DOI: 10.11992/tis.201609017
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20170626.1740.020.html

强化学习的地–空异构多智能体协作覆盖研究

张文旭, 马磊, 贺荟霖, 王晓东
（西南交通大学 电气工程学院, 四川 成都 610031）

摘　要: 以无人机 (unmanned aerial vehicle, UAV) 和无人车 (unmanned ground vehicle, UGV) 的异构协作任务为背景, 通过 UAV 和 UGV 的异构特性互补, 为了扩展和改进异构多智能体的动态覆盖问题, 提出了一种地–空异构多智能体协作覆盖模型。在覆盖过程中, UAV 可以利用速度与观测范围的优势对 UGV 的行动进行指导; 同时考虑智能体的局部观测性与不确定性, 以分布式局部可观测马尔可夫 (decentralized partially observable Markov decision processes, DEC-POMDPs) 为模型搭建覆盖场景, 并利用多智能体强化学习算法完成对环境的覆盖。仿真实验表明, UAV 与 UGV 间的协作加快了团队对环境的覆盖速度, 同时强化学习算法也提高了覆盖模型的有效性。

关键词: 异构多智能体; 覆盖问题; 地–空; UAV/UGV; DEC-POMDPs; 强化学习
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)02−0202−06

中文引用格式: 张文旭, 马磊, 贺荟霖, 等. 强化学习的地–空异构多智能体协作覆盖研究[J]. 智能系统学报, 2018, 13(2): 202–207.
英文引用格式: ZHANG Wenxu, MA Lei, HE Huilin, et al. Air-ground heterogeneous coordination for multi-agent coverage based on reinforced learning[J]. CAAI transactions on intelligent systems, 2018, 13(2): 202–207.

Air-ground heterogeneous coordination for multi-agent coverage based on reinforced learning

ZHANG Wenxu, MA Lei, HE Huilin, WANG Xiaodong
(School of Electrical Engineering, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: With the heterogeneous coordinate task of unmanned aerial vehicles (UAVs) and unmanned ground vehicle (UGVs) as the background to this study, a novel air-ground heterogeneous coverage model for a coordinated multi-agent is proposed by the complementation between UAV and UGV heterogeneity, in order to extend and improve the dynamic coverage of a heterogeneous multi-agent system. During the coverage process, the advantages of mobility and the observation scope of the UAV were used in order to guide the actions of the UGV. Moreover, in view of the partial agent observability and uncertainty, decentralized and partially observable Markov decision processes (DEC-POMDPs) were applied as the model in order to establish the coverage environment. Additionally, the reinforced learning algorithm of multi-agents was utilized in order to complete the coverage of the environment. The simulation results revealed that the coverage process was accelerated by the cooperation of the UAV and UGV. Additionally, the reinforced learning algorithm also improved the effectiveness of the coverage model.

Keywords: heterogeneous multi-agent system; coverage; air-ground; UAV/UGV; DEC-POMDPs; reinforced learning

近年来, 多智能体覆盖问题得到了越来越多的关注[1], 并作为多智能体协调控制的一个重要研究方向, 有着重要的理论和应用价值, 在服务保障、工业制造、军事侦察、安全保卫、灾后搜救、星球探索、资源勘察等方面都有着广阔的应用前景[2], 其主要研究包括路径规划、动态避障、任务分配等方面[3-4]。

对于一个多智能体系统, 智能体的异构特性可以更大地发挥多智能体的优势, 更好地完成协作任务[5]。目前, 大多数的覆盖研究都基于智能体为同

收稿日期: 2016−09−21.　**网络出版日期:** 2017−06−26.
基金项目: 国家自然科学基金青年基金项目 (61304166).
通信作者: 张文旭. E-mail: wenxu_zhang@163.com.

