DOI:10.16356/j.1005-2615.2025.05.004

基于强化学习的多机协同超视距空战决策算法

王志刚[1], 龚华军[1], 尹 逸[2], 刘小雄[2]
(1.南京航空航天大学自动化学院, 南京 211106; 2.西北工业大学自动化学院, 西安 710072)

摘要: 现代战争中的空战态势复杂多变,因此探索一种快速有效的决策方法十分重要。本文对多架无人机协同对抗问题展开研究,提出一种基于长短期记忆(Long and short-term memory, LSTM)和多智能体深度确定策略梯度(Multi-agent deep deterministic policy gradient,MADDPG)的多机协同超视距空战决策算法。首先,建立无人机运动模型、雷达探测区模型和导弹攻击区模型。然后,提出了多机协同超视距空战决策算法。设计了集中式训练LSTM-MADDPG分布式执行架构和协同空战系统的状态空间来处理多架无人机之间的同步决策问题;设计了学习率衰减机制来提升网络的收敛速度和稳定性;利用LSTM网络改进了网络结构,增强了网络对战术特征的提取能力;利用基于衰减因子的奖励函数机制加强无人机的协同对抗能力。仿真结果表明所提出的多机协同超视距空战决策算法使无人机具备了协同攻防的能力,同时算法具备良好的稳定性和收敛性。

关键词: 协同空战决策;多智能体强化学习;混合奖励函数;长短期记忆网络
中图分类号: V249.1 **文献标志码:** A **文章编号:** 1005-2615(2025)05-0831-11

Multi-aircraft Collaborative Beyond-Visual-Range Air Combat Decision-Making Algorithm Based on Reinforcement Learning

WANG Zhigang[1], *GONG Huajun*[1], *YIN Yi*[2], *LIU Xiaoxiong*[2]
(1. Collage of Automation Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China;
2. School of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: As the modern air combat environment grows increasingly complex and dynamic, the need for rapid and effective decision-making methods has become urgent. This paper proposes a multi-aircraft cooperative beyond-visual-range air combat decision-making algorithm based on long and short-term memory (LSTM) and multi-agent deep deterministic policy gradient (MADDPG) to address the challenge of collaborative confrontation of multiple unmanned aerial vehicles (UAVs). First, a beyond-visual-range air combat environment is established, including the UAV movement model, the radar detection zone model, and the missile attack zone model. Second, the multi-aircraft collaborative beyond-visual-range air combat decision-making algorithm is proposed. This algorithm includes a centralized-training distributed-execution framework and a state space of the collaborative air combat system to handle synchronous decision-making across multiple UAVs, a learning rate decay mechanism to enhance network convergence speed and stability, an improved network based on LSTM to strengthen tactical feature extraction, and a decay-factor-based reward function to improve cooperative confrontation performance. Experimental results demonstrate that the proposed algorithm equips UAVs with effective collaborative attacking and defensive capabilities, while exhibiting strong stability and convergence.

收稿日期: 2024-05-25;**修订日期:** 2024-11-26;**网络出版时间:** 2025-07-16
通信作者: 王志刚,男,研究员,E-mail:zgwang@nuaa.edu.cn。
网络出版地址: link.cnki.net/urlid/32.1429.V.20250715.1510.004

引用格式: 王志刚,龚华军,尹逸,等. 基于强化学习的多机协同超视距空战决策算法[J]. 南京航空航天大学学报(自然科学版),2025,57(5):831-841. WANG Zhigang, GONG Huajun, YIN Yi, et al. Multi-aircraft collaborative beyond-visual-range air combat decision-making algorithm based on reinforcement learning[J]. Journal of Nanjing University of Aeronautics & Astronautics(Natural Science Edition),2025, 57(5):831-841.

Key words: cooperative air combat decision making; multi-intelligence reinforcement learning; hybrid reward functions; long and short-term memory (LSTM) networks

无人机具备无人员伤亡、成本低、持续作战能力强等特点。随着人工智能技术的蓬勃发展,无人机参与的未来空战必将进入智能化时代[1]。多架高性能无人机组成的战术编队体系在空战数据链的加持下,相互补充、相互协调,可进行协同超视距空战。超视距空战是指在空中作战中,对抗双方在目视距离(一般为8 km)之外探测到对方位置,使用导弹进行攻击的战斗。其特点为作战距离远、多目标攻击、体系对抗、攻防一体;超视距空战与近距空战的主要区别在于作战距离和作战方式。随着空战环境的复杂多变,飞行员需要处理大量信息,难以及时感知态势变化,对飞机进行飞行决策,并在飞行过程中根据战场环境做出实时调整。因此,空战智能决策技术成为当前研究的重要问题[2]。

超视距空战智能决策是指在飞行员目视距离外,使用人工智能辅助或替代飞行员进行空战机动及载荷调度决策的技术[3]。该技术始于20世纪60年代,以美国国家航空航天局(National Aeronautics and Space Administration, NASA)兰利研究中心资助的自适应机动逻辑(Adaptive maneuvering logic, AML)[4-5]为代表,其核心为由空战专家总结的空战规则组成的知识驱动型专家系统。20世纪90年代,NASA兰利研究中心在AML的基础上,进一步开发了PALADIN系统[6-7]。相比AML,PALADIN的最大创新点在于其规则是基于空战仿真自动生成。2016年,基于美国空军研究实验室(Air Force Research Laboratory, AFRL)的AFSIM仿真系统开发的"阿尔法空战"系统[8]战胜了美国退役空军上校基恩·李,与它原理相似的还有波音公司开发的双边对抗学习系统。这两个系统都是首先基于人类经验设计策略结构,然后基于对抗博弈实现参数演进,只是后者的环境适应性更强。2010年之后,基于机器学习的空战决策技术逐渐得到了发展,其中一项广为人知的项目是美国国防高级研究计划局主导的"阿尔法狗斗"智能近距空中格斗项目。该项目中,苍鹭系统公司的智能决策系统最终以5:0的优势大胜F-16飞行教官[9]。

