计算机科学
COMPUTER SCIENCE

基于深度确定性策略梯度与注意力Critic的多智能体协同清障算法

王宪伟, 冯翔, 虞慧群

引用本文

王宪伟, 冯翔, 虞慧群. 基于深度确定性策略梯度与注意力Critic的多智能体协同清障算法[J]. 计算机科学, 2024, 51(7): 319-326.

WANG Xianwei, FENG Xiang, YU Huiqun. Multi-agent Cooperative Algorithm for Obstacle Clearance Based on Deep Deterministic PolicyGradient and Attention Critic [J]. Computer Science, 2024, 51(7): 319-326.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600129

基于深度确定性策略梯度与注意力 Critic 的多智能体协同清障算法

王宪伟[1]　冯　翔[1,2]　虞慧群[1,2]
1 华东理工大学计算机科学与工程系　上海 200237
2 上海智慧能源工程技术研究中心　上海 200237
(y30211041@mail.ecust.edu.cn)

摘　要　动态障碍物一直是阻碍智能体自主导航发展的关键因素,而躲避障碍物和清理障碍物是两种解决动态障碍物问题的有效方法。近年来,多智能体躲避动态障碍物(避障)问题受到了广大学者的关注,优秀的多智能体避障算法纷纷涌现。然而,多智能体清理动态障碍物(清障)问题却无人问津,相对应的多智能体清障算法更是屈指可数。为解决多智能体清障问题,文中提出了一种基于深度确定性策略梯度与注意力 Critic 的多智能体协同清障算法(Multi-Agent Cooperative Algorithm for Obstacle Clearance Based on Deep Deterministic Policy Gradient and Attention Critic,MACOC)。首先,创建了首个多智能体协同清障的环境模型,定义了多智能体及动态障碍物的运动学模型,并根据智能体和动态障碍物数量的不同,构建了 4 种仿真实验环境;其次,将多智能体协同清障过程定义为马尔可夫决策过程(Markov Decision Process,MDP),构建了多智能体 t 的状态空间、动作空间和奖励函数;最后,提出一种基于深度确定性策略梯度与注意力 Critic 的多智能体协同清障算法,并在多智能体协同清障仿真环境中与经典的多智能体强化学习算法进行对比。实验证明,相比对比算法,所提出的 MACOC 算法清障的成功率更高、速度更快,对复杂环境的适应性更好。

关键词:强化学习算法;马尔可夫决策过程;多智能体协同控制;动态障碍物清除;注意力机制
中图分类号　TP183

Multi-agent Cooperative Algorithm for Obstacle Clearance Based on Deep Deterministic Policy Gradient and Attention Critic

WANG Xianwei[1],FENG Xiang[1,2] and YU Huiqun[1,2]
1 Department of Computer Science and Engineering,East China University of Science and Technology,Shanghai,200237,China
2 Shanghai Engineering Research Center of Smart Energy,Shanghai,200237,China

Abstract　Dynamic obstacles have always been a key factor hindering the development of autonomous navigation for agents. Obstacle avoidance and obstacle clearance are two effective methods to address the issue. In recent years,multi-agent obstacle avoidance(collision avoidance) has been an active research area,and there are numerous excellent multi-agent obstacle avoidance algorithms. However,the problem of multi-agent obstacle clearance remains relatively unknown,and the corresponding algorithms for multi-agent obstacle clearance are scarce. To address the issue of multi-agent obstacle clearance,a multi-agent cooperative algorithm for obstacle clearance based on deep deterministic policy gradient and attention Critic(MACOC) is proposed. Firstly,the first multi-agent cooperative environment model for obstacle clearance is created,and the kinematic models of the agents and dynamic obstacles are defined. Four simulation environments are constructed based on different numbers of agents and dynamic obstacles. Secondly,the process of obstacle clearance cooperatively by multi-agent is defined as a Markov decision process(MDP) model. The state space,action space,and reward function for multi-agent are constructed. Finally,a multi-agent cooperative algorithm for obstacle clearance based on deep deterministic policy gradient and attention critic is proposed,and it is compared with classical multi-agent algorithms in the simulated environments for obstacle clearance. Experimental results show that,the pro-

到稿日期:2023-06-16　返修日期:2023-11-16
基金项目:国家自然科学基金面上项目(62276097);国家自然科学基金重点项目(62136003);国家重点研发计划(2020YFB1711700);上海市经信委"信息化发展专项资金"(XX-XXFZ-02-20-2463);上海市科技创新行动计划(21002411000)
This work was supported by the National Natural Science Foundation of China(62276097),Key Program of National Natural Science Foundation of China(62136003),National Key Research and Development Program of China(2020YFB1711700),Special Fund for Information Development of Shanghai Economic and Information Commission(XX-XXFZ-02-20-2463) and Scientific Research Program of Shanghai Science and Technology Commission(21002411000).
通信作者:冯翔(xfeng@ecust.edu.cn)

posed MACOC algorithm has a higher success rate in obstacle clearance, faster speed, and better adaptability to complex environments compared to the compared algorithms.

Keywords Reinforcement learning algorithm, Markov decision process, Multi-agent cooperative control, Dynamic obstacle clearance, Attention mechanism

1 引言

随着人工智能技术的不断发展，自动驾驶、物流配送机器人等无人自主导航应用纷纷涌现。然而，动态障碍物一直是阻碍智能体自主导航发展的关键因素，而躲避障碍物和清理障碍物是两种解决动态障碍物问题的有效方法。近年来，优秀的多智能体避障算法层出不穷[1-4]。然而，多智能体清理动态障碍物(清障)问题却无人问津，相对应的多智能体清障算法更是屈指可数。

多智能体清障问题指多智能体探测并且清除环境或者道路中的动态障碍物以确保路径的安全性和平滑性。然而目前存在两个问题，严重阻碍了多智能体清障问题的解决。

(1)多智能体协同清障仿真环境缺失。优秀的多智能体协同清障仿真环境可为清障算法提供实验的平台，是解决多智能体清障问题的先决条件。为解决这一问题，本文创建了首个多智能体协同清障的环境模型，定义了多智能体及动态障碍物的运动学模型，并根据动态障碍物和智能体的数量不同，构造了4种仿真实验环境。

