Machine Learning. Beijing, China, 2014: 387-395

[42] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France, 2015: 1889-1897

[43] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms//Proceedings of the Conference on Neural Information Processing Systems, California, USA, 2017: 1-12

[44] Littman M L. Markov games as a framework for multi-agent reinforcement learning//Proceedings of the Eleventh International Conference on International Conference on Machine Learning, New Brunswick, USA, 1994: 157-163

[45] Amato C, Chowdhary G, Geramifard A, et al. Decentralized control of partially observable Markov decision processes//Proceedings of the 52nd IEEE Conference on Decision and Control, Firenze, Italy, 2013: 2398-2405

[46] Oliehoek F A, Amato C. A concise introduction to decentralized POMDPs. Cham, Switzerland: Springer Press, 2016

[47] Wu T, Zhou P, Liu K, et al. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. IEEE Transactions on Vehicular Technology, 2020, 69(8): 8243-8256

[48] Xu J, Kang X, Zhang R, et al. Optimization for master-UAV-powered auxiliary-aerial-IRS-assisted IoT networks: An option-based multi-agent hierarchical deep reinforcement learning approach. IEEE Internet of Things Journal, 2022, 9(22): 22887-22902

[49] Yang J, Wang E, Trivedi R, et al. Adaptive incentive design with multi-agent meta-gradient reinforcement learning//Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. New Zealand, 2022: 1436-1445

[50] Zhang S Q, Lin J, Zhang Q. A multi-agent reinforcement learning approach for efficient client selection in federated learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence, 2022, 36(8): 9091-9099

[51] Christianos F, Papoudakis G, Rahman M A, et al. Scaling multi-agent reinforcement learning with selective parameter sharing//Proceedings of the 38th International Conference on Machine Learning. 2021: 1989-1998

[52] Avalos R. Exploration and communication for partially observable collaborative multi-agent reinforcement learning//Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, New Zealand, 2022: 1829-1832

[53] Perrin S, Lauriere M, Pérolat J, et al. Mean field games flock! the reinforcement learning way//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada, 2021: 356-362

[54] Sun H, Han L, Yang R, et al. Optimistic curiosity exploration and conservative exploitation with linear reward shaping//Proceedings of the Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 1-16

[55] Wang Y, Zhong F, Xu J, et al. Tom2c: Target-oriented multi-agent communication and cooperation with theory of mind//Proceedings of the International Conference on Learning Representations. 2022: 1-17

[56] Foerster J N, Assael Y M, de Freitas N, et al. Learning to communicate with deep multi-agent reinforcement learning//Proceedings of the International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 2145-2153

[57] Sukhbaatar S, Szlam A, Fergus R. Learning multiagent communication with backpropagation//Proceedings of the International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 2252-2260

[58] Das A, Gervet T, Romoff J, et al. Tarmac: Targeted multi-agent communication//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 1538-1546

[59] Liu Y, Wang W, Hu Y, et al. Multi-agent game abstraction via graph attention neural network//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 7211-7218

[60] Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(05): 7211-7218

[61] Singh A, Jain T, Sukhbaatar S. Learning when to communicate at scale in multiagent cooperative and competitive tasks//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-15

[62] Niu Y, Paleja R R, Gombolay M C. Multi-agent graph-attention communication and teaming//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. 2021: 964-973

[63] Wang T, Wang J, Zheng C, et al. Learning nearly decomposable value functions via communication minimization//Proceedings of the International Conference on Learning Representation. UK, 2020: 1-15

[64] Jiang J, Lu Z. Learning attentional communication for multi-agent cooperation//Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 7265-7275

[65] Sheng J, Wang X, Jin B, et al. Learning structured communication for multi-agent reinforcement learning. Autonomous Agents and Multi-Agent Systems, 2022, 36(2): 1-31

[66] Kim D, Moon S, Hostallero D, et al. Learning to schedule communication in multi-agent reinforcement learning// Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-17

[67] Jiang J, Dun C, Huang T, et al. Graph convolutional reinforcement learning//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-13

[68] Tianshu Chu, Sandeep Chinchali and Sachin Katti. Multi-agent reinforcement learning for networked system control// Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-17

[69] Yali Du, Bo Liu, Vincent Moens, Ziqi Liu, Zhicheng Ren, Jun Wang, Xu Chen and Haifeng Zhang. Learning correlated communication topology in multi-agent reinforcement learning// Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. 2021:456-464

