Computer Science and Application 计算机科学与应用, 2022, 12(12), 2825-2833
Published Online December 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1212287

Hans汉斯

基于强化学习QMIX的多机器人区域覆盖策略

段磊磊

天津工业大学，天津

收稿日期：2022年11月17日；录用日期：2022年12月17日；发布日期：2022年12月26日

摘　要

未知环境下的多机器人区域覆盖是指多个机器人遍历环境中每个无障碍物的区域。机器人区域覆盖作为多机器人系统研究的重要组成部分，在灾后救援、野外勘测、森林防火等众多领域有着广泛的应用，具有十分重要的研究意义。传统的多机器人覆盖方法需要考虑区域分割、任务分配等问题，且没有协同策略的覆盖方法只是单个机器人方法的简单叠加。而在强化学习中机器人可以通过自主学习的方式求得问题可行解。本文将多机器人区域覆盖问题转换为多机器人强化学习中团队奖励值最大化的求解问题，搭建了基于Actor-Critic结构的多机器人强化学习网络，考虑到机器人个体行为对环境造成的不平稳问题，选择考虑了全局信息的QMIX网络作为多机器人行为的评价网络。最后设计了强化学习与仿真环境端到端的数据交互接口，简化了训练数据交互过程。算法训练结果表明本文提出的算法能达到较高的覆盖率，验证了该算法解决区域覆盖任务问题的有效性和可行性。

关键词

区域覆盖，多机器人，强化学习

Multi-Robot Area Coverage Strategy Based on Reinforcement Learning QMIX

Leilei Duan

Tiangong University, Tianjin

Received: Nov. 17th, 2022; accepted: Dec. 17th, 2022; published: Dec. 26th, 2022

Abstract

Multi-robot area coverage in unknown environment refers to multiple robots traversing every obstruction-free area in the environment. As an important part of multi-robot system research, robot area coverage has been widely used in many fields, such as post-disaster rescue, field survey, forest fire prevention, and so on, and has very important research significance. Traditional mul-

文章引用: 段磊磊. 基于强化学习 QMIX 的多机器人区域覆盖策略[J]. 计算机科学与应用, 2022, 12(12): 2825-2833.
DOI: 10.12677/csa.2022.1212287

ti-robot coverage methods need to consider regional segmentation, task allocation and other problems, and the coverage method without collaborative strategy is just a simple superposition of a single robot method. In reinforcement learning, the robot can obtain feasible solutions through autonomous learning. In this paper, the multi-robot area coverage problem is transformed into a solution problem of maximizing team reward value in multi-robot reinforcement learning, and a multi-robot reinforcement learning network based on Actor-Critic structure is built. Considering the instability of individual robot behavior on the environment, QMIX network considering global information is selected as the evaluation network of multi-robot behavior. Finally, the end-to-end data interaction interface between reinforcement learning and simulation environment is designed to simplify the training data interaction process. The algorithm training results show that the proposed algorithm can achieve a higher coverage rate, and verify its effectiveness and feasibility in solving the problem of regional coverage task.

Keywords

Area Coverage, Multiple Robots, Reinforcement Learning

1. 引言

随着计算机技术的多元化发展以及硬件成本的低廉化，机器人的智能化程度不断被提升，正加速赋能人们的生产生活。单个机器人在执行任务时存在耗时长、鲁棒性差等缺点，且单个机器人的电池容量对于环境复杂、作业范围广的任务而言也是一个挑战，因此，研究者开始将目光集中于单机智能向多机智能转变的研究方向上，通过多个机器人共同工作来满足实际的生产需要。