(2)目前对于智能体清障算法的研究大多聚焦于单智能体和静态障碍物领域，例如Nayyar等[5]提出了一种单智能体清除静态障碍物的方法并将其应用于协助紧急疏散。然而，在现实生活中，动态障碍物的存在更加普遍，并且单智能体清障的效率较低。而对于效率更高、应用更广的多智能体清除动态障碍物算法的研究却屈指可数。为了拓展清障算法的应用领域并且提高清障效率，本文提出了一种基于深度确定性策略梯度与注意力Critic的多智能体协同清障算法。

多智能体系统指一组共享环境的自主和交互式实体[6]。与单个智能体相比，多智能体在未知环境、仅知部分信息、计算和分布式控制中执行困难任务的性能更佳[7]。多智能体协同控制通过控制智能体协同完成复杂任务，可提高完成任务的效率并且获得更高的奖励，并可应用于众多领域，如无人机编队控制[8]、城市交通控制[9]和资源分配[10]等领域。然而，目前并没有将多智能体协同控制应用清障问题的先例。因此，本文提出了首个基于多智能体协同控制的清障问题解决方案，该方案有助于促进智能体自主导航的发展与应用。

综上所述，本文的主要贡献如下：

(1)创建了首个多智能体协同清障的环境模型，定义了多智能体及动态障碍物的运动学模型，并根据智能体和动态障碍物数量的不同，构造了若干仿真实验环境。

(2)将多智能体协同清障过程定义为马尔可夫决策过程，构建了多智能体的状态空间、动作空间和奖励函数。

(3)提出了一种基于深度确定性策略梯度与注意力Critic的多智能体协同清障(MACOC)算法。

2 相关工作

深度Q网络(Deep Q-Network, DQN)算法[11]和AlphaGo[12]的横空出世，吸引了众多学者纷纷踏上深度强化学习的研究之旅。在单智能体强化学习算法中，OpenAI[13]提出的近似策略优化(Proximal Policy Optimization, PPO)算法具有样本效率更高、稳定性和收敛性更好、可以处理连续动作空间的问题等优点，同时已被OpenAI列为单智能体强化学习算法的默认推荐算法。为了提高仿真环境中动态障碍物的智能性和真实性，本文将使用PPO算法训练得到的智能躲避策略应用于动态障碍物。

在高维连续空间的多智能体系统领域，深度强化学习同样得到了广泛应用。例如，Li等[14]提出了一种基于渐进式互信息协作的多智能体强化学习(Progressive Mutual Information Collaboration, PMIC)算法；Foerster等[15]提出将团队回报拆分为独立回报的对抗性多智能体计数事实学习(Counterfactual Multi-Agent Learning, COMA)算法；Yu[16]等提出了可以高效训练和学习的多智能体近端策略优化(Multi-Agent Proximal Policy Optimization, MAPPO)算法。与上述算法相比，Lowed等[17]提出的多智能体深度确定性策略梯度学习(Multi-Agent Deep Deterministic Policy Gradient, MADDPG)算法具有利用局部信息输出最优动作，以及无特殊的通信需求等优点。然而，由于不同的智能体在协作清障环境中具有不同程度的关联，Critic网络不能很好地利用信息的关联性，导致MADDPG算法控制的多智能体协同清障效率较低。为了提升Critic网络的性能，本文提出一种基于深度确定性策略梯度与注意力Critic的多智能体协同清障算法，在Critic网络中引入了一种注意力机制，使其能够分辨不同智能体信息的重要程度，以帮助智能体采取可获取更高团队奖励的动作。同时本文将选取MAPPO算法和MADDPG算法作为本文的对比算法。

3 多智能体协同清障环境

本章首先创建了多智能体协同清障的环境模型，然后定义了多智能体及动态障碍物的运动学模型，最后根据动态障碍物和智能体数量的不同，构造了4种仿真实验环境。

3.1 环境模型

多智能体协同清障环境模型如图1所示，包括基础环境、运动学模型、多智能体和动态障碍物4种基础元素。运动学模型是多智能体协同清障环境中物体运动的几何模型，它描述了物体的位置、速度和加速度等几何特征；多智能体是清障过程中的主体，其运动受多智能体协同清障模型控制，通过智能体间的协同合作，共同清理环境中的动态障碍物；动态障碍

物是清障过程中的客体，为提高仿真清障环境的真实性，根据现实世界中的动态障碍物运动策略的不同，我们将运动策略为 PPO 算法训练得到的智能躲避策略的动态障碍物定义为智能障碍物，将运动策略为随机运动策略的动态障碍物定义为非智能障碍物。

在多智能体协同清障环境模型中，多智能体和动态障碍物在每回合的初始位置是随机的。首先，多智能体感知环境的初始状态集合S_t，并将其作为多智能体协同清障模型的输入，多智能体协同清障模型输出动作集合A_t。与此同时，若动态障碍物为智能障碍物，则感知环境的初始状态S_t'，输入智能躲避模型，得到输出的动作A_t'；若动态障碍物为非智能障碍物，则输出随机动作A_t''。然后环境状态将由S_e转化为新状态S_e'，并且返回多智能体的奖励集合A_t。当环境中所有的障碍物均被清除后，该回合结束，重新随机初始化后，开始新的回合。

