山西大学学报(自然科学版)47(2):333—345,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023042

基于匹配博弈和通信机制的主从协作群智感知算法

胡治国[1,2*],杜瑞芳[1],秦雪健[1],刘郭庆[1]

(1.山西大学 计算机与信息技术学院,山西 太原 030006;
2.同济大学 嵌入式系统与服务计算教育部重点实验室,上海 201804)

摘　要:利用强化学习制定多智能体的移动群智感知策略是学术界普遍采用的技术手段,但是其普遍存在训练过程不加处理地利用所有智能体的状态和动作信息,学习更新效率较低,且对被采集对象重要程度的差异性缺少考虑等问题。本文以指挥员-无人机主从协作场景为研究对象,提出了一种基于匹配博弈和通信机制的主从协作群智感知算法。首先,通过引入Gale-shapely匹配博弈算法思想,建立无人机能量属性与待采集目标数据质量属性之间的最优稳定匹配,实现基于数据重要程度优先的采集策略。为保证无人机对高质量目标的持续特定关注,本文结合了当前较为流行的通信规则型多智能体强化学习算法MAAC(Multi-Actor-Attention-Critic)框架,引入了多注意力机制模块,实现了数据采集过程中主-从智能体间高效的信息交流与共享。实验表明,我们提出的c-MGCM(Crowd-sensing Method based on Matching Game and Communication Mechanism)方法在奖励值、匹配对的距离值等多个评价指标上都优于MADDPG(Multi-agent Deep Deterministic Policy Gradient)、DDPG(deep Deterministic Policy Gradient)等经典算法,在奖励值方面有2~3倍的提升,在数据质量方面有至少14%的提升。该结果表明了c-MGCM方法的高效性和稳定性。

关键词:群智感知;数据采集;匹配博弈;通信机制
中图分类号:O436　　**文献标志码**:A　　**文章编号**:0253-2395(2024)02-0333-13

A Master-slave Collaborative Mobile Crowd-sensing Algorithm Based on Matching Game and Communication Mechanism

HU Zhiguo[1,2*], DU Ruifang[1], QIN Xuejian[1], LIU Guoqing[1]

(1.*School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China;*
2. *Key Laboratory of Embedded System and Service Computing of Ministry of Education (Tongji University), Shanghai* 201804*, China*)

Abstract: The use of reinforcement learning to formulate mobile swarm intelligence perception strategies for multiple intelligences is a common technical approach in academia, but it commonly suffers from problems such as the training process using the state and action information of all intelligences without processing, low learning update efficiency, and lack of consideration of the variability in the importance of the collected objects. In this paper, a master-slave collaborative swarm intelligence perception algorithm based on matching game and communication mechanism is proposed for the commander-UAV master-slave collaboration scenario. Firstly, by introducing the idea of Gale-shapely matching game algorithm, an optimal and stable matching between UAV energy attributes and data quality attributes of the target to be collected is established, and an acquisition strategy based on the priority of data importance is realized. In addition, to ensure the UAV's continuous and specific attention to high-quality targets, we combine the MAAC

收稿日期:2022-11-10;**接受日期**:2023-03-08
基金项目:山西省基础研究计划(20210302123455);嵌入式系统与服务计算教育部重点实验室开放课题(同济大学)(ES-SCKF 2021-04)
* **通信作者**:胡治国(1977-),男,山西灵石人,博士,副教授,主要研究方向为计算机网络。E-mail:huzhiguotj@sxu.edu.cn
引文格式:胡治国,杜瑞芳,秦雪健,等.基于匹配博弈和通信机制的主从协作群智感知算法[J].山西大学学报(自然科学版),2024,47(2):333-345. DOI:10.13451/j.sxu.ns.2023042

(Multi-Actor-Attention-Critic) framework, which is currently a more popular communication rule-based multi-agent reinforcement learning algorithm, and introduces a multi-attention mechanism module to achieve efficient information exchange and sharing between master-slave intelligences in the data acquisition process. Experiments show that our proposed Crowd-sensing Method based on Matching Game and Communication Mechanism (c-MGCM) method outperforms classical algorithms, such as MADDPG and DDPG, in several evaluation metrics such as reward value and distance value of matched pairs, with a 2-3 times improvement in reward value and at least 14% improvement in data quality. The results demonstrate the efficiency and stability of the c-MGCM method.

Key words: mobile crowd-sensing; data collection; matching game; communication mechanism

0 引言

移动群智感知(Mobile Crowd-sensing,MCS)是指利用移动设备的感知能力来获取数据的一种新型数据采集模式[1]。与传统基于传感器的MCS方式相比,基于无人机群(Unmanned Aerial Vehicles,UAVs)的MCS方式,凭借其部署的灵活性、快捷的移动性及可搭载设备的多样性,为移动群智感知提供泛在服务赋予了更大的潜力[2-3]。

基于无人机群智感知场景下的数据收集问题,当前研究大致分为两类:一类是基于凸优化的方法,其核心思想是将多约束数据收集问题(通常为非凸问题)分解成若干个子问题,来降低问题求解的复杂度,进而转化为凸优化问题进行求解[4-6];另一类是基于强化学习的方法,这其中又以深度强化学习和多智能强化学习方法最具代表性,并已在复杂约束下的数据采集问题中取得了重大的进展[7-9]。通过对相关研究成果的分析,我们发现存在以下不足:1)现在方法通常以待采集目标的数量多少作为策略优化目标,没有考虑待采集对象在数据质量上的差异(即数据重要程度不同、数据生存时间不同等方面);2)基于强化学习的数据采集策略大多基于中心化训练的方式,在更新学习策略的参数时,使用的是来自其他所有智能体的状态和动作信息,对相关信息没有进行有效的甄别、筛选,导致方法学习效率较低。

针对上述问题,我们将诺贝尔经济学奖得主劳埃德·沙普利提出的Gale-shapely算法中的匹配博弈思想[10]引入到数据采集领域,以期建立智能体能量属性与待采集数据重要程度属性之间稳定的匹配关系;采用多头注意力及通信交流机制将每个智能体的状态-动作进行编码,得出了智能体之间相互影响作用的权重大小,以期实现智能体间信息的有效传递和融合。基于上述思想,我们提出了c-MGCM(Crowd-sensing Method based on Matching Game and Communication Mechanism)算法,与其他研究相比,c-MGCM算法不仅可完成数据差异性环境中的数据采集任务,而且方法具有高效性和稳定性。

本文第1节讨论了本领域相关工作进展;第2节部分讨论了本论文所针对的具体问题,及完成了相关问题的形式化建模;第3节详细介绍本论文提出的c-MGCM算法主要步骤;第4节对比验证了c-MGCM算法性能;第5节是总结与展望。

1 相关工作

利用移动感知设备实现数据的有效采集是群智感知领域的基础性问题。早期研究将该问题转化为传统的旅行推销员问题(Traveling Salesman Problem,TSP),这种方法比较适用于简单约束条件下的智能体采集数据轨迹规划问题的求解,通常可给出针对不同约束下TSP问题的精确解。如文献[11]提出了Lin-Kernighan-Helsgaun算法,其算法复杂度为$O(n^{2.2})$(n为拓扑图节点数)。文献[12]提出了EARTH(efficient path planning for reliable data gathering)算法,其算法复杂度为$O(n^3)$,等等。但上述方法通常不适用于多目标及多智能体的群智感知环境。

近年来,多约束条件下群智感知领域成为研究热点,其研究方法可分为基于凸优化方法及基于强化学习的方法两类。一般来说,复杂环境中的多约束群智感知问题通常是非凸,难以直接求解,因此须首先将非凸优化任务分解为若干单独的凸目标问题求解,最终得到原问题

的近似解。如,文献[13]将非凸MCS任务解耦为无人机轨迹和唤醒时间分配规划两个优化子问题。文献[14]使用近似算法和分层分解方法将非凸MCS任务转换为数据速率、流路由和逗留时间分配子任务。目前,连续凸近似(Successive Convex Approximation)方法是常用的求解策略。但是,基于凸优化方法通常要求智能体面对的外部环境是可预测或已知的。显然,上述方法与群智感知现实场景还存在差异。