由于多机器人团队在任务执行过程中表现出效率高、鲁棒性好等特点，因此越来越多的研究者将其应用在不同领域，机器人区域覆盖任务便是其一。区域覆盖机器人的工作是集体实现一个目标——通过传感器感知周围环境，在避开障碍物的情况下遍历整个环境区域。区域覆盖技术已经被广泛应用于多种领域，比如灾后搜索和救援任务[1]；敏感设施或重点区域环境监测[2]、安全监控；工业检查[3] [4]；自然灾害监测[5]等场景。传统算法在解决非线性、特征复杂的问题时难以求得最优解，深度学习与强化学习技术的结合为多机器人解决复杂问题提供了新的研究方向，机器人可以通过自主学习来获得一个较好的策略，之后可直接应用于工作场景，这在很大程度上降低了传统方法中设计系统的复杂性。Heydari J 等人[6]对覆盖问题进行了系统的分析，并将其表述为一个最优停止时间问题，明确地考虑了覆盖性能与其代价之间的权衡，通过强化学习技术计算解决这个问题，但解决方法只是针对单个机器人的情况。

多机器人区域覆盖任务的目标是多个机器人共同完成一片未知区域的探索，机器人之间属于合作式关系。对于单个机器人来说，以学习到最优策略为目标，其回报函数只与自身有关，但在多机器人系统中，每个成员的动作都会影响整个系统的效果。本文以单元格的表示方式离散化整个场景，机器人覆盖每一个新单元格都会得到相应的奖励，对于多个机器人共同解决的覆盖任务而言，奖励值越大则覆盖率越高。因此本文将多机器人区域覆盖问题转换为奖励值最大化求解问题。Rashid T 等人提出了考虑全局状态信息的混合式网络结构 QMIX [7]，其整合每个机器人的局部值函数得到联合动作值函数，并以最大回报值为优化目标。因此本文基于 QMIX 设计了多机器人强化学习网络解决多机器人区域覆盖问题。

2. 基于强化学习 QMIX 网络的多机器人区域覆盖方法

2.1. 系统总体框架

基于 Actor-Critic 的集中式训练分布式执行策略得到了广泛应用，本文同样采用该结构设计多机器人强化学习网络研究区域覆盖问题，本文的算法框架图如图 1 所示。

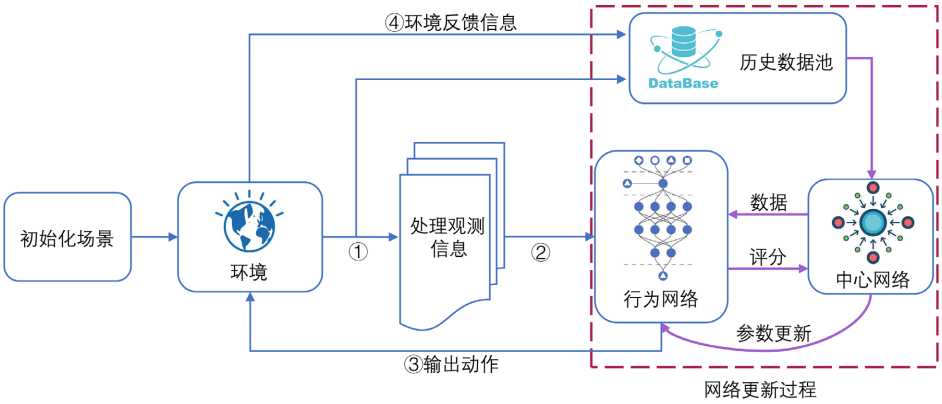

Figure 1. Algorithm overall framework diagram
图 1. 算法总体框架图

2.2. 机器人观测信息表示方法

对于某些覆盖区域虽然我们缺少对其内部环境的了解，但可以预测出其范围大小。因此本文根据区域大小选择四边形作为未知区域的离散化方式，四边形的面积大小与机器人覆盖范围相关。通过这种表示方式机器人以周围四个邻居的方向作为机器人的动作空间，在决策下一个目标位置时可以通过雷达信息标注周围邻居状态。Kan X [8]等人提出用三维坐标来表示单元格位置，在机器人实际运动时再将其转换为世界坐标系，通过这种方式简化了目标位置选择、路径规划过程中数据复杂难以处理的问题。本文同样采用两种坐标系的表示方法，用三维坐标表示地图中的每一个单元格位置，当计算周围邻居的坐标位置时，以机器人当前所在位置的三维坐标为基准，根据邻居相对于当前机器人位置的方向修改坐标即可。如图 2 中当前四边形的坐标为(0, 0, 0)，则它右边的邻居就是(1, 0, −1)，坐标之和为零。只有机器人在实际运动时才将单元格坐标转换为世界坐标系，其余情况都采用三维坐标，这样设计有利于将当前任务完成情况以张量的表现形式作为行为网络的输入。