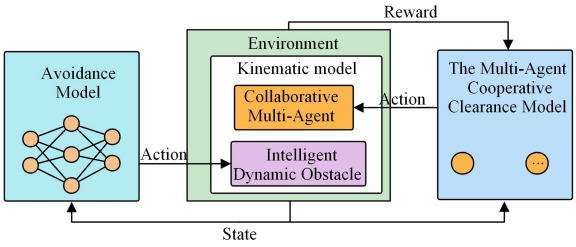


图 1　多智能体协同清障环境模型

Fig. 1　Environment model of obstacle clearance cooperatively by multi-agent

3.2　运动学模型

图 2 所示的运动学模型是多智能体协同清障环境中物体运动的几何模型，它描述了物体的位置、速度和加速度等几何特征。

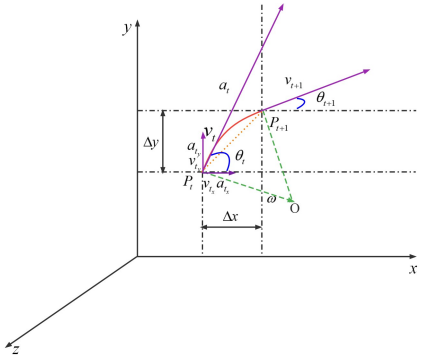


图 2　运动学模型

Fig. 2　Kinematic model

图 2 中，t 时刻，P_t 表示清障环境中物体的位置，θ_t 表示物体相对于 X 轴的运动方向，v_t 表示物体的运动速度并且基于 X-Y 轴分解为 v_{t_x} 和 v_{t_y}，a_t 表示物体的切向加速度并且分解为 a_{t_x} 和 a_{t_y}，ω_t 表示物体的法向加速度，红色曲线 $\overset{\frown}{P_tP_{t+1}}$ 表示物体 t 时刻至 $t+1$ 时刻的运动轨迹。

本文使用五元组 $(x_t, y_t, \theta_t, v_{t_x}, v_{t_y})$ 表示 t 时刻物体的运动学状态，由图 2 中的运动学和物理学关系，可得式(1)的运动学方程：

$$\begin{cases} x_{t+1} = x_t + v_{t_x} + \dfrac{1}{2}a_{t_x} \\ y_{t+1} = y_t + v_{t_y} + \dfrac{1}{2}a_{t_y} \\ \theta_{t+1} = \theta_t + \omega_t \\ v_{t+1_x} = v_{t_x} + a_{t_x} \\ v_{t+1_y} = v_{t_y} + a_{t_y} \end{cases} \tag{1}$$

若时刻 t 的运动学状态 $(x_t, y_t, \theta_t, v_{t_x}, v_{t_y})$、切向加速度 a_t 和法向加速度 ω_t 已知，则 $t+1$ 时刻的运动学状态可知。同时，由几何学知识可知，$\theta_{t+1} = \arctan(v_{t+1_x}, v_{t+1_y})$，因此本文可通过控制两个切向加速度分量 a_{t_x} 和 a_{t_y} 来控制物体的运动。

3.3　仿真清障环境

根据智能体和障碍物的数量不同，本文构建了以图 3 为例的 4 种仿真实验环境。仿真清障环境由 4 种元素构成：①代表动态障碍物、②代表多智能体、③代表已清障区和其余白色区域表示未清障区。

清障环境为 200 * 200 的正方形区域，黑色为动态障碍物，最大速度为 1，红色为智能体，最大速度为 2。图 3 中 20 * 20 的正方形红色区域为已清障区，其余白色区域为未清障区。已清障区表示已完成清理工作并且无障碍物存在的区域。在清障过程中，智能体需防止动态障碍物进入已清障区，若进入，则清障任务失败；若未进入，且清理完所有障碍物，则任务成功。该规则设定了两个目的：提高智能体的清障效率并促使其在最短时间内完成清障任务；模拟现实生活中分区域清障的实际场景。

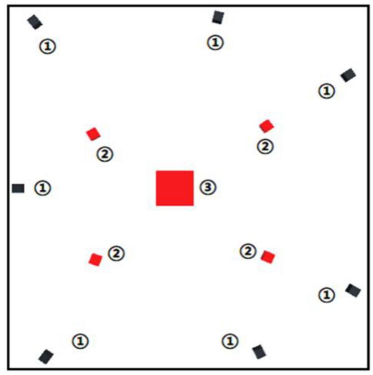


图 3　仿真清障环境(电子版为彩图)

Fig. 3　Simulate environment of obstacle clearance

根据动态障碍物和智能体的数量不同，本文构造了 4 个仿真清障环境，分别是 2 智能体 vs. 2 障碍物、3 智能体 vs. 3 障碍物、3 智能体 vs. 5 障碍物、4 智能体 vs. 7 障碍物。

4　多智能体协同清障过程

本章首先公式化定义了多智能体协同清障问题，然后将多智能体协同清障过程定义为马尔可夫决策过程(MDP)，最后构建了多智能体的状态空间、动作空间和奖励函数。

4.1　问题定义

强化学习的目标是学习得到一个最大化期望回报的策略

π_θ[18]，智能体的目标函数通过折扣回报的方式定义为式(2)：

$$J^R(\theta)=\mathop{\mathbb{E}}_{(s_t,a_t)\sim\rho_{\pi_\theta}}[\sum_t\gamma^t r(s_t,a_t)] \tag{2}$$

其中，ρ_{π_θ}是由策略π_θ决定的轨迹分布；$\gamma\in[0,1]$为折扣率，表示长期回报的权重。

在多智能体协同清障问题中，多智能体的目标是在清理所有动态障碍物的基础上，得到一个最大化团队累积回报的策略π_{θ^*}[19]。本文将多智能体协同清障问题的目标定义为式(3)：

