计算机科学
COMPUTER SCIENCE

基于MADDPG的无人机群空中拦截作战决策研究

蔺向阳, 邢清华, 邢怀玺

引用本文

蔺向阳, 邢清华, 邢怀玺.基于MADDPG的无人机群空中拦截作战决策研究[J]. 计算机科学, 2023, 50(6A): 220700031-7.

LIN Xiangyang, XING Qinghua, XING Huaixi. Study on Intelligent Decision Making of Aerial Interception Combat of UAV Group Based onMADDPG [J]. Computer Science, 2023, 50(6A): 220700031-7.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

干扰环境下无人机群动态频谱决策方法

Dynamic Spectrum Decision-making Method for UAV Swarms in Jamming Environment

计算机科学, 2022, 49(12): 326-331. https://doi.org/10.11896/jsjkx.220400228

基于多智能体强化学习的端到端合作的自适应奖励方法

Adaptive Reward Method for End-to-End Cooperation Based on Multi-agent Reinforcement Learning

计算机科学, 2022, 49(8): 247-256. https://doi.org/10.11896/jsjkx.210700100

基于定向天线的飞行自组网定向路由协议综述

Review of Directional Routing Protocols for Flying Ad-Hoc Networks Based on Directional Antennas

计算机科学, 2021, 48(11): 334-344. https://doi.org/10.11896/jsjkx.210400182

多智能体强化学习综述

Overview on Multi-agent Reinforcement Learning

计算机科学, 2019, 46(8): 1-8. https://doi.org/10.11896/j.issn.1002-137X.2019.08.001

0—1整数规划在科技项目管理中的应用

计算机科学, 2003, 30(7): 77-79.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700031

基于 MADDPG 的无人机群空中拦截作战决策研究

蔺向阳[1]　邢清华[2]　邢怀玺[2]
1 中国人民解放军军事科学院　北京 100091
2 空军工程大学防空反导学院　西安 710051

摘　要　基于未来现代化作战需求,构建作战想定,研究在此想定条件下,使用强化学习解决关于红蓝双方无人机编队空中拦截任务的多目标智能决策问题。根据作战模式和应用需求,选择多智能体确定性梯度算法,并对算法原理进行简要介绍;按照想定,编程搭建了完备的模拟作战训练平台;设计智能体网络模型、网络参数和训练方法;经过训练,初步达到预期效果。实验证明了所选用算法能够有效地解决该类问题,不仅为该类问题的现实应用提供了技术支撑,也为更复杂作战场景和作战任务条件下智能决策的研究提供了理论基础和实验参考。

关键词:MADDPG;无人机群;智能决策;空中拦截作战;多智能体强化学习
中图法分类号　TP181

Study on Intelligent Decision Making of Aerial Interception Combat of UAV Group Based on MADDPG

LIN Xiangyang[1],XING Qinghua[2] and XING Huaixi[2]
1 Academy of Military Sciences,Beijing 100091,China
2 Air Defense and Anti-Missile College,Air Force Engineering University,Xi'an 710051,China

Abstract　Based on the requirements of future modern operations,a combat scenario is built. Under this scenario,reinforcement learning is used to solve the multi-target intelligent decision-making problem about aerial interception mission of UAVs. The multi-agent reinforcement learning algorithm is selected according to the operational mode and application requirements,and the algorithm principle and process are briefly introduced. The simulated combat system is developed. Design network model,network parameters and training methods. After training,the expected results have been achieved. The effectiveness of the experiment is proved,which not only provides technical support for practical application of this kind of problem,but also provides theoretical basis and experimental reference for the study of intelligent decision making in more complex combat scenarios and combat mission conditions.

Keywords　MADDPG,UAV group,Intelligent decision,Air interception combat,Multi-agent reinforcement learning

21 世纪以来,以人工智能(Artificial Intelligence,AI)为代表的一系列高新科技快速发展,催发了一系列新型的武器装备,战场环境和战争形态逐步趋于信息化、网络化和智能化,与之匹配的作战模式、作战思想和作战理论也在迅速变革。传统的以人为主的决策战法已经难以满足新时代国防和军队建设要求,智能算法在现代战争中扮演着越来越重要的角色。

目前在军事智能化研究方面,郑少秋综述了智能化作战的基本概念和关键技术[1]。姜广顺和孙祁研究了外军人工智能的发展现状[2-3]。在一般军事应用方面,付翔研究汇总了智能化空战的各方面能力体现[4],王闯研究了防空反导的智能战场态势感知[5],蔺向阳研究了要点防空模型的智能兵力优化[6]。房霄使用深度强化学习研究了舰艇的空中威胁行为[7],刘亚杰研究了舰炮武器的智能化应用[8]。李高云和黄巍研究了智能化在电子对抗装备方面的应用[9-10]。在无人机和智能决策方面,赵伟对其发展现状、应用和未来进行了深入分析[11]。丁振林使用 ε-贪心算法研究动态目标分配[12]。董康生整理研究了美军无人空战装备的发展动态[13]。郑凯元分别使用人工鱼群算法和 DDPG 算法研究了无人机的路径规划[14]。李波使用 MADDPG 算法解决了无人机的简单协同任务决策[15]。在国外研究中,Galan 综述了机器学习在军事数据处理和决策制定等应用中的基本概况[16],Sharma 研究了如何用 AI 辅助电子战[17],Lei 使用 AI 解决了无人地面战车的使用[18],Hodicky 研究基于 AI 的兵棋推演辅助[19],等。

在目前众多研究中,较大一部分仍停留在战略构想和概念层面,实际落于技术实现和应用层面的相对较少,尤其在群无人机的全智能化作战方面。在此背景下,本文选择无人机群的智能化目标选择方面进行研究。根据未来现代化作战需求,合理构建作战假想,并搭建相适配的模拟智能体训练环境;使用目前解决群智能体最有效的多智能强化学习算法;

