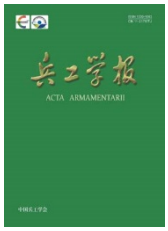


兵工学报
Acta Armamentarii
ISSN 1000-1093,CN 11-2176/TJ

《兵工学报》网络首发论文

题目：稀疏奖励下基于强化学习的无人集群自主决策与智能协同
作者：李超，王瑞星，黄建忠，江飞龙，魏雪梅，孙延鑫
收稿日期：2022-03-21
网络首发日期：2022-08-26
引用格式：李超，王瑞星，黄建忠，江飞龙，魏雪梅，孙延鑫．稀疏奖励下基于强化学习的无人集群自主决策与智能协同[J/OL]．兵工学报．
https://kns.cnki.net/kcms/detail/11.2176.TJ.20220825.1755.011.html

网络首发时间：2022-08-26 10:39:08
网络首发地址：https://kns.cnki.net/kcms/detail/11.2176.TJ.20220825.1755.011.html

稀疏奖励下基于强化学习的无人集群自主决策与智能协同

李超[1,2]，王瑞星[1]，黄建忠[1]，江飞龙[3]，魏雪梅[1]，孙延鑫[1]

（1. 中国兵器工业试验测试研究院 技术中心，陕西 西安 710116；2. 南京理工大学 机械工程学院，江苏 南京 210094；
3. 哈尔滨工业大学 航天学院，黑龙江 哈尔滨 150001）

摘要：无人集群将深刻地塑造战争样式，为提升无人集群自主决策算法能力，对异构无人集群攻防对抗自主决策方法进行研究。对无人集群对抗模型设计进行总体概述，并对无人集群攻防对抗场景进行模型设计；针对无人集群自主决策采用强化学习技术广泛存在的稀疏奖励问题，提出基于局部回报重塑的奖励机制设定方法；在此基础上叠加优先经验回放，有效地改善稀疏奖励问题；通过程序仿真和演示系统设计，验证该方法的优越性。该方法的研究将加速基于强化学习技术的无人集群自主决策算法网络收敛过程，对无人集群自主决策算法研究具有重要意义。

关键词：多智能体；无人智能；博弈对抗；强化学习；稀疏奖励

中图分类号：TP18　　**文献标志码：**A　　**文章编号：**1000-1093(202X)XX -XX-XXX

DOI：10.12382/bgxb.2022.0177

Autonomous Decision-making and Intelligent Collaboration of Unmanned Clusters Based on Reinforcement Learning under Sparse Rewards

LI Chao[1,2], WANG Ruixing[1], HUANG Jianzhong[1], JIANG Feilong[3], WEI Xuemei[1], SUN Yanxin[1]

（1. Technology Center, Norinco Group Test and Measuring Academy, Xi'an 710116, Shaanxi, China；2. School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, China；3. School of Astronautics, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China）

Abstract: Unmanned swarms will profoundly shape the pattern of warfare. In order to improve the autonomous decision-making algorithm capability of unmanned swarms, the autonomous decision-making method of heterogeneous unmanned swarm attack and defense confrontation was studied. An overview of the design of the unmanned swarm confrontation model and the model design of the unmanned swarm attack-defense confrontation scenario are carried out. Aiming at the sparse reward problem that the reinforcement learning technology widely exists in the autonomous decision-making of the unmanned swarm, a reward based on Local Reward Reshaping is proposed. And on this basis, the Prioritized Experience Replay is superimposed, which effectively improves the sparse reward problem. Finally, the superiority of this method is verified by program simulation and demonstration system design. This research will accelerate the network convergence process of the autonomous decision-making algorithm for unmanned swarms based on reinforcement learning technology, which is of great significance to the research on autonomous decision-making algorithms of unmanned swarms.

Keywords: multi-agent; unmanned intelligent; game confrontation; reinforcement learning; sparse reward

收稿日期：2022-03-21
基金项目：国防科技创新特区重点项目(20-163-30-ZT-004-015-01)
作者简介：李超(1986—)，男，副研究员，博士研究生。E-mail: lichao24215@126.com
通信作者：王瑞星(1997—)，男，助理研究员，硕士。E-mail: hitwangruixing@163.com

0 引言

二战结束以来，尽管大规模世界战争未有发生，但局部性战争却从未停止，从朝鲜战争到阿富汗战争再到纳卡战争以及硝烟弥漫的俄乌战场，科技力量带来的加成逐渐显现，尤其是新世纪发生的几次战争中，无人智能装备发挥了重要的作用[1]。

未来，无人智能集群作战将会是典型的作战模式。而无人集群的最终应用离不开无人集群对抗建模及群体智能演化机理、无人集群探测识别与态势感知、无人集群通信、无人集群导航、无人集群自主决策、无人集群运动控制以及无人集群对抗策略迁移与泛化[2]、无人集群试验与评估[3]等技术研究。其中在无人集群自主决策研究领域[4]，强化学习技术被广泛使用。

多智能体系统，由一系列相互作用的智能体构成，多个智能体之间通过相互通信、合作、竞争等方式，完成单个智能体不能完成的，大量而又复杂的工作。目前，结合多智能体系统和强化学习方法形成的多智能体强化学习正逐渐成为研究热点[5]。

如图 1 所示，多智能体强化学习技术框架包含环境、智能体两部分，智能体 n 感知环境状态，输出状态矩阵 s_n，输出的状态组成联合状态集$(s_1,s_2,s_3,\cdots,s_n)$，并依据策略网络选择动作 a_n，输出的动作组成联合动作集$(a_1,a_2,a_3,\cdots,a_n)$，作用于环境，环境依据联合动作给予对应奖励 r_n 组成总奖励集$(r_1,r_2,r_3,\cdots,r_n)$并更新状态[6]。在多智能体与环境交互过程中，奖励为多智能体策略迭代的重要依据。丰富的奖励反馈，可以有效引导多智能体学习到最优动作策略，但在强化学习技术的应用领域中，奖励稀疏性问题广泛存在。尤其随着深度学习技术与强化学习技术的深度融合，深度神经网络被应用到强化学习应用领域之后，因网络训练过程需要大量样本支撑，稀疏奖励问题也就愈加凸显[7]。