$$\theta^*=\mathop{\text{argmax}}_{\theta_i}\sum_i^N J^R(\theta_i) \tag{3}$$

其中，N 表示智能体的数量。

4.2 马尔可夫决策过程

本文将多智能体协同清障过程中多智能体与环境的交互过程建模为如图 4 所示的马尔可夫决策过程(MDP)，同时，重构了状态空间、动作空间和奖励函数等。

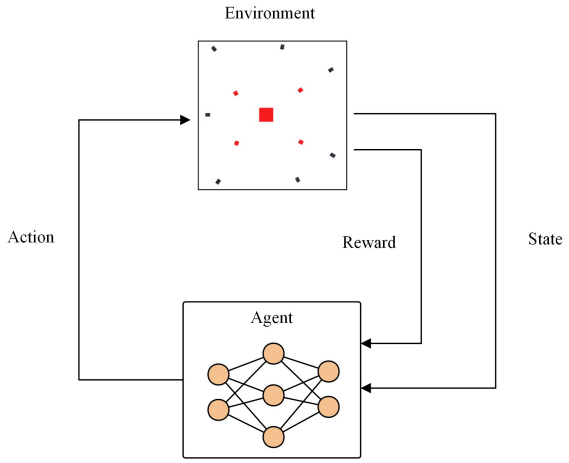


图 4 马尔可夫决策过程

Fig. 4 Markov decision process

本文使用五元组$\langle S,A,R,P,\gamma\rangle$表示该马尔可夫决策过程，每个元素的含义如下。

S：状态空间，t 时刻智能体的状态集合$S_t\in S$。

A：动作空间，t 时刻智能体的动作集合$A_t\in A$。

R：奖励函数，$S\times A\to R$。

P：状态转移函数，$P(s'|s,a)$代表智能体采取动作 a 后，由状态 s 转移到状态s'的概率。

γ：折扣率，$\gamma\in[0,1]$，代表当前奖励和未来奖励的重要性权重。

4.2.1 状态空间

状态集合中包含 4 个特征向量：(1)动态障碍物的特征向量$\boldsymbol{S}_o=[[x,y]_o,[v_x,v_y]_o]$；(2)多智能体的特征向量$\boldsymbol{S}_a=[[x,y]_a,[v_x,v_y]_a]$；(3)已清障区的特征向量$\boldsymbol{S}_d=[[x,y]_d,[v_x,v_y]_d]$；(4)未清障区的特征向量$\boldsymbol{S}_u=[[x,y]_u,[v_x,v_y]_u]$。第$i$个智能体在时刻 t 的状态集合被定义为式(4)：

$$S_i^t=[S_o,S_a,S_d,S_u]_t \tag{4}$$

4.2.2 动作集合

根据 3.2 节的运动学模型可知，通过切向加速度$[a_{t_x},a_{t_y}]$可以控制智能体的运动，因此我们将第i个智能体时刻 t 的动作集合定义为式(5)：

$$A_i^t=[a_{i_x},a_{i_y}]_t \tag{5}$$

4.2.3 奖励函数

强化学习中，智能体的学习依赖于环境的奖惩机制，稀疏奖励将导致智能体接收的环境反馈信号过少而无法得到优秀的决策模型[20]。本文的奖励函数设计采取了连续奖励与终止奖励相结合的模式。

如 3.3 节所述，多智能体协同清障过程中，多智能体的目标是在保证没有障碍物进入已清障区的前提下，以最快的速度清理未清障区内的所有障碍物。因此，本文的奖励函数由两部分组成，即智能体与障碍物间的连续距离奖励以及清障任务成功或失败的终止奖励。

第 i 个智能体在时刻 t 的连续距离奖励函数的定义如式(6)所示：

$$r_{i,1}^t=-\alpha\sqrt{((x_i^t-x_t^n)^2+(y_i^t-y_t^n)^2)} \tag{6}$$

其中，α 代表奖励因子。为了避免智能体在多个障碍物间左右摇摆，我们通过式(7)选择距离智能体最近的障碍物。

$$n=\mathop{\text{argmin}}_{j=1,2,3,\cdots,n}\sqrt{((x_{i,t}^c-x_t^j)^2+(y_{i,t}^c-y_t^j)^2)} \tag{7}$$

第 i 个智能体的终止奖励函数如式(8)所示：

$$r_{i,2}=\begin{cases}r_{\text{fail}}, & \text{if 清障任务失败}\\ r_{\text{success}}, & \text{if 清障任务成功}\end{cases} \tag{8}$$

其中，r_{fail}为清障任务失败的终止奖励，本文取$r_{\text{fail}}=-50$；r_{success}为清障成功的终止奖励，本文取$r_{\text{success}}=20$。

智能体的奖励函数由式(6)定义的连续距离奖励函数和式(8)定义的终止奖励函数组成，第 i 个智能体在时刻 t 的奖励函数定义如式(9)所示：

$$r_i^t=r_{i,1}^t+r_{i,2} \tag{9}$$

5 多智能体协同清障算法

本章提出了一种基于深度确定性策略梯度与注意力 Critic 的多智能体协同清障(MACOC)算法。MACOC 算法使用了集中式训练分布式执行的多智能体强化学习算法框架。训练时，具有全局状态信息输入的 Critic 网络指导 Actor 网络的训练；执行时，仅由具有局部状态信息的 Actor 网络输出动作。

由于输入 Critic 的状态信息是全局且高维的，Critic 网络无法分辨来自不同智能体的信息的重要程度，从而降低了多智能体的协同清障效率。因此，我们在 Critic 网络中引入注意力机制，通过提高 Critic 网络对其他智能体重要信息的关注程度，来提高多智能体协同清障效率。

图 5 给出了本文提出的注意力 Critic 网络框架图，其中！i 表示第 i 个智能体以外的其他智能体。注意力 Critic 网络的输入为动作状态对(s,a)，其中状态集合$s=(s_1,\cdots,s_N)$，动作集合$a=(s_1,\cdots,s_N)$，输出为 Actor 网络输出动作状态对(s,a)的动作价值$Q_i^c(s,a)$。$Q_i^c(s,a)$的定义如下：

$$Q_i^c(s,a)=g_i(e_i,x_i)\ \text{where}\ e_i=\varphi_i(s_i,a_i) \tag{10}$$

其中，e_i表示第 i 个智能体动作状态对的嵌入值，φ_i表示多层感知器嵌入函数，g_i表示由多层神经感知器构成的动作价值

函数，x_i 表示其他智能体对第 i 个智能体动作价值的累积贡献。x_i 的定义如下：

$$x_i = Concat_{j\in !i} x_{head_j} = \sum_{j\in !i} \alpha_j h_j = \sum_{j\in !i} \alpha_j LU(M_\varphi^V \varphi_j(s_j, a_j))$$

$$\text{where } \alpha_j \propto \exp(e_j^T \boldsymbol{M}_k^T \boldsymbol{M}_q e_i) \tag{11}$$

其中，h_j 表示第 $j(j\in !i)$ 个智能体状态动作的嵌入函数。$\boldsymbol{M}_\varphi^V$ 是一个共享线性转化矩阵，LU 代表非线性激活函数 Leaky Rule，α_j 是注意力权重，表示第 i 个智能体对第 $j(j\in i)$ 个智能体的关注程度。$\boldsymbol{M}_k^T$ 与 $\boldsymbol{M}_q$ 均为共享矩阵，$\boldsymbol{M}_k^T$ 将 e_j 转化为“查询(Query)”，$\boldsymbol{M}_q$ 将 e_i 转化为“键(Key)”，然后使用双线性映射来计算并且归一化 Query 和 Key 之间的相关性，并将两个嵌入之间的相似度值转化为 Softmax 函数的输出。匹配结果会根据这两个矩阵的维度进行缩放，以防止梯度消失。