空战智能决策技术已经发展到了基于强化学习的算法的阶段。通过强化学习,无人机可以在仿真环境中不断调整空战策略,模拟演练空战中可能遇到的各种情况,形成空战战术。针对空空导弹对空战系统状态多维复杂的问题,张强等[10]设计了基于Q网络的强化学习和与导弹攻击区相关的奖励函数,形成了一套超视距空战决策系统。为解决在复杂环境中,飞机进行自主态势决策的问题,Li等[11]应用改进的深度强化学习算法进行空战任务决策。Sun等[12]则通过多智能体强化学习方法训练出了超越专家水平的超视距空战智能决策系统。强化学习算法为超视距空战决策问题提出了一种新的解决方案。但是,目前基于强化学习的超视距空战决策方法多聚焦1V1空战决策,对于多机协同空战决策算法研究较少,且现有协同超视距空战决策算法训练效率较低,花费成本较高[13-15]。

针对上述问题,本文提出了一种基于多智能体强化学习的多机协同超视距空战算法。针对多架无人机同步决策的问题,设计了集中式训练、分布式执行架构和协同空战系统的状态空间。利用学习率衰减机制提升神经网络的收敛性能。针对敌方战术特征提取的问题,采用长短期记忆(Long and short-term memory, LSTM)网络处理具有时序特征的空战数据,提取空战战术特征。利用基于衰减因子的奖励函数机制来加强无人机的协同对抗能力。最后,通过仿真分析验证了本文所提算法的有效性。

1 超视距空战环境建模

根据多架飞机协同超视距空战决策要求,本文设计了强化学习总体架构。该框架由智能体模型、无人机运动控制模型、空战环境、空战态势感知、态势评估等组成。

智能体决策部分利用强化学习机制建立了由战场环境信息到无人机控制量的映射,使得无人机可以根据战场实时态势调整飞行轨迹,完成作战任务。无人机运动控制模型执行智能体决策给出的无人机运动控制量,完成无人机的状态更新。敌我双方的无人机、各自的机载火控雷达探测区和空空导弹攻击区共同构成了战场环境。态势感知和评估部分能够感知战场环境的变化,根据环境状态信息,利用奖励函数对无人机执行的控制量进行评价。

1.1 无人机运动控制模型

为了准确描述无人机的飞行特征,在航迹坐标系下建立无人机的模型

$$\begin{cases} \mathrm{d}v/\mathrm{d}t = g(n_x - \sin\theta) \\ \mathrm{d}\psi/\mathrm{d}t = gn_z \sin\varphi/(v\cos\theta) \\ \mathrm{d}\theta/\mathrm{d}t = (g/v)(n_z\cos\varphi - \cos\theta) \\ \mathrm{d}x/\mathrm{d}t = v\cos\theta\cos\psi \\ \mathrm{d}y/\mathrm{d}t = v\cos\theta\sin\psi \\ \mathrm{d}z/\mathrm{d}t = v\sin\theta \end{cases} \tag{1}$$

式(1)包含6个一阶微分方程，定义了无人机的速度v、航迹方位角ψ、航迹倾斜角θ和三轴位置(x,y,z)。本文将切向过载n_x、法向过载n_z和航迹滚转角φ组合为三元组$[n_x,n_z,\varphi]$，作为智能体决策模块的输出，用于控制无人机完成一系列的飞行动作。

根据载机和目标的速度与位置信息可以得到载机和目标的相对关系，两者的相对关系可以为无人机的机载火控雷达探测区和导弹攻击区提供判断条件。

$$\begin{cases} \boldsymbol{v}_{\mathrm{self}} = [\cos\psi_{\mathrm{self}}\cos\theta_{\mathrm{self}},\ \sin\psi_{\mathrm{self}}\cos\theta_{\mathrm{self}},\ \sin\theta_{\mathrm{self}}] \\ \boldsymbol{v}_{\mathrm{relative}} = \\ [\cos\psi_{\mathrm{relative}}\cos\theta_{\mathrm{relative}},\ \sin\psi_{\mathrm{relative}}\cos\theta_{\mathrm{relative}},\ \sin\theta_{\mathrm{relative}}] \\ \boldsymbol{d} = [x_{\mathrm{self}} - x_{\mathrm{relative}},\ y_{\mathrm{self}} - y_{\mathrm{relative}},\ z_{\mathrm{self}} - z_{\mathrm{relative}}] \\ d_{\mathrm{d}} = |\boldsymbol{d}| \\ A = \arccos \boldsymbol{v}_{\mathrm{relative}} \cdot \boldsymbol{d}/d_{\mathrm{d}} \\ A_{\mathrm{T}} = \arccos \boldsymbol{v}_{\mathrm{self}} \cdot \boldsymbol{d}/d_{\mathrm{d}} \\ \beta = \arccos(\boldsymbol{v}_{\mathrm{self}} \cdot \boldsymbol{v}_{\mathrm{relative}}) \\ \Delta h = z_{\mathrm{self}} - z_{\mathrm{relative}} \end{cases} \tag{2}$$

式中：$\boldsymbol{v}_{\mathrm{self}}$为载机的速度方向；$\boldsymbol{v}_{\mathrm{relative}}$为目标的速度方向；$\theta_{\mathrm{self}}$和$\psi_{\mathrm{self}}$为载机的航迹倾斜角和航迹方位角；$\theta_{\mathrm{relative}}$和$\psi_{\mathrm{relative}}$为目标的航迹倾斜角和航迹方位角；$x_{\mathrm{self}}$、$y_{\mathrm{self}}$和$z_{\mathrm{self}}$为载机的三轴位置；$x_{\mathrm{relative}}$、$y_{\mathrm{relative}}$和$z_{\mathrm{relative}}$为目标的三轴位置；$\boldsymbol{d}$为载机到目标的距离矢量，$d_{\mathrm{d}}$为载机与目标的距离；$A$为脱离角，即目标的飞行速度与敌我双方距离矢量的夹角；A_{T}为偏离角，即载机的飞行速度与敌我双方距离矢量的夹角；β为两机速度矢量的夹角；Δh为载机和目标的高度差。通过式(2)可以计算得到如图1所示的相对关系。