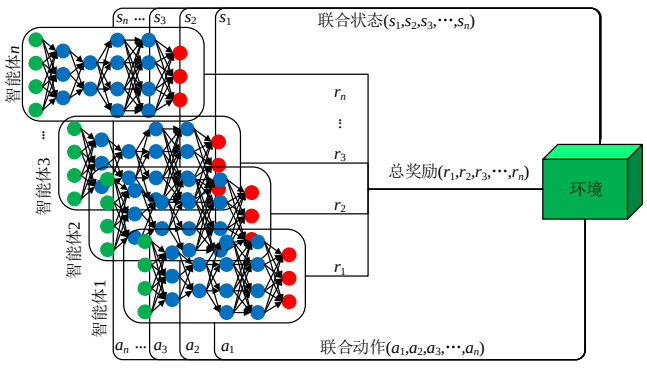


图 1 多智能体强化学习原理图

Fig.1 Multi-agent reinforcement learning schematic

针对广泛存在的奖励稀疏性，奖励塑造利用经验知识人工设计奖励函数以扩充奖励体系引导智能体进行最优策略学习[8-9]。课程学习通过不断增加任务难度以改善奖励稀疏造成的网络收敛缓慢问题[10]。事后经验回放是一种从失败经历中提取有效信息的强化学习方法，通过对失败经历进行处理产生奖励信息，解决奖励的稀疏性问题[11]。分层强化学习通过缩小各层策略的动作序列空间，提高解决稀疏奖励问题的能力[12]。现有奖励体制机制研究多针对于单智能体策略学习问题，且仿真或试验的场景设定较为简单，状态-动作空间维度较低[13-15]。

针对基于强化学习的无人集群自主决策与智能协同策略学习这一多智能体问题存在的奖励稀疏性，建立了无人集群攻防对抗任务场景模型，并提出了基于局部回报重塑的奖励机制设定方法，在此基础上叠加优先经验回放(PER)，通过程序仿真及演示系统验证，本研究有效地改善了奖励稀疏性，极大提升了策略学习的效率。

1 无人集群对抗模型设计

针对无人集群对抗问题特点，设计的模型框架应包含以下 3 层内容：

1) 场景层：该层主要对无人集群对抗的场景类别和场景特点进行设计。明确场景目标、场景构成，为后续无人集群对抗模型设计奠定基础。

2) 单元层：该层主要对对抗场景下单元数量及单元属性进行设计。其中异构无人集群对抗还需对单元种类进行设计。通常，异构无人集群对抗可包含探测单元、防御单元和攻击单元等。另外在单元属性方面，可设计生命属性、移动属性、探测属性、攻击属性、防御属性等。

3) 规则层：该层应明确集群对抗双方在具体对抗场景下的博弈策略及胜负判别规则。

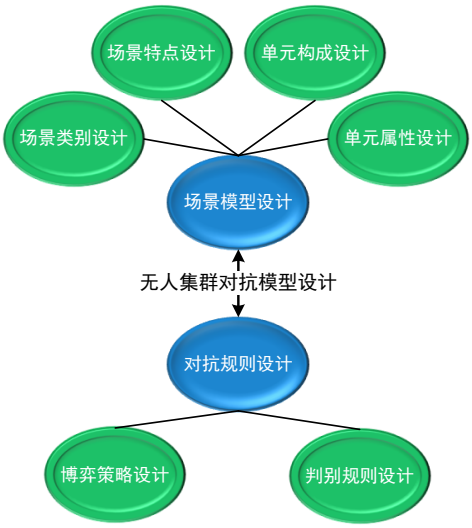


图 2 无人集群对抗模型构成

Fig.2 Composition of the unmanned cluster confrontation model

在模型设计时，场景层、单元层、规则层设计可以划分为两大过程，如图 2 所示。场景模型设计，包含场景类别设计、场景特点设计、单元构成设计及单元属性设计。对抗规则设计，包含无人集群对抗双方的博弈策略设计以及对抗过程的判别规则设计。

1.1 无人集群攻防对抗场景模型设计

基于图 2 所示模型框架，本研究设计了无人集群攻防对抗场景模型，攻防对抗为无人集群对抗典型任务场景。在单元种类方面，设计攻击、探测、防御 3 种单元。在单元数量方面，攻击单元、防御单元、探测单元分别为 6 个、4 个、2 个。在任务目标方面，基于蓝方采用设定策略的前提，通过基于强化学习的自主决策与智能协同技术，使得红方单元学习到比蓝方更优的博弈策略。

图 3 为无人集群攻防对抗仿真模型红蓝初始站位图，其中正方形框线表示红蓝对抗区域，双方无人集群智能单元在对抗区域两侧一字排开。在仿真示意方面，双方攻击单元、防御单元、探测单元及生命值、防御范围、探测范围等单元属性示意如图所示[16-18]。图 3 中，AT、DE、DT 分别表示红蓝双方攻击单元、防御单元、探测单元存活数。

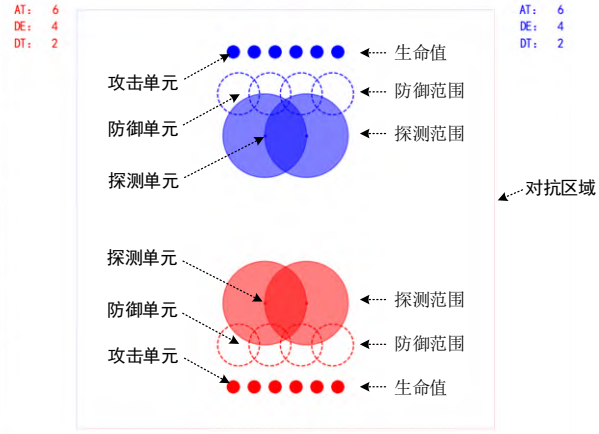


图 3 红蓝无人集群攻防对抗仿真模型初始站位图

Fig.3 The initial site map of the red and blue unmanned cluster attack and defense confrontation simulation model

1.2 无人集群攻防对抗规则设计

无人集群对抗规则设计包含博弈策略设计及判别规则设计，其中博弈策略设计包含集群对抗双方在任务场景下的博弈对抗策略。针对本研究所设计的无人集群攻防对抗任务场景，红蓝对抗双方的博弈策略如图 4 所示。