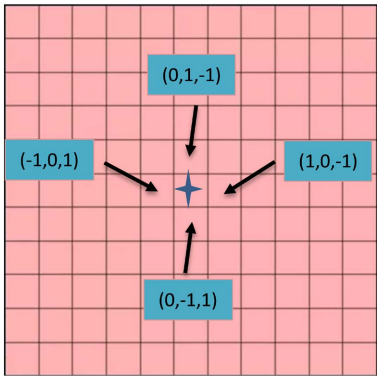

Figure 2. Scene representation method
图 2. 场景表示方法

在本文中，没有采用先划分场景再分配给机器人的方式，而是由多个机器人共同负责整个区域，在多机器人系统中，每个机器人的行为都会对环境造成变化，进而会影响其他机器人的选择行为，因此机器人需要学习全局性的策略。因此对于机器人的观测信息分为三部分内容，如图 3 所示，包括当前时刻场景中已经覆盖完成的全局状况、障碍物分布的全局状况以及自己所在位置。所有机器人的共同目标是探索整个区域，因此区域访问情况、障碍物分布情况都被共享给所有机器人，只有自己位置是唯一的。

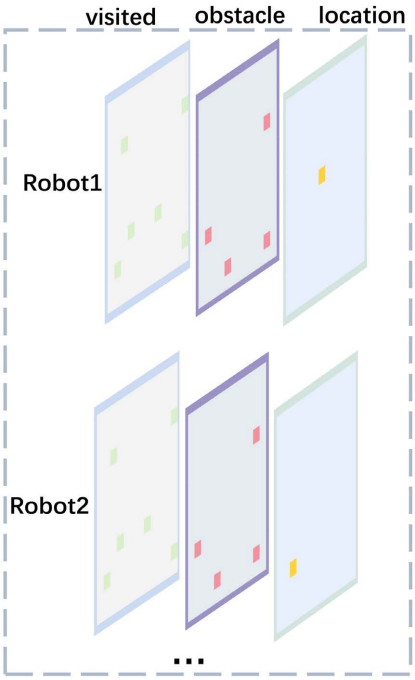

Figure 3. Robot observation information
图 3. 机器人观测信息

2.3. 网络结构

在基于 Actor-Critic 结构的强化学习方法中，每个机器人可以独立学习一个 Actor 网络负责处理从环境获得的信息，然后根据输出的动作空间概率选择动作，Critic 网络再根据动作的好坏进行评分，通过这种训练方式可以学习到适应于个体的策略。但在训练时如果缺乏信息共享，很难学习依赖于多个个体之间交互的策略，而集中式 Critic 网络是将所有信息收集后给出评分，然后反馈给 Actor 网络，这种方式能够考虑到个体行为造成的环境非平稳性问题。在本文中研究的区域覆盖任务中，机器人的目标一致，因此可以选择同一个 Actor 网络，共享同一套网络参数，对于 Critic 网络而言则需要考虑到全局状态信息。

图 4 是本文设计的 Actor 网络模型，输入张量包括三部分，第一个通道记录了整个场景当前时刻所有已经访问过的信息，第二个通道记录整个场景的障碍物，这两部分内容是所有机器人共同包含并维持更新的。第三个通道是机器人自己所在位置。卷积神经网络具有一种所谓的人工神经元，能够在一定的范围之内和其他的单元进行相应，从而很好地去处理图像以及一些其他领域的特征，本文将观测信息转换为三通道的张量，使用卷积层处理数据。机器人当前时刻动作并非与上一时刻没有关系，因此将 CNN 的输出内容与上一时刻的动作编码和机器人编号拼接起来作为循环神经网络 RNN 的输入，采用 RNN 的原因是可以通过中间隐藏层保留上一时刻的信息，最后输出动作空间分别对应的概率。