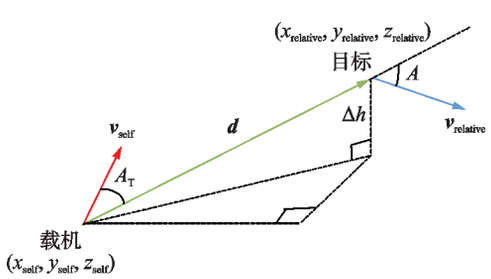


图1　载机和目标的相对关系图

Fig.1　Relative relationship between the carrier and the target

1.2　机载火控雷达探测区模型

本文根据雷达的属性和特征建立了机载火控雷达的探测区。机载火控雷达在搜索、确认、跟踪3个工作阶段的转换关系、每个阶段的主要功能和雷达建模如下。

(1) 空域搜索

在无人机进入空战战场后，进入空域搜索阶段，向指定空域内发射周期性的电磁波进行探测。当电磁波遇到目标时会反射回来，若无人机接收到电磁波回波，则可以确认目标的信息。能否探测到目标与当前我机的航迹姿态、敌我双方的高度和偏离角有关。假设红方为我方，蓝方为敌方，建立雷达探测区

$$\begin{cases} \Delta h_{\min} \leqslant |\Delta h| \leqslant \Delta h_{\max} \\ d_t \leqslant d_{\mathrm{rmax}} \\ -\theta_{\mathrm{rmax}} \leqslant A_{\mathrm{T}} \leqslant \theta_{\mathrm{rmax}} \\ -\psi_{\mathrm{rmax}} \leqslant A_{\mathrm{T}} \leqslant \psi_{\mathrm{rmax}} \end{cases} \tag{3}$$

式中：$\Delta h_{\min}$和$\Delta h_{\max}$为机载火控雷达(Airborne fire-control radar，AFR)的最小搜索高度差和最大搜索高度差；Δh为敌我双方的高度差，$\Delta h = h_t^{\mathrm{r}} - h_t^{\mathrm{b}}$，$h_t^{\mathrm{r}}$和$h_t^{\mathrm{b}}$为$t$时刻红方和蓝方的高度；$d_t$为$t$时刻红方和蓝方的距离；$d_{\mathrm{rmax}}$为AFR的最大探测距离；$\theta_{\mathrm{rmax}}$和$\psi_{\mathrm{rmax}}$为AFR的最大搜索俯仰角和最大搜索偏航角。

(2) 确认目标

当目标满足式(3)的4个条件时，则符合了被我方发现的先决条件，我方雷达开始确认目标状态。若目标被确认，则其雷达告警系统会报警，表明自身被我方雷达发现，目标越靠近雷达准线，越容易被我方雷达发现；敌我双方的距离越近，越易被我方雷达发现。具体关系为

$$P_{\mathrm{r}} = \left(1 - \frac{A_{\mathrm{T}}}{\theta_{\mathrm{rmax}}}\right) \cdot \left(1 - \frac{A_{\mathrm{T}}}{\psi_{\mathrm{rmax}}}\right) \cdot \mathrm{e}^{-\sigma\frac{d_t}{d_{\mathrm{rmax}}}} \tag{4}$$

式中：σ为与AFR的散射截面相关的参数，当散射截面积为5 m^{2}时，σ为0.162 5。若$P_{\mathrm{r}} > 0.2$，则成功确认目标。若未能成功确认目标，雷达将重新进入空域搜索阶段；若成功确认目标，雷达将转入跟踪目标阶段。在探测并确认目标后，雷达进入跟踪状态，此时机载火控系统会计算导弹攻击区和相应的击毁概率，一旦满足打击条件立即发射导弹击毁目标。

(3) 跟踪阶段

在探测并确认目标后，雷达会进入跟踪状态，此时机载火控系统会计算导弹攻击区和相应的击毁概率，一旦满足打击条件立即发射空空导弹击毁目标。

1.3 导弹攻击区模型

考虑到导弹性能对无人机作战能力的限制,使用衡量中远程空空导弹的战斗能力的6个要素,分别为导弹的最大离轴发射角φ_{mmax}、最大攻击距离d_{mmax}、最小攻击距离d_{mmin}、不可逃逸的圆锥角φ_{memax}、不可逃逸的最大距离d_{memax}和不可逃逸的最小距离d_{memin}来完成攻击区的建模,具体划分为

$$R=\begin{cases}R_{\mathrm{attack}} & d_{\mathrm{mmin}}\leqslant d_t\leqslant d_{\mathrm{mmax}},\ A_{\mathrm{T}}\leqslant\varphi_{\mathrm{mmax}}\\ R_{\mathrm{noescape}} & d_{\mathrm{memin}}\leqslant d_t\leqslant d_{\mathrm{memax}},\ A_{\mathrm{T}}\leqslant\varphi_{\mathrm{memax}}\end{cases}\tag{5}$$

当目标进入导弹攻击区后,根据所处位置不同,被击毁的概率也不同。无人机具备了通过一些连续机动规避导弹攻击的能力,所以在计算导弹击毁概率时,需要考虑目标此时能否通过连续机动规避导弹。通过载机的偏离角A_{T}和目标的脱离角A来描述空战双方的规避优势从而定量分析导弹的击毁概率:A_{T}越小,载机的导弹离轴发射角越小,导弹更容易命中目标;而A越小,则目标的飞行方向越接近雷达准线和攻击区中线,越难通过连续机动躲避雷达跟踪和导弹攻击。结合无人机的规避优势,将攻击区进一步划分为5个部分,每部分的毁伤概率不同。

目标的毁伤概率为

$$P_{\mathrm{ad}}=\tau_{\mathrm{a}}\cdot P_{\mathrm{a}}+\tau_{\mathrm{d}}\cdot P_{\mathrm{d}}\tag{6}$$