[70] Peng P, Wen Y, Yang Y, et al. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. arXiv preprint, arXiv: 1703.10069, 2017

[71] Li P, Tang H, Yang T, et al. PMIC: Improving multi-agent reinforcement learning with progressive mutual information collaboration//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 12979-12997

[72] Yuan L, Wang J, Zhang F, et al. Multi-agent incentive communication via decentralized teammate modeling//Proceedings of the Conference on Artificial Intelligence. Harvard, USA, 2022: 9466-9474

[73] Zhu C, Dastani M, Wang S. A Survey of Multi-agent reinforcement learning with communication. arXiv preprint, arXiv:2203.08975, 2022

[74] Kim W, Park J, Sung Y. Communication in multi-agent reinforcement learning: Intention sharing//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2021: 1-15

[75] Zhang K, Yang Z, Başar T. Handbook of Reinforcement Learning and Control: Multi-agent reinforcement learning: A selective overview of theories and algorithms. Cham, Switzerland: Springer, 2021

[76] Foerster J, Nardelli N, Farquhar G, et al. Stabilising experience replay for deep multi-agent reinforcement learning// Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 1146-1155

[77] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments//Proceedings of Conference and Workshop on Neural Information Processing Systems. Long Beach, USA, 2017: 6379-6390

[78] Gupta J K, Egorov M, Kochenderfer M. Cooperative multi-agent control using deep reinforcement learning//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. Sao Paulo, Brazil, 2017: 66-83

[79] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2974-2982

[80] Zhang T, Li Y, Wang C, et al. Fop: Factorizing optimal joint policy of maximum-entropy multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. 2021: 12491-12500

[81] Mao W, Yang L, Zhang K, et al. On improving model-free algorithms for decentralized multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 15007-15049

[82] Tan M. Multi-agent reinforcement learning: Independent vs. cooperative agents//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 1993: 330-337

[83] de Witt C S, Gupta T, Makoviichuk D, et al. Is independent learning all you need in the starcraft multi-agent challenge? . arXiv preprint arXiv:2011.09533, 2020

[84] Pham H X, La H M, Feil-Seifer D, et al. Cooperative and distributed reinforcement learning of drones for field coverage. arXiv preprint arXiv:1803.07250, 2018

[85] Nguyen M T, La H M, Teague K A. Collaborative and compressed mobile sensing for data collection in distributed robotic networks. IEEE Transactions on Control of Network Systems, 2017, 5(4): 1729-1740

[86] Zhang K, Yang Z, Liu H, et al. Fully decentralized multi-agent reinforcement learning with networked agents//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 5872-5881

[87] Zhang K, Yang Z, Basar T. Networked multi-agent reinforcement learning in continuous spaces//Proceedings of the IEEE Conference on Decision and Control. Florida, USA, 2018: 2771-2776

[88] Wang J, Ren Z, Han B, et al. Towards understanding cooperative multi-agent q-learning with value factorization// Proceedings of the Conference on Neural Information Processing Systems, 2021, 34: 29142-29155

[89] Rashid T, Samvelyan M, Schroeder C, et al. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning//Proceedings of International Conference on Machine Learning. Stockholm, Sweden, 2018: 4295-4304

[90] Son K, Kim D, Kang W J, et al. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 5887-5896

[91] Wang J, Ren Z, Liu T, et al. Qplex: Duplex dueling multi-agent q-learning//Proceedings of the International Conference on Learning Representations. 2022: 1-27

[92] Huang W, Li K, Shao K, et al. Multiagent q-learning with sub-team coordination//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. 2022: 1630-1632

[93] Hong Y, Jin Y, Tang Y. Rethinking individual global max in cooperative multi-agent reinforcement learning//Proceedings of

the Conference on Neural Information Processing Systems, New Orleans, USA, 2022: 1-17

[94] Christianos F, Schäfer L, Albrecht S. Shared experience actor-critic for multi-agent reinforcement learning//Proceedings of the Conference on Neural Information Processing Systems, 2020, 33: 10707-10717

[95] Wang Z, Zhang Y, Yin C, et al. Multi-agent deep reinforcement learning based on maximum entropy//Proceedings of the IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, Chongqing, China, 2021, 4: 1402-1406