构的假设前提, 而异构多智能体与覆盖问题的结合相对薄弱, 比如, 文献[6]在一阶动态异构覆盖问题中, 考虑不同的速度对应不同的控制输入, 设计了一种分布式覆盖控制策略; 文献[7]研究了非凸环境下的覆盖问题, 提出了一种梯度环境分割算法; 文献[8]在异构无线传感器网络中研究了覆盖与消耗的控制算法; 文献[9]介绍了一种基于加权 Voronoi 图的异构机器人覆盖框架, 根据异构覆盖代价进行加权, 实现代价最小的覆盖任务。针对异构多智能体的覆盖问题, 目前多智能体的异构性多体现在传感器的异构上, 即感知范围的不同, 少有研究从智能体运动方式的异构性上进行考虑。另一方面, 无人机 (unmanned aerial vehicle, UAV) 和无人车 (unmanned ground vehicle, UGV) 的异构特性协作是多智能体的前沿性研究课题[10], 它们在速度、负载、通信、观测能力等方面具有很强的互补性, 二者协作可以有效拓宽应用范围, 其应用价值受到了世界各国学者的广泛关注[11], 现有的工作主要集中在路径规划、搜索定位、跟踪追逃等方面, 比如, 文献[12]提出了一种 UAV 和 UGV 的合作导航策略, 利用 UAV 的大视野特性引导 UGV 避障; 文献[13]研究了多 UAV 和 UGV 的合作监控, 通过二者的观测数据融合完成对目标的侦查; 文献[14]基于 UAV 和 UGV 的合作框架研究了人群跟踪的决策和监控。但是, 针对 UAV 和 UGV 互补特性的协作覆盖问题尚未得到研究。

本文提出了一种地–空异构多智能体的协作覆盖模型, 针对未知环境下的动态覆盖问题, 依靠 UAV 机动性能与观测范围的优势, 在覆盖过程中对 UGV 的动作进行指导, 同时考虑了智能体的观测局部性和不确定性, 基于分布式局部可观测马尔可夫 (DEC-POMDPs) 模型建立栅格地图覆盖环境, 根据 UAV 和 UGV 的异构特性设计覆盖场景, 并利用多智能体强化学习算法完成对地图的覆盖。

1　问题描述

1.1　多智能体覆盖问题

覆盖问题大体上可分为静态与动态覆盖两类, 静态覆盖主要关注传感器位置的优化, 动态覆盖则要求智能体群组遍历区域内所有兴趣点。动态覆盖包含了导航与避障的研究内容, 目的是利用移动机器人或固定传感器, 在物理接触或传感器感知范围内遍历目标环境区域, 并尽可能地满足时间短、重复路径少和未遍历区域小的优化目标[2]。

本文考虑带有观测不确定性的异构多智能体动态覆盖问题, 以栅格地图为覆盖环境, UGV 作为覆盖执行者, UAV 则作为引导者。利用 UAV 观测范围广和移动速度快的优势对 UGV 的覆盖行动进行指导, 以扩大 UGV 的观测视野和提高团队对位置环境的获取准确性, UGV 不断移动直到栅格被覆盖到指定的程度。智能体的路径以栅格序号进行表示, 便于算法中地图信息和智能体状态的更新。

1.2　分布式马尔可夫模型

分布式控制是多智能体系统的一个重要特性, 由于智能体携带的传感器存在精度误差, 且覆盖环境复杂多变, 智能体的局部观测性和环境的不确定性将难以避免[12]。针对以上问题, 考虑采用分布式局部可观测马尔可夫模型 (DEC-POMDPs)[13], 其由一个八元组构成:

$$\left\langle I,S,\{A_i\},P,\{\Omega_i\},O,R,b^0\right\rangle \tag{1}$$

式中: I表示有限的智能体集合; S表示一个有限的系统状态集合; $\{A_i\}$表示智能体i可采取的动作的集合; P表示系统的转移; $\{\Omega_i\}$表示智能体i的观测集合; O表示观测函数; R表示回报函数; b^0为初始状态分布。求解 DEC-POMDPs 的目的是找到一个联合策略$\vec{\pi}=(\pi_1,\pi_2,\cdots,\pi_n)$使回报函数$R$最大化。