式中:P_{a}为与目标规避优势相关的毁伤概率;P_{d}为与敌我距离相关的毁伤概率;τ_{a}和τ_{d}分别为规避优势和距离毁伤概率的权重因子,具体根据双方偏离角、脱离角和距离决定。

$$P_{\mathrm{a}}=\begin{cases}A/\pi+1 & \mathrm{position(aircraft\_aim)}\in 1\\ -(A/(\pi/6+A_{\mathrm{T}}))\times A+0.5 & \\ \qquad \mathrm{position(aircraft\_aim)}\in 2,\ \arctan v_y/v_x\geqslant 0,\ A<5\pi/6-A_{\mathrm{T}} & \\ -(0.5/(5\pi/6-A_{\mathrm{T}}))\times A-0.5\times\pi/6+A_{\mathrm{T}}/(5\pi/6-A_{\mathrm{T}}) & \\ \qquad \mathrm{position(aircraft\_aim)}\in 2,\ \arctan v_y/v_x\geqslant 0,\ A>=5\pi/6-A_{\mathrm{T}} & \\ (0.3/(\pi/6+A_{\mathrm{T}})\times A+0.5 & \\ \qquad \mathrm{position(aircraft\_aim)}\in 2,\ \arctan v_y/v_x<0,\ A<5\pi/6-A_{\mathrm{T}} & \\ (0.3/(A_{\mathrm{T}}-5\pi/6))\times A+(0.5\times A_{\mathrm{T}}-43/60)/(A_{\mathrm{T}}-5\pi/6) & \\ \qquad \mathrm{position(aircraft\_aim)}\in 2,\ \arctan v_y/v_x<0,\ A\geqslant 5\pi/6-A_{\mathrm{T}} & \\ (0.3/(\pi/6+A_{\mathrm{T}}))\times A+0.5 & \\ \qquad \mathrm{position(aircraft\_aim)}\in 3,\ \arctan v_y/v_x\geqslant 0,\ A<5\pi/6-A_{\mathrm{T}} & \\ (0.3/(A_{\mathrm{T}}-5\pi/6))\times A+(0.5\times A_{\mathrm{T}}-43/60)/(A_{\mathrm{T}}-5\pi/6) & \\ \qquad \mathrm{position(aircraft\_aim)}\in 3,\ \arctan v_y/v_x\geqslant 0,\ A>=5\pi/6-A_{\mathrm{T}} & \\ -(A/(\pi/6+A_{\mathrm{T}}))\times A+0.5 & \\ \qquad \mathrm{position(aircraft\_aim)}\in 3,\ \arctan v_y/v_x<0,\ A<5\pi/6-A_{\mathrm{T}} & \\ -(0.5/(5\pi/6-A_{\mathrm{T}}))\times A-0.5\times(\pi/6+A_{\mathrm{T}})/(5\pi/6-A_{\mathrm{T}}) & \\ \qquad \mathrm{position(aircraft\_aim)}\in 3,\ \arctan v_y/v_x<0,\ A>=5\pi/6-A_{\mathrm{T}} & \\ A/\pi & \mathrm{position(aircraft\_aim)}\in 4\end{cases}$$

$$P_{\mathrm{d}}=1-((d_{\mathrm{t}}-d_1)/d_2)^2$$

式中:d_1和d_2为与导弹攻击区相关的距离参数;d_{t}为载机和目标的距离。

综上所述,当目标毁伤概率大于阈值时,发射导弹将目标击毁。

2 多机协同超视距空战决策算法

本文设计多智能体深度确定策略梯度(Multi-agent deep deterministic policy gradient, MADDPG)算法对无人机协同超视距空战决策进行研究,将集中式训练和分布式执行架构与超视距空战决策算法结合,以2V2超视距空战为例,设计了基于LSTM-MADDPG多机协同超视距空战决策算法。

2.1 LSTM-MADDPG算法

LSTM-MADDPG算法由多个LSTM-DDPG网络组成,算法的特点如下。

(1)本文采用集中训练和分布执行的策略。应用集中学习方式训练Critic网络;执行动作时,每个个体用独立的Actor网络选择动作。Critic用于全局观测,Actor网络用于局部观测。

(2)改进经验回放记录数据。为了让网络适用环境,训练中的信息由$(\boldsymbol{x},x',a_1,a_2,\cdots,a_n,r_1,r_2,\cdots,r_n)$组成,$\boldsymbol{x}=(o_1,o_2,\cdots,o_n)$表示每个智能体的观测信息。

（3）利用所有策略的整体效果对网络进行优化，以提高算法的稳定性和收敛速度。

集中式训练分布式执行架构会对LSTM-DDPG算法进行改进，每一个智能体都会对其他智能体的策略进行函数逼近。将智能体策略网络参数设置为$\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_n)$，智能体的策略为$\boldsymbol{\pi}=(\pi_1,\pi_2,\cdots,\pi_n)$，第$n$个智能体的累积期望奖励为

$$J(\theta_n)=E_{s\sim\rho^{\pi},a_n\sim\pi_{\theta_n}}\sum_{t=0}^{\infty}\gamma^t r_{n,t} \tag{7}$$

针对DDPG算法的确定性策略μ_{θ_n}，对累积期望奖励求梯度

$$\nabla_{\theta_i}J(\mu_i)=$$
$$E_{x,a\sim D}[\nabla_{\theta_i}\mu_i(a_i|o_i)\nabla_{a_i}Q_i^{\mu}(\boldsymbol{x},a_1,a_2,\cdots,a_n)|_{a_i=\mu_i(o_i)}] \tag{8}$$

式中：$Q_i^{\mu}(\boldsymbol{x},a_1,a_2,\cdots,a_n)$表示第$i$个智能体的状态-动作函数；$D$表示网络训练的经验池，经验池中存放训练数据$(\boldsymbol{x},x',a_1,a_2,\cdots,a_n,r_1,r_2,\cdots,r_n)$。采用集中方式的Critic网络其更新计算方式

$$\begin{cases}L(\theta_i)=E_{x,a,r,x'}[(Q_i^{\mu}(\boldsymbol{x},a_1,a_2,\cdots,a_n)-y)^2]\\ y_i=r_i+\gamma\bar{Q}_i^{\mu'}(x',a_1',a_2',\cdots,a_n')|_{a_j'=\mu_j'(o_j)}\end{cases} \tag{9}$$

式中：$\bar{Q}_i^{\mu'}$表示目标网络；$\boldsymbol{\mu}'=[\mu_1',\mu_2',\cdots,\mu_n']$为目标策略具有滞后更新的参数$\theta_j'$。