由于强化学习是通过与环境的交互来达成策略回报最大化,其对感知环境的适用性远远优于基于凸优化的方法,因此在群智感知领域获得了广泛的应用。如文献[7]结合了DDPG与Actor-Critic框架,提出了"e-Divert"方法,实现能源消耗最小化和数据采集数量的最大化。文献[8]利用MADDPG方法,提出了"Edics"方法,实现了数据采集、公平率、能源消耗三者之间的均衡,等等。但是目前基于强化学习的方法,数据采集策略主要是基于中心化训练的方式得出,训练过程不加处理地利用所有智能体的状态和动作信息,效率较低;且通常仅以被采集对象的数量多少作为优化目标,而对被采集对象重要程度的差异性缺少针对性设计。

近期,基于通信规则的强化学习兴起[15-17],其核心思想是通过智能体间发送和接收抽象的通信信息来优化智能体的策略、协调智能体间动作,为在复杂环境下多智能体协同完成任务提供新的方法手段,并已在战争游戏[18]、红蓝帽子问题[19]等领域获得了巨大的成功。

2 问题建模与描述

2.1 问题描述

本论文以无人机群智感知领域中最常用多智能"主-从"协作场景为研究对象,场景要素由指挥者、无人机、待采集目标组成,其中指挥者与无人机之间构成"主-从"关系,即指挥者处于主导、指挥位置,无人机处于服从、行动角色。本论文设置环境区域存在 $\mathcal{P}\triangleq\{p=1,2,\cdots,P\}$ 个指挥者,每个指挥者在单次数据采集过程中的位置固定;存在 $\mathcal{U}\triangleq\{u=1,2,\cdots,U\}$ 个无人机,每个无人机在数据采集过程中处于自由运动状态,用 $\beta(u)$ 来表示无人机 u 用于采集数据的初始能量;存在 $\mathcal{M}\triangleq\{m=1,2,\cdots,M\}$ 类目标,每类目标位置固定,$h(m)$ 表示目标 m 的数据重要程度;v^u 表示无人机 u 速度。

假设完成全部采集任务需要 $\mathcal{T}\triangleq\{t=1,2,\cdots,T\}$ 时间步,每一次采集任务开始时,指挥者、无人机、目标位置是随机的。本论文重点讨论目标数据重要程度存在差异的环境中,无人机如何依靠其与指挥者的通信信息到达目标数据位置,实现无人机与采集对象的优化匹配,实现采集策略的优化高效。不失一般性,本论文假设数据的重要程度越高,则数据的质量越大。

2.2 问题建模

本论文将无人机数据采集过程构建为马尔可夫决策过程,定义为 $M=<S,\mathcal{A},F,R,\gamma>$。

状态空间:$S=\{s=(s_p,s_u)|p\in\mathcal{P},u\in\mathcal{U}\}$。$s_p$ 表示指挥者 p 的状态空间,包括三个部分:$s_p=(i_u,Poi_t^u,Poi_t^m)$,$i_u$ 表示被指挥无人机的 id 号,Poi_t^u 表示被指挥无人机 u 在时刻 t 的位置信息,Poi_t^m 表示目标 m 在时刻 t 的位置信息。s_u 表示无人机的状态空间,包括四个部分:$s_u=(i_p,C_p,Poi_t^u,Vel_t^u)$,$i_p$ 表示无人机 u 所对应的指挥者 p 的 id 号,C_p 表示无人机 u 所对应的指挥者 p 的通信信息,Poi_t^u 表示无人机 u 在时刻 t 所处的位置信息,Vel_t^u 表示无人机 u 在时刻 t 的速度信息。

动作空间:$\mathcal{A}=\{a_t=(a_t^p,a_t^u)\}$,$a_t^p=(a_t^{p_1},a_t^{p_2},a_t^{p_3},a_t^{p_4},a_t^{p_5})$ 为指挥者 p 的动作空间,表示指挥者 p 所要传递的通信信息;$a_t^u=(a_t^{u_1},a_t^{u_2},a_t^{u_3},a_t^{u_4},a_t^{u_5})$ 为无人机 u 的动作空间,表示无人机 u 所要采取的动作行为,其中 $a_t^{u_1}$、$a_t^{u_2}$、$a_t^{u_3}$、$a_t^{u_4}$、$a_t^{u_5}$ 分别代表无人机上、下、左、右、停留五个动作。

状态转移概率:$F:S\times\mathcal{A}\times S\to[0,1]$ 表示状态转移概率矩阵,$F(s_t,a_t,s_{t+1})$ 表示智能体在当前状态 s_t 下采取行为 a_t 后随机转移到下一状态 s_{t+1} 的概率。

奖励函数:$R:S\times\mathcal{A}\times S\to\mathbb{R}$,表示无人机在当前状态 s_t 下采取行为 a_t 后转移到下一状态 s_{t+1} 所获得的即时奖励值。本实验中奖励值包括两部分:

$$r_t=r_t^p+r_t^d。\tag{1}$$

(a)$r_t^p=\lambda_1 l(u,m)$,其中 $l(u,m)$ 为无人机与

目标数据之间的距离,$\lambda_1<0$。r_t^p表示奖励值与无人机与目标数据之间的距离呈负相关,即距离越大,奖励值越小,它可以促进无人机向目标数据的逐渐靠近。

(b) $r_t^d=\begin{cases}C_1 & l(u,m)<(r_u+r_m)\lambda_2\\ 0 & \text{otherwise}\end{cases}$,其中$r_u$、$r_m$分别表示无人机与目标数据的半径,$C_1>0$,$\lambda_2>0$。它表示当无人机与目标数据之间的距离小于某个常数$\lambda_2$时,便会获得一个正奖励$C_1$,促进无人机与目标数据的靠近。

本文实验中,设定指挥者与被指挥无人机所获得奖励值是一致的。

折扣因子:γ,表示未来奖励对当前智能体动作选择的影响程度,$\gamma\in[0,1]$。

智能体在学习最优策略的过程就是最大化累计回报的过程:

$$J_i(\pi_i)=E_{a\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_i(s_t,a_t)\right]。\tag{2}$$

3 c-MGCM算法

针对上述问题场景和模型,本论文提出c-MGCM算法来实现无人机对目标数据的高效采集。c-MGCM算法主要由两部分构成。其一,通过引入Gale-shapely匹配博弈算法思想,建立无人机能量属性与待采集目标数据质量属性之间的最优稳定匹配,实现基于数据重要程度优先的采集策略;其二,为保证无人机对高质量目标的持续、特定关注,结合了当前较为流行的通信规则型多智能体强化学习算法MAAC框架,引入了多注意力机制模块,实现了数据采集过程中主-从智能体间高效的信息交流与共享。

3.1 无人机-目标数据匹配博弈策略

本文借鉴诺贝尔经济学奖得主劳埃德·沙普利提出的Gale-shapely匹配博弈算法思想,实现了无人机与目标数据间的稳定匹配,具体过程如下。

设无人机集合为$U=\{u_1,u_2,\cdots,u_i,\cdots,u_U\}$,无人机的能量集合为$\beta=\{\beta_1,\beta_2,\cdots,\beta_i,\cdots,\beta_U\}$,其中$\beta_i$表示无人机$u_i$的能量。设$SP=\{sp_1,sp_2,\cdots,sp_j,\cdots,sp_M\}$表示数据点集合,数据点的数据质量集合为$Q=\{q_1,q_2,\cdots,q_j,\cdots,q_M\}$,其中$q_j$代表数据点$sp_j$的数据质量,每个无人机和数据点都有各自的偏好列表,偏好列表中的排序代表他们对于对方的每一个参与者的匹配意愿。用μ表示一个匹配集合,$\mu(sp_j)$表示与sp_j匹配的无人机,$\mu(u_i)$表示与u_i匹配的数据点。$\mu(sp_j)=u_i$和$\mu(u_i)=sp_j$都可以表示无人机u_i与数据点sp_j构成的匹配,如果无人机u_i没有匹配,表示为$\mu(u_i)=\varnothing$。用$\succ$符号表示“优于”,用$sp_j\succ\mu(u_i)$表示在u_i的偏好列表中,sp_j排在$\mu(u_i)$的前面,也就是说,sp_j是u_i比当前已有匹配更想要的匹配对象,任何非空匹配都优于空匹配。如果数据点sp_j与无人机u_i没有匹配,但是双方都是对方比当前匹配更好的匹配选择,即:$(u_i,sp_j)\notin\mu$,$sp_j\succ\mu(u_i)$并且$u_i\succ\mu(sp_j)$,我们就称(u_i,sp_j)是一个阻碍稳定对。当一个匹配中不存在阻碍稳定对时,就说这个匹配已经达到了稳定。具体流程见算法1。