基金项目:国家自然科学基金(71771216,72071209,72001214)
This work was supported by the National Natural Science Foundation of China(71771216,72071209,72001214).
通信作者:蔺向阳(95014052@qq.com)

设计网络结构、奖励函数和训练方法训练智能体，得到预期效果。

1 作战想定与问题描述

想定作战背景为在限定正方形区域内进行的空中拦截作战，区域如图 1 所示，边长为 400 km。红蓝双方主要作战力量为无人机群组成作战编队，A 和 B 两位置点为红蓝双方机场，双方无人机编队由此起飞。作战任务为双方无人机编队根据战场力量布局，实时进行目标选择，并根据目标调整飞行策略，完成既定拦截与突防任务。P_1，P_2 两位置点为蓝方欲轰炸目标点。

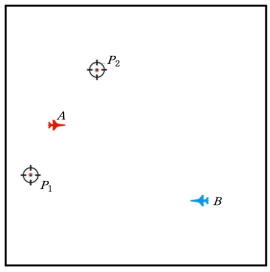


图 1 假想作战区域示意图

Fig. 1 Schematic diagram of imaginary combat area

作战相关假设如下，鉴于保密要求，所有装备不涉及具体型号和真实参数信息。

(1)蓝方作战目的为躲避红方机群拦截，对目标点 P_1 和 P_2 造成最大杀伤。

(2)蓝方作战力量为 4 组无人机编队，每组为一个作战单位，编队最高航速为 720 km/h，最低航速为 430 km/h，最大加速度为 20 m/s^2。

(3)蓝方每组编队携带一组近距空地弹，有效攻击距离为 50 km，即蓝方单位与目标点距离小于此距离时，可对目标点进行单次成功轰炸，完成轰炸后，编队失去作战能力。目标点受到 2 次轰炸后，完全摧毁。

(4)红方作战目的为拦截所有蓝方编队，保卫目标点 P_1 和 P_2。

(5)红方作战力量为 4 组无人机编队，每组为一个作战单位，编队最高航速为 800 km/h，最低航速为 430 km/h，最大加速度为 25 m/s^2。

(6)红方每组编队携带一组近距空空导弹，有效攻击距离为 10 km，红方单位与目标距离小于此距离时，可对目标进行单次成功拦截，拦截后双方皆失去作战能力。

(7)红蓝双方同时从各自机场起飞，作战区域仅限于图上区域。有效距离仅计算二维平面距离，暂不考虑高度。任务时间从起飞后算起最多 30 min，不考虑返航时间。

(8)假设电磁干扰较弱，双方可通过雷达侦测敌方实时坐标信息，也能与友方通信共享友方所有信息。

使用强化学习算法研究智能决策问题，研究内容分为 3 个部分。

(1)研究如何在作战想定和假设基础上，对作战行动进行抽象化和模型化处理，提取重要作战信息，并以此为依据编程开发模拟作战系统作为实验平台，用于后期实验。

(2)研究如何将强化学习的基本框架应用于多无人机组，使每一个无人机组单元映射为模拟作战系统中的一个智能体，并合理设计对应的神经网络结构和参数。

(3)研究如何实现智能体神经网络与模拟作战平台信息的对接，快速有效地训练多智能体，使得训练后的多智能体在模拟作战中能够具备较高的智能决策水平，精准选择攻击目标，完成作战任务。同时，完成对整个研究内容的验证。

2 模拟作战系统开发

模拟作战平台即智能体的训练平台，用于对代表无人机组的多智能体进行训练，开发主要分 3 个部分。

2.1 战场坐标化处理

首先，根据作战假想，对战场信息进行坐标化处理。如图 2 所示，以战场中心为原点，分别以正东、正北方向为 X 轴、Y 轴正方向，以 1 km 为单位长度建立平面直角坐标系。此时，各重要点的位置坐标大致为：$A(-120,20)$，$B(100,-100)$，$P1(-160,-60)$，$P2(-60,100)$。

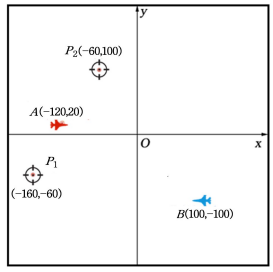


图 2 坐标化处理后的假想战场示意图

Fig. 2 Schematic diagram of imaginary battlefield after coordinate processing

2.2 模拟作战系统搭建

在坐标建立的基础上，使用 Pyglet 开发强化学习的模拟作战系统作为训练环境。系统的刷新频率为 5 s/step，即对应假想作战中，每 5 s 系统刷新一步。系统的初始化界面如图 3 所示。界面窗口为 1 600 * 1 600 像素，代表 400 km * 400 km 的假想战场。

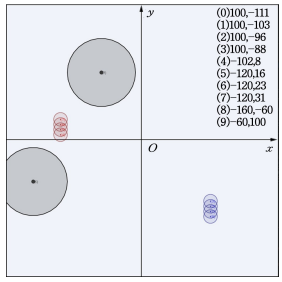


图 3 训练环境的初始化界面

Fig. 3 Initialization interface of training environment

红蓝方各单位以示意点的形式显示在界面上。各点上有相应的序号。其中，0 到 3 号代表 4 组已起飞蓝方无人机编队，4 到 7 号代表 4 组已起飞红方无人机编队，红色外圈指示其攻击范围，蓝方无人机进入后将被拦截。8 号和 9 号代表两个蓝方欲轰炸目标点，不可移动，黑色外圈指示出蓝方的有效轰炸距离，蓝方无人机进入后将进行有效轰炸。