在算法设计中Critic网络采用全局信息进行网络学习，Actor网络采用局部观测量进行网络学习。代价函数为

$$\begin{cases}L(\phi_i^j)=-E_{o_j,a_j}[\log\hat{\mu}_{\phi_i^j}(a_j|o_j)+\lambda H(\hat{\mu}_{\phi_i^j})]\\ L(\theta_i)=E_{x,a,r,x'}[(Q_i^{\mu}(\boldsymbol{x},a_1,a_2,\cdots,a_n)-y)^2]\\ y_i=r_i+\gamma\bar{Q}_i^{\mu'}(x',a_1',a_2',\cdots,a_n')|_{a_j'=\mu_j'(o_j)}\end{cases} \tag{10}$$

只要最小化代价函数，就能得到其他智能体策略的逼近，因此式(10)的y可以替换为

$$\begin{cases}L(\theta_i)=E_{x,a,r,x'}[(Q_i^{\mu}(\boldsymbol{x},a_1,a_2,\cdots,a_n)-y)^2]\\ y_i=r_i+\gamma\bar{Q}_i^{\mu'}(x',\hat{\mu}'^1_{\phi_i'}(o_1),\hat{\mu}'^1_{\phi_i'}(o_2),\cdots,\hat{\mu}'^{n}_{\phi_i'}(o_n))\end{cases} \tag{11}$$

针对强策略很难去适应新的对手策略问题，LSTM-MADDPG应用了一种策略集合的思想，即第i个智能体的策略μ_i由1个具有k个子策略的集合构成，在每一个训练episode中只用1个子策略（简写为$\mu_i^{(k)}$）。对每一个智能体，最大化其策略集合的整体奖励为

$$J_e(\mu_i)=E_{k\sim\mathrm{unif}(1,K),s\sim\rho^{\mu},a\sim\mu_i^{(k)}}\sum_{t=0}^{\infty}\gamma^t r_{i,t} \tag{12}$$

可以为每一个子策略构建1个记忆存储单元$D_i^{(k)}$，去优化策略集合的整体效果，因此对每一个子策略的更新梯度为

$$\nabla_{\theta_i^{(k)}}J_e(\mu_i)=$$
$$\frac{1}{K}E_{[x,a]_K\sim D_i^{(k)}}[\nabla_{\theta_i^{(k)}}\mu_i^{(k)}(a_i|o_i)\nabla_{a_i}Q^{\mu_i}(\boldsymbol{x},a_1,a_2,\cdots,a_n)]|_{a_i=\mu_i^{(k)}(o_i)} \tag{13}$$

计算时序差分误差

$$\delta_t^i=Q(S_t,[\boldsymbol{\mu}(o_t^1|\theta_1^{\mu}),\boldsymbol{\mu}(o_t^2|\theta_2^{\mu}),\boldsymbol{\mu}(o_t^3|\theta_3^{\mu}),\boldsymbol{\mu}(o_t^4|\theta_4^{\mu})]|w_i^Q)-\hat{y}_t^i \tag{14}$$

然后通过梯度下降算法更新参数w_i^Q使得价值网络的预测值更接近TD目标值

$$w_i^Q=w_i^Q-\alpha\cdot\delta_t^i\cdot\nabla_{w_i^Q}Q(S_t,[\boldsymbol{\mu}(o_t^1|\theta_1^{\mu}),\boldsymbol{\mu}(o_t^2|\theta_2^{\mu}),\boldsymbol{\mu}(o_t^3|\theta_3^{\mu}),\boldsymbol{\mu}(o_t^4|\theta_4^{\mu})]|w_i^Q) \tag{15}$$

最后可以通过软更新的方法更新每一架无人机的目标策略网络和目标价值网络

$$\begin{cases}\theta_i^{\mu'}=\tau\theta_i^{\mu}+(1-\tau)\theta_i^{\mu'}\\ w_i^{Q'}=\tau w_i^Q+(1-\tau)w_i^{Q'}\end{cases} \tag{16}$$

2.2　协同空战系统的状态空间设计

由于协同空战中智能体大幅增加，根据多机协同超视距空战和集中式训练分布式执行架构的特点，设计了协同空战系统的全局状态量S和私有观测状态量o_i。以2V2协同超视距空战为例，智能体有4个，分别为本机、友机、敌1号机和敌2号机，其全局状态量S如表1所示。

表1　全局状态量设计

Table 1　Global state design

状态分量	变量	描述	维数
本机	state_self	本机信息，包括状态、速度、高度、航迹俯仰角和航迹偏航角	5
	state_relative	与友机和敌机的相对信息，包括存活状态、脱离角、偏离角、相对距离和两机速度矢量的夹角	5×3
	state_radar	雷达对两架敌机的状态和雷达告警器对两架敌机的状态	4
	state_missile	导弹对两架敌机的状态和导弹告警器对两架敌机的状态	4
友机		与本机类似，但是不含state_relative信息	13
敌1号机		与本机类似，但是不含state_relative信息	13
敌2号机		与本机类似，但是不含state_relative信息	13

在集中式训练、分布式执行框架中,每一个智能体无人机都有自己的价值网络,价值网络输入即全局状态量S均按本机、友机、敌1号机和敌2号机的排序。私有观测状态量o_i也是由本机、友机、敌1号机和敌2号机的顺序组成,因为友机之间可以通过数据链进行通信,所以私有观测状态量o_i中可以包含全部的友机信息。在实际空战中,敌机的雷达和导弹攻击区状态均可以通过本机的雷达告警器和导弹告警器得到,所以在私有观测状态量o_i中敌1号机和敌2号机仅保留自身的状态信息。最后,私有观测状态量o_i总共51维。

2.3 学习率衰减机制

对于学习率的设置,本文希望在训练前期通过较大的学习率来加快网络收敛速度,在训练后期则通过较小的学习率来找到全局最优点。为实现这一目标,设计了学习率衰减机制

$$l_{\text{decayed}} = l_0 \cdot \left(m^{\frac{N_{\text{episode}}}{N_{\text{decay}}}} \right) \tag{17}$$