算法1 匹配博弈算法流程

输入:无人机的能量列表β,目标数据的质量列表Q,无人机的配额c_u与数据点的配额c_s,一个空列表*result*。

输出:一个稳定匹配列表。

```
初始化一个匹配表格μ,设置其中的所有值为0
flag = True
While(flag):
    flag = False
    for all β1∈β:
        for all qj∈Q:
            如果(βi,qj)构成一个阻碍稳定对:
                如果βi的匹配数量已经达到cu:
                    在μ(βi)中寻找一个最差的匹配下标j',并设置μi,j'=0
                如果qj的匹配数量已经达到cs
                    在μ(qj)中寻找一个最差的匹配下标i',并设置μi',j=0
                设置μi,j=1
                flag = True
遍历匹配表格μ:
    如果μi,j=1:
        result=result∪[ui,spj]
返回列表result
```

无人机和数据点的偏好列表是由匹配博弈机制生成建立的,即通过匹配博弈机制建立无人机与待采集目标之间的稳定的匹配关系。本文设定的偏好列表含义:建立起无人机与待采集目

标重要程度的匹配关系(偏好列表),即能量最充足的无人机去采集质量最高的数据,能量次充足的无人机去采集质量次高的数据,以此类推,以此来实现数据采集的优先采集和最大采集。

本论文中无人机能量初值与待采集目标质量初值均随机产生,稳定的匹配关系的建立是依据 Gale-shapely 匹配博弈算法思想完成。Gale-shapely 匹配博弈理论已表明:无论初值如何取值,最终均会建立起稳定的匹配关系;只要初值不变,无论初值以何种排列组合出现,则匹配关系不变。在本论文中,我们建立的是无人机与待采集目标之间稳定的匹配关系,具体而言,实验中无人机能量与待采集目标质量初值不变,无论其以何种排列组合出现,则建立起的匹配关系将不会变化,对最终的实验结果不产生影响;如实验中无人机能量与待采集目标质量在每次实验中均不相同,则显然匹配关系需重新计算,则会对实验结果产生影响。通过多组不同初值的实验验证,c-MGCM 算法的实验结果均优于其他对比算法,由于篇幅有限,我们仅列出一种情况。

3.2 基于MAAC框架的数据采集策略

通过 3.1 节无人机—目标数据匹配博弈方法,获得了无人机与目标数据之间的最优稳定匹配,指挥者依据掌握的无人机与目标之间匹配关系、位置关系、距离关系等环境信息,进而引导无人机具体动作。本论文将 MAAC 算法框架中多头注意力模块引入到指挥者与无人机的评论网络中,多头注意力模块可使得指挥者与无人机在接收其他智能体的观测信息与动作信息时,有选择性地对信息进行处理,重点关注可获取更大回报的信息。具体实现如下。

3.2.1 多注意力机制引入与实现

构建如图 1 网络模型,智能体 i 的评论网络 $Q_i^{\phi}(o,a)$ 由两部分组成:1)智能体 i 的局部状态 o_i 与动作值 a_i 的编码值 e_i;2)其余智能体的贡献值 X_i;

$$Q_i^{\phi}(o,a)=f_i(g_i(o_i,a_i),x_i), \tag{3}$$

$$x_i=\sum_{j\neq i}\alpha_j v_j=\sum_{j\neq i}\alpha_j h(V g_j(o_j,a_j))。 \tag{4}$$

v_j 由智能体 j 的编码 $e_j=g_j(o_j,a_j)$ 通过线性共享矩阵 $\boldsymbol{V}$ 变换,再由一个非线性变换函数 h 处理得到。α_j 为权重值,通过将 e_j 与 $e_i=g_i(o_i,a_i)$ 的相似值进行 softmax 处理得到:

$$\alpha_j \propto \exp(e_j{}^{\mathrm{T}}\boldsymbol{W}_k{}^{\mathrm{T}}\boldsymbol{W}_q e_i), \tag{5}$$

其中,$\boldsymbol{W}_q$ 将 e_i 转化为 *query*,$\boldsymbol{W}_k$ 将 e_j 转化为 *key*,根据这两个矩阵的维数进行匹配,以防止梯度消失。本论文中,每一个注意力头使用一套独立的参数($\boldsymbol{W}_k$, $\boldsymbol{W}_q$, $\boldsymbol{V}$),并且每一个注意力头都可以从不同角度关注指挥者、无人机的信息。

3.2.2 模型的执行与训练

如图 2 所示,每一个智能体(指挥者、无人机)i 都拥有一个策略网络(Policy network)$\pi^{\psi}(\cdot)$,一个评论网络(Critic network)$Q^{\phi}(\cdot)$,以及各自所对应的目标策略网络(Target policy network)$\pi^{\psi'}(\cdot)$、目标评论网络(Target critic network)$Q^{\phi'}(\cdot)$。当对指挥者与无人机“主-从”模型进行训练时,采用中心化训练的方式,设指挥者智能体的数量为 P,无人机智能体的数量为 U。从经验缓冲池 D 中随机取出 B 组经验用