1.3　Q 学习

文献[14]提出了一类通过引入期望的延时回报, 求解无完全信息的马尔可夫决策过程的方法, 称为 Q-学习 (Q-learning)。Q-学习是一种与模型无关的基于瞬时策略的强化学习方法, 通过对状态–动作对的值函数进行估计, 以求得最优策略。Q-学习算法的基本形式如下:

$$Q^*(s,a)=R(s,a)+\gamma\sum_{s'\in S}P(s,a,s')\max Q^*(s',a') \tag{2}$$

式中: $Q^*(s,a)$表示智能体在状态s下采用动作a所获得的奖赏折扣总和, γ为折扣因子, $P(s,a,s')$表示概率函数。对于一个动态覆盖问题而言, 强化学习算法的优势在于, 智能体无需提前了解环境模型, 它可以通过与环境的交互来获得状态信息, 并通过反馈的覆盖效果对所采取的行动进行评价, 利用不断的试错和选择, 逐步改进和完善覆盖策略, 达到覆盖重复路径少、覆盖时间短等优化目标。

2　覆盖问题设计

2.1　异构多智能体设计

对于异构多智能体系统, 首先需要对单个智能体的特性进行分析。UGV 能够装载大容量动力装置和大型精密仪器, 具备较高的数据处理运算能力, 但移动速度慢, 视野范围小, 在障碍物密集的区域, 行动能力受到极大限制; 相比之下, UAV 具有较高的移动速度和空间灵活性, 移动过程中不需要考

虑地面复杂的障碍环境，然而它的实时运算能力、负载能力和电量荷载受到较大限制。

根据 UGV 和 UAV 的上述特性，在地–空异构多智能体覆盖问题中，如图 1 所示，UAV 以五角星表示，定义 UAV 采取类似于摄像头抽象环境扫描算法，在环境中的观测范围为一个扫描半径为 2 个栅格的圆形区域，如虚线区域所示，其中 12 个阴影栅格为 UAV 的观测，并据此获得相关观测矩阵。

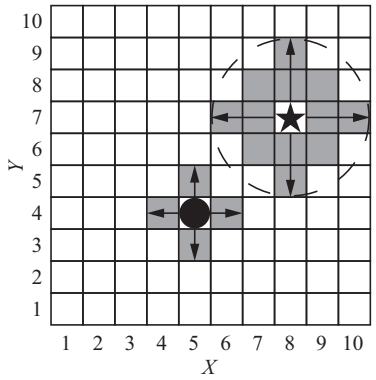

图 1　UAV and UGV 的异构观测

Fig. 1　The heterogeneous observation of UAV and UGV

UAV 获得的观测信息不仅用于决策 UAV 的下一步移动动作，还需要向 UGV 提供额外的地图环境信息。其次，考虑到 UAV 的速度异构特性，定义其移动速度为每步 2 个栅格，图中箭头表示智能体的移动方向。UGV 以圆圈表示，不同于 UAV 具备广阔的高空视野，UGV 的观测范围较小，定义其观测为前、后、左、右 4 个栅格，即周围的阴影栅格，设定移动速度为每步 1 个栅格。UGV 的优势在于对环境信息的测量精度要高于 UAV。

2.2　覆盖场景设计

定义 1　基于 DEC-POMDPs 的覆盖环境需要体现出多智能体的异构性、分布式和不确定性，其组成类似于式 (1)，可以抽象为一个 8 元组 $\langle I,\boldsymbol{S},\{A_i\},P,\{\Omega_i\},O,R,b^0\rangle$ 结构，其中：

I 为智能体数量集合：$I=\{1,2,3\}$，异构多智能体系统包含 3 个智能体，其中编号为 1 和 2 的智能体为 UGV，编号为 3 的智能体是 UAV。

$\boldsymbol{S}$ 为状态矩阵：用来描述整张地图上各栅格被访问的情况，即各智能体自身的状态，状态集合为 $\boldsymbol{S}:J\times L$。其中 J 表示地图被覆盖的情况，地图上每个栅格的状态信息又可表示为尚未访问状态 s_1、已访问状态 s_2、障碍物状态 s_3 3 类情况。L 包含第 i 个智能体在地图上的位置 P_i。