式中:l_{decayed}为衰减后的学习率;l_0为初始设置的学习率;m为衰减比例;N_{episode}为总的训练回合;N_{decay}为衰减步数。

在超视距空战决策算法训练过程中,根据设计的学习率衰减机制,学习率可以随着训练回合而变化。在算法训练初期,学习率较大,可以让无人机的空战策略迅速向最优策略靠拢;随着训练的进行,学习率逐渐减小,降低网络梯度变化的速度,使得无人机学习的空战策略不会错过最优策略。学习率衰减机制既保证了超视距空战决策算法的训练速度,又避免了算法陷入局部最优。

2.4 混合奖励函数及其校正机制

设计的混合奖励包括状态奖励和事件奖励,并且根据多机协同超视距空战的特点,增加了回合奖励和衰减因子。

(1) 状态奖励

状态奖励的目的是使无人机能够学会安全飞行,并且引导无人机向目标靠近,产生与目标对抗的经验。所以,通过引导奖励来实现这一目的。由于在多机协同超视距空战中存在多个目标,所以设计了一个目标分配方法:先计算我方两架无人机与敌方两架无人机之间的相对距离,然后选择相对距离最短的两架无人机相互作为目标,而剩下的两架无人机则互为目标。引导奖励r_{state}为

$$r_{\text{state}} = \begin{cases} 0.01 & \Delta d > 50 \\ 0 & -50 \leqslant \Delta d \leqslant 50 \\ -0.01 & \Delta d < -50 \end{cases} \tag{18}$$

式中:Δd为一个决策周期内本机与对应目标之间相对距离的变化量,可计算为

$$\Delta d = d_{\text{last}} - d_{\text{current}} \tag{19}$$

式中:d_{last}为上一个决策时刻本机与对应目标之间相对距离;d_{current}为当前决策时刻本机与对应目标之间相对距离。

(2) 事件奖励

事件奖励通过给超视距空战中涉及的标志性事件单独设计奖励,从而引导无人机在攻击目标的同时,能够规避敌方的导弹攻击区。主要事件和对应的奖励如表2所示。

表2 事件奖励

Table 2 Event rewards

事件类型	奖励函数	事件类型	奖励函数
目标进入己方AFR搜索区	$r_{\text{event\_afr1}}$	己方进入目标AFR搜索区	$-r_{\text{event\_afr1}}$
己方AFR确认目标	$r_{\text{event\_afr2}}$	己方被目标AFR确认	$-r_{\text{event\_afr2}}$
己方AFR跟踪目标	$r_{\text{event\_afr3}}$	己方被目标AFR跟踪	$-r_{\text{event\_afr3}}$
目标进入己方导弹攻击区	$r_{\text{event\_missile1}}$	己方进入目标导弹攻击区	$-r_{\text{event\_missile1}}$
己方导弹未命中目标	$r_{\text{event\_missile2}}$	敌方导弹未命中己方	$-r_{\text{event\_missile2}}$
己方导弹命中目标	$r_{\text{event\_missile3}}$	敌方导弹命中己方	$-r_{\text{event\_missile3}}$
超出战场边界	$r_{\text{event\_battlefield}}$	飞机速度限制	$r_{\text{event\_}v}$

(3) 回合奖励

多机协同超视距空战中,需要将敌方所有无人机全部击毁才能结束本回合的空战。设计了回合奖励r_{epsiode},即在空战回合结束时,获胜的一方智能体可以获得奖励,即$r_{\text{epsiode}}=10$;而战败的一方智能体则给予惩罚,即$r_{\text{epsiode}}=-10$;在规定的时间内,空战双方未能分出胜负,双方智能体既不获得奖励也不给予惩罚,$r_{\text{epsiode}}=0$。综上所述,混合奖励函数可计算为

$$r = r_{\text{epsiode}} + \max(0, (1 - \text{epsiode}/\lambda_{\text{state}})) \cdot r_{\text{state}} + \max(0, (1 - \text{epsiode}/\lambda_{\text{event}})) \cdot r_{\text{event}} \tag{20}$$

式中:λ_{state}为状态奖励衰减因子;λ_{event}为事件奖励的衰减因子。

衰减因子的引入,可以在多机协同超视距空战算法前期引导智能体无人机快速学会安全飞行,并积攒与敌方交战的经验,加速算法的收敛。随着训练的增加,获得的奖励只剩下回合奖励,以防止算

法陷入局部最优，鼓励智能体无人机尝试更多的战术配合，使多机协同超视距空战涌现出更多的战术。

2.5　多机协同超视距空战决策算法框架

基于多机超视距空战的任务场景，本文提出基于LSTM-MADDPG的多智能体空战决策框架，该框架主要由3个部分组成：深度强化学习模块、环境模块和数据处理模块。具体结构如图2所示（以2V2情况为例）。强化学习模块主要分为两部分，中央控制器和分布式执行。中央控制器主要负责评估智能体无人机执行的动作；分布式执行主要输出红方无人机和蓝方无人机需要执行的动作。环境模块主要由无人机模型、导弹攻击区和奖励模块组成；奖励模块根据战场态势进行态势评估；空战数据处理模块的功能是处理环境模块发送的空战数据，对空战数据进行掩码处理、归一化和打包，并将其发送到共享经验池。

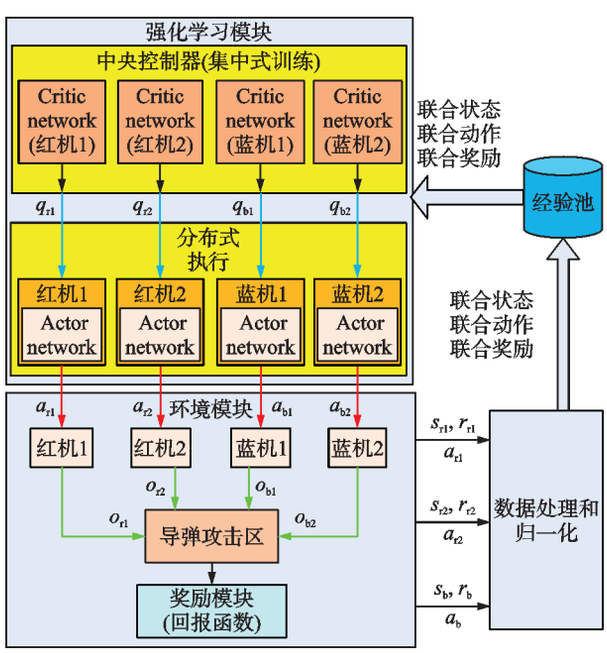


图2　多机协同超视距空战决策框架图

Fig.2　Multi-aircraft cooperative beyond-visual-range air combat decision-making framework