[96] Mahajan A, Rashid T, Samvelyan M, et al. Maven: Multi-agent variational exploration//Proceedings of the Conference on Neural Information Processing Systems, 2019, 32: 7613-7624

[97] Wang Y, Han B, Wang T, et al. Dop: Off-policy multi-agent decomposed policy gradients//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-24

[98] McNally L, Brown S P, Jackson A L. Cooperation and the evolution of intelligence. Proceedings of the Royal Society B: Biological Sciences, 2012, 279(1740): 3027-3034

[99] Wang Z, Zhu H, He M, et al. Gan and multi-agent drl based decentralized traffic light signal control. IEEE Transactions on Vehicular Technology, 2021, 71(2): 1333-1348

[100] Wang Bo-Han, Wu Ting-Yu, Li Wen-Hao, et al. Large-scale UAVs confrontation based on multi-agent reinforcement learning. Journal of System Simulation, 2021, 33(8): 1739-1753 (in Chinese)
(王泊涵，吴婷钰，李文浩等. 基于多智能体强化学习的大规模无人机集群对抗. 系统仿真学报，2021, 33(8): 1739-1753)

[101] Wang L, Zhang Y, Hu Y, et al. Individual reward assisted multi-agent reinforcement learning//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 23417-23432

[102] Rahmattalabi A, Chung J J, Colby M, et al. D++: Structural credit assignment in tightly coupled multiagent domains//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Daejeon, South Korea, 2016: 4424-4429

[103] Mannion P, Devlin S, Duggan J, et al. Reward shaping for knowledge-based multi-objective multi-agent reinforcement learning. The Knowledge Engineering Review, 2018, 33(e23): 1-29

[104] Marthi B. Automatic shaping and decomposition of reward functions//Proceedings of the International Conference on Machine learning. Corvallis, USA, 2007: 601-608

[105] Majumdar S, Khadka S, Miret S, et al. Evolutionary reinforcement learning for sample-efficient multiagent coordination//Proceedings of the International Conference on Machine Learning. 2020: 6651-6660

[106] Li G, Gomez R, Nakamura K, et al. Human-centered reinforcement learning: A survey. IEEE Transactions on Human-Machine Systems, 2019, 49(4): 337-349

[107] MacGlashan J, Ho M K, Loftin R, et al. Interactive learning from policy-dependent human feedback//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 2285-2294

[108] Knox W B, Stone P. Interactively shaping agents via human reinforcement: The TAMER framework//Proceedings of the International Conference on Knowledge Capture. Redondo Beach, USA, 2009: 9-16

[109] Loftin R, MacGlashan J, Peng B, et al. A strategy-aware technique for learning behaviors from discrete human feedback//Proceedings of the AAAI Conference on Artificial Intelligence. Quebec City, Canada, 2014: 937-943

[110] Loftin R, Peng B, MacGlashan J, et al. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. Autonomous Agents and Multi-agent Systems, 2016, 30(1): 30-59

[111] Cederborg T, Grover I, Isbell C L, et al. Policy shaping with human teachers//Proceedings of the International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 3366-3372

[112] Griffith S, Subramanian K, Scholz J, et al. Policy shaping: Integrating human feedback with reinforcement learning//Proceedings of the International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2625-2633

[113] Xu M, Liu Z, Huang P, et al. Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability. arXiv preprint, arXiv:2209.08025, 2022

[114] Kumar A, Zhou A, Tucker G, et al. Conservative q-learning for offline reinforcement learning//Proceedings of the International Conference on Neural Information Processing Systems. 2020, 33: 1179-1191

[115] Fujimoto S, Gu S S. A minimalist approach to offline reinforcement learning//Proceedings of the International Conference on Neural Information Processing Systems. 2021, 34: 20132-20145

[116] 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

[117] Pan L, Huang L, Ma T, et al. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 17221-17237

[118] Hu J, Hu S, Liao S. Policy regularization via noisy advantage values for cooperative multi-agent actor-critic methods//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. London, UK, 2022: 1-10

[119] Li S, Wu Y, Cui X, et al. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33(01): 4213-4220

[120] Zhang K, Sun T, Tao Y, et al. Robust multi-agent reinforcement learning with model uncertainty//Proceedings of the International Conference on Neural Information Processing