Ω_i 为观测集合：表示第 i 个智能体的观测集合。对 UGV 而言，$\Omega_i(t)=\{\mathrm{env}_i(t),\mathrm{pos}_i(t),\mathrm{pos}_j(t)\}$，$i\in[1,2]$，依次描述 t 时刻第 i 个 UGV 的自身局部观测信息 $\mathrm{env}_i(t)$、自身的位置信息 $\mathrm{pos}_i(t)$ 以及根据通信获得的其他 UGV 位置信息 $\mathrm{pos}_j(t)$，设定 UGV 无法观测到 UAV 的位置；对于 UAV 而言，观测集合 $\Omega_3(t)=\{\mathrm{env}_3(t),\mathrm{pos}_1(t),\mathrm{pos}_2(t),\mathrm{pos}_3(t)\}$，依次描述 t 时刻 UAV 自身对环境的观测、各个 UGV 的相对位置 (当 UGV 处于 UAV 传感范围时)、UAV 自身的位置。在环境观测矩阵 $\mathrm{env}_i(t)$ 中，包含智能体观测范围内 n 个点的环境观测信息集合 $j(t)=\{\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4\}$，其中 ε_1 表示尚未访问的栅格，ε_2 表示已经访问的栅格，ε_3 表示障碍物栅格，ε_4 表示 UGV。智能体的传感器感应范围内出现其他 UGV 时，能根据该 UGV 对环境的状态感知获取其相对位置，对于一辆 UGV 而言，其观测范围内的其他 UGV，作为障碍物进行考虑。

O 为观测概率函数：体现了智能体观测的不确定性，$O(o|s,a)$ 表示智能体执行 a 后转移到状态 s 时获得观测 o 的概率。智能体从所处环境中获取观测信息可以用概率矩阵进行表示，假设栅格地图上观测点的观测函数相同，定义其观测–状态概率分布函数如表 1 所示。

表 1　观测-状态概率分布函数

Table 1　The probability distribution function of observation-state

概率分布	s_1	s_2	s_3
ε_1	0.9	0.1	0
ε_2	0.1	0.8	0.1
ε_3	0	0.1	0.9

$O(s_1|\varepsilon_1,a)=0.9$ 表示在执行 a 后，到达真实状态 s_1 时，s_1 为 ε_1 的观测概率为 0.9。

A_i 为动作集合：表示第 i 个智能体的动作。对于 UGV 和 UAV，t 时刻可能产生的动作为 $A_i(t)=\{\mathrm{up},\mathrm{down},\mathrm{left},\mathrm{right}\}$。

R 为回报函数：表示环境对智能体的行动给出的评价。对于 UGV，执行一次行动后存在着“没走过”、“走过”和“障碍物”3 个状态，分别对应着 30、–5 和–10 的回报值，栅格的边界作为障碍物考虑。在覆盖问题中，UAV 对 UGV 的观测起指导作用，所以 UAV 的回报由两部分组成，第 1 部分为 UAV 自身的回报，和 UGV 的回报定义相同，第 2 部分为 UGV 反馈的回报，其定义为

$$R_{\mathrm{UAV}}=\mu\cdot R_{\mathrm{UAV}}+(1-\mu)\cdot\sum R_{\mathrm{UGV}} \tag{3}$$

式中：μ 为权重系数，当 UAV 的观测范围里没有 UGV 时 $\mu=1$。

b^0 为初始信念状态：智能体根据初始信念状态和初始 Q 值函数获取当前应选择的动作向量。其更新公式为

$$b^t(s(t+1)) = O(s_1|\varepsilon,a)\sum_{s\in S} b^t(s(t))P(s(t+1)|s(t),a) \quad (4)$$

3　基于强化学习的覆盖算法

3.1　异构多智能体学习决策

在覆盖场景中, 我们将 UGV 设定为任务执行者, 负责访问地图上尚未被探索的栅格, 而将 UAV 设定为作团队中的督导者, 通过通信向 UGV 提供更广阔的视野信息, 配合 UGV 建立更精确的信念状态, 实现更高效的覆盖。

考虑到智能体的结构异构性和局部观测性, 假设 UAV 可以向观测范围内的 UGV 进行单向通信, 并发送 UAV 的观测信息, 而 UGV 之间不能进行通信。UAV 的强化学习一步策略更新的流程如图 2 所示。