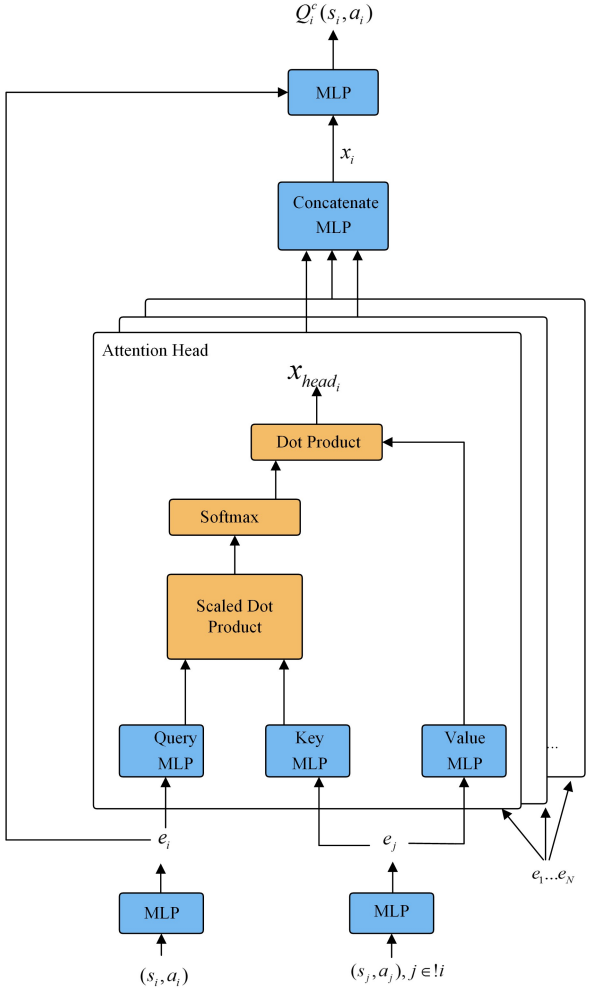


图 5 注意力 Critic 网络框架

Fig. 5 Framework of attention Critic network

本文中，计算第 $j(j\in !i)$ 个智能体对第 i 个智能体的价值贡献时，均使用独立的参数集合($\boldsymbol{M}_j^k$，$\boldsymbol{M}_j^q$，$\boldsymbol{M}_j^v$)，并且使用 x_i 表示多个注意力头的累积贡献。同时，每个注意力头的参数集合($\boldsymbol{M}_j^k$，$\boldsymbol{M}_j^q$，$\boldsymbol{M}_j^v$)在所有智能体中共享，这有助于本文方法在智能体具有不同奖励但有共同特征的环境中有效地学习。

如图 6 的 MACOC 算法框架所示，每个智能体都有单独的 Actor 网络和 Critic 网络。Actor 网络是一个策略网络，输入智能体的局部状态，输出动作。我们期望 Actor 网络输出具有最大动作价值的动作。Actor 网络目标函数的定义如式(12)所示：

$$J(\theta) = \mathbb{E}_{(s_i,a_i,s_j,a_j)\sim D}(Q_i^c(s,a|\varphi)) \tag{12}$$

$$\text{where } Q_i^c(s,a|\varphi) = Q_i^c(s_i,a_i,s_j,a_j|\varphi)|_{a_i=\pi_\theta(s_i)}(j\in !i)$$

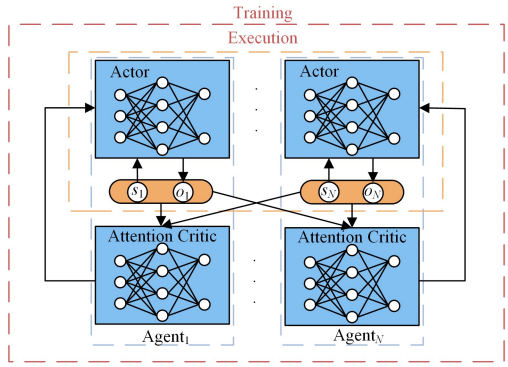


图 6 MACOC 算法框架

Fig. 6 Framework of MACOC algorithm

根据式(12)可得目标函数的策略梯度的定义：

$$\nabla J(\theta) = \mathbb{E}_{(s_i,a_i,s_j,a_j)\sim D}(\nabla \pi_\theta(s_i) \nabla Q_i^c(s,a|\varphi)) \tag{13}$$

其中，网络参数 θ 将采用梯度上升的方式进行更新。

注意力 Critic 网络是一个输出动作价值并评估 Actor 网络输出动作优劣的网络。注意力 Critic 网络的目标是最小化网络输出的预估动作价值与真实动作价值的 TD 误差。TD 误差的定义如式(14)所示：

$$L(\varphi) = \mathbb{E}_{(s_i,a_i,s_j,a_j)\sim D}[(Q_i^c(s,a|\varphi) - (r + Q_i^c(s',a'|\varphi')))^2] \tag{14}$$

其中，网络参数 φ 将采用梯度下降的方式进行更新。算法伪代码如算法 1 所示。

算法 1 MACOC 算法

输入：智能体的个数 N，初始化 Actor 网络参数 θ_i，注意力 Critic 网络参数 φ_i(i=1，2，…，N)，以及经验回放池 D

输出：θ_i 和 φ_i

1. while not stop do
2. for k=1 to max-episode-length do
3. for i=1，2，…，N do
4. 感知当前环境局部状态 s_i，根据当前策略和探索率得到动作 a_i：

$$a_i = \pi(s_i|\theta_i) + \epsilon$$

5. 多智能体执行动作集合 a，获得环境奖励 r，感知当前环境的全局状态 s，将四元组(s，a，r，s′)存放在回放池 D 中更新环境状态：s←s′
6. end for
7. end for
8. 从经验回放池 D 中随机抽取小批量样本 B
9. for i=1，2，…，N do
10. 根据式(10)计算动作价值 $Q_i^c(s,a|\varphi)$
11. 更新注意力 Critic 网络参数 φ_i：