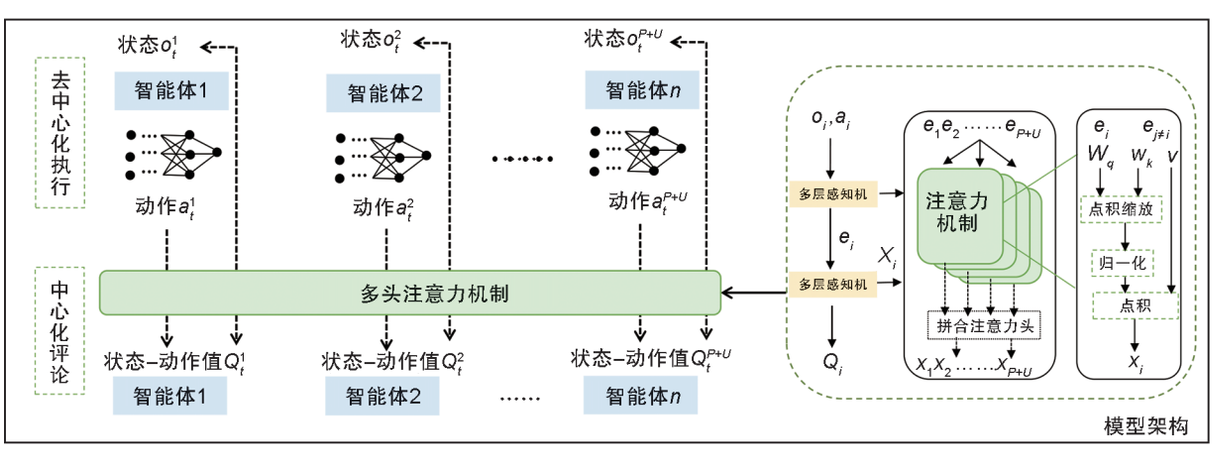

图 1 智能体的网络模型架构

Fig. 1 The network model architecture of agents

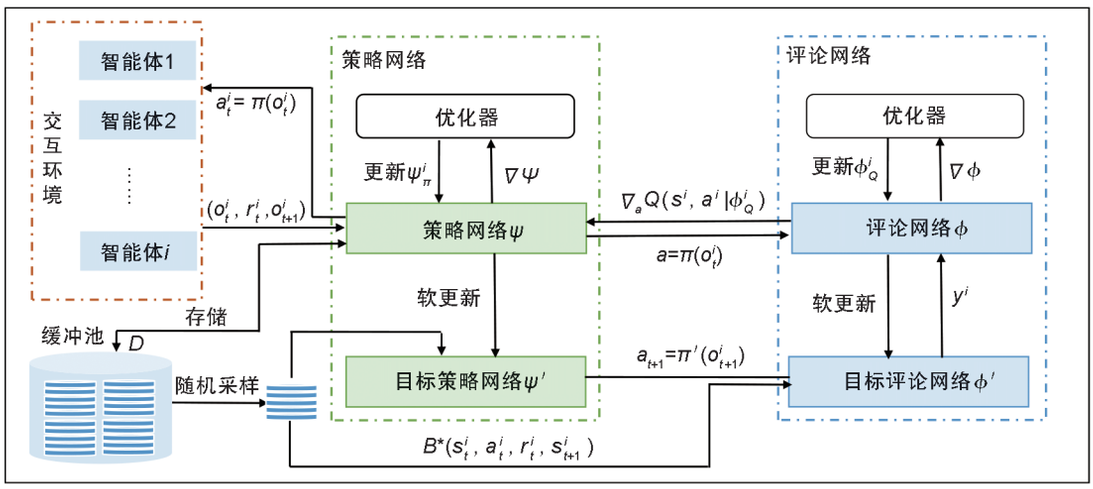

图2 智能体的模型训练结构图

Fig. 2 The structural diagram of the model training of the agent

于更新(见算法2),而目标策略网络根据经验状态o_{t+1}给出目标行为值a_{t+1}。由于参数共享,所有指挥者与无人机的评论网络一起更新,因此采用最小化联合回归损失函数进行更新[19](见算法3),公式如下:

$$L_Q(\phi)=\sum_{i=1}^{N}E_{(o,a,r,o')\sim D}[(Q_i^\phi(o,a)-y_i)^2],\tag{6}$$

$$y_i=r_i+\gamma E_{a'\sim\pi(o')}[Q_i^{\phi'}(o',a')-\kappa\log(\pi_{\psi_i'}(a'_i|o'_i))],\tag{7}$$

式(7)中,ϕ'和ψ'分别为目标评论网络和目标策略网络的参数。κ为一个常参数,决定熵最大化和奖励最大化之间的平衡。指挥者与无人机的策略以如下梯度进行上升更新[19]:

$$\nabla_{\psi_i}J(\pi_\psi)=E_{a\sim\pi_\psi}[\nabla_{\psi_i}\log(\pi_{\psi_i}(a_i|o_i))(\kappa\log(\pi_{\psi_i}(a_i|o_i))-(Q_i^\phi(o,a)-b(o,a_{\setminus i})))]。\tag{8}$$

在指挥者与无人机的策略网络更新中,本论文MAAC框架中的反事实基线来解决多智能体中的“信用”分配问题。式(8)中$Q_i^\phi(o,a)-b(o,a_{\setminus i})$表示智能体$i$特定的动作值与所有智能体的平均动作值的差值。一般情况下,$b(o,a\setminus i)=E_{a_i\sim\pi_i(o_i)}[Q_i^\phi(o,(a_i,a_{\setminus i}))]$。引入该反事实基线,可以得出单个智能体增加的奖励是否归因于其他智能体的行为,进而可以促进指挥者与无人机学习到更优秀的策略。

在本论文中,c-MGCM算法及对比算法均采用式(2)的最大化作为优化求解目标,即算法策略的优化过程就是依据式(2)追求最大化累计回报(奖励)的过程。

算法2 c-MGCM算法流程

1:初始化数据采集环境并行数量E,策略网络学习率lr_P,评论网络学习率lr_Q,折扣因子γ,缓冲池D,批量采样数目B,更新次数n_update
2:$T_{update}=0$
3:*for* $i=1,\cdots,n\_episodes$ *do*
初始化每个并行环境e,初始化全局状态s_0
4: *for* $t=1,\cdots,timesteps$ *do*
5: 在每一个并行环境e中,每一个指挥者、无人机i输出行为动作a_i^t
6: 所有并行环境共同执行指挥者、无人机所输出的动作$a^t=\{a_1^t\cdots a_p^t,a_1^t\cdots a_U^t\}$
7: 获得奖励值$r^t=\{r_1^t\cdots r_p^t,r_1^t\cdots r_U^t\}$,下一时刻状态$s^{t+1}=\{o_1^{t+1}\cdots o_P^{t+1},o_1^{t+1}\cdots o_U^{t+1}\}$
8: 对于所有并行环境,将经验$\{s^t,a^t,r^t,s^{t+1}\}$存入缓冲池D中
9: $T_{update}=T_{update}+E$
10: *if* $len(D)>B$ *and* $T_{update}\%100<E$
11: *for* $j=1,\cdots,n\_update$ *do*
12: 从缓冲池D中随机采取数目为B的经验样本
13: 更新指挥者与无人机的评论网络与策略网络参数
14: 更新目标网络参数
15: $\phi'=\tau\phi+(1-\tau)\phi'$
16: $\psi'=\tau\psi+(1-\tau)\psi'$
17: end for
18: end for
19:end for

算法设计中,以最大化累计回报为目标驱动、引导无人机进行自主学习。从强化学习的基本原理来说,累计回报越大,就可说明算法的性能越好,即c-MGCM算法的匹配博弈机制

越有效、数据采集质量越高。

算法3 更新评论网络和策略网络参数

1:从缓冲池D中采样B组经验$(o^{B}_{1\ldots P+U}, a^{B}_{1\ldots P+U}, r^{B}_{1\ldots P+U}, o'^{B}_{1\ldots P+U})$
2:对于每个智能体(指挥者、无人机)i计算$Q_i^{\phi}(o^{B}_{1\ldots P+U}, a^{B}_{1\ldots P+U})$
3:利用目标策略网络计算下一时刻行为动作值$a'^{B}_i \sim \pi_i^{\phi'}(o'^{B}_i)$
4:利用目标评论网络对每个智能体(指挥者、无人机)i计算$Q_i^{\phi'}(o'^{B}_{1\ldots P+U}, a'^{B}_{1\ldots P+U})$
5:利用公式(6)(7)更新评论网络参数
6:利用策略网络计算行为值$a^{B}_{1\ldots N} \sim \pi_i^{\phi}(o^{B}_{1\ldots N})$
7:利用评论网络对每个智能体(指挥者、无人机)i计算$Q_i^{\phi}(o^{B}_{1\ldots P+U}, a^{B}_{1\ldots P+U})$
8:利用公式(8)更新策略网络参数

4 实验与结果分析

4.1 实验场景

本文以“主-从”协作模式下基于无人机的数据采集任务为具体应用场景,其中指挥员负责判断环境、指示无人机实施行动,无人机通过指挥者传递的信息到达目标数据的位置。单次数据采集任务中指挥员位置固定,而无人机可在目标区域自由活动。实验中设定指挥者为$P \triangleq \{p_1, p_2, p_3, p_4\}$,无人机为$\mathcal{U} \triangleq \{u_1, u_2, u_3, u_4\} \beta(u_1) \neq \beta(u_2) \neq \beta(u_3) \neq \beta(u_4))$,目标数据为

$$\mathcal{M} \triangleq \{m_1, m_2, m_3, m_4\},$$
$$(h(m_1) \neq h(m_2) \neq h(m_3) \neq h(m_4)),$$

智能体的数量为$P+U$。本论文假设智能体间通信信道完全满足智能体通信的需求,无网络限速,丢包等不利因素的影响。

4.2 评价指标与对比算法

本文整体实验结构如下:4.4.1节实验验证c-MGCM算法在奖励值方面的性能、4.4.2节实验验证c-MGCM算法在距离值方面的性能、4.4.3节实验验证c-MGCM算法在数据采集质量方面的性能、4.4.4节实验验证c-MGCM算法在能量消耗方面的性能。主要考虑如下:奖励值、距离值的大小作为算法在理论层面的评价指标;数据采集质量和能量消耗值作为算法在实施层面的评价指标;4.4.1节和4.4.2节实验中,我们将匹配博弈机制均引入到所有的对比算法之中,来说明c-MGCM算法引入通信机制的有效性;4.4.3节和4.4.4节实验中,对比算法均采用标准版本(即不引入匹配博弈机制),来说明c-MGCM算法引入匹配博弈机制的有效性。因此,本文选择奖励值、距离值、数据质量、能源消耗作为评价c-MGCM算法优劣程度的指标。