Systems. 2020, 33: 10571-10583
[121] Chow Y, Ghavamzadeh M, Janson L, et al. Risk-constrained reinforcement learning with percentile risk criteria. The Journal of Machine Learning Research, 2017, 18(1): 6070-6120
[122] Tessler C, Mankowitz D J, Mannor S. Reward constrained policy optimization. arXiv preprint, arXiv:1805.11074, 2018
[123] Pan L, Rashid T, Peng B, et al. Regularized softmax deep multi-agent q-learning//Proceedings of the Conference on Neural Information Processing Systems, 2021, 34: 1365-1377
[124] Achiam J, Held D, Tamar A, et al. Constrained policy optimization//Proceedings of the International Conference on Machine Learning, Sydney. Australia, 2017: 22-31
[125] Yang Q, Simão T D, Tindemans S H, et al. WCSAC: Worst-case soft actor critic for safety-constrained reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 10639-10646
[126] Liu Z, Cen Z, Isenbaev V, et al. Constrained variational policy optimization for safe reinforcement learning//Proceeding of the International Conference on Machine Learning. Baltimore, USA, 2022: 13644-13668
[127] Saunders W, Sastry G, Stuhlmueller A, et al. Trial without error: Towards safe reinforcement learning via human intervention//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. Stockholm, Sweden, 2018:2067-2069
[128] Xu Y, Liu Z, Duan G, et al. Look before you leap: Safe model-based reinforcement learning with human intervention//Proceedings of the Conference on Robot Learning. Auckland, New Zealand, 2022: 332-341
[129] Hüttenrauch M, Adrian S, Neumann G. Deep reinforcement learning for swarm systems. Journal of Machine Learning Research, 2019, 20(54): 1-31
[130] Zaheer M, Kottur S, Ravanbakhsh S, et al. Deep sets//Proceedings of the International Conference on Neural Information Processing Systems. Long Beach, USA, 2017, 3394-3404
[131] Long Q, Zhou Z, Gupta A, et al. Evolutionary population curriculum for scaling multi-agent reinforcement learning//Proceedings of International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020:1-18
[132] Liu I J, Yeh R A, Schwing A G. PIC: Permutation invariant critic for multi-agent deep reinforcement learning//Proceedings of the 3rd Conference on Robot Learning. Osaka, Japan, 2020: 590-602
[133] Wang W, Yang T, Liu Y, et al. From few to more: Large-scale dynamic multiagent curriculum learning//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(05): 7293-7300
[134] Peng H, Du B, Liu M, et al. Dynamic graph convolutional network for long-term traffic flow prediction with reinforcement learning. Information Sciences, 2021, 578: 401-416
[135] Kiran B R, Sobh I, Talpaert V, et al. Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(6): 4909-4926
[136] Nazari M, Oroojlooy A, Snyder L, et al. Reinforcement learning for solving the vehicle routing problem//Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 9861-9871
[137] Zhu R, Li L, Wu S, et al. Multi-agent broad reinforcement learning for intelligent traffic light control. Information Sciences, 2023, 619: 509-525
[138] Hu H, Yang X, Xiao S, et al. Anti-conflict AGV path planning in automated container terminals based on multi-agent reinforcement learning. International Journal of Production Research, 2021, 61(1): 1-16
[139] Hachiya D, Mas E, Koshimura S. A reinforcement learning model of multiple UAVs for transporting emergency relief supplies. Applied Sciences, 2022, 12(20): 10427
[140] Jeon S, Lee H, Kaliappan V K, et al. Multiagent reinforcement learning based on fusion-multi-actor-attention-critic for Multiple-Unmanned-Aerial-Vehicle Navigation Control. Energies, 2022, 15(19): 7426
[141] Pane Y P, Nageshrao S P, Kober J, et al. Reinforcement learning based compensation methods for robot manipulators. Engineering Applications of Artificial Intelligence, 2019, 78: 236-247
[142] Yu C, Wang X, Feng Z. Coordinated multiagent reinforcement learning for teams of mobile sensing robots//Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. Montreal, Canada, 2019: 2297-2299
[143] Zhu R, Li S, Wang P, et al. Energy-efficient deep reinforced traffic grooming in elastic optical networks for cloud - fog computing. IEEE Internet of Things Journal, 2021, 8(15): 12410-12421
[144] Hu X, Liao X, Liu Z, et al. Multi-agent deep reinforcement learning-based flexible satellite payload for mobile terminals. IEEE Transactions on Vehicular Technology, 2020, 69(9): 9849-9865
[145] Ling M H, Yau K L A, Qadir J, et al. A reinforcement learning-based trust model for cluster size adjustment scheme in distributed cognitive radio networks. IEEE Transactions on Cognitive Communications and Networking, 2018, 5(1): 28-43
[146] NguyenA. C., PamukluT., SyedA., KennedyW. S. and Erol-KantarciM., Reinforcement learning-based deadline and battery-aware offloading in smart farm IoT-UAV networks//Proceedings of the IEEE International Conference on Communications. Foshan, China, 2022: 189-194
[147] Din A, Ismail M Y, Shah B, et al. A deep reinforcement learning-based multi-agent area coverage control for smart agriculture. Computers and Electrical Engineering, 2022, 101: 108089
[148] Pamuklu T, Nguyen A C, Syed A, et al. IoT-Aerial base station task offloading with risk-sensitive reinforcement learning for smart agriculture. IEEE Transactions on Green Communications and Networking, 2022, 7(1): 171-182
[149] Wiberg V, Wallin E, Nordfjell T, et al. Control of rough terrain vehicles using deep reinforcement learning. IEEE