$$\varphi_i = \varphi_i - \nabla_\varphi \frac{1}{|B|} L(\varphi)$$

12. 更新 Actor 网络参数 θ_i：

$$\theta_i = \theta_i + \nabla_\theta \frac{1}{|B|} J(\theta)$$

13. end for
14. end while

6 实验

本章将 MACOC 算法应用于本文构建的多智能体协同清障仿真环境中，并回答了下列问题。

问题 1 本文构建的多智能体协同清障仿真环境是否可以正常运行？

问题 2 本文提出的 MACOC 算法在清障仿真环境中能否有效解决多智能体协同清障问题？

问题 3 与对比算法相比，本文提出的 MACOC 算法能否获得更高的清障率以及整体回报？

问题 4 本文提出的 MACOC 算法针对智能障碍物和非智能障碍物是否均具备优秀的清障性能？

6.1 环境设置

本实验算法的超参数设置如表 1 所列。每类实验均选取 3 个随机种子，实验基于 40 核 CPU 的 Linux 服务器平台。本文选取 MAPPO 算法和 MADDPG 算法作为实验对比算法。MAPPO 算法和 MADDPG 算法均为目前最受欢迎的基于 Actor-Critic 网络架构的高效多智能体强化学习算法[21]。

表 1 超参数设置

Table 1 Hyperparameter settings

Hyperparameter	Value
Total Number of Steps	2×10^7
Max Stepsof Each Epoch	30000
Hidden Layers In Actor And Critic	(256:256)
Discount Factors	0.99
GAE Parameter	0.95
Gradient Descent Algorithm	ADAM
Actor Learning Rate	3×10^{-4}
Critic Learning Rate	1×10^{-4}
Activation Function	Tanh

6.2 评价标准

本文采取平均回合奖励、平均回合步长以及平均回合清障成功率这 3 种评价准则来评估算法清障的实验效果。

平均回合奖励(Average Episodic Reward，AER)的定义如式(15)所示：

$$ACR_k=\frac{1}{n_{ik}}\sum_{i=1}^{M}\sum_{j=1}^{n_{ik}}r_{ij} \tag{15}$$

其中，k 为迭代轮次，本文中取 60000 步长为一个迭代轮次；n_{ik} 代表智能体在一个迭代轮次中经历的回合数；M 为智能体个数；r_{ij} 代表第 i 个智能体在第 j 回合获得的累积奖励。

平均回合步长(Average Episodic Length，AEL)的定义如式(16)所示：

$$AEL_k=\frac{1}{n_{ik}}\sum_{i=1}^{M}\sum_{j=1}^{n_{ik}}L_{ij} \tag{16}$$

其中，L_{ij} 代表第 i 个智能体在第 j 回合的步长。

平均回合清障率(Average Success Rate，ASR)的定义如式(17)所示：

$$DSR_k=\frac{1}{n_{ik}}\sum_{i=1}^{M}\sum_{j=1}^{n_{ik}}S_{ij} \tag{17}$$

当一个回合中多智能体清除掉所有障碍物时，$S_{ij}=1$，否则 $S_{ij}=0$。

6.3 实验分析

本节实验分析的实验场景为智能障碍物的仿真清障环境场景，原因在于，智能障碍物的运动策略为 PPO 算法训练得到的智能躲避策略时，清除难度更高，更能体现本文提出的 MACOC 算法的清障性能。同时表 3 列出了实验场景为非智能障碍物的仿真环境场景的算法实验结果对比，并针对问题 4 进行了相关的实验结果分析。

MACOC 算法在 4 智能体 vs. 7 智能障碍物仿真环境的清障过程如图 7 所示。第①帧为环境的初始状态，红色物体为清障智能体，轨迹为红色，黑色物体为动态智能障碍物，轨迹为黑色；第②－⑤帧中，智能体根据多智能体协同清障策略模型分别向自己的清障目标前进，由黑色轨迹可知，在智能体接近智能障碍物时，智能障碍物会表现出明显的躲避行为；第⑥－⑧帧中，某一智能体清除目标智能障碍物后，多智能体协同清障模型将为其分配新的智能障碍物；第⑨帧中，多智能体清除掉所有智能障碍物，完成清障任务。上述过程表明本文构建的多智能体协同清障仿真环境可以正常运行，回答了问题 1。

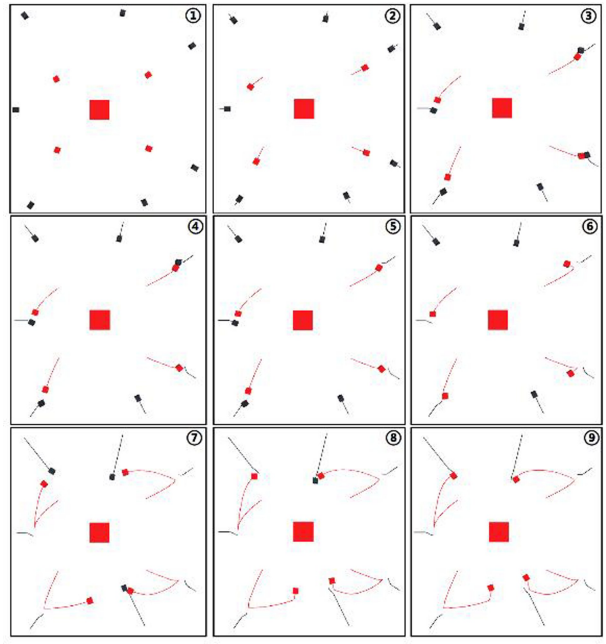


图 7 4 智能体 vs. 7 智能障碍物仿真环境清障过程(电子版为彩图)

Fig. 7 Obstacle clearance process in the simulation environment of 4 agents vs. 7 intelligent obstacles

实验数值结果如表 2 所列。从表中可以观察到，随着实验环境中的智能障碍物数量的增加，AER 逐渐降低，AEL 逐渐增加。其原因在于，智能障碍物越多，多智能体清除所有智能障碍物所需要的回合步长就越多，因此 AEL 逐渐增加；同时根据式(6)的奖励函数，回合步长越多，回合累积奖励就越小，因此，AER 也会越低。但可以注意到，随着环境中智能障碍物的增多，MAPPO 和 MADDPG 算法的 AES 都逐渐降低，尤其是在 3 VS 5 实验环境中，MAPPO 的 AES 仅为75.86%，