汇总假想战场和模拟作战系统中各参数的对应关系，如表 1 如列。

表 1　假想战场与模拟作战系统中各参数的对比

Table 1　Comparison of parameters in imaginary battlefield and simulated combat system

参数类别	假想战场参数	模拟系统参数
作战范围	400 km * 400 km	1 600 pt * 1 600 pt
单场最长作战时间	30 min	360 step
蓝方速度	430～720 km/h	2.4～4.0 pt/step
红方速度	430～800 km/h	2.4～4.4 pt/step
蓝方最大加速度	20 m/s^2	2 pt/step^2
红方最大加速度	25 m/s^2	2.5 pt/step^2
蓝方空地弹杀伤距离	50 km	200 pt
红方空空弹拦截距离	10 km	40 pt

2.3　智能体运动规则设计

本模型主要研究智能目标分配,故智能体的动作为选择离散目标。红方动作空间是 4 维,蓝方动作空间是 2 维。当选定最优目标后,智能体会在不低于最小巡航速度的前提下,尽快调整战机速度,指向目标方向,并尽可能保持最大航度向目标方向前进。

具体实现方法为:

(1)设目标组为

$$\boldsymbol{A}=\begin{cases}[A_1,A_2,A_3,A_4], & \text{如果 } \boldsymbol{A} \text{ 为蓝方战机}\\ [A_1,A_2], & \text{如果 } \boldsymbol{A} \text{ 为红方阵地}\end{cases}$$

设最优目标为 A_i,其坐标为(x_a,y_a)。

(2)设智能体的初始坐标为(x,y),速度为(v_x,v_y),最大速度为 mv,最大加速度为 ma。

(3)令智能体在单位时刻 t 后的速度为(v_x',v_y'),有:

$$\begin{cases}v_x'=ma\cdot(x_a-x)/\sqrt{(x_a-x)^2+(y_a-y)^2}\\ v_y'=ma\cdot(y_a-y)/\sqrt{(x_a-x)^2+(y_a-y)^2}\end{cases}\tag{1}$$

(4)求速度变化 $\Delta v=(v_x'-v_x,v_y'-v_y)$,计算其模 $\|\Delta v\|=\sqrt{(v_x'-v_x)^2+(v_y'-v_y)^2}$,如果有 $\|\Delta v\|>ma$,则对其进行修正:

$$\begin{cases}\Delta v_x=ma\cdot(v_x'-v_x)/\|\Delta v\|\\ \Delta v_y=ma\cdot(v_y'-v_y)/\|\Delta v\|\end{cases}\tag{2}$$

之后得到修正后 t 时刻的速度:

$$\begin{cases}v_x'=v_x+\Delta v_x\\ v_y'=v_y+\Delta v_y\end{cases}\tag{3}$$

(5)由智能体初始坐标(x,y)、初始速度(v_x,v_y)和 t 时刻速度(v_x',v_y'),得 t 时刻坐标(x',y'):

$$\begin{cases}x'=t\cdot(v_x+v_x')/2+x\\ y'=t\cdot(v_y+v_y')/2+y\end{cases}\tag{4}$$

3　算法与训练模型设计

3.1　多智能体强化学习算法分析

根据作战假想,在红蓝双方中,同一阵营下各无人机相互协作,而不同阵营下各无人机处于相互竞争的关系,因此选择目前对解决此类“协作-竞争”关系最有效的群智能体强化学习算法——多智能体深度确定性策略梯度算法[20](Multi-Agent Deep-Deterministic Policy Gradient Algorithms,MADDPG)。该算法也是目前最为流行的多智能体强化学习算法之一。同时考虑到未来此类研究必将从模拟平台推广至现实演练,彼时每一批数据都意味着巨额的时间和经济代价。而 MADDPG 算法继承于单智能体的深度确定性策略梯度[21](Deep-Deterministic Policy Gradient,DDPG),具备离线学习的能力,能有效节省数据,这对未来现实作战应用意义重大。

3.1.1　DDPG 算法

DDPG 算法使用经典的演员-评论家(Actor-Critic,AC)框架结构[22]。AC 框架分两个网络,演员 A 网络基于智能体的当前状态输出策略,为智能体选择最优动作,并通过策略梯度算法[23](Policy Gradient,PG)不断优化策略网络。评价 C 网络对智能体每个状态-动作组的理论价值即 Q 值进行评估,并根据智能体的实际收益,通过时序差分(Temporal-Difference,TD)调整网络参数,提高评估的准确性。

AC 结构通过使用两个深度神经网络分别实现了输出动作空间的连续化和输入状态空间的连续化,完美地将低维、离散空间上的强化学习拓展到高维、连续空间。DDPG 在 AC 的框架上,借用了深度 Q 学习[24](Deep Q-learning,DQN)的在线-目标双网络结构,实现了由在线训练向离线训练的拓展,省却了如重要性采样(Importance Sampling)等一系列复杂操作,同时也提高了数据的利用率,符合现实生产生活的需要,得到了广泛应用。

图 4 展示了 DDPG 的数据流结构。左边为策略网络,通过策略梯度优化在线策略网络参数,将状态映射到最优策略,之后根据策略输出确定的动作,并将其送入右边的在线价值网络评估状态-动作价值。在线价值网络使用价值梯度更新优化,提高评估准确度,将状态-动作组映射为价值函数。两个在线网络更新若干次后,使用滑动平均更新法对目标网络进行更新。