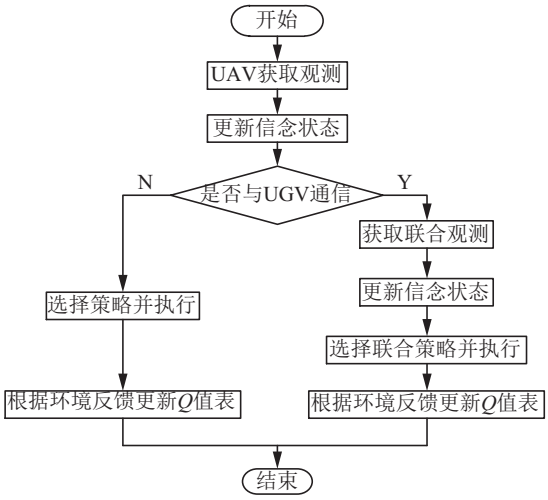

图 2　UAV 强化学习一步策略更新流程

Fig. 2　The one-step strategy update flow of reinforcement learning of UAV

UGV 获得的观测能够被分为两类: 1) 根据智能体自身传感器获得的局部观测信息Ω_{local}; 2) 依赖通信行为获得的 UAV 的观测信息Ω_{other}, 则联合观测表示为$\Omega_{\text{joint}} = \{\Omega_{\text{local}}, \Omega_{\text{other}} \cup \varnothing\}$, $\Omega_{\text{local}} \in \Omega_{\text{joint}}$。

由于局部观测性的存在, UGV 不一定在所有时刻都能获得 UAV 的观测信息, 本文用类似文献[15]所提通讯受限的多智能体在线规划算法的思想, 将学习过程分为可以通信与不能通信两种情况。在 DEC-POMDPs 模型中嵌入多个局部可观察马尔可夫决策过程 (partially observable Markov decision processes, POMDP) 模型作为辅助学习单元, 在 POMDP 模型中使用最大似然算法, 如表 1 所示, 并将局部状态近似看作全局状态。当执行策略更新时, 依照观测来源将观测划分为局部观测Ω_{local}和联合观测Ω_{joint}两类, 强化学习框架如图 3 所示。

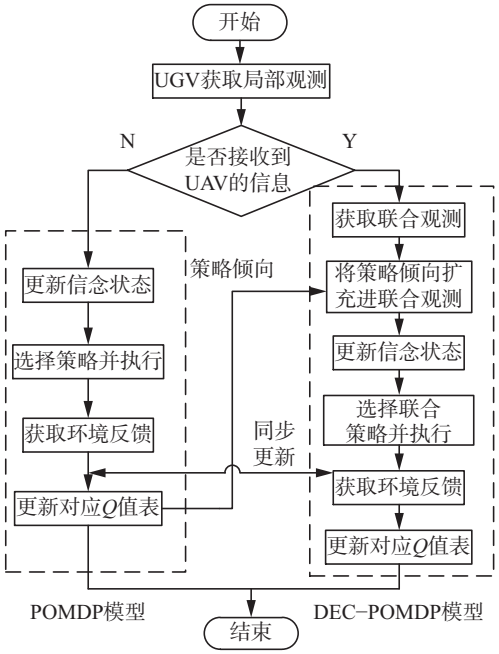

图 3　异构多智能体强化学习框架

Fig. 3　The frame of reinforcement learning of heterogeneous multi-agent

当智能体团队执行联合行动, 并获取联合观测Ω_{joint}后, 也获得相应的局部观测Ω_{local}信息, 此时从 POMDP 对应的 Q 值表中获取局部观测Ω_{local}对应的动作a_k, 并将其作为策略倾向在联合观测中扩充观测矩阵。另外, 在智能体获取环境反馈后, 更新 DEC-POMDPs 模型相应的 Q 值表的同时, 由于$\Omega_{\text{local}} \in \Omega_{\text{joint}}$, 同步更新 POMDP 模型 Q 值表中与Ω_{local}对应的键值。当 UAV 和 UGV 的观测范围出现重叠时, 考虑到智能体观测精度的异构特性, 栅格地图的联合观测状态为$O_{\text{joint}} = \beta \cdot O_{\text{UAV}} + (1-\beta)\cdot O_{\text{UGV}}$, 其中, β 为权重系数。