而本文提出的 MACOC 算法的 AES 均保持在 98%以上，这说明 MACOC 算法具有优异的稳定性和环境适应性，同时也证明本文提出的 MACOC 算法在清障仿真环境中可以高效解决多智能体协同清障问题，回答了问题 2。

根据 AER 评价标准，在 4 种实验环境中，本文提出的 MACOC 与 MAPPO 和 MADDPG 相比，平均回合奖励分别提高了 7.01%，12.5%，16.78%，15.69%和 26.75%，21.03%，29.20%，35.79%。

表 2　4 种仿真清障环境（智能障碍物）的算法实验结果

Table 2　Experimental results in four simulate environments of obstacle clearance(intelligent obstacles)

Scenario	Algorithm	AER	AEL	AES
2 VS 2	MACOC	**−85.08±0.2**	**42.22±0.25**	**99.78%±0.18%**
2 VS 2	MAPPO	−91.49±1.74	46.15±2.58	98.93%±1.06%
2 VS 2	MADDPG	−106.37±4.17	50.58±2.15	96.53%±0.57%
3 VS 3	MACOC	**−107.23±2.67**	**51.33±1.16**	**99.52%±0.09%**
3 VS 3	MAPPO	−122.64±5.96	57.32±3.09	97.51%±0.67%
3 VS 3	MADDPG	−131.61±6.68	61.28±2.51	97.00%±1.52%
3 VS 5	MACOC	**−156.40±4.20**	**76.13±0.78**	**98.67%±0.12%**
3 VS 5	MAPPO	−187.94±6.84	82.71±1.53	92.60%±3.31%
3 VS 5	MADDPG	−221.15±4.64	90.98±1.70	75.86%±8.27%
4 VS 7	MACOC	**−196.41±2.33**	**83.86±0.44**	**98.26%±0.30%**
4 VS 7	MAPPO	−232.97±4.57	91.80±1.15	93.27%±2.16%
4 VS 7	MADDPG	−247.01±20.2	95.69±4.16	89.98%±4.79%

根据 AEL 评价标准，在 4 种实验环境中，本文提出的 MACOC 与 MAPPO 和 MADDPG 相比，平均回合步长分别减少了 8.52%，10.45%，7.96%，8.65%和 16.03%，12.28%，15.08%，18.01%。根据 AES 评价标准，在 4 种实验环境中，本文提出的 MACOC 算法的平均回合成功率均高于 98%，优于 MACOC 算法和 MADDPG 算法。由 AER 和 AES 可知，在 4 种实验环境中，MACOC 算法较 MAPPO 和 MADDPG 算法获得了更高的清障率和整体回报，回答了问题 3。

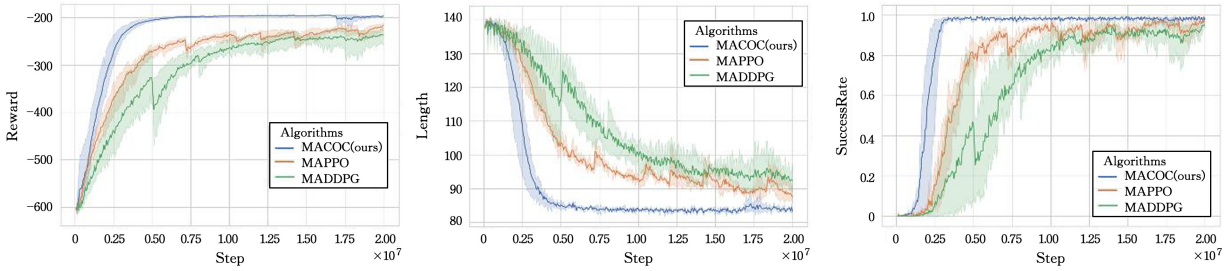
对比表 2 和表 3 中的实验结果可知，与智能障碍物的仿真清障环境相比，在非智能障碍物的仿真清障环境中，3 种算法具有更高的 AER 和更低的 AEL。原因在于，相比具有躲避策略的智能障碍物，非智能障碍物的清除难度较小，算法模型所需的步长更少，即 AEL 更低，根据式(6)的连续奖励函数，算法的 AER 更高。同时，在非智能障碍物的仿真清障环境中，本文提出的 MACOC 算法的 AES 也均高于对比算法，并且均保持在 98%以上。这说明本文提出的 MACOC 算法对智能障碍物和非智能障碍物均具备优秀的清障性能，回答了问题 4。

表 3　4 种仿真清障环境（非智能障碍物）的算法实验结果

Table 3　Experimental results in four simulate environments of obstacle clearance(non-intelligent obstacles)

Scenario	Algorithm	AER	AEL	AES
2 VS 2	MACOC	**−65.18±0.63**	**36.29±0.15**	**99.55%±0.38%**
2 VS 2	MAPPO	−76.64±2.25	40.23±2.52	97.21%±1.01%
2 VS 2	MADDPG	−87.33±3.73	43.22±1.23	94.49%±3.39%
3 VS 3	MACOC	**−72.57±1.08**	**40.95±1.27**	**99.83%±0.23%**
3 VS 3	MAPPO	−88.66±11.99	44.35±4.25	98.75%±1.82%
3 VS 3	MADDPG	−93.72±7.69	46.68±2.62	98.42%±1.53%
3 VS 5	MACOC	**−136.12±4.86**	**64.33±0.44**	**98.47%±0.29%**
3 VS 5	MAPPO	−148.63±15.75	66.24±4.61	97.74%±2.03%
3 VS 5	MADDPG	−175.76±4.64	75.75±6.02	87.45%±7.15%
4 VS 7	MACOC	**−149.25±8.13**	**68.91±0.14**	**98.24%±0.31%**
4 VS 7	MAPPO	−167.89±20.53	71.57±0.40	96.18%±2.45%
4 VS 7	MADDPG	−195.72±15.34	84.05±4.37	89.52%±4.76%