本文选择DDPG、MADDPG、MAAC(不含网络参数更新)作为对比方法,这是因为DDPG是解决连续动作控制问题的经典强化学习算法;而MADDPG是多智能强化学习算法中利用局部信息就能给出最优动作的代表性算法;此外,我们将MAAC基础算法作为对比对象,为展示本文算法在网络更新方面的性能,我们将MAAC基础算法的网络参数不做更新处理,本文将该方法标记为MAAC-NU。

4.3 参数设置与参数验证

本论文设置学习率{0.01, 0.05, 0.10, 0.50}、折扣因子{0.05, 0.10, 0.5, 0.90}、策略网络学习率{0.01, 0.1, 0.50, 0.90}、评论网络学习率{0.01, 0.05, 0.50, 0.90},组成若干组合进行参数验证,以30 000回合的奖励均值进行分析(表1—表2)。实验结果显示,在学习率和折扣因子为{0.01, 0.09}时,策略网络和评论网络学习率为{0.01, 0.01}时,算法性能较好,因此本论文验证实验采用上述参数组合。参数设置实验结果如表3。

4.4 实验结果

本论文设定数据采集过程持续35个时间步,进行30 000次回合训练。本论文中无人机能量初值、待采集目标数据质量初值均随机产生,利用Gale-shapely匹配博弈理论建立匹配关系。实验中对比方法为c-MGCM、DDPG、MADDPG、MAAC-NU,评价指标为奖励值、距离值、数据采集质量和能量消耗值。实验中无人机每执行一个时间步消耗0.01单元能量,无人机的能量为{9, 3, 1, 5},目标数据的质量为{6, 2, 1, 7}。本文训练30 000回合下的奖励值、距离值、数据质量、能源消耗值,且每1 500回合取平均值,得到了图3—图6中的数据。

4.4.1 奖励值对比

本实验设置4个指挥者、4个无人机、4个目标,构成4个匹配对,图3是部分实验结果。

通过图3,可以发现指挥者在c-MGCM、MADDPG、DDPG三种算法引导下奖励值逐渐增加,但是在c-MGCM的引导下奖励值远远大

表 1 不同学习率和折扣因子对奖励回报的影响

Table 1 The impact of different learning rates and discount factors on reward returns

学习率	对象	折扣因子			
		0.05	0.1	0.5	0.99
0.01	指挥者-1	−22.90	−9.34	−2.46	213.50
	无人机-4	−24.33	−7.08	−3.63	211.34
0.05	指挥者-1	−21.26	−14.33	−1.55	212.70
	无人机-4	−23.23	−21.23	36.96	214.29
0.10	指挥者-1	−18.27	−20.61	0.13	81.92
	无人机-4	−7.09	−8.11	9.50	62.61
0.50	指挥者-1	−15.13	−7.72	11.12	−22.89
	无人机-4	−2.95	−7.81	33.90	−29.08

表 2 不同网络参数对奖励回报的影响

Table 2 The impact of different network parameters on reward returns

策略网络	对象	评论网络			
		0.01	0.05	0.5	0.9
0.01	指挥者-1	205.03	181.06	−86.51	−25.84
	无人机-4	202.70	178.56	−76.95	−39.92
0.10	指挥者-1	204.52	167.66	−84.43	−77.93
	无人机-4	198.06	114.55	−83.55	−55.33
0.50	指挥者-1	40.17	−19.41	−57.04	−56.10
	无人机-4	14.86	−16.17	−56.20	−45.67
0.90	指挥者-1	−27.50	−28.62	−47.20	−43.72
	无人机-4	−26.30	−26.19	−46.23	−44.35

表 3 参数设置

Table 3 Settings of parameters

参数	含义	大小
γ	衰减因子	0.99
lr_P	策略网络学习率	0.01
lr_Q	评论网络学习率	0.01
τ	更新率	0.01
memory-size	经验池大小	1 000 000
batch-size	采样数量	1024
n_episodes	训练回合数	30 000
episode_length	回合步数	35

于其他两种算法。如图 3(a)所示,在 30 000 回合数时,指挥者 1 在 c-MGCM 引导下获得的奖励均值为 225.07,而在 MADDPGA 下为仅为 97.00,二者相差 128.08;图 3(b)中指挥者 2 在 c-MGCM 引导下获得的奖励均值为 225.33,而 MADDPG 为 107.13,相差 118.20;如图 3(c)中,在 30 000 回合数时,无人机 3 在 c-MGCM 引导下获得的奖励均值为 229.03,而 DDPG 为仅为 93.24,二者相差 135.79。

从图 3 中可以发现智能体在 c-MGCM 算法的引导下不仅获得较高的奖励值均值,而且在 6 000 回合附近就基本达到稳定。这是因为 c-MGCM 不仅采用的是中心化训练、去中心化执行的架构,而且使用了多头注意力机制,使得每个智能体可以有选择性地处理来自其他智能体的状态与动作信息,因此表现优异。而 MADDPG 虽然有中心化训练、去中心化的架构,但是每个智能体更新的时候是不加处理地利用了其他所有智能体的状态与动作信息,因此表现效果差于 c-MGCM。而 DDPG 算法下的智能体拥有自己的网络,且更新网络参数时仅使用自己的状态动作信息,无法得知其他智能体的信息,因此效果较差。