3.2　基于强化学习的覆盖算法

解决强化学习问题主要是找到一个策略使智能体团队最终达到最大的奖励信号。如果在所有状态下, 策略π都大于或等于策略π'的期望回报值, 那么称这个策略为最优策略, 记作π^*。而最优策略对应的状态–联合动作对$(s,\boldsymbol{a})$也有相同的最优值函数, 记作Q^*。在 POMDP 模型下, 智能体i在s状态下执行行动a获得的 Q 值为

$$Q_i(s(t),a) = R(s(t),a) + \sum_{s\in S}\sum_{o\in\Omega} P(s(t+1)|s(t),a)\,O(o|s(t+1),a)\,V(s(t+1),a) \quad (5)$$

Q 学习更新公式为

$$Q_t(s(t),a) = (1-\alpha)Q_t(s(t),a) + \alpha\left[R(s(t),a) + \max_a\{Q_t(s(t+1),a)\}\right] \quad (6)$$

DEC-POMDPs 与 POMDP 的唯一区别在于智能体的数量由单个变为多个, 其 Q-学习迭代表达式与 POMDP 类似, 智能体的行动由单独行动a变为联

合行动$\boldsymbol{a}$：

$$Q_t(s(t),\boldsymbol{a})=(1-\alpha)Q_t(s(t),\boldsymbol{a})+\alpha\left[R(s(t),\boldsymbol{a})+\max_a\{Q_t(s(t+1),\boldsymbol{a})\}\right] \tag{7}$$

4 仿真结果

仿真实验考虑一个20×20大小的栅格地图环境，如图 4 所示，最外围是地图边界，黑色区域表示障碍物，智能体的初始位置固定，五角星表示 UAV，圆圈表示 UGV。智能体团队的任务为尽可能多地访问到所有栅格，即完成对格子世界的覆盖。当走过的栅格超过 95% 以上时，认为此次覆盖任务成功；当智能体在 1 500 步仍不能完成 95% 的覆盖时，认为此次任务失败。定义学习率为 0.6，折扣因子为 0.2，μ=0.4，β=0.3。仿真实验在 MATLAB 2012b 环境下进行，并利用 Mysql 数据库存储 Q 值表。

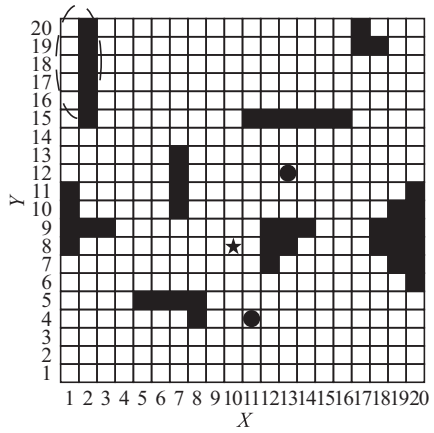

图 4　多智能体覆盖环境

Fig. 4　The coverage environment of multi-agent

执行 1 000 幕覆盖实验后的学习效果如图 5 所示，可以看出随着学习幕数的增加，经过 700 幕左右学习后，智能体团队完成地图覆盖所需步数逐渐收敛到较稳定的值，其中虚线为覆盖步数拟合曲线，图中覆盖步数存在的毛刺原因为智能体的观测带有不确定性，当观测信息出现错误时，可能使智能体当前学习幕的覆盖完成步数出现波动。

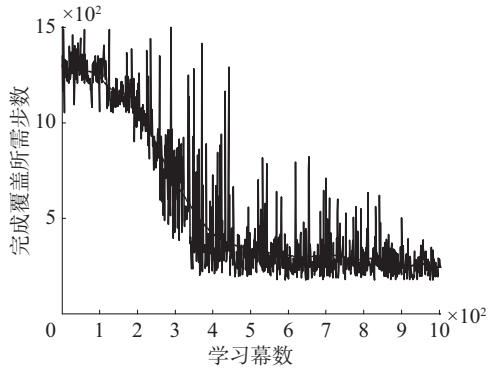

图 5　异构多智能体覆盖完成步数

Fig. 5　The coverage steps of heterogeneous multi-agent

图 6 对比了 UAV 加入任务时的覆盖成功率，图中实线表示一个 UAV 和两个 UGV 组成的异构多智能体团队，虚线表示只有两个 UGV 组成的团队。从图中可以看出，两种智能体团队对地图的覆盖成功率都随着强化学习算法的迭代不断得到提高。但是，在存在 UAV 的团队中，因为 UAV 可以对 UGV 的覆盖行动进行指导，所以在经过 700 幕左右学习时，团队覆盖成功率就开始趋于稳定，而只有 UGV 的团队，需要 900 幕左右的学习才开始趋于稳定，因此体现出 UAV 与 UGV 协作覆盖的优势。