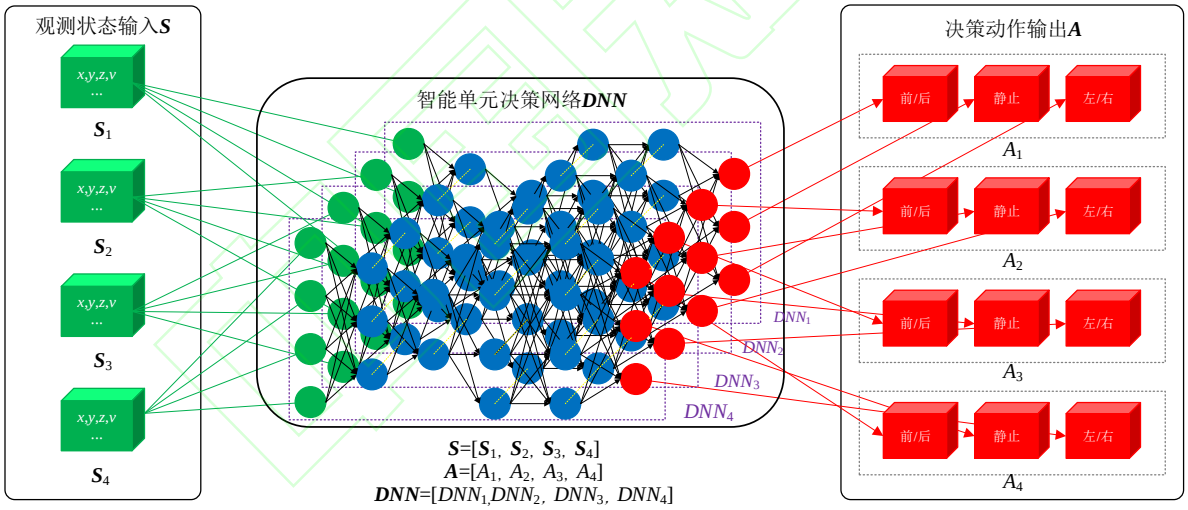


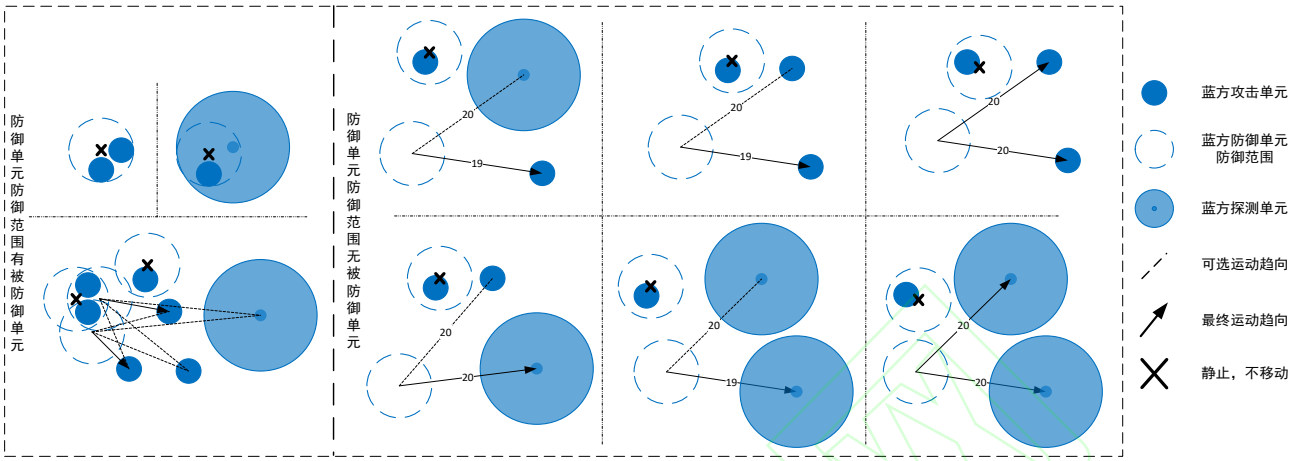
图 4 红方智能单元自主决策原理示意图

Fig.4 Schematic diagram of the autonomous decision-making principle for the red intelligent units

红方为基于深度神经网络的自主决策单元，以图 4 为示意，若红方无人集群包含 4 个智能单元，智能单元决策网络 ***DNN*** 则相应设计 4 个智能体的自主决策 DNN_1、DNN_2、DNN_3、DNN_4，在网络设计过程中，每个智能体的自主决策在结构上独立，但在参数上存在耦合，从而使得多类多个智能单元具备自主决策能力的同时具备协同能力涌现的潜能。在决策网络输入输出方面，以智能体观测状态如自身/对方位置、速度、数量等参数为状态输入，$\boldsymbol{S}_n$为第 n 个智能体观测状态，包含位置(x,y,z)、速度 v 等状态量，多个智能体观测状态组成智能体联合状态集 $\boldsymbol{S}$。以智能体移动/静止状态选择、移动方向为动作输出[19-21]，多个智能体动作输出组成智能体联合动作集 $\boldsymbol{A}$。

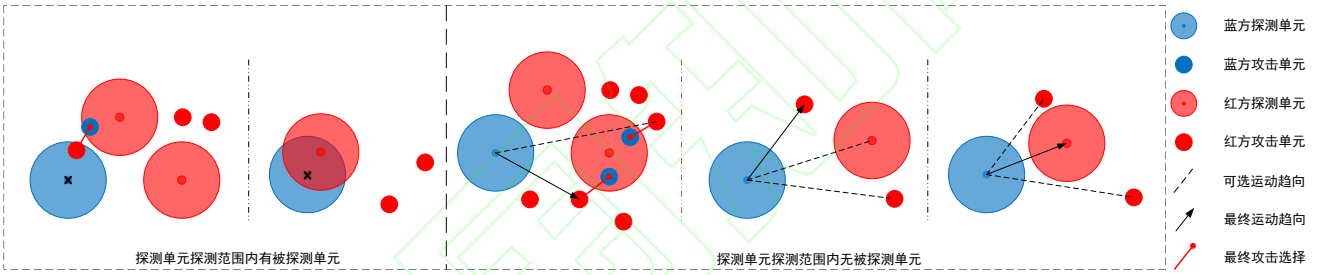
蓝方单元采用既定博弈策略，通过对无人智能单元集群作战战术战法的深入了解，设计了蓝方单元博弈策略。针对防御属性、探测属性、攻击属性，依据各属性范围内有无被防御、被探测、被攻击单元将 3 种智能单元的移动策略分别归为两类。其中针对防御单元，如图 5(a)所示，当单个单元防御范围有被防御单元时，防御单元静止；当多个防御单元防御范围重叠处有被防御单元，其中一个防御单元静止，其余单元被认定为防御范围内无被防

御单元；当防御范围内无被防御单元时，趋向最近的需被防御单元。针对探测单元，如图 5(b)所示，当探测范围内有被探测单元，探测单元静止；当探测范围内无被探测单元时，如存在未被探测到攻击己方单元的敌方攻击单元，则趋向最近的该类型单元；否则趋向最近的被探测单元。针对进攻单元，如图 5(c)所示，当攻击范围内无处于探测单元探测视角下的被攻击单元，则趋向最近的该类型单元；否则，攻击单元静止并转为攻击状态。