从图 3 中可以看到 MAAC-NU 算法所获得奖励均值接近 0,这是因为此方法虽有 MAAC 模型,但是设定不更新网络参数,因此表现远远差于其他三种算法。

4.4.2 距离值对比

本部分讨论无人机在数据采集训练过程中与匹配数据之间的距离值。

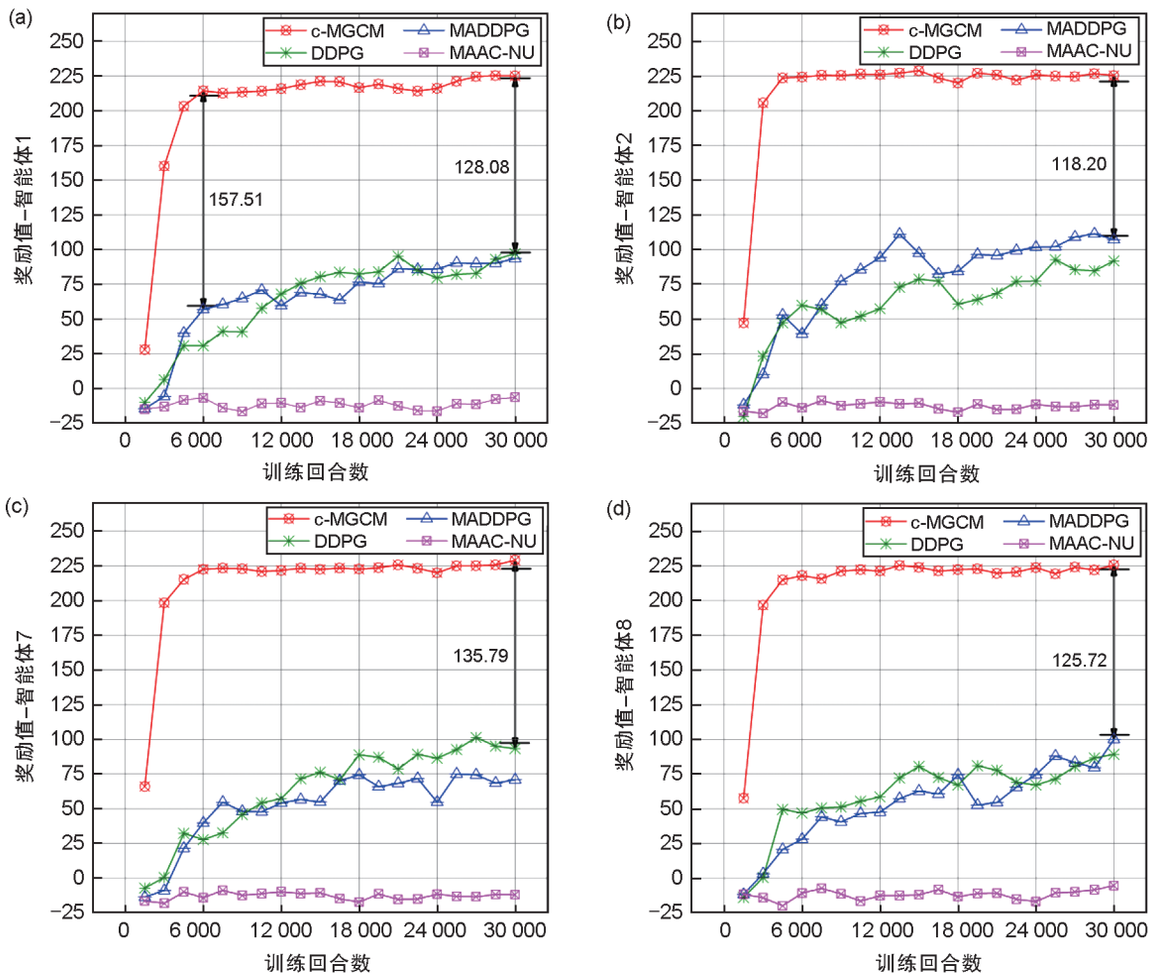

图3 不同智能体的奖励值对比

(a) 指挥者1获得的奖励值对比;(b) 指挥者2获得的奖励值对比;(c) 无人机3获得的奖励值对比;(d) 无人机4获得的奖励值对比

Fig. 3 The comparison of reward obtained by agents

(a) The reward obtained by cammander1; (b) The reward obtained by cammander2; (c) The reward obtained by UAV3; (d) The reward obtained by UAV4

从图4可以看出无人机在c-MGCM、MADDPG、DDPG三种方法的引导下与匹配数据的距离值是不断缩短的,但是在c-MGCM方法的引导距离值是最短的。如图4(a),在30 000回合数时,无人机1在c-MGCM引导下的距离均值为0.03,而在MADDPG的引导下距离均值为0.35,二者相差0.32;在图4(c)中,无人机3在c-MGCM引导下的距离均值为0.015,而在DDPG下为0.39,二者相差0.37;在图4(d)中,在训练回合数为30 000时,无人机4在c-MGCM引导下的距离均值为0.02,而在MADDPG下为0.31,二者相差0.29。

从图4可以看到无人机在c-MGCM下的引导下与匹配数据之间的距离是最短的,且在6 000回合附近达到稳定。这是因为c-MGCM中的多头注意力机制使得指挥者与无人机之间进行信息交流,无人机能较快较准确地得知匹配数据的位置信息,因此能较快到达数据周围,并采集数据。而MADDPG、DDPG算法中没有通信机制,因此表现远远差于c-MGCM算法。

从图4可以看到无人机在MAAC-NU算法下距离值是不断波动的,由于网络参数的不更新,无人机没有学到与匹配数据之间的交流机制,所以距离值远远高于其他三种算法。

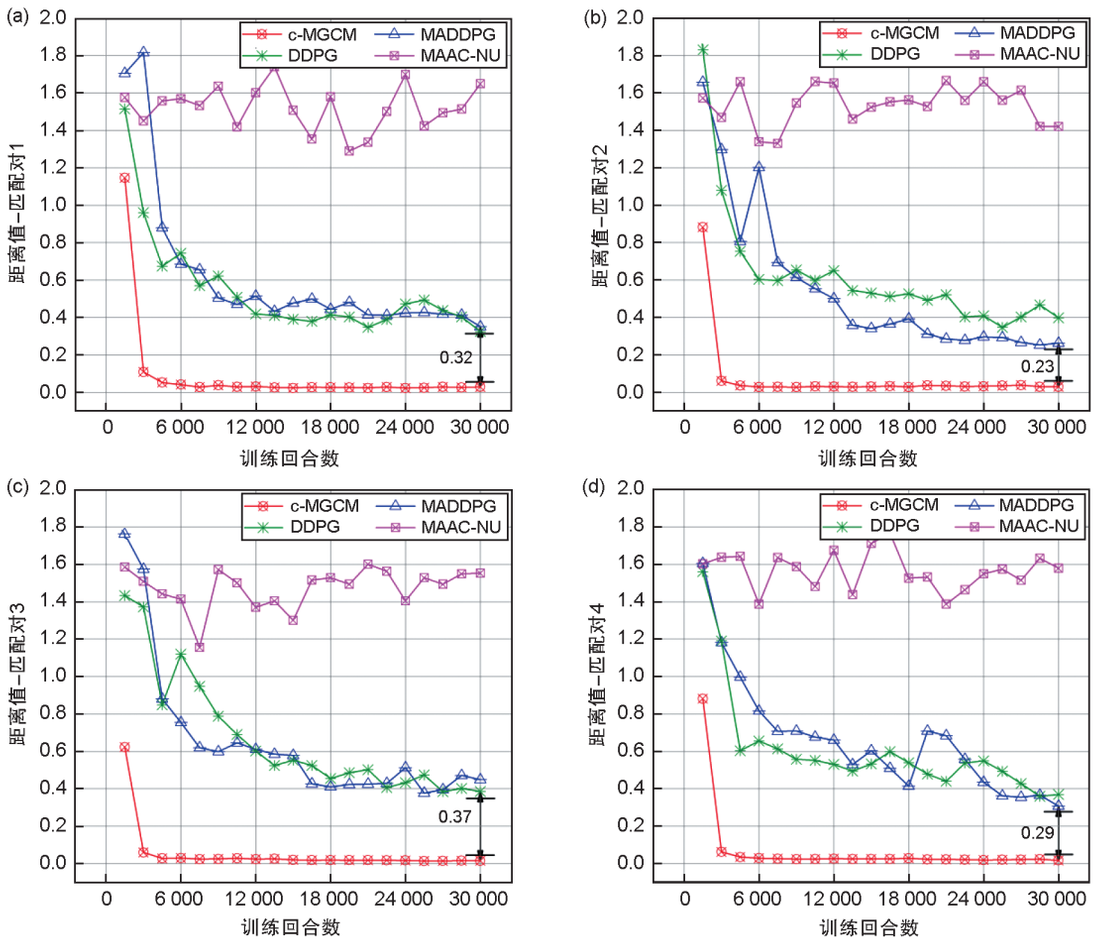

图4 不同匹配对的距离值对比

(a) 匹配对1的距离值对比;(b) 匹配对2的距离值对比;(c) 匹配对3的距离值对比;(d) 匹配对4的距离值对比

Fig. 4 The comparison of distance for different matched pairs

(a) The comparison of distance for matched pair 1; (b) The comparison of distance for matched pair 2; (c) The comparison of distance for matched pair 3; (d) The comparison of distance for matched pair 4

4.4.3 数据质量对比

图5实验显示,无人机在c-MGCM、MADDPG、DDPG三种方法下采集的数据质量均呈增加趋势;无人机在c-MGCM算法的引导下获得的数据质量均值高于其他方法,并在7 000回合左右达到稳定;如,在30 000回合时,无人机在c-MGCM下获得的数据质量均值为15.98,而MADDPG为13.65,二者相差2.33。

这是因为:(1)匹配博弈机制根据无人机能量属性与目标数据质量属性实现无人机与目标数据的最优匹配,促进重要数据的优先采集;(2)中心化训练、去中心化执行的架构使得无人机能够利用其余智能体的信息进行参数学习,提升了学习能力;(3)多头注意力机制使得无人机从多角度关注目标数据,并进行信息处