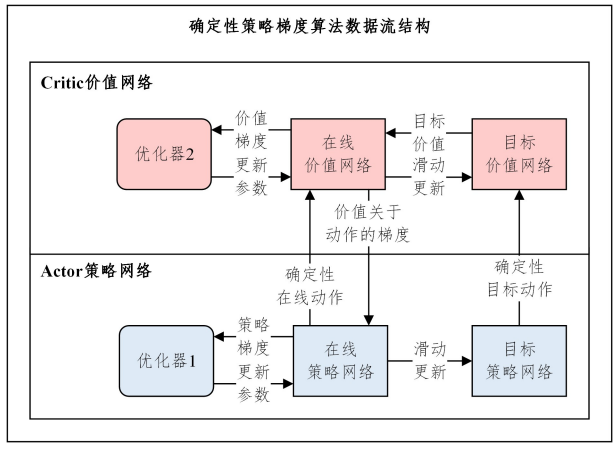


图 4　DDPG 数据流结构示意图

Fig. 4　Schematic diagram of DDPG data flow structure

3.1.2　MADDPG 算法

MADDPG 在 DDPG 算法基础上改进。不同于单智能体相对单一稳定的训练环境,多智能体面临的情况复杂得多,因此需要解决如维度爆炸、环境的不稳定性、奖励函数设定困难等更为复杂的问题。

因此,为了适应多智能体的训练任务,MADDPG 使用了集中式训练、分布式执行的方法,将训练和执行两个过程分开,在这两个独立的过程中分别使用较少的训练量,最终完成了一个整体的训练工作。网络中数据流的结构示意图如图 5 所示。图 5 中外圈蓝色区域表示每个智能体在训练价值网络参数时使用全局的信息,保证了训练环境的稳定性;内圈红色部分表示在实际运行时,每一个智能体仅使用局部的信息得出策略,一定程度上减少了维度爆炸的复杂度。除此之外,由于以 DDPG 为基础算法,MADDPG 还具备离线学习能力,

能够更好地使用经验回放等技巧，提高了数据利用率。

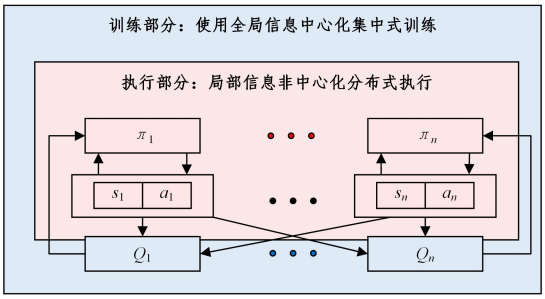


图 5 MADDPG 数据流结构示意图

Fig. 5 Schematic diagram of MADDPG data flow structure

MADDPG 算法流程如算法 1 所示。

算法 1 MADDPG 算法

输入：初始化所有个智能体的价值网络 $Q(s,a|\theta_Q)$，策略网络 $\mu(s|\theta_\mu)$ 的网络参数 $\theta_Q=(\theta_{1,Q},\cdots,\theta_{N,Q})$ 和 $\theta_\mu=(\theta_{1,\mu},\cdots,\theta_{N,\mu})$；将两类在线网络参数复制传递给对应目标网络的参数：$\theta_{Q'}\leftarrow\theta_Q$，$\theta_{\mu'}\leftarrow\theta_\mu$；初始化记忆库 D

输出：最优网络参数 θ_Q，θ_μ 和最优策略

循环执行 M 条轨迹，对于每一条轨迹 for episode=1，…，M：

获得初始化状态 $s^1=(s_1^1,\cdots,s_N^1)$

执行步进操作，对于当前轨迹中的每一步 for t=1，…，T：

1. 输入 s^t 根据在线策略，获得每一个智能体的行为 $a_i^t=\mu(s_i^t|\theta_{i,\mu})$
2. 执行每一个智能体的行为 $a^t=(a_1^t,\cdots,a_N^t)$，得到收益 r^t 和新的状态 s^{t+1}
3. 将状态转变序列 $D^t=(s^t,a^t,r^t,s^{t+1})$ 存入记忆库 D 中，且复制状态 $s^t\leftarrow s^{t+1}$
4. 当 D 库中数据足够多时，随机从中采样 S 组序列[D_1，…，D_S]作为一个批次对在线网络进行训练。对于第 j 组序列 $D^j=(s^j,a^j,r^j,s^{j+1})$：

5. 对每一个智能体 for i=1，…，N：

5.1. 令：$y^j=r_i^j+\gamma Q_i'(s^{j+1},\mu'(s^{j+1}|\theta_\mu')|\theta_{Q_i'})$ 作为目标 Q 价值

5.2. 定义损失函数 $L_i=1/S\cdot\sum_j(y^j-Q_i(s^j,a^j|\theta_{Q_i}))^2$

5.3. 最小化 L_i 更新在线 Critic 网络参数 θ_{Q_i}

5.4. 计算样本策略梯度，负梯度下降更新在线 Actor 网络参数 $\theta_{i,\mu}$：

$$\nabla_{\theta_{i,\mu}}J=1/S\cdot\sum_j\nabla_{a_i}Q_i(s^j,a_1^j,\cdots,a_i^j,\cdots,a_N^j|\theta_{Q_i})|_{a_i=\mu(s_i^j)}\nabla_{i,\mu}\mu_i(s_i^j|\theta_{i,\mu})$$

5.5. 滑动法更新目标网络参数 $\theta_{Q'}$，$\theta_{\mu'}$：

$\theta_{Q'}\leftarrow\tau\theta_Q+(1-\tau)\theta_{Q'}$，$\theta_{\mu'}\leftarrow\tau\theta_\mu+(1-\tau)\theta_{\mu'}$

循环步进至当前单条轨迹结束

循环执行至 M 条轨迹结束，停止训练

3.2 深度神经网络模型设计

根据作战想定，本实验中共使用 8 个智能体，分两组，每组 4 个，其中组内为合作关系，组间为竞争关系。共需建立 4 类深度神经网络，即在线策略网络 P 网络、在线价值网络 Q 网络和分别与之对应的目标策略网络 TP 网络、目标价值网络 TQ 网络。每个智能体都有这 4 个网络，同类网络对应结构相同。