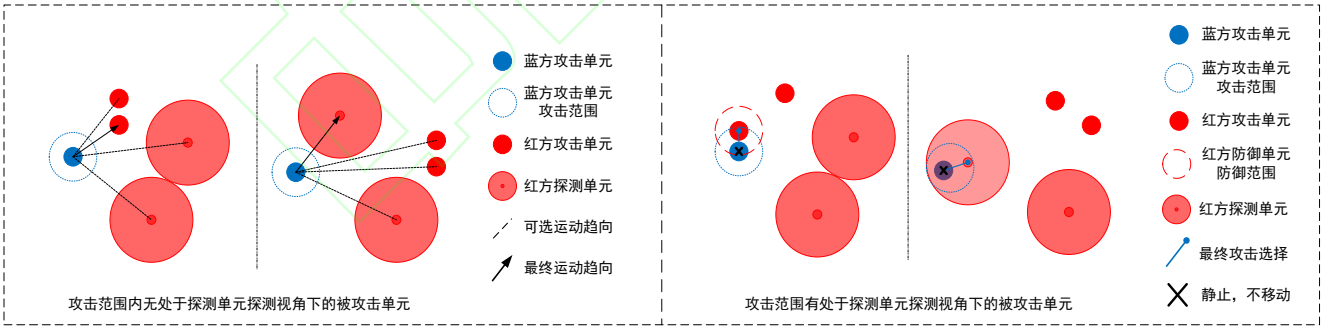
(a) 防御单元移动策略示意图

(a) Schematic diagram of defense unit movement strategy



(b) 探测单元移动策略示意图

(b) Schematic diagram of detection unit movement strategy



(c) 攻击单元移动策略示意图

(c) Schematic diagram of attack unit movement strategy

图 5 蓝方智能单元博弈对抗策略示意图

Fig.5 Schematic diagram of the game confrontation strategy of the blue intelligent units

判别规则设计包含对抗过程有效性判别及对抗终局胜负性判别。有效性判别方面，双方应在对抗区域内、设定属性限制下进行对抗。胜负性判别方面，为考察智能单元自主决策算法的学习效率，在双方单次对局中，设置最大仿真步。未达到最大仿真步时，若一方智能单元中探测或攻击单元被全部消灭，判定该方对局失败。达到最大仿真步时，对局结束，对所有智能单元的剩余总血量进行比较，总血量大的一方对局胜利，相同则判定平局。

2 稀疏奖励解决方法

2.1 基于局部回报重塑的无人攻防集群对抗奖励工程设定

无人集群对抗领域问题在应用强化学习技术时，现有的奖励体系依据对局是否胜利进行奖励反

馈，对局胜利给予奖励，对局失败无奖励。上述为稀疏奖励的一种极端形式，名为二元奖励。在该奖励机制下，策略网络训练过程会被严重滞缓甚至策略网络根本无法收敛。

本研究提出基于局部回报重塑的奖励工程设计方法。即首先将任务分解为多个子任务，对应明确任务目标及子目标，在细分的过程中确定任务执行主体与子目标之间的逻辑关系。以异构无人集群对抗场景为例，因不同种类的智能体特点属性不同，在任务中扮演的角色也有所不同。因此可以有针对性地对任务目标进行分解，适配不同种类智能体的功能属性[22]。

针对本文研究的无人集群攻防对抗任务场景，在设定奖励机制的过程中，依据不同种类智能单元的属性特点分别设计奖励函数。针对攻击单元，鼓励攻击、低血量退避、躲避敌方探测单元等行为。针对防御单元，鼓励有效防御行为。针对探测单元，鼓励有效探测、躲避敌方攻击、躲避敌方探测单元等行为。针对所有单元，鼓励尽快结束回合行为，以上各种奖励引导项的最终目标为短时间内消灭对方智能单元从而在单次对局中获胜。上述定性设计结果如图 6 所示。

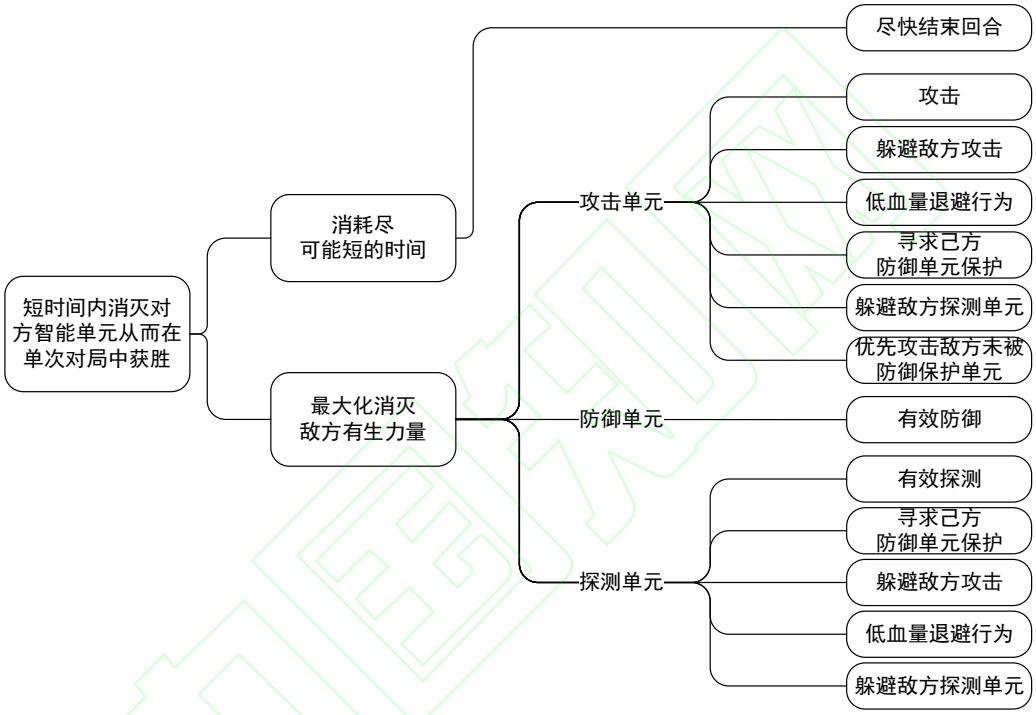


图 6 无人集群攻防对抗场景下基于局部回报重塑的奖励工程设定

Fig.6 Reward engineering setting based on Local Reward Reshaping in unmanned swarm attack and defense confrontation scenarios

定量奖励函数设计方面，目前奖励函数中奖惩数值设计主要依靠经验。本研究场景下的奖励函数为

$$r=\sum_{i}^{N} r_i \tag{1}$$

$$r_i=\begin{cases}100，\text{第}i\text{回合任务完成}\\ \sum_{j=1}^{m} r_{\mathrm{AT}j}+\sum_{k=1}^{o} r_{\mathrm{DE}k}+\sum_{l=1}^{p} r_{\mathrm{DT}l}，\text{其他}\\ -100，\text{第}i\text{回合任务失败}\end{cases} \tag{2}$$