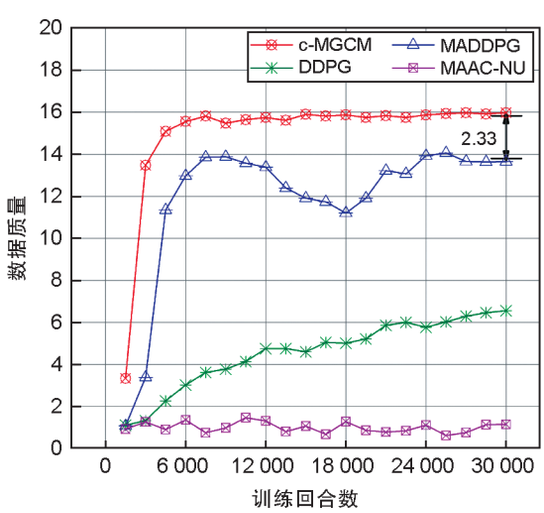

图5 不同方法的数据质量对比

Fig. 5 The comparison of data quality among different methods

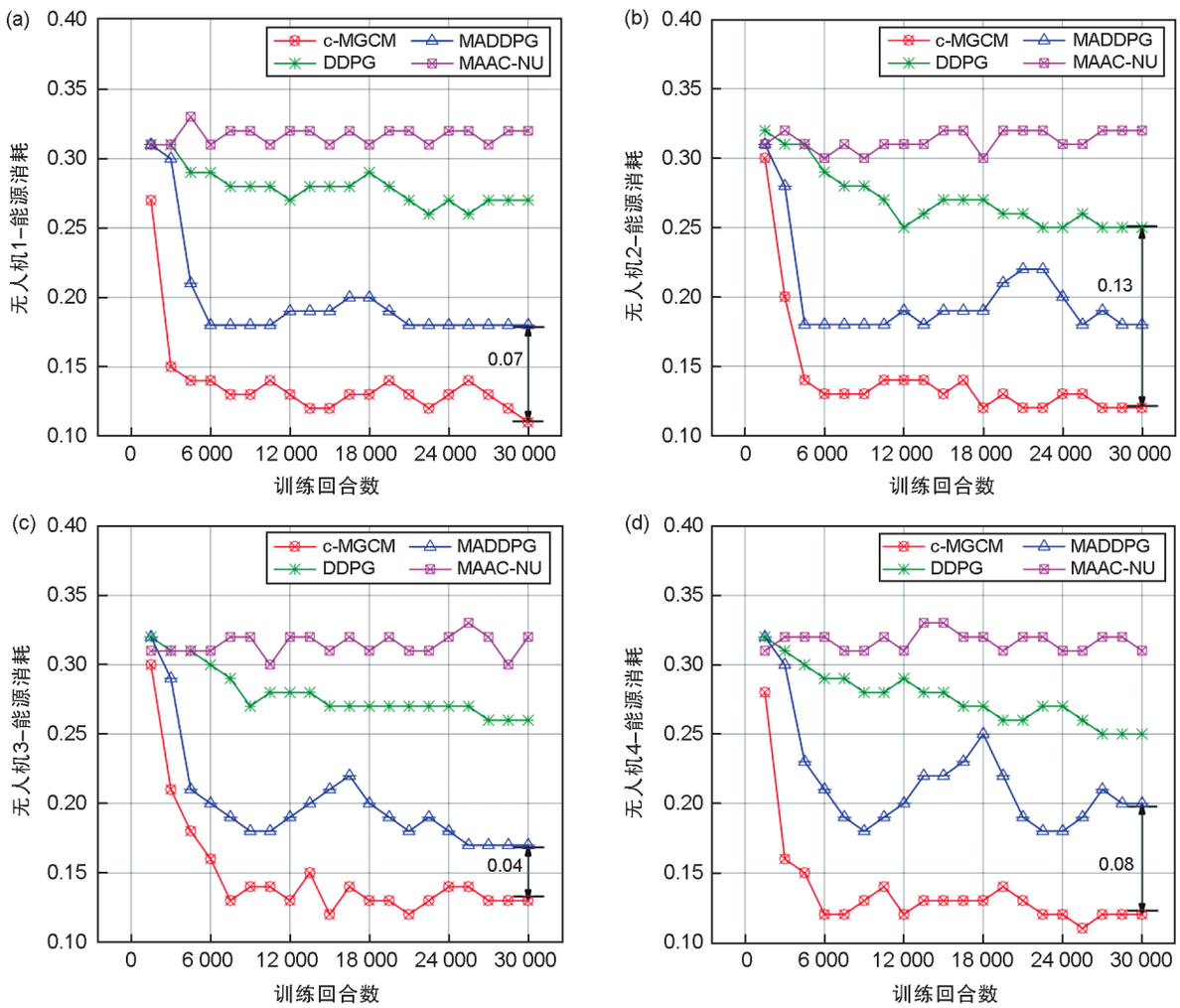

图6 不同方法下不同无人机的能源消耗对比

(a) 无人机1的能源消耗对比;(b) 无人机2的能源消耗对比;(c) 无人机3的能源消耗对比;(d) 无人机4的能源消耗对比

Fig. 6 The comparison of the energy consumption of different drones under different methods

(a) The comparison of energy consumption of drone 1; (b) The comparison of energy consumption of drone 2;
(c) The comparison of energy consumption of drone 3; (d) The comparison of energy consumption of drone 4

理,提升了数据采集的效率。而对比方法MADDPG、DDPG、MAAC-NU缺少这三方面的优势,因此表现不如c-MGCM算法。

4.4.4 能源消耗对比

显然,无人机在到达目标之前,算法执行的时间步越少则运动过程中消耗的能源越少,本论文设置无人机在到达目标之前,每执行一个时间步消耗0.01单元能量。

图6实验显示:随着训练回合的增加,无人机学会了以更为快速的方式到达目标数据,因此,相同训练回合的情况下,无人机的能量消耗呈递减趋势;无人机在c-MGCM算法的引导下消耗的能源均值低于其他方法。如,在30 000回合,无人机在c-MGCM引导下所消耗的能源均值为0.12,而MADDPG为0.2,DDPG为0.25,MAAC-NU为0.31,c-MGCM与MADDPG二者相差0.08。

其具体原因与4.3.3相同,即匹配博弈、多头注意力等机制的引入,使得无人机以更高的效率到达目标数据。

4.4.5 注意力头数目的讨论

注意力机制是通信机制强化学习方法中的重要组成部分,不同注意力头目的数量会对实验结果产生直接的影响。遗憾的是,我们没有找到如何设置算法中注意力头目数量的理论依据。为了更好地确定c-MGCM算法参数,本论

文从奖励值、距离值这两个理论层面的评价指标初步分析了如何选取注意力头目数量的问题。从实验结果来看,注意力头目数量增加时通常数据采集算法的性能越好,但算法的复杂程度不可避免会增加。综合考虑注意力头目数量对算法性能和计算复杂度的影响,本论文在实验时将注意力头目的数量设置为4。

本部分验证了注意力头的数目分别为1、3、4以及6时智能体1的奖励值与匹配数据对1的距离值。表4与表5分别为图7(a)与图7(b)所对应的数据。

表4 不同注意力头数目下的奖励值

Table 4 The rewards under different number of attention heads

训练回合数	head=1	head=3	head=4	head=6
6 000	89.32	161.41	151.38	185.95
12 000	211.94	222.50	214.01	219.02
18 000	223.98	223.13	219.33	220.27
24 000	224.63	223.16	216.29	221.91
30 000	223.11	224.55	224.03	222.89

表5 不同注意力头数目下的距离值

Table 5 The distances under different number of attention heads

训练回合数	head=1	head=3	head=4	head=6
6 000	0.508 0	0.257 0	0.252 0	0.106 0
12 000	0.053 0	0.030 6	0.025 3	0.020 6
18 000	0.034 3	0.031 3	0.026 0	0.016 8
24 000	0.034 7	0.026 9	0.021 4	0.015 6
30 000	0.027 5	0.023 4	0.020 7	0.016 6

从图7(a)可以看到无论注意力头的数目为多少,随着训练回合数的增加,智能体所获得的奖励均值是逐渐增加的。刚开始6 000回合时,注意力头数目为1的效果是最差的,注意力头数目为6的效果是最好的;当训练到30 000回合时,注意力头数目分别为1、3、4、6所获得的奖励值基本是接近的,分别为223.17、224.55、224.03、222.89,因此随着训练回合数的增加,注意力头的数目对智能体所获得的奖励均值的影响是比较小的。

从图7(b)可以看到无论注意力头的数目为多少,随着训练回合数的增加,智能体与匹配数据之间的距离是逐渐减小的。我们可以看到注意力头数目为6时表现效果是最好的,而注意力头数目为1时表现效果较差,因此注意力头的数目对无人机与匹配数据之间的距离还是有一定影响的。

5 结论

本文提出了一种基于匹配博弈和通信机制的主从协作群智感知c-MGCM算法,算法引入MAAC框架和多头注意力机制,实现了训练过程中智能体间状态和动作信息有效的甄别和利用,解决了传统方法学习效率较低的问题;引入Gale-shapely匹配博弈算法思想和设计新的奖励函数,实现了无人机能量属性与待采集目