P 网络共分 4 层，包含两个隐藏层，相邻层之间全连接，具体结构如表 2 所列。

表 2 智能体 P 网络结构

Table 2 P network structure of agent

输入层	隐藏层 1	隐藏层 2	输出层
46 维	64 维 Relu 激活	64 维 Relu 激活	2 维或 4 维 Softmax 激活

P 网络的输出是目标选择的分类动作，当智能体为红方时，输出为 4 维；当智能体为蓝方时，输出为 2 维。P 网络的输入为单个智能体可知的局部状态信息，46 维数据的具体意义如表 3 所列。

表 3 智能体 P 网络输入数据的具体意义

Table 3 Specific significance of P network input data of agent

输入信息	自身速度	自身位置	自身生命	保卫目标点的位置	保卫目标点的生命	其他智能体的速度	其他智能体的位置	其他智能体的生命
信息维度	2	2	1	4	2	14	14	7

Q 网络共分 4 层，包含两个隐藏层，相邻层之间全连接，具体结构如表 4 所列。

Q 网络的输出为状态-动作组的价值，属于 1 维数值标量。Q 网络的输入为全局状态和动作信息，70 维数据的具体意义如表 5 所列。

表 4 智能体 Q 网络结构

Table 4 Q network structure of agent

输入层	隐藏层 1	隐藏层 2	输出层
70 维	64 维 Relu 激活	64 维 Relu 激活	1 维 无激活

表 5 智能体 Q 网络输入数据的具体意义

Table 5 Specific significance of Q network input data of agent

输入信息	所有智能体的速度	所有智能体的位置	所有智能体的生命	保卫目标点的位置	保卫目标点的生命	红方智能体的动作	蓝方智能体的动作
信息维度	16	16	8	4	2	16	8

TP 网络和 TQ 网络分别是 P 网络和 Q 网络的目标网络，网络结构形式和数据意义与之相同，不再重复赘述。

3.3 智能体奖励函数设计

智能体的奖励函数对智能体的训练效果起着决定性作用，不仅需要根据双方的作战任务总体设定，使得智能体学会“协作-竞争”，达到最终作战目的；还要增加辅助中途引导，奖励避免长时间任务过程中奖励过于稀疏导致的训练缓慢问题；同时还需要注意与神经网络的结构和训练参数相适应，以防出现训练过程中参数爆炸导致训练失败的情形。

在本实验中，蓝方任务单一，即轰炸目标点，根据实验经验，设置固定的奖励函数。

(1)总目标奖励：若在规定作战时间内，蓝方将任意一个目标点摧毁，整体各获得正向奖励 20；若两目标点全部摧毁，则整体各获得正向奖励 40。

(2)分目标奖励：当蓝方单位对任意目标点轰炸成功，则本单位获得正向奖励 10。

(3)辅助引导奖励：只要蓝方单位动作选择的目标点还未摧毁，本单位获少量正向奖励 0.01，否则获得负向奖励－1。

对比蓝方，红方的总任务是保卫目标点，但要引导其主动去拦截蓝方，因此设置如下奖励函数。

(1)总目标奖励：若在规定作战时间内，蓝方将任意一个目标点摧毁，则整体各获得负向奖励－20；若两目标点全部摧毁，则整体各获得负向奖励－40；只有作战结束时目标点未受轰炸，才整体各获得正向奖励 10。

(2)分目标奖励：当红方单位对任意蓝方单位拦截成功时，本单位获得正向奖励 10。

(3)辅助引导奖励：只要红方单位动作选择的蓝方目标还未失去作战能力，则本单位获得少量正向奖励 0.01，否则获得负向奖励－1。

辅助奖励中，负向奖励数值较大，以抑制智能体选择“死亡”目标；正向奖励极小，是为了避免淹没总目标和分目标奖励，使智能体在“生存”目标中选择最适合自己的。

4 实验与结果分析

设定网络训练样本记忆库 D 的容量为 100 000 组。训练抽样每批的规模 batch_size＝256。神经网络参数按标准正态分布进行初始化。

首先按照初始化参数进行探索实验：初始探索率为 $\varepsilon=0.01$。将每一步的状态转移组(s,a,r,s')存入记忆库。当记忆库存满后，开始训练网络。同时随着实验的继续运行，记忆库中样本被不断覆盖更新。

实验每运行 40 步，对所有智能体训练 1 轮。由 MADDPG 训练原理可知，对于每个智能体，需要训练的仅为图 2 中两个在线网络。损失函数分别是通过策略负梯度不断降低的策略损失函数 P_loss，和通过价值梯度不断降低的价值损失函数 Q_loss。

每训练 200 轮，记录一次各智能体网络的损失函数 P_loss 与 Q_loss 的值。训练的学习率初始化为 $lr=0.0001$。根据长延时收益强化模型折扣系数选择法[25]，选择折扣因子为$\gamma=0.99$。为了提高训练效率，打乱了样本之间的相关性，采用随机抽取的方法从样本库中获得训练样本。记录整个训练过程中各智能体的任务完成情况，以及每个智能体对应损失函数的变化。

通过实验可以发现，开始训练时，双方智能体都是随机选择目标，但由于蓝方的可选目标较少且固定，因此任务成功率较高。

图 6 给出了训练初期某局对战记录中抽取的 6 个典型帧。图中会出现多个红方单位拦截一个蓝方目标，或者拦截已经消失的目标，导致蓝方其他单位失去拦截，最终战果是蓝方进行了两次有效轰炸，红方上侧目标点被完全摧毁。