式中：r 为单次对局总奖励；r_i 为第 i 回合总奖励；$r_{\mathrm{AT}j}$ 为单回合内第 j 个攻击单元奖惩差；$r_{\mathrm{DE}k}$ 为单回合内第 k 个防御单元奖惩差；$r_{\mathrm{DT}l}$ 为单回合内第 l 个探测单元奖惩差。$r_{\mathrm{AT}j}$ 计算公式为

$$\begin{cases}r_{\mathrm{AT}j}=r_1+r_2+r_3-r_4-r_5-r_6-r_7-r_8\\ r_1=r_1+x_1，\text{攻击敌方单元一次}\\ r_2=r_2+x_2，\text{击毁敌方单元}\\ r_3=r_3+x_3，\text{被防御单元保护一次}\\ r_4=r_4+y_1，\text{被攻击一次}\\ r_5=r_5+y_2，\text{被击毁}\\ r_6=r_6+y_3，\text{攻击被防御一次}\\ r_7=r_7+y_4，\text{被探测一次}\\ r_8=r_8+a，\text{每一回合}\\ r_9=r_9+b，\text{触碰区域边界一次}\end{cases} \tag{3}$$

式中：x_i 为奖励项，值为正；y_i、a、b 为惩罚项，值为负。$r_{\mathrm{DE}k}$ 及 $r_{\mathrm{DE}l}$ 奖励函数计算方式同上。

2.2 基于局部回报重塑及 PER 的无人集群攻防对抗方法框架

通过局部回报重塑的奖励工程设计方法对无人

集群攻防对抗场景已有的二元奖励进行扩充，奖励体系得到了丰富。将单次对局下基于局部回报重塑的无人集群对抗所有回合奖励信号进行输出，结果如图 7 所示。在总数约 800 个状态动作序列对样本中，只有约 12%的状态动作序列对存在奖励信号，其余均无奖励信号。这意味着约 88%状态下采取动作的有效性无法进行评判。因此，通过局部回报重塑方法设计的奖励机制奖励稀疏性依旧严峻。

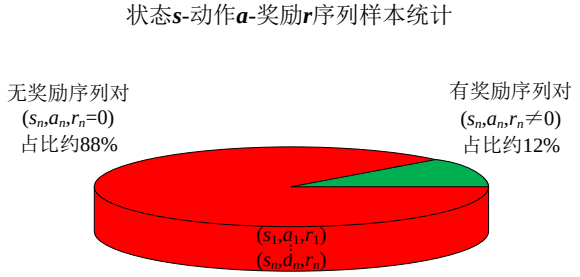


图 7 局部回报重塑方法下奖励稀疏性示意

Fig.7 Reward sparsity under the Local Reward Reshaping

依据是否存在奖励信号，智能单元与环境交互产生的状态动作序列样本具备不同的重要性。存在奖励信号的状态动作序列对有学习价值，优先级高。而无奖励信号的状态动作序列对无学习价值，优先级低。因此，可采用 PER[23]实现对经验样本的差别利用。在回放训练的过程中，通过对有效经验样本进行优先级排序，实现对高价值样本的优先利用，以实现对抗策略的快速有效学习。在算法方面，因本研究仿真模型为离散输出，故选择深度强化学习(DQN)算法[24]。采用回放记忆单元存储(s,a,r,s')序列，基于 PER 算法对回放记忆单元进行优先级采样，作为动作值函数逼近网络与目标值网络的训练样本，通过 DQN 误差函数计算迭代，进行网络参数更新。

综上，若将局部回报重塑方法称为稀疏奖励问题下奖励信号的“开源”，PER 的使用则是从“节流”的角度对样本进行了高效利用。通过奖励信号“开源”、“节流”两种手段实现了对稀疏奖励问题的有效解决，最终形成基于局部回报重塑及 PER 的无人集群对抗自主决策与智能协同策略学习方法框架，如图 8 所示。首先通过局部回报重塑方法对基于强化学习技术的无人集群对抗问题所固有的二元奖励进行扩充，局部回报重塑很大程度上改善了奖励的稀疏性，但奖励稀疏性依旧严峻，然后依据样本是否存在奖励信号及奖励信号数值进行优先级排序，在样本回放学习过程中进行优先级采样，对高价值样本进行优先学习。最终通过两种方法组合实现稀疏奖励下无人集群自主决策与智能协同对抗策略的高效学习。