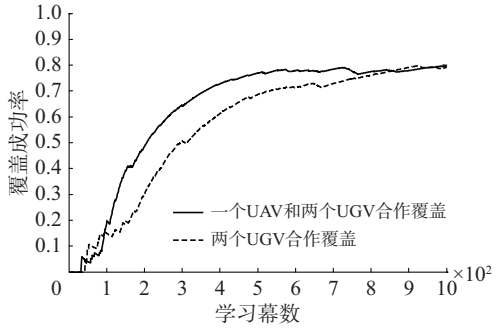

图 6　覆盖试验成功率

Fig. 6　The success rate of coverage

最后，在地图左上角设置了一个影响整体覆盖效果的“陷阱”区域，用以进一步的验证在强化学习过程中，UAV 对 UGV 的引导效果。陷阱区域如图 4 中虚线圈区域所示，为边界与障碍物所夹的 6 个栅格，访问此区域的回报$R=0.3$，低于访问其他空旷区域的回报。当覆盖率达到 95% 时，认为本次覆盖任务成功，但陷阱区域属于不应该访问的 5% 部分，每幕覆盖实验结束后，记录陷阱区域被访问的次数，每 20 个学习幕进行一次采样。

图 7 对比了 UAV 加入覆盖任务时对陷阱区域的访问效果，由图中可以看出，两种智能体团队对陷阱区的访问次数，都将随着学习幕数的增加而逐渐减少，最终将不再访问陷阱区，体现了强化学习算法对于覆盖问题的有效性。但是，在只有两个 UGV 组成的团队进行覆盖任务时，由于 UGV 的观测范围较小，团队需要更多的学习幕数后，才能减少对陷阱区域的访问。

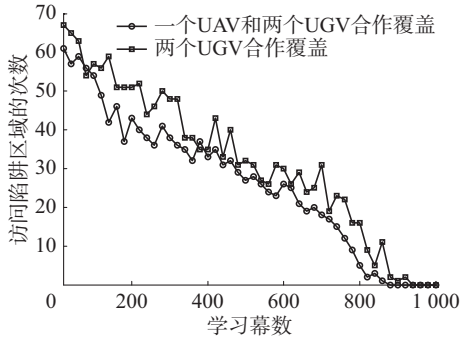

图 7　陷阱区域访问次数统计

Fig. 7　The count of visits to the trap

5 结束语

本文探讨了异构多智能体与动态覆盖问题的结合, 以 UAV 和 UGV 的异构协作任务为背景, 提出了一种地–空异构多智能体协作覆盖模型。根据 UAV 和 UGV 的异构特性, 设计了 UAV 和 UGV 互补的覆盖观测方法, 同时考虑到智能体观测的局部性和不确定性, 以 DEC-POMDPs 为模型建立覆盖场景, 并利用多智能体强化学习算法完成了对环境的覆盖。进一步工作主要包括: 1) 在强化学习动作选择中考虑 UAV 和 UGA 的动力学模型; 2) 在 UAV 与 UGV 的互补特性中考虑分布式系统的信息融合问题, 以提高学习收敛速度。

参考文献:

[1] KANTAROS Y, ZAVLANOS M M. Distributed communication-aware coverage control by mobile sensor networks [J]. Automatica, 2016, 63: 209–220.

[2] 蔡自兴, 崔益安. 多机器人覆盖技术研究进展[J]. 控制与决策, 2008, 23(5): 481–486, 491.
CAI Zixing, CUI Yi'an. Survey of multi-robot coverage[J]. Control and decision, 2008, 23(5): 481–486, 491.