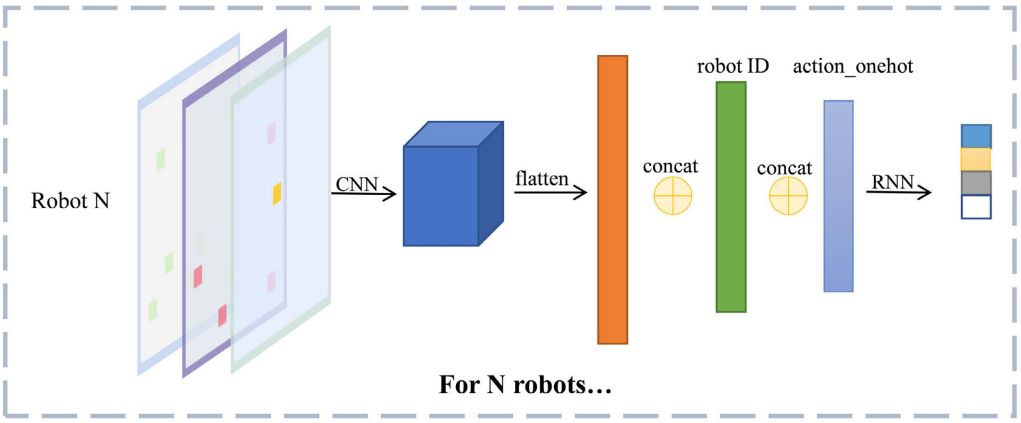

Figure 4. Actor network structure
图 4. Actor 网络结构

对于 Critic 网络来说，它是通过评分的方式来引导 Actor 选择能获得更高奖励值的动作，每个机器人在覆盖一个新的单元格后会获得奖励，覆盖重复的单元格会得到惩罚，因此在每一步中都可以明确其奖励值。而 QMIX 作为集中式网络，通过一个混合网络对单机器人局部值函数进行合并，并在训练学习过程中加入全局状态信息辅助，能够解决值最大化问题。在集中训练时候，首先从经验池中取一批数据，计算出所有评分，然后结合全局状态信息。推理过程如图 5 所示，接收所有机器人的行为效用值 Q 后，再将权重和偏差赋值到网络自身，从而推理出全局效用值 Q_{tot}。其中权重和偏置是由全局状态信息经过超网络生成。

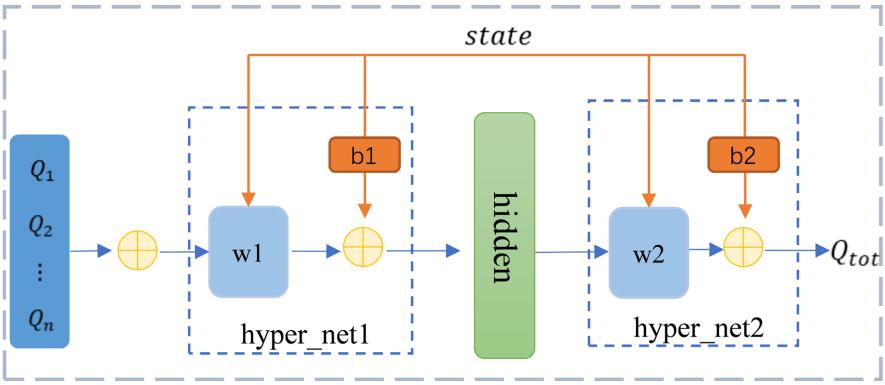

Figure 5. Critic network structure
图 5. Critic 网络结构

设置评估网络(evaluate)与目标网络(target)采用 TD-error 算法对模型进行更新。接收在状态 S 下 Actor 网络所选行为的 Q 值作为输入，输出 Q_{tot} (evaluate)。接收在状态 S_{next} 下 Actor 网络所有行为中最大的 Q 值作为输入，输出 Q_{tot} (target)。