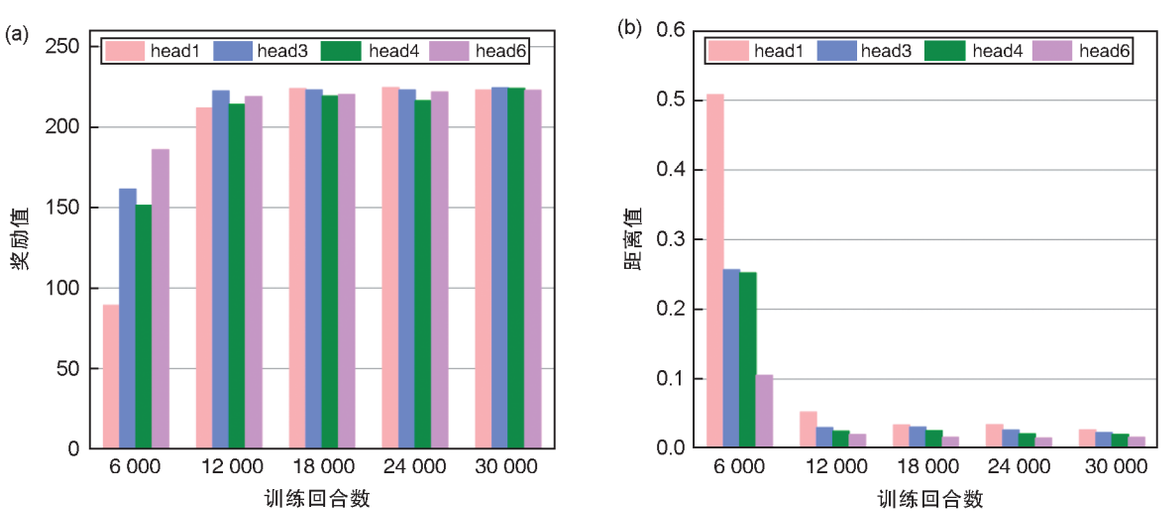

图7 不同注意力头数目对奖励值与距离值的影响

(a) 不同注意力头数目下的奖励值;(b) 不同注意力头数目下的距离值

Fig. 7 Effects of different number of attention heads on rewards and distances

(a) The reward under different number of attention heads; (b) The distance under different number of attention heads

标数据质量属性之间的最优稳定匹配,解决了如何基于数据重要程度进行采集的问题。实验表明,c-MGCM算法保证了对高价值目标的持续、特定关注,在算法奖励值、数据采集质量等多个评价指标上都优于其他经典方法。然而,本文方法采用的依然是中心化训练、去中心化执行的强化学习思想。因此,完全去中心化的群智感知方法探索,以及c-MGCM方法在更加复杂的匹配场景(如一对多最优匹配)的应用与完善将是我们下一步工作的方向。

参考文献:

[1] ZHU X Y, LUO Y Y, LIU A F, *et al*. A Deep Learning-based Mobile Crowdsensing Scheme by Predicting Vehicle Mobility[J]. *IEEE Trans Intell Transp Syst*, 2021, **22**(7): 4648-4659. DOI: 10.1109/TITS.2020.3023446.

[2] BAYOUMY K, GABER M, ELSHAFEEY A, *et al*. Smart Wearable Devices in Cardiovascular Care: Where we are and how to Move Forward[J]. *Nat Rev Cardiol*, 2021, **18**(8): 581-599. DOI: 10.1038/s41569-021-00522-7.

[3] PAN Z X, YU H, MIAO C Y, *et al*. Crowdsensing Air Quality with Camera-enabled Mobile Devices[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. New York: ACM, 2017: 4728-4733. DOI: 10.5555/3297863.3297880.

[4] SAMIR M, SHARAFEDDINE S, ASSI C M, *et al*. UAV Trajectory Planning for Data Collection from Time-constrained IoT Devices[J]. *IEEE Trans Wirel Commun*, 2020, **19**(1): 34-46. DOI: 10.1109/TWC.2019.2940447.

[5] WANG W, ZHAO N, CHEN L, *et al*. UAV-assisted Time-efficient Data Collection *via* Uplink NOMA[J]. *IEEE Trans Commun*, 2021, **69**(11): 7851-7863. DOI: 10.1109/TCOMM.2021.3106134.

[6] ZHANG R, PANG X W, LU W D, *et al*. Dual-UAV Enabled Secure Data Collection with Propulsion Limitation[J]. *IEEE Trans Wirel Commun*, 2021, **20**(11): 7445-7459. DOI: 10.1109/TWC.2021.3084140.

[7] LIU C H, DAI Z P, ZHAO Y N, *et al*. Distributed and Energy-efficient Mobile Crowdsensing with Charging Stations by Deep Reinforcement Learning[J]. *IEEE Trans Mob Comput*, 2021, **20**(1): 130-146. DOI: 10.1109/TMC.2019.2938509.

[8] LIU C H, CHEN Z, ZHAN Y. Energy-efficient Distributed Mobile Crowd Sensing: A Deep Learning Approach[J]. *IEEE Journal on Selected Areas in Communications*, 2019, **37**(6): 1262-1276. DOI: 10.1109/JSAC.2019.2904353.

[9] LIU C H, PIAO C Z, TANG J. Energy-efficient UAV Crowdsensing with Multiple Charging Stations by Deep Learning[C]//IEEE INFOCOM 2020 - IEEE Conference on Computer Communications. Toronto: IEEE, 2020: 199-208. DOI: 10.1109/INFOCOM41043.2020.9155535.

[10] GALE D, SHAPLEY L S. College Admissions and the Stability of Marriage[J]. *Am Math Mon*, 1962, **69**(1): 9-15. DOI: 10.1080/00029890.1962.11989827.

[11] HELSGAUN K. Solving the Equality Generalized Traveling Salesman Problem Using the Lin-Kernighan-Helsgaun Algorithm[J]. *Math Prog Comp*, 2015, **7**(3): 269-287. DOI: 10.1007/s12532-015-0080-8.

[12] WANG Y C, CHEN K C. Efficient Path Planning for a Mobile Sink to Reliably Gather Data from Sensors with Diverse Sensing Rates and Limited Buffers[J]. *IEEE Trans Mob Comput*, 2019, **18**(7): 1527-1540. DOI: 10.1109/TMC.2018.2863293.

[13] ZHAN C, ZENG Y. Aerial-Ground Cost Tradeoff for Multi-UAV-enabled Data Collection in Wireless Sensor Networks [J]. *IEEE Trans Commun*, 2020, **68**(3): 1937-1950. DOI: 10.1109/TCOMM.2019.2962479.

[14] WANG C, GUO S T, YANG Y Y. An Optimization Framework for Mobile Data Collection in Energy-harvesting Wireless Sensor Networks[J]. *IEEE Trans Mob Comput*, 2016, **15**(12): 2969-2986. DOI: 10.1109/TMC.2016.2533390.

[15] FOERSTER J N, ASSAEL Y M, DE FREITAS N, *et al*. Learning to Communicate with Deep Multi-agent Reinforcement Learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 2145-2153. DOI: 10.5555/3157096.3157336.

[16] IQBAL S, SHA F. Actor-attention-critic for Multi-agent Reinforcement Learning[EB/OL]. arXiv Preprint: 1810.02912, 2018. https://arxiv.org/abs/1810.02912.

[17] SUKHBAATAR S, SZLAM A, FERGUS R. Learning Multiagent Communication with Backpropagation[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 2252-2260. DOI: 10.5555/3157096.3157348.

[18] PENG P, WEN Y, YANG Y, *et al*. Multiagent Bidirectionally-coordinated Nets: Emergence of Human-level Coordination in Learning to Play StarCraft Combat Games[J]. arXiv Preprint:1703.10069, 2017.

[19] FOERSTER J N, ASSAEL Y M, DE FREITAS N, *et al*. Learning to Communicate to Solve Riddles with Deep Distributed Recurrent Q-networks[J]. arXiv preprint: 1602.02672, 2016.