如图 8 所示，本文以 4 智能体 vs. 7 智能障碍物仿真环境为例展示了算法模型的训练收敛过程。MAPPO 算法和 MADDPG 算法的收敛速度较慢，收敛过程波动较大，而 MACOC 算法的收敛速度更快，收敛过程更稳定。



(a) AER　(b) AEL　(c) AES

图 8　4 智能体 vs. 7 智能障碍物仿真清障环境中的 3 种算法收敛图

Fig. 8　Convergence graphs of three algorithms in the simulation environment of 4 agents vs. 7 intelligent obstacles

取得上述实验结果的原因在于，MAPPO 使用近端策略优化算法来优化策略，通过限制更新幅度来保持策略的稳定性，但这导致算法在探索和利用之间的平衡方面受到一定的限制，从而影响了回报的表现。而本文的 MACOC 算法使用深度确定性策略梯度算法来进行策略优化，通过最大化动作值函数来更新策略，提高了回报表现。MADDPG 的 Critic 网络接收的其他智能体的状态信息的重要性程度不同，但是它们并没有进行信息重要性的区分，导致 Critic 对 Actor 网络的指导性能较差，无法训练得到优异的清障策略模型。而本文提出的 MACOC 算法利用 Attention Critic 网络来区分其他智能体的状态信息的重要性程度，从而更好地指导 Actor 网络的训练，得到更加优异的清障策略模型。

结束语　为解决多智能体协同清障问题，本文首先创建了多智能体协同清障模型，并建立了 4 类清障环境。其次，本文将多智能体协同清障过程定义为马尔可夫决策过程，并构造了状态空间、动作空间和奖励函数。最后，提出了一种基于深度确定性策略梯度与注意力 Critic 的多智能体协同清障（MACOC）算法。实验证明，本文提出的 MACOC 算法在 AER，AEL 和 AES 这 3 个实验评价标准中均优于 MAPPO 和 MADDPG 算法，可训练得到优异的多智能体协同清障

模型，为多智能体协同清障问题提供了一个基于强化学习的解决方案，有助于推动多智能体自主导航的发展。

然而，本文算法仅在小规模理想实验环境中进行了测试，现实环境中尚存在很多影响清障过程的其他因素。在未来的工作中，我们将继续完善所提出的多智能体协同清障算法MACOC，以解决大规模协同清障问题，并将其用于解决实际生活中影响智能体自主导航的障碍物问题中。

参 考 文 献

[1] NTAKOLIA C, MOUSTAKIDIS S, SIOURAS A. Autonomous path planning with obstacle avoidance for smart assistive systems[J]. Expert Systems with Applications, 2023, 213: 119049.

[2] CORNO M, GIMONDI A, PANZANI G, et al. A non-optimization-based dynamic path planning for autonomous obstacle avoidance[J]. IEEE Transactions on Control Systems Technology, 2022, 31(2): 722-734.

[3] DING J, GAO L, LIU W, et al. Monocular camera-based complex obstacle avoidance via efficient deep reinforcement learning [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(2): 756-770.

[4] LI Z, LI J, WANG W. Path Planning and Obstacle Avoidance Control for Autonomous Multi-Axis Distributed Vehicle Based on Dynamic Constraints[J]. arXiv: 1312. 7572, 2013.

[5] NAYYAR M, WAGNER A R. Aiding Emergency Evacuations Using Obstacle-Aware Path Clearing[C]// 2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO). IEEE, 2021: 7-14.

[6] LU X, JIA Y. Scaled Event-Triggered Resilient Consensus Control of Continuous-Time Multi-Agent Systems Under Byzantine Agents[J]. IEEE Transactions on Network Science and Engineering, 2022, 10(2): 1157-1174.

[7] ORR J, DUTTA A. Multi-Agent Deep Reinforcement Learning for Multi-Robot Applications: A Survey [J]. Sensors, 2023, 23(7): 3625-3625.

[8] YU Y, GUO J, CHADLI M, et al. Distributed adaptive fuzzy formation control of uncertain multiple unmanned aerial vehicles with actuator faults and switching topologies[J]. IEEE Transactions on Fuzzy Systems, 2022, 31(3): 919-929.

[9] DENG Z, YANG K, SHEN W, et al. Cooperative Platoon Formation of Connected and Autonomous Vehicles: Toward Efficient Merging Coordination at Unsignalized Intersections [J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(5): 5625-5639.

[10] HAO Q, XU F, CHEN L, et al. Hierarchical Multi-agent Model for Reinforced Medical Resource Allocation with Imperfect Information[J]. ACM Transactions on Intelligent Systems and Technology, 2022, 14(1): 1-27.

[11] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control throughdeep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[12] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[13] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv: 1707. 06347, 2017.

[14] LI P Y, TANG H Y, YANG T P, et al. Improving Multi-Agent Reinforcement Learning with Progressive Mutual Information Collaboration[C] //2022 International Conference on Machine Learning. 2022: 12979-12997.

[15] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[16] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of ppo in cooperative multi-agent games[J]. Advances in Neural Information Processing Systems, 2022, 35: 24611-24624.

[17] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). 2017: 6382-6393.

[18] FAFSAR M M, CRUMP T, FAR B. Reinforcement learning based recommender systems: A survey [J]. ACM Computing Surveys, 2022, 55(7): 1-38.

[19] ZHAO F, WANG Z, WANG L, et al. A multi-agent reinforcement learning driven artificial bee colony algorithm with the central controller[J]. Expert Systems with Applications, 2023, 219: 119672.

[20] REN J, GUO S, CHEN F. Orientation-preserving rewards' balancing in reinforcement learning [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(11): 6458-6472.

[21] OROOJLOOY A, HAJINEZHAD D. A review of cooperative multi-agent deep reinforcement learning [J]. Applied Intelligence, 2023, 53(11): 13677-13722.

WANG Xianwei, born in 1999, postgraduate, is a member of CCF (No. P2627G). His main research interests include reinforcement learning and robot navigation.

FENG Xiang, born in 1977, Ph. D, professor, is a member of CCF (No. 16665M). Her main research interests include distributed swarm intelligence and evolutionary computing, reinforcement learning, and big data intelligence.

(责任编辑：杨雪敏)