红方无人机和蓝方无人机根据其策略网络的初始参数选择初始动作，红蓝双方的无人机执行此动作与环境交互，以获得新的状态和奖励。此时，数据处理模块对双方的状态、动作和奖励等数据进行打包、归一化、掩码处理，将其处理成联合动作集合、联合奖励集合和联合状态集合。待共享经验池满后，强化学习模块开始对数据进行采样，并将采样的联合状态、联合动作和联合奖励发送给中央控制器，对价值网络进行更新。在价值网络更新完成之后，会对红方和蓝方执行的动作进行评价，指导策略网络完成更新。策略网络更新后，红方和蓝方飞机将自身的观测结果输入网络，策略网络输出相应的动作。无人机在收到强化学习模块的动作决策后，执行相应的动作，与空战环境进行交互，产生新一轮的状态、动作和奖励。数据处理模块将新的数据处理完成后，将其发送到共享经验池中。如此循环往复，直至满足多机协同超视距空战训练结束的要求。

3　实验结果与分析

为验证本章提出的多机协同超视距空战决策算法的有效性，在2V2超视距空战的背景下，进行LSTM-MADDPG与混合网络（Q mixing network，QMIX）以及价值分解网络（Value-decomposition network，VDN）两种多智能体强化学习算法的对比实验，验证所设计的优化机制对多机协同超视距空战决策算法训练效果的提升作用。设计2V2的超视距空战实验场景，对训练过程中涌现的战术进行分析。

3.1　仿真环境设置

多机协同超视距空战决策算法的神经网络参数如表3所示，算法训练参数如表4所示。

表3　神经网络参数设置

Table 3　Neural network parameter settings

神经网络名称	感知层细胞数量	拟合层神经网络结构	激活函数	输出层激活函数
Critic网络	512	(512,256,64,1)	ReLu	—
Actor网络	512	(512,256,64,3)	ReLu	tanh

表4　训练参数设置

Table 4　Training parameter settings

训练参数	数值	训练参数	数值
Actor网络的学习率	0.01	Critic网络的学习率	0.01
折扣率	0.95	批大小	128
软更新权重	0.01	探索初始能力因子	1
探索能力衰减因子	0.000 02	学习率衰减比例	e
学习率衰减步数	2 000	状态奖励的衰减因子	3 000
事件奖励的衰减因子	6 000		

3.2　对比实验

设计仿真2V2超视距空战，进行LSTM-MADDPG与QMIX和VDN算法的对比实验。在2V2超视距空战中有4架智能体无人机，由于每架无人机的平均奖励收敛情况基本一致，所以选取其中1架无人机的平均奖励来展示算法的收敛情况。下面将对3种算法的训练情况进行分析，如图3所示。

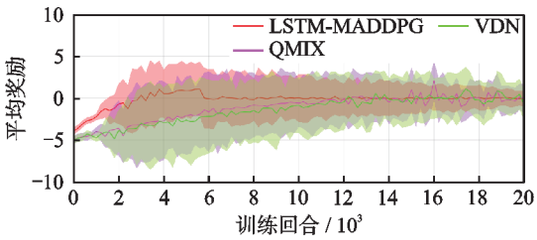


图3 多智能体强化学习算法平均奖励对比图

Fig.3 Comparison of average rewards of multi-intelligent body reinforcement learning algorithms

首先分析LSTM-MADDPG平均奖励的变化情况。在算法训练初期,作为智能体的无人机尚未学会安全飞行,会出现坠毁和飞出空战区域的情况,此时200个回合奖励的标准差还不是很大。随着训练回合的增加,无人机学会了安全飞行,所获得的平均奖励逐渐增加。而双方又在状态奖励的作用下开始接近,随着双方距离的接近,无人机触发了空战中的事件奖励,所以1 600回合后的200个回合奖励的标准差开始逐渐变大。6 000回合后,为了防止无人机陷入专家经验,状态奖励和事件奖励衰减为0,无人机获得的奖励只有回合奖励。随着训练回合的继续增加,无人机开始学会相互协同、进攻脱离或者引诱迂回等战术,平均奖励开始趋于稳定,200个回合奖励的标准差逐渐减少,LSTM-MADDPG算法逐渐收敛。

VDN算法训练收敛趋势与LSTM-MADDPG算法类似,但是VDN算法的收敛速度远不如LSTM-MADDPG算法。QMIX算法训练收敛趋势也与LSTM-MADDPG算法类似,其收敛速度和效果介于LSTM-MADDPG算法和VDN算法之间。

算法训练完成后,以2V2超视距空战为任务场景,调用VDN和QMIX分别与LSTM-MADDPG进行100次对抗,对抗结果如图4所示。

若对抗双方同时使用LSTM-MADDPG算法,训练过程中胜率变化如图5所示。

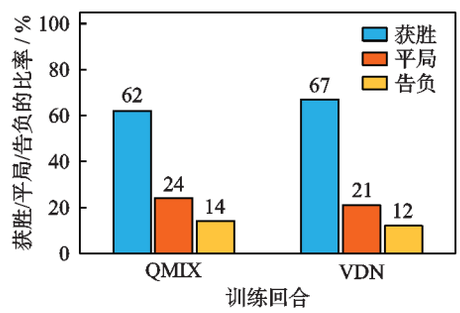


图4 LSTM-MADDPG分别与QMIX和VDN空战对抗胜率统计图

Fig.4 Statistical graph of air combat confrontation win rates between LSTM-MADDPG and QMIX or VDN