[3] MAHBOUBI H, MOEZZI K, AGHDAM A G, et al. Distributed deployment algorithms for improved coverage in a network of wireless mobile sensors[J]. IEEE transactions on industrial informatics, 2014, 10(1): 163–174.

[4] TAO Dan, WU T Y. A survey on barrier coverage problem in directional sensor networks[J]. IEEE sensors journal, 2015, 15(2): 876–885.

[5] TIAN Yuping, ZHANG Ya. High-order consensus of heterogeneous multi-agent systems with unknown communication delays[J]. Automatica, 2012, 48(6): 1205–1212.

[6] SONG Cheng, LIU Lu, FENG Gang, et al. Coverage control for heterogeneous mobile sensor networks on a circle[J]. Automatica, 2016, 63: 349–358.

[7] KANTAROS Y, THANOU M, TZES A. Distributed coverage control for concave areas by a heterogeneous robot-swarm with visibility sensing constraints[J]. Automatica, 2015, 53: 195–207.

[8] WANG Xinbing, HAN Sihui, WU Yibo, et al. Coverage and energy consumption control in mobile heterogeneous wireless sensor networks[J]. IEEE transactions on automatic control, 2013, 58(4): 975–988.

[9] SHARIFI F, CHAMSEDDINE A, MAHBOUBI H, et al. A distributed deployment strategy for a network of cooperative autonomous vehicles[J]. IEEE transactions on control systems technology, 2015, 23(2): 737–745.

[10] CHEN Jie, ZHANG Xing, XIN Bin, et al. Coordination between unmanned aerial and ground vehicles: a taxonomy and optimization perspective[J]. IEEE transactions on cybernetics, 2016, 46(4): 959–972.

[11] ZHOU Yi, CHENG Nan, LU Ning, et al. Multi-UAV-aided networks: aerial-ground cooperative vehicular networking architecture[J]. IEEE vehicular technology magazine, 2015, 10(4): 36–44.

[12] PAPACHRISTOS C, TZES A. The power-tethered UAV-UGV team: a collaborative strategy for navigation in partially-mapped environments[C]//Proceedings of 22nd Mediterranean Conference of Control and Automation. Palermo, Italy, 2014: 1153–1158.

[13] GROCHOLSKY B, KELLER J, KUMAR V, et al. Cooperative air and ground surveillance[J]. IEEE robotics and automation magazine, 2006, 13(3): 16–25.

[14] KHALEGHI A M, XU Dong, WANG Zhenrui, et al. A DDDAMS-based planning and control framework for surveillance and crowd control via UAVs and UGVs[J]. Expert systems with applications, 2013, 40(18): 7168–7183.

[15] 马磊, 张文旭, 戴朝华. 多机器人系统强化学习研究综述[J]. 西南交通大学学报, 2014, 49(6): 1032–1044.
MA Lei, ZHANG Wenxu, DAI Chaohua. A review of developments in reinforcement learning for multi-robot systems[J]. Journal of southwest Jiaotong university, 2014, 49(6): 1032–1044.

[16] PUTERMAN M L. Markov decision processes: discrete stochastic dynamic programming[M]. New York: John Wiley and Sons, 1994.

[17] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine learning, 1992, 8(3/4): 279–292.

[18] WU Feng, ZILBERSTEIN S, CHEN Xiaoping. Online planning for multi-agent systems with bounded communication[J]. Artificial intelligence, 2011, 175(2): 487–511.

作者简介:

张文旭, 男, 1985 年生, 博士研究生, 主要研究方向为多智能体系统、机器学习, 发表学术论文 4 篇, 其中被 EI 检索 4 篇。

马磊, 男, 1972 年生, 教授, 博士, 主要研究方向为控制理论及其在机器人、新能源和轨道交通系统中的应用等, 主持国内外项目 14 项, 发表学术论文 40 余篇, 其中被 EI 检索 37 篇。

贺荟霖, 女, 1993 年生, 硕士研究生, 主要研究方向为机器学习。

王晓东, 男, 1992 年生, 硕士研究生, 主要研究方向为机器学习, 获得国家发明型专利 3 项, 发表学术论文 4 篇。