2.4. 仿真环境设计

Gazebo 是广泛用于机器人仿真实验的 3D 平台，本文借助此平台完成算法训练。除此之外，本文根据机器人的运动特点和环境表示方法构建了强化学习算法与仿真环境之间进行数据交换的接口。图 6 列出了包含的一些功能模块，核心内容包括接收机器人从环境中获取的观测信息、转换观测信息为适合于行为网络输入的维度、将行为网络输出的动作传达给仿真环境中的机器人模型、计算执行动作后获得的

奖励。所有机器人观测信息都由三个通道的张量组成，由于任务一致性，机器人之间共享同一个行为网络，因此需要将所有机器人观测信息进行整合。在将处理后的观测信息输入到行为网络产生下一步动作时，应避免把被障碍物占据的候选者考虑在内，因此，不仅需要处理观测信息，还需要将机器人所在位置的邻居状态信息处理后一并输入行为网络。行为网络在确定了所有机器人的下一个目标位置后，中间接口需要做的工作是将动作指令传达给仿真环境中的机器人，在仿真环境中多个机器人需要被分别控制，使用命名空间的方式可以使得机器人模型互相独立，并使用 ROS 提供的控制组件和话题通信机制实现机器人的控制。

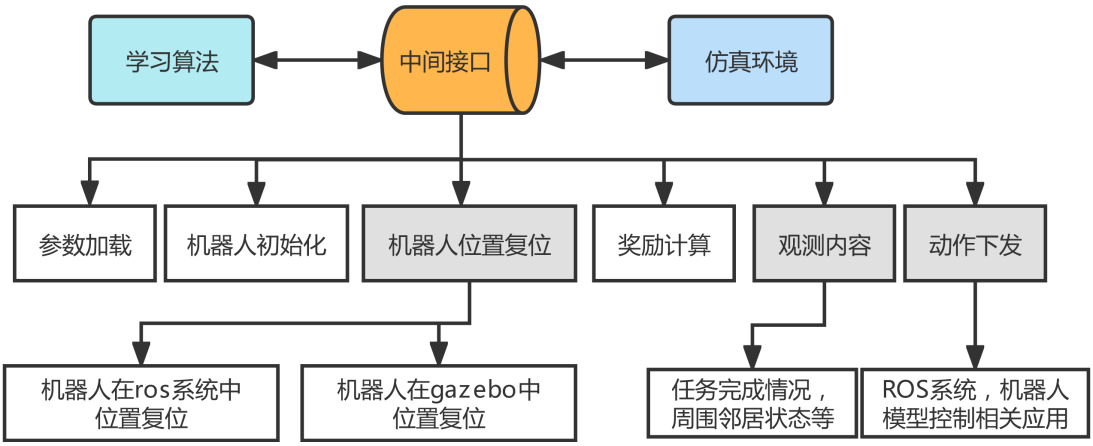

Figure 6. Simulation environment framework
图 6. 仿真环境框架

本文利用这些功能模块实现了算法训练与实验仿真的结合，行为网络推测出的结果可以直接被作用到仿真环境中的机器人模型，同时机器人与环境交互后的结果及环境状态经过处理后直接反馈给学习算法。通过集中处理数据信息，简化了数据交互过程。

3. 实验结果分析

本文首先创建了随机摆放障碍物的环境，然后利用 Gazebo 仿真工具完成策略的学习。所有实验都是在一个 8 GB 内存的单一 GPU 1080Ti 机器上完成。

在 3D 仿真环境中训练策略算法时，机器人执行每一步前都需要先获取周围的观测信息以及环境状态信息，然后组合成张量的形式输入网络，机器人执行网络模型输出的动作并将奖励等内容反馈给算法，以此往复。由于执行每一步都需要与环境交互，因此学习效率较慢。本文采用经验重放的方法充分利用数据加快学习效率，将每一回合的动作、观测值、奖励数据都保存下来。在每回合结束时更新网络，更新网络时从经验池中一次选择多批数据，通过这种方式来弥补仿真环境下训练速度慢的不足。

在训练过程中，本文通过统计机器人探索完成的单元格数量来确定任务的完成程度，如果在每回合学习中单个机器人执行步数多，那么单个机器人在经过长期探索过程后，也能独立完成任务，但这将会导致大量的重叠率，并且也无法学习一个有利于多机器人协同工作的策略。因此，本文设置了机器人单回合最大步数，限制了最大重叠率，当步数超过阈值则终止当前任务，然后进入下一回合的学习中。COMA [9]算法也是一种考虑了回报问题的多机器人网络学习算法，因此，本文最后与该算法进行了对比。