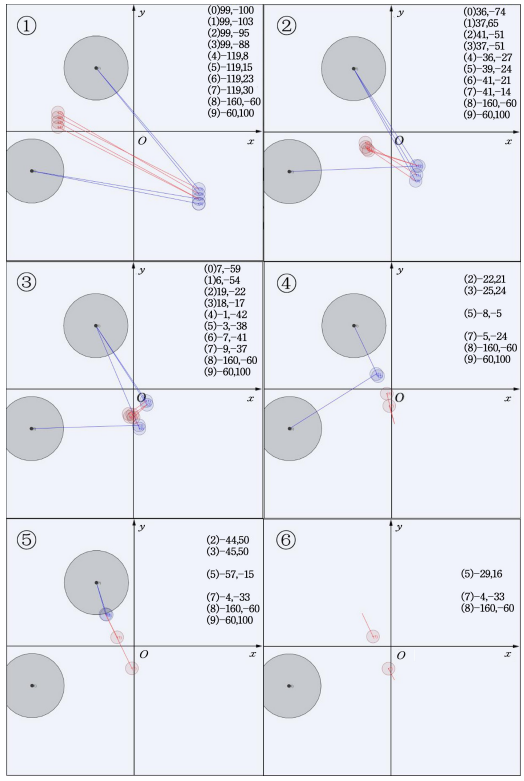


图 6 未经训练时某轮实验记录中抽取的 6 个典型帧

Fig. 6 6 typical frames extracted from anepisode of experimental records without training

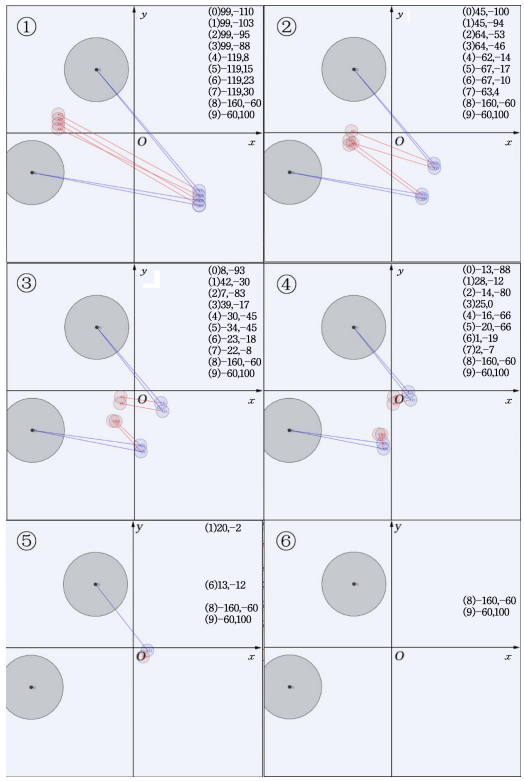
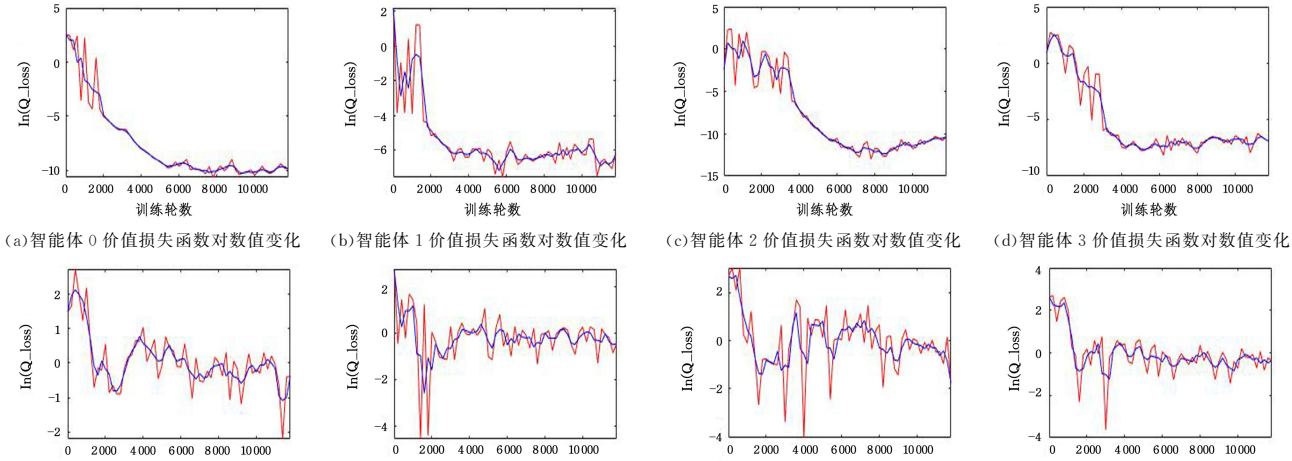


图 7 经过训练后某轮实验记录中抽取的 6 个典型帧

Fig. 7 6 typical frames extracted from an episode of experimental records after training

继续实验发现，红方逐渐学会了分散拦截目标，且选择目标更为固定，这对其接近并成功拦截蓝方极其重要。因此，红方拦截率逐渐提高。图 7 给出了经大量训练后某局对战记录

中抽取的 6 个典型帧。图中红方基本实现了合理的一对一目标分配，最终以红方完全拦截蓝方，实现保卫目标地的作战任务而告终。整个训练过程中各智能体的价值损失函数和策略损失函数的变化如图 8 和图 9 所示。其中，红色为实际函数变化曲线，参考意义不大；蓝色为平滑处理后的变化曲线，是主要的分析对象。发现两图中所有曲线的大致趋势都是下降的，说明训练在整个过程中是收敛的，达到了较好的训练效果。