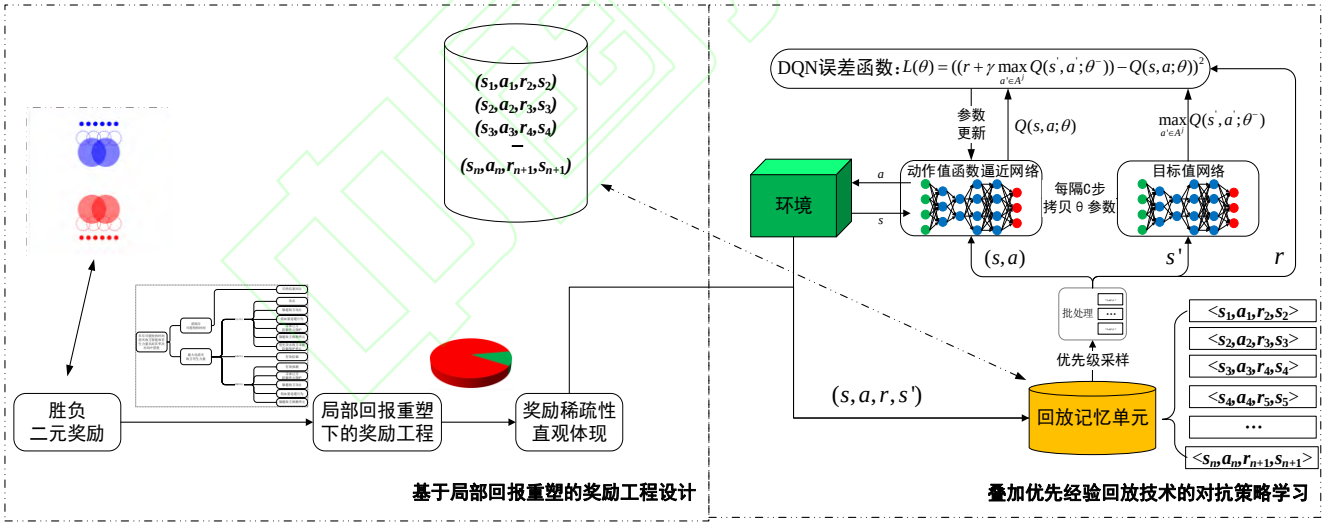


图 8 基于局部回报重塑及 PER 的无人集群对抗自主决策与智能协同策略学习方法框架

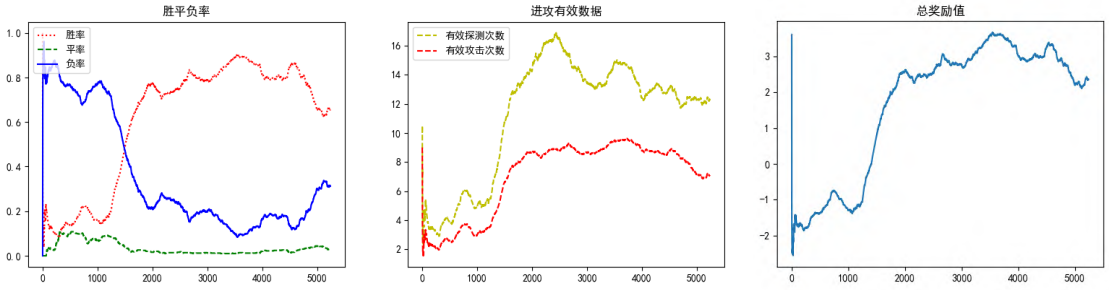
Fig.8 The method framework for autonomous decision-making and intelligent collaborative strategy learning for unmanned swarm confrontation based on local reward reshaping and prioritized experience replay

3　无人集群攻防对抗仿真及演示系统设计

3.1 无人集群攻防对抗程序仿真

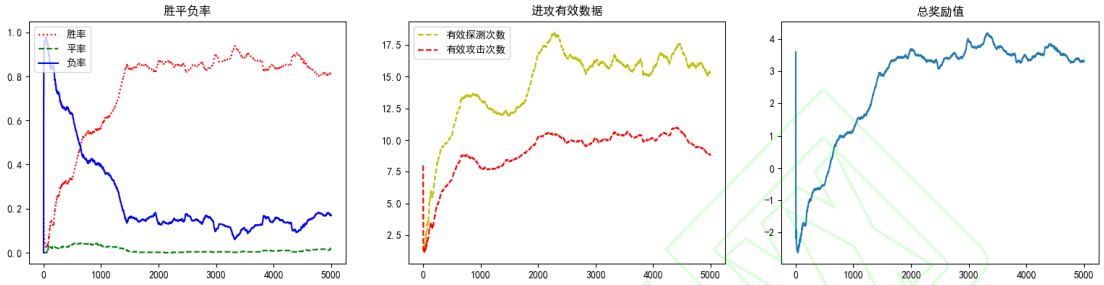
在无人集群攻防对抗场景下，将强化学习算法与稀疏奖励方法组合设计进行了程序仿真。

在 DQN 及局部回报重塑组合算法(简称 DQN+局部回报重塑算法)下，通过 2000 代训练，红方智能单元策略收敛，胜率约 80%，如图 9(a)所示。采用 DQN 改进算法即 Double DQN 及局部回报重塑组合算法(简称 Double DQN+局部回报重塑算法)，通过缓解策略学习过程中价值高估问题，训练 1500 代后，策略实现了收敛，如图 9(b)所示。在上述方法基础上，叠加 PER(简称 Double DQN+局部回报重塑+PER 算法)，通过对有效样本的高效利用，训练 700 代后策略实现收敛，如图 9(c)所示。



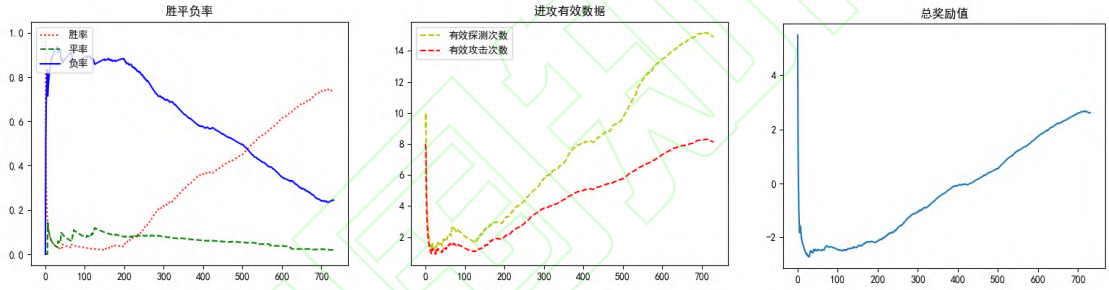
(a) DQN+局部回报重塑算法对抗训练曲线

(a) The training curve based on DQN and Local Reward Reshaping algorithm



(b) Double DQN+局部回报重塑算法训练曲线

(b) The training curve based on Double DQN and Local Reward Reshaping algorithm



(c) Double DQN+局部回报重塑+PER 算法训练曲线

(c) The training curve based on Double DQN and local reward reshaping and prioritized experience replay algorithm

图 9 无人集群攻防对抗算法效率对比

Fig.9 Efficiency comparison for attacking and defending confrontation algorithms of unmanned cluster

此外，在进攻有效数据及防御有效数据方面，仿真曲线图均呈现逐渐提升的趋势，证明了算法的有效性，3 种算法效率对比如表 1 所示。

表 1 无人集群攻防对抗算法效率对比

Table 1 Efficiency comparison for attacking and defending confrontation algorithms of unmanned cluster

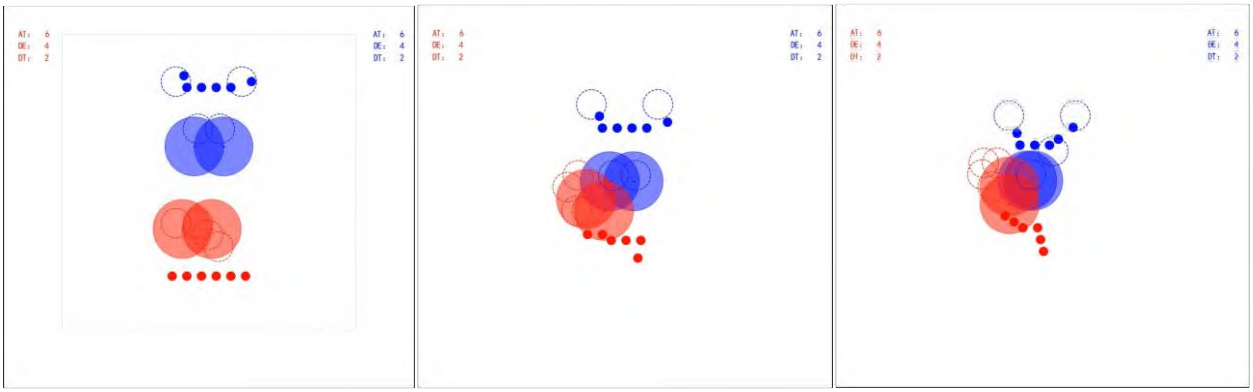
算法	算法效果	性能提升
DQN +局部回报重塑算法	训练 2000 代，策略收敛，胜率约 80%。	
Double DQN+局部回报重塑算法	训练 1500 代，策略收敛，胜率约 80%。	提升 25%
Double DQN+局部回报重塑+PER 算法	训练 700 代，策略收敛，胜率约 80%。	提升 65%