Robotics and Automation Letters, 2021, 7(1): 390-397

[150] Gao X, Jin Y, Dou Q, et al. Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search//Proceedings of the IEEE International Conference on Robotics and Automation. Paris, France, 2020: 8440-8446

[151] Pore A, Corsi D, Marchesini E, et al. Safe reinforcement learning using formal verification for tissue retraction in autonomous robotic-assisted surgery//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. 2021: 4025-4031

[152] Lee H R, Lee T. Multi-agent reinforcement learning algorithm to solve a partially-observable multi-agent problem in disaster response. European Journal of Operational Research, 2021, 291(1): 296-308

[153] Xue Jun-Xiao, Kong Xiang-Yan, Dong Bo-Wei, et al. Obstacle avoidance and simulation of carrier-based aircraft on the deck of aircraft carrier. Journal of System Simulation, 2021, 35(3):592-603 (in Chinese)
(薛均晓，孔祥燕，董博威，陶浩，管海洋，石磊，徐明亮. 航母甲板上舰载机的混合避障和仿真. 系统仿真学报，2021，35(3)：592-603)

[154] Xue Jun-Xiao, Kong Xiang-Yan, Guo Yi-Bo, et al. Dynamic obstacle avoidance method for carrier aircraft based on deep reinforcement learning. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(7): 1102-1112 (in Chinese)
(薛均晓，孔祥燕，郭毅博等. 基于深度强化学习的舰载机动态避障方法. 计算机辅助设计与图形学学报，2021，33(7)：1102-1112)

[155] Lloyd S P, Witsenhausen H S. Weapons allocation is NP-complete//Proceedings of the Summer Computer Simulation Conference. Reno, USA, 1986: 1054-1058

[156] Luo W, Lü J, Liu K, et al. Learning-based policy optimization for adversarial missile-target assignment. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 52(7): 4426-4437

[157] Wang Liang, Wang Wen, Wang Yu-You, et al. Feasibility of reinforcement learning for UAV-based target searching in a simulated communication denied environment. Scientia Sinica Informationis, 2020, 50(3): 375-395 (in Chinese)
(汪亮，王文，王禹又等. 强化学习方法在通信拒止战场仿真环境中多无人机目标搜寻问题上的适用性研究. 中国科学：信息科学，2020，50(3)：375-395)

[158] Melo F S, Ribeiro M I. Q-Learning with linear function approximation//Proceedings of International Conference on Computational Learning Theory. San Diego, USA, 2007: 308-322

[159] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning//Proceedings of International Conference on Machine Learning. New York, USA, 2016: 1928-1937

[160] Abbeel P, Schulman J. Deep reinforcement learning through policy optimization//Proceedings of Conference of Neural Information Processing Systems. Barcelona, Spain, 2016: 1-120

[161] Fu Q, Qiu T, Yi J, et al. Concentration network for reinforcement learning of large-scale multi-agent systems//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36: 9341-9349