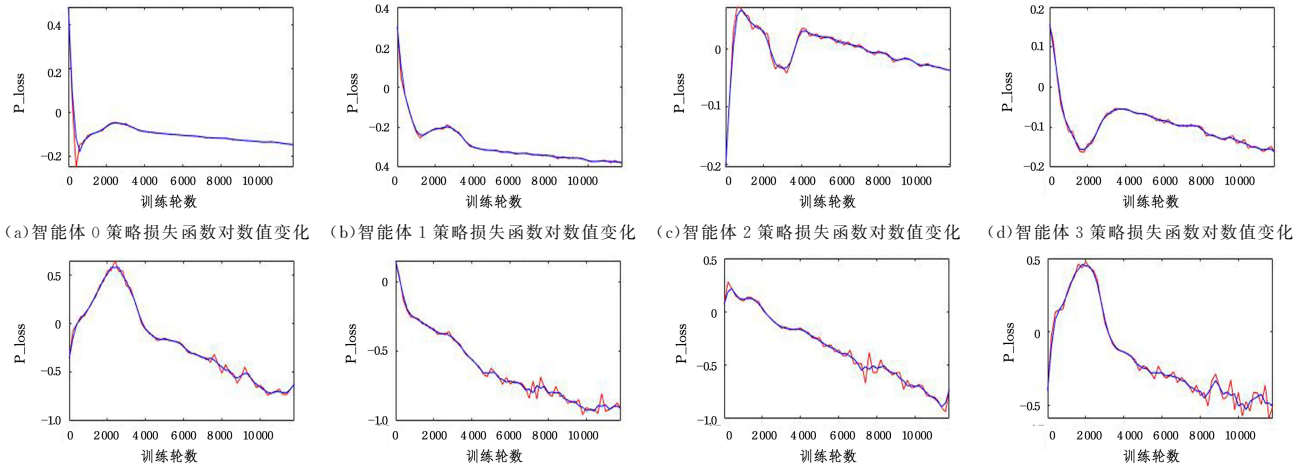


(a)智能体 0 价值损失函数对数值变化 (b)智能体 1 价值损失函数对数值变化 (c)智能体 2 价值损失函数对数值变化 (d)智能体 3 价值损失函数对数值变化

(e)智能体 4 价值损失函数对数值变化 (f)智能体 5 价值损失函数对数值变化 (g)智能体 6 价值损失函数对数值变化 (h)智能体 7 价值损失函数对数值变化

图 8 各智能体价值损失函数对数值的变化

Fig. 8×Changes ofvalue loss function logarithm value of each agent



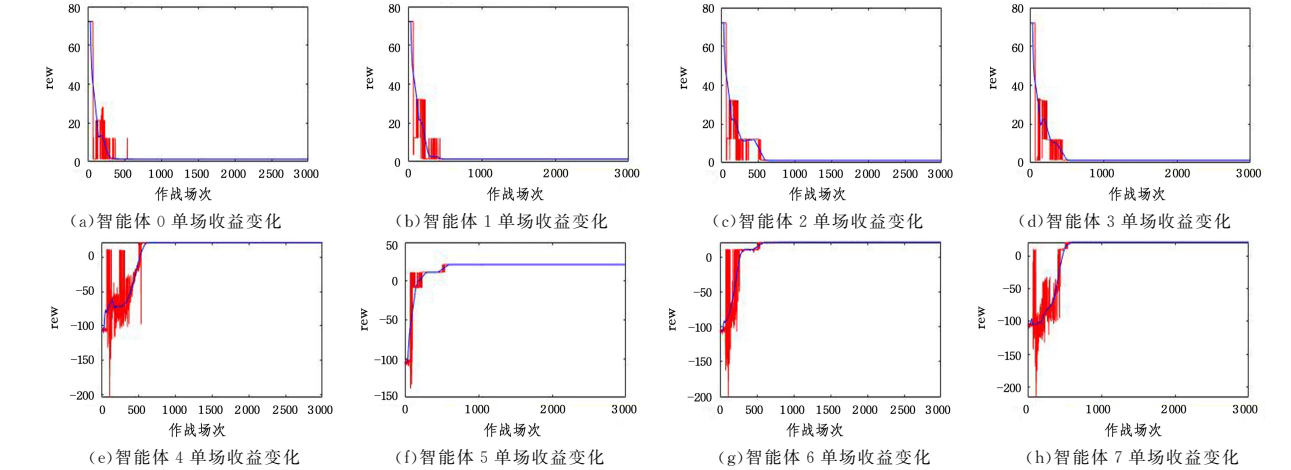
(a)智能体 0 策略损失函数对数值变化 (b)智能体 1 策略损失函数对数值变化 (c)智能体 2 策略损失函数对数值变化 (d)智能体 3 策略损失函数对数值变化

(e)智能体 4 策略损失函数对数值变化 (f)智能体 5 策略损失函数对数值变化 (g)智能体 6 策略损失函数对数值变化 (h)智能体 7 策略损失函数对数值变化

图 9 各智能体策略损失函数值的变化

Fig. 9 Change of strategy loss function value of each agent

图 10 给出前 3000 局内，各智能体在每局对局中所得 收益之和的变化曲线。



(a)智能体 0 单场收益变化 (b)智能体 1 单场收益变化 (c)智能体 2 单场收益变化 (d)智能体 3 单场收益变化

(e)智能体 4 单场收益变化 (f)智能体 5 单场收益变化 (g)智能体 6 单场收益变化 (h)智能体 7 单场收益变化

图 10 各智能体单局收益的变化

Fig. 10 Changes of the single episodereward of each agent

可以发现，前 500 局对局中，蓝方智能体的单场收益逐渐降低至最小值，而后稳定；而此时红方智能体的单场收益则逐渐增至最大值，而后稳定。说明 500 局左右时智能体的基本行动策略就已经确定，之后就是不断地“强化”策略，收敛网络参数。而造成随着训练的进行胜利逐渐一边倒的原因在于红方距离目标点更近，当红蓝双方同时按最优策略分配目标时，红方必定能够将对方拦截，正如图 7 所示。