在同样达约 80%胜率的对抗能力前提下，DQN +局部回报重塑算法训练了 2000 代，Double DQN+局部回报重塑算法训练了 1500 代，算法提升 25%，Double DQN+局部回报重塑+PER 算法训练 700 代，算法提升 65%。

上述为算法在对抗策略宏观层面的表现。在微观层面，即策略收敛后的单次对局中，红方无人智能集群呈现协同对抗态势，如图 10 所示。攻击单元整体居中，在己方防御单元的防御保护下对处于己方探测单元探测视角内的敌方单元进行饱和攻击；防御单元居阵型前方，集中防御，保护己方攻击单元和探测单元；探测单元居阵型后方，向后退避、向前冲锋行为动态切换，为己方攻击单元提供探测视角的同时，最大化保证自身的生存。不同类型单元根据自身属性特点实现了行为协同、功能互补，同类型单元也呈现出明显的群集优势。

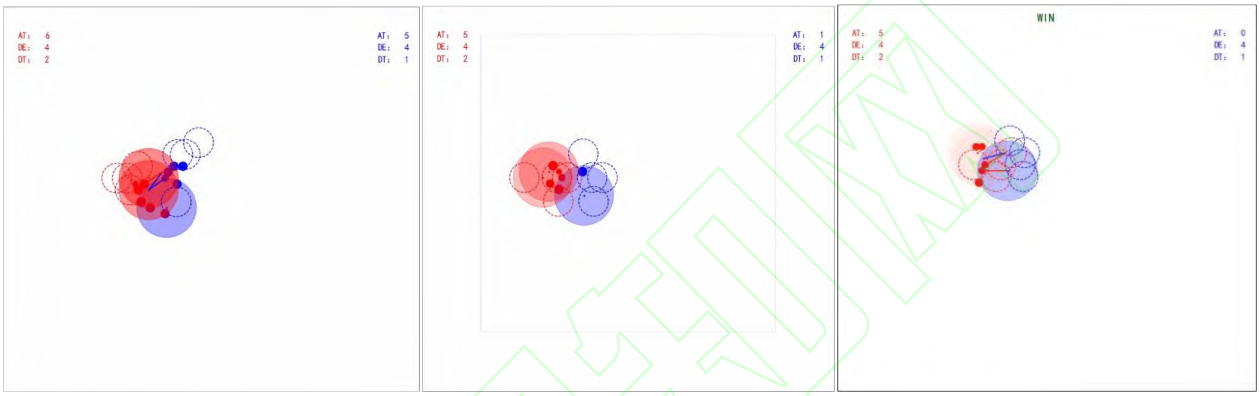
3.2 无人集群对抗演示系统

为了直观展示红蓝双方集群对抗过程，设计了无人集群对抗实时演示系统，如图 11 所示。演示系统中针对无人集群攻防对抗任务场景设计了 5 个模



(a) 时间帧 1　　(b) 时间帧 2　　(c) 时间帧 3

(a) Time frame 1　　(b) Time frame 2　　(c) Time frame 3



(d) 时间帧 4　　(e) 时间帧 5　　(f) 时间帧 6

(d) Time frame 4　　(e) Time frame 5　　(f) Time frame 6

图 10 红蓝无人集群攻防对抗仿真对局态势图

Fig.10 The situation for attacking and defending confrontation simulation of red and blue unmanned cluster

块，其中实时攻防对抗态势演示模块位于演示面板中央，双方实时对抗过程以回合步为单位进行更新。攻防对抗双方实时胜率演示模块、攻防对抗双方实时有效进攻/防御数据位于演示面板左侧；双方各类型单元实时存活数、双方各类型单元实时总血量位于演示面板右侧；左右侧四大演示模块除攻防

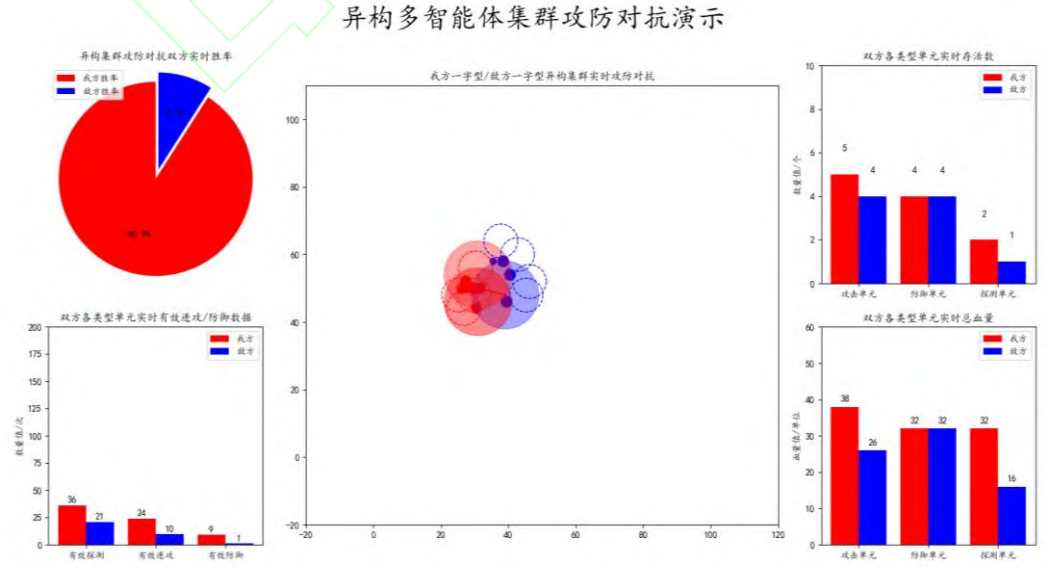


图 11 无人集群攻防对抗任务场景演示面板

Fig.11 The demonstration panel for attacking and defending confrontation scenario of unmanned cluster

对抗双方实时胜率以对局为单位更新外均以回合步为单位更新。

4 结论

无人集群为无人系统与群体智能的结合，意图通过群体智能算法使多数量无人系统具备自组织能力并实现协同能力涌现。在这一过程中，强化学习技术被广泛采用，稀疏奖励问题广泛存在。本文构建了无人集群对抗模型框架并以无人集群攻防对抗为具体场景进行了模型设计，通过分析奖励函数机理机制，设计了局部回报重塑方法，并叠加 PER 方法，最后进行了程序仿真与演示系统设计。经对比证明该方法有效提升了算法效率，后续将在以下方面展开进一步研究。