训练结束后，本文算法覆盖率结果如图 7(a)所示，与图 7(b) COMA 算法相比较而言，本文算法覆盖率可以达到百分之九十，表现出较高的覆盖率。因为在本文任务中，场景已被栅格化，可以通过标记单元格状态来计算奖惩值，机器人覆盖一个新的单元格会获得各自的奖励，当机器人重复覆盖单元格时就

会得到惩罚，它们的共同目标是在完成一个未知区域的覆盖任务时争取奖励最大化。因此，本文将多机器人区域覆盖任务转换为团队奖励值最大化的求解问题，而 QMIX 网络可以整合每个机器人的局部值函数而得到联合动作值函数，并以回报最大化为目标，将其作为本算法中的评价网络可以获得很好的效果。图 8 是两种算法的奖励值对比结果，图 8(a)中奖励值不断上升的原因是机器人在达到某一覆盖率的过程中，在不断优化所需要的最少执行步数。图 9 记录了实验过程中的 Loss 变化，Loss 的波动说明机器人探索到了新的学习空间。本文中的评价网络是联合所有机器人的局部值函数然后求其最大效用值，因此在更新模型时，将 Actor 网络与 Critic 网络的参数一起更新，如图 9(a)所示。对于 COMA 算法，两个网络模型分别更新，如图 9(b)，图 9(c)所示。

为了能够学到多机器人协同策略，本文在训练算法过程中限定了单回合中机器人最大执行步数，单个机器人在执行最大步数的情况下只能完成所有机器人任务总和的平均值，因此只有机器人之间能够考虑对方行为对环境的影响并且相互协作时才能达到较高的覆盖率。通过本实验结果说明了本文提出的算法能够为多机器人区域覆盖问题提高一个可行解。