结束语 本文从实际需求出发，构建多无人机空中智能拦截作战的作战想定。根据假想作战环境和作战力量，基于 Pyglet 搭建模拟作战平台；根据作战任务特点和未来应用需求选择 MADDPG 多智能体训练算法，设计深度神经网络结构；根据具体作战任务，设计既满足任务需要，又能提高训练效率的奖励函数；基于 python 建立智能体。在 python 中，智能体与模拟作战平台对接，根据经验设置训练参数，实现智能体的训练，直至智能体损失函数逐渐下降收敛。对应红方拦截成功率逐渐上升，蓝方在突防成功率下降的同时，目标选择更加稳定。实验总体达到了预期的训练效果，证明所搭建的训练系统实用性高，训练算法的选择、神经网络的构造和奖励函数的设计合理有效，采用的训练方法和训练参数也具有一定的参考性，为未来实战的应用提供了技术支撑，也为更广泛作战场景、更复杂作战任务的智能决策提供了理论基础和实验参考。

参 考 文 献

[1] ZHENG Q, WU H, LIANG R P, et al. Intelligent warfare and its demand for intelligent command and control technology[J]. Fire Control & Command Control, 2022, 47(2): 1-6.

[2] JIANG G S, HAN Z Q, WANG F. Analysis of the application status and development prospect of foreign military artificial intelligence [C]// Unmanned Systems Summit Forum 2021 (USS 2021). 2021.

[3] SUN Q, ZHANG B C. Russia: Prioritizing AI research and development[J]. Prosecutorial View, 2021(24): 56-57.

[4] FU X, YE Y K, ZHANG P, et al. Research on Characteristics of air Combat capability for military intelligence[J]. Winged Missiles Journal, 2021(9): 73-79.

[5] WANG C, LI S, JIANG H B, et al. Study on intelligent battlefield Situation Estimation of air defense and missile defense[J]. Fire Control & Command Control, 2020, 45(3): 7-13.

[6] LIN X Y, XING Q H, LIU F X. Research on Optimization of Combat Force for Key Air Defense Model[J]. Systems Engineering and Electronics, 2022, 44(3): 921-928.

[7] FANG X, ZENG B, SONG X X, et al. Warship air threat behavior modeling based on deep reinforcement learning[J]. Modern Defense Technology, 2020, 48(5): 59-66.

[8] LIU Y J, ZHANG Y. Intelligent Thinking of Naval Gun Weapons[J]. Armory Automation, 2022, 41(3): 21-24.

[9] LI G Y, KUANG S Y, JIANG G, et al. Discussion on the development path of intelligent Electronic warfare equipment [J]. Journal of China Academy of Electronics. 2022, 17(1): 7-11.

[10] HUANG W, HE X Z, WANG B X. Application of artificial intelligence technology in Army electronic countermeasures equipment[J]. Defense Science and Technology, 2022, 43(1): 26-31.

[11] ZHAO W, YE J, WANG B. Intelligent Command Decision and control based on artificial intelligence[J]. Information Security and Communication Confidentiality, 2022(2): 2-8.

[12] DING Z L, LIU G L, XIE Y, et al. Dynamic target assignment algorithm based on reinforcement learning and Neural network [J]. Electronic Design Engineering, 2020, 28(13): 54-60.

[13] DONG K S, HU W B, SHEN Y M, et al. Intelligent Development of Unmanned Aerial Combat Equipment of American Army and its Enlightenment [J]. Modern Defense Technology, 2022, 50(4): 28-37.

[14] ZHENG K Y. Research on UAV Track Planning Algorithm Based on Intelligent Cognition[D]. Harbin: Harbin Engineering University, 2021.

[15] LI B, YUE K Q, GAN Z G, et al. Multi-uav cooperative mission Decision making based on MADDPG[J]. Journal of Astronautics, 2021, 42(6): 757-765.

[16] GALAN J, CARRASCO R, LATORRE A. Military Applications of Machine Learning: A Bibliometric Perspective[J]. Mathematics, 2022, 10(9): 1397.

[17] SHARMA P, SARMA K K, MASTORAKIS N E. Artificial Intelligence Aided Electronic Warfare Systems-Recent Trends and Evolving Applications[J]. IEEE Access, 2020, 8(99): 1.

[18] LEI L. Automatic driving technology using artificial intelligence [J]. Agro Food Industry Hi Tech, 2017, 28(1): 570-574.

[19] HODICK J, PROCHÁZKA D, BAXA F, et al. Computer Assisted Wargame for Military Capability-Based Planning[J]. Entropy, 2020, 22(8): 861.

[20] LOWE R, WU Y, TAMAR A, et al. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments[J]. arXiv: 1706.02275, 2017.

[21] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [J]. arXiv: 1509.02971, 2015.

[22] DEGRIS T, WHITE M, SUTTON R S. Off-Policy Actor-Critic [J]. arXiv: 1205.4839, 2012.

[23] SUTTON R S, MCALLESTER D, SINGH S, et al. Policy Gradient Methods for Reinforcement Learning with Function Approximation[J]. Submitted to Advances in Neural Information Processing Systems, 1999, 12.

[24] VOLODYMYR M, KORAY K, DAVID S, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[25] LIN X, XING Q, LIU F. Choice of discount rate in reinforcement learning with long-delay rewards[J]. Systems Engineering and Electronics, 2022, 33(2): 12.

LIN Xiangyang, born in 1994, Ph.D candidate. His main research interests include military systems operations research optimization and reinforcement learning.