1) 针对稀疏奖励问题，当前方法在智能性、泛化性、设计耗时方面具备提升空间，可进一步研究智能性更强、泛化性更好、设计耗时更短的稀疏奖励算法，促进强化学习技术从理论研究迈向工程应用。

2) 当前研究关注自主决策算法，后续可对基于实际动力学模型及态势感知下的自主决策算法展开研究，进一步提升自主决策算法验证过程置信度。

参考文献（References）

[1] 王莉. 人工智能在军事领域的渗透与应用思考[J]. 科技导报, 2017, 35(15):15-19.
WANG L. The penetration and application of artificial intelligence in the military field[J]. Science & Technology Review, 2017, 35(15): 15-19. (in Chinese)

[2] 罗德林, 徐扬, 张金鹏. 无人机集群对抗技术新进展[J]. 科技导报, 2017, 35(7): 26-31.
LUO D L, XU Y, ZHANG J P. New progresses on UAV swarm confrontation[J]. Science & Technology Review, 2017, 35(7): 26-31. (in Chinese)

[3] 梁晓龙, 侯岳奇, 胡利平, 等. 无人集群试验评估研究现状分析及理论方法[J]. 南京航空航天大学学报, 2020, 52(6): 846-854.
LIANG X L, HOU Y Q, HU L P, et al. Review on evaluation and theoretical methods of un-manned swarm test[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2020, 52(6): 846-854.(in Chinese)

[4] 朱建文, 赵长见, 李小平, 等. 基于强化学习的集群多目标分配与智能决策方法[J]. 兵工学报, 2021, 42(9): 2040-2048.
ZHU J W, ZHAO C J, LI X P, et al. Multi-target assignment and intelligent decision based on reinforcement learning[J]. Acta Armamentarii, 2021, 42(9): 2040-2048.(in Chinese)

[5] 杜威, 丁世飞. 多智能体强化学习综述[J]. 计算机科学, 2019, 46(8):1-8.
DU W, DING S F. Overview on multi-agent reinforcement learning[J]. Computer Science, 2019, 46(8): 1-8.(in Chinese)

[6] 郭宪, 方勇纯. 深入浅出强化学习[M]. 北京:电子工业出版社, 2018:1-10.
GUO X, FANG Y C. Reinforcement learning in a simple and in-depth way[M]. Beijing: Publishing House of Electronics Industry, 2018:1-10.(in Chinese)

[7] 陈智超. 基于深度强化学习的无人潜航器智能对抗决策[D]. 哈尔滨:哈尔滨工业大学, 2020.
CHEN Z C. UUV intelligent countermeasure decision making based on deep reinforcement learning[D]. Harbin:Harbin Institute of Technology, 2020. (in Chinese)

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

[9] HARE J. Dealing with sparse rewards in reinforcement learning: arXiv:1910.09281v2[R]. Ithaca, NY, US:Cornell University, 2019.

[10] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]//Proceedings of the 26th annual international conference on machine learning. Montreal, Canada: International Machine Learning Society, 2009:41-48.

[11] ANDRYCHOWICZ M, WOLSKI F, RAY A, et al. Hindsight experience replay: arXiv:1707.01495v3[R]. Ithaca, NY, US:Cornell University, 2018. .

[12] RAFATI J, NOELLE D C. Learning representations in model-free hierarchical reinforcement learning: arXiv:1810.10096v3[R]. Ithaca, NY, US:Cornell University, 2019.

[13] 杨瑞, 严江鹏, 李秀. 强化学习稀疏奖励算法研究——理论与实验[J]. 智能系统学报, 2020, 15(5): 888-899.
YANG R, YAN J P, LI X. Summary of sparse reward algorithms in reinforcement learning -- theory and experiment[J]. CAAI Transactions on Intelligent Systems, 2020, 15(5): 888-899. (in Chinese)

[14] 方嘉良. 基于强化学习的稀疏奖励问题研究[D]. 北京:中国地质大学, 2020:29-39.
FANG J L. Research on Sparse Reward Based on Reinforcement Learning[D]. Beijing:China University of Geosciences , 2020: 29-39.(in Chinese)

[15] 杨惟轶, 白辰甲, 蔡超, 等. 深度强化学习中稀疏奖励问题研究综述[J]. 计算机科学, 2020, 47(3):182-191.
YANG W Y, BAI C J, CAI C, et al. Survey on sparse reward

in deep reinforcement learning[J]. Computer Science, 2020, 47(3):182-191.(in Chinese)

[16]王瑞星. 含有稀疏奖励的异构多智能体强化学习对抗方法研究[D]. 哈尔滨:哈尔滨工业大学, 2021.
WANG R X. Research on reinforcement learning countermeasures for heterogeneous multi-agents with sparse rewards[D]. Harbin: Harbin Institute of Technology, 2021.(in Chinese)

[17]王瑞星, 董诗音, 江飞龙, 等. 稀疏奖励下基于强化学习的异构多智能体对抗[J]. 信息技术, 2021(5):12-20.
WANG R X, DONG S Y, JIANG F L, et al. Heterogeneous multi-agent confrontation based on reinforcement learning under the sparse reward[J]. Information Technology, 2021(5):12-20. (in Chinese)

[18]李理, 李旭光, 郭凯杰,等.国产化环境下基于强化学习的地空协同作战仿真[J].兵工学报, 2022, 43(增刊 1): 74-81.
LI L，LI X G，GUO K J，et al. Simulation of ground-air cooperative combat based on reinforcement learning in localization environment[J]. Acta Armamentarii, 2022, 43(s1): 74-81. (in Chinese)

[19]HE Y M, XING L N, CHEN Y W, et al. A generic Markov decision process model and reinforcement learning method for scheduling agile earth observation satellites[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020.

[20]LIU H, LI X M, WU G H, et al. An iterative two-phase optimization method based on divide and conquer framework for integrated scheduling of multiple UAVs[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(9): 5926-5938.

[21]LI B J, WU G H, HE Y M, et al. An overview and experimental study of learning-based optimization algorithms for vehicle routing problem: arXiv:2107.07076v2[R]. Ithaca, NY, US:Cornell University, 2022.

[22]WANG R X, LI Y Q, ZHANG H L, et al. Satellite mission support efficiency evaluation based on cascade decomposition and Bayesian network[C]//Proceedings of International Conference on Wireless and Satellite Systems. Nanjing, China:Springer, 2020: 46-60.

[23]SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay: arXiv:1511.05952v4[R]. Ithaca, NY, US:Cornell University, 2016.

[24]MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning: arXiv:1312.5602v1[J]. Ithaca, NY, US:Cornell University, 2013.