[162] Shah S, Dey D, Lovett C, et al. Airsim: High-fidelity visual and physical simulation for autonomous vehicles//Proceedings of the Conference on Field and Service Robotics. Zurich, Switzerland, 2018: 621-635

[163] Vemprala S, Mian S, Kapoor A. Representation learning for event-based visuomotor policies//Proceedings of the Conference on Neural Information Processing Systems, 2021, 34: 4712-4724

[164] Dai X, Ke C, Quan Q, et al. RFlySim: Automatic test platform for UAV autopilot systems with FPGA-based hardware-in-the-loop simulations. Aerospace Science and Technology, 2021, 114: 106727

[165] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An open urban driving simulator//Proceedings of the Conference on Robot Learning. Mountain View, USA, 2017: 1-16

[166] Chen J, Li S E, Tomizuka M. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(6): 5068-5078

[167] Lopez N G, Nuin Y L E, Moral E B, et al. Gym-gazebo2, a toolkit for reinforcement learning using ROS 2 and Gazebo. arXiv preprint, arXiv:1903.06278, 2019

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

[169] Berndt J. JSBSim: An open source flight dynamics model in C++//Proceedings of the AIAA Modeling and Simulation Technologies Conference and Exhibit. Providence, USA, 2004: 4923

[170] Lopez P A, Behrisch M, Bieker-Walz L, et al. Microscopic traffic simulation using sumo//Proceedings of the Conference on intelligent transportation systems. Maui, USA, 2018: 2575-2582

[171] Zhang H, Feng S, Liu C, et al. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario//Proceedings of the World Wide Web Conference. San Francisco, USA, 2019: 3620-3624

[172] Samvelyan M, Rashid T, Schroeder de Witt C, et al. The StarCraft multi-agent challenge//Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. Montreal, Canada, 2019: 2186-2188

[173] Mordatch I, Abbeel P. Emergence of grounded compositional language in multi-agent populations//Proceedings of the 31th AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 1495-1502

[174] Zheng L, Yang J, Cai H, et al. Magent: A many-agent reinforcement learning platform for artificial collective intelligence//Proceedings of the 31th AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 8222-822

LI Lu-Lu, Ph. D. candidate. Her research interests include reinforcement learning, collective intelligence system, etc.

ZHU Rui-Jie, Ph. D., His research interests include collective intelligence system, machine learning, etc.

SUI Lu-Yao, M. S. candidate. Her research interests include reinforcement learning, collective intelligence system, etc.

LI Ya-Fei, Ph. D., associate professor. His research interests include machine learning, urban computing, collective intelligence system, etc.

XU Ming-Liang, Ph. D., professor. His current research interests include machine learning, collective intelligence system, etc.

FAN Hui-Tao, Ph. D., researcher, member of the Chinese Academy of Engineering. His current research interests include aircraft design, collective intelligence system, etc.

Background

Intelligent Collective System (ICS) is a system composed of agents gathered in space and time, and agents complete tasks in a collaborative way. The emerging intelligence form ICS is Collective Intelligence (CI). The assignments of ICS are complex, and the environment is dynamic.

Reinforcement Learning (RL) is a kind of end-to-end method that combines perception and decision-making. It has an autonomous learning capability through trial and error iterative optimization. To match the complexity and diversity of the natural world, RL develops from Single-Agent Reinforcement Learning methods (SARL) to Multi-Agent Reinforcement Learning (MARL). The research direction of RL evolved from targeted decision-making with a single objective to cooperative multiple agents. RL provides a new way to improve the performance of ICS. Specifically, the agents of ICS need joint decision-making to achieve CI, and obtain feedback of reward mechanism, then use the feedback to optimize the policy. RL proposes the generalized paradigm to construct the process for making decisions.

In this paper, we first analyze the RL method in ICS from communication, cooperation, reward assessment and policy optimization, then introduce the typical applications, and list the open-source platforms. Finally, the future directions of RL methods for ICS have been discussed.

This project is supported by the Key Project of the National Natural Science Foundation of China (Grant No. 62036010), the National Natural Sciences Foundation of Youth Project (Grant No. 62001422), the National Natural Science Foundation of China (Grant No. 61972362 62372416), and the National Key Research and Development Program Project (Topic No. 2021YFB3301504).