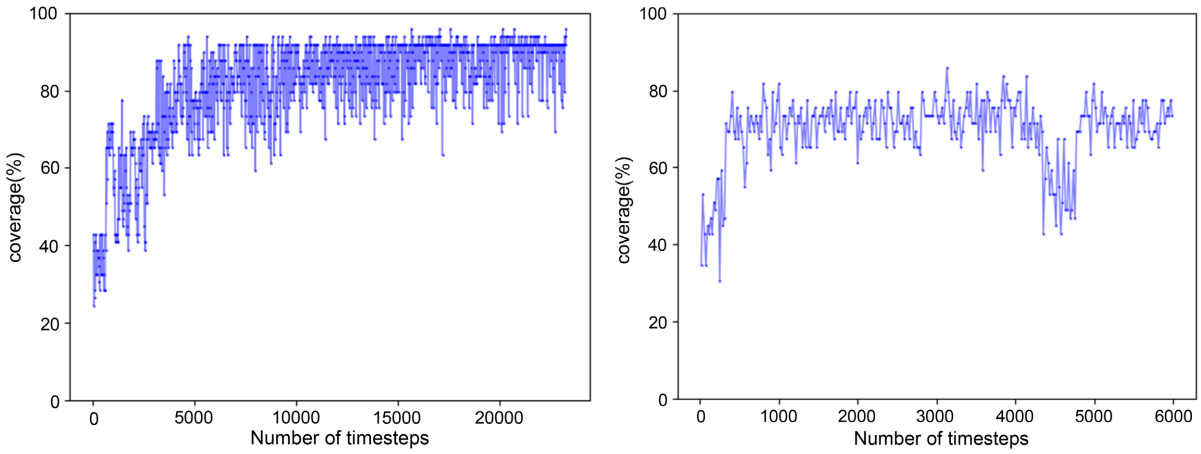

Figure 7. Comparison of coverage rate
图 7. 覆盖率对比结果

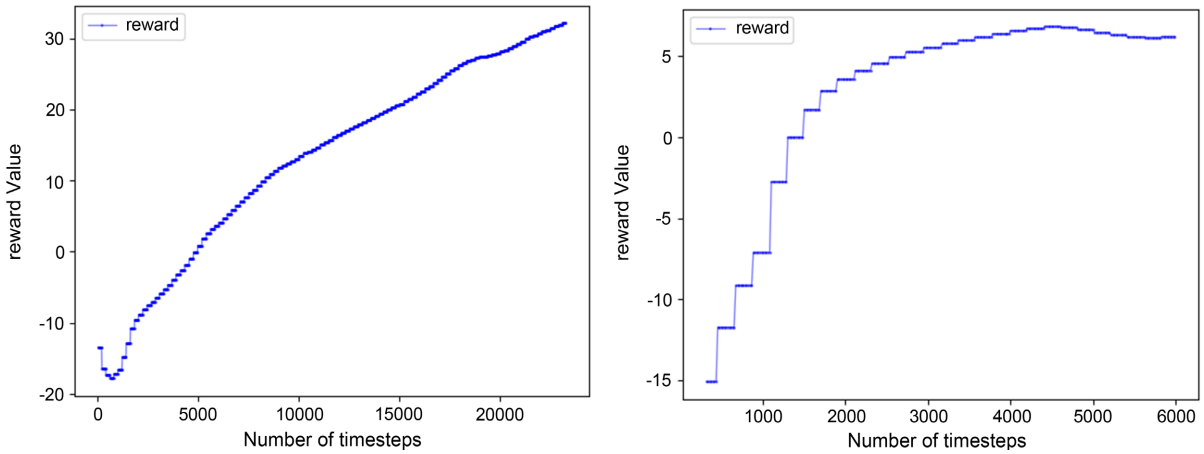

Figure 8. Reward value comparison results
图 8. 奖励值对比结果

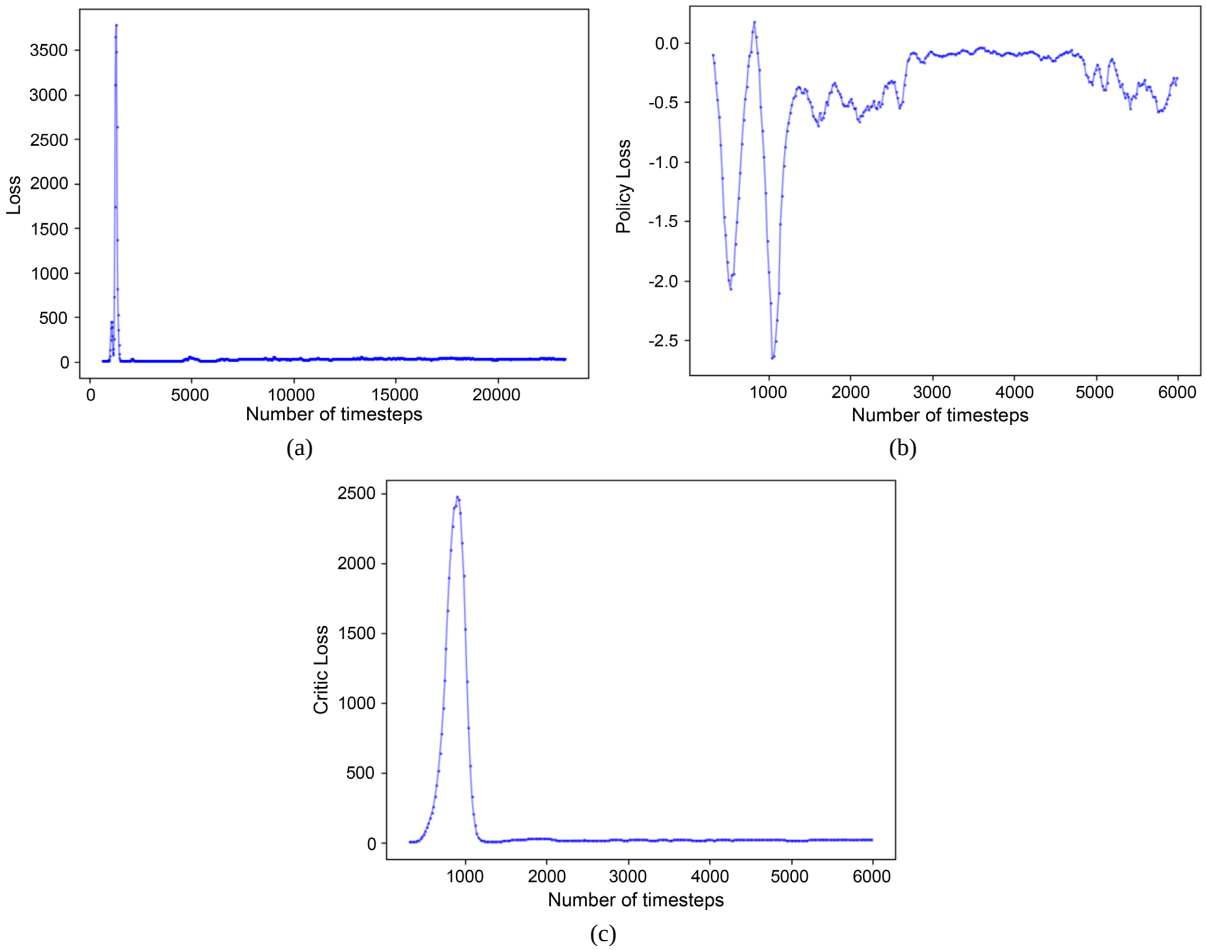

Figure 9. Variation of Loss
图 9. Loss 变化

4. 结论

本文研究了多机器人强化学习算法在区域覆盖任务中的应用，由机器人自主学习策略。首先将多机器人区域覆盖问题转换为团队奖励值最大化的求解问题。其次搭建了基于 Actor-Critic 结构的学习网络。在多机器人系统中，机器人个体的行为会引起环境的变化，因此采用基于 QMIX 的集中式网络考虑所有机器人的状态信息。最后，构建了强化学习算法与仿真环境的中间接口，实现强化学习与仿真环境端到端的数据交换。最后经过算法训练验证了本文所提出算法的有效性。

参考文献

[1] Walker, J. (2019) Search and Rescue Robots–Current Applications on Land, Sea, and Air.

[2] Merino, L., Caballero, F., Martínez-de-Dios, J.R., *et al*. (2012) An Unmanned Aircraft System for Automatic Forest Fire Monitoring and Measurement. *Journal of Intelligent & Robotic Systems*, **65**, 533-548. https://doi.org/10.1007/s10846-011-9560-x

[3] Breitenmoser, A., Tâche, F., Caprari, G., *et al*. (2010) MagneBike: Toward Multi Climbing Robots for Power Plant Inspection. In: *Proceedings of the* 9*th International Conference on Autonomous Agents and Multiagent Systems*: *Industry Track*, International Foundation for Autonomous Agents and Multiagent Systems, Richland, 1713-1720.

[4] Knight, W. (2017) Drones and Robots Are Taking over Industrial Inspection. MIT Technology Review.

[5] DeBusk, W.M. (2010) Unmanned Aerial Vehicle Systems for Disaster Relief: Tornado Alley. AIAA Infotech@Aerospace

2010. https://doi.org/10.2514/6.2010-3506

[6] Heydari, J., Saha, O. and Ganapathy, V. (2021) Reinforcement Learning-Based Coverage Path Planning with Implicit Cellular Decomposition. arXiv:2110.09018.

[7] Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J. and Whiteson, S. (2018, July). Qmix: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *International Conference on Machine Learning*, 4295-4304.

[8] Kan, X., Teng, H. and Karydis, K. (2020) Online Exploration and Coverage Planning in Unknown Obstacle-Cluttered Environments. *IEEE Robotics and Automation Letters*, **5**, 5969-5976. https://doi.org/10.1109/LRA.2020.3010455

[9] Foerster, J., Farquhar, G., Afouras, T., *et al*. (2018) Counterfactual Multi-Agent Policy Gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 2974-2982. https://doi.org/10.1609/aaai.v32i1.11794