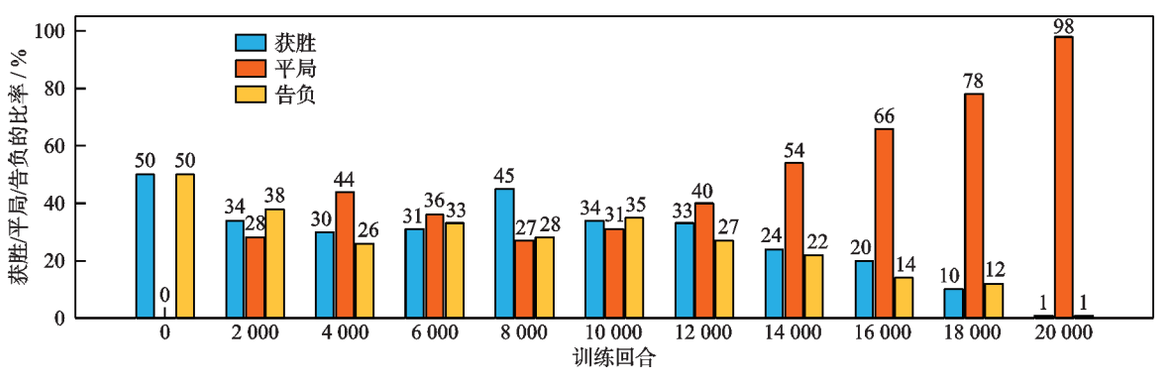


图5 采用LSTM-MADDPG算法的空战对抗胜率统计图

Fig.5 Statistical graph of air combat confrontation win rates using LSTM-MADDPG

在训练初期,红方和蓝方均没有学会安全飞行,此时的告负概率基本是超出作战区域,且红方的获胜和告负概率均为50%。随着训练的增加,红方和蓝方学会安全飞行,开始尝试对抗,所以红方有胜有负,此时告负中有被蓝方击落的情况。随着持续训练,红方和蓝方激烈对抗,平局的占比减少,获胜和告负的情况增加。在12 000回合后,红方和蓝方开始学习如何进攻、逃脱对方雷达锁定和导弹攻击区的策略,所以平局的情况开始增加,获胜和告负的情况减少。最后,红方和蓝方在规定的空战时间内,哪一方都无法消灭对方,所以平局的占比达到98%,通过仿真表明本文设计的空战决策算法满足要求。

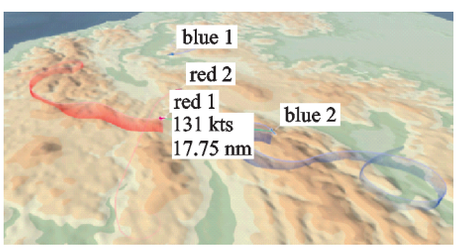
3.3 2V2空战涌现的战术分析

设计2V2超视距空战实验场景,展示多机协同超视距空战决策算法涌现的空战战术。根据红蓝双方的对抗轨迹、雷达状态和导弹攻击区状态,分析对抗双方决策行为的合理性,验证多架飞机进行协同空战决策的有效性。具体空战场景和红蓝方初始状态设置如表5所示。场景1和场景2的红蓝双方对抗三维轨迹如图6所示。

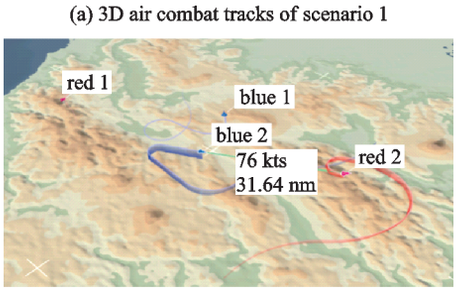
表 5　红蓝方初始状态设置

Table 5　Initial state settings of two sides

场景	阵营	无人机状态			
		v/(m·s^{-1})	ψ/(°)	θ/(°)	(x,y,z)/(km,km,km)
场景 1：红蓝方均势，双方迎头飞行	红方 1 号机	100	0	0	(−55,50,3)
	红方 2 号机	100	0	0	(−55,50,3)
	蓝方 1 号机	100	0	180	(55,50,3)
	蓝方 2 号机	100	0	180	(55,−50,3)
场景 2：红蓝方迎头飞行（不同高度）	红方 1 号机	100	0	0	(−50,75,2)
	红方 2 号机	100	0	0	(−50,−50,2)
	蓝方 1 号机	100	0	180	(50,50,3)
	蓝方 2 号机	100	0	180	(50,75,3)



(a) 3D air combat tracks of scenario 1



(b) 3D air combat tracks of scenario 2

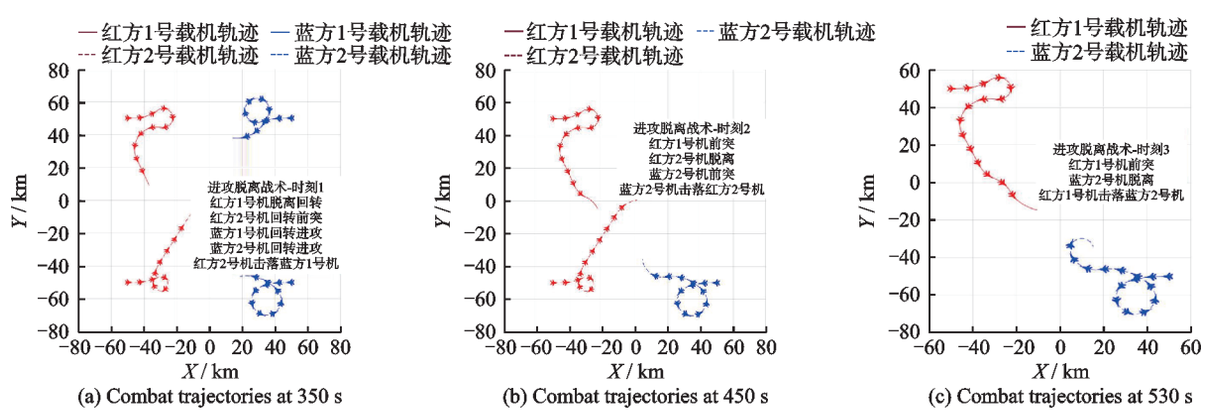
图 6　空战轨迹曲线

Fig.6　Aircraft combat trajectories

以场景 1 为例，将涌现出的进攻脱离战术分解为 3 个时刻，空战轨迹、红蓝双方的雷达和导弹攻击区状态具体如图 7 所示。

如图 7(a)所示，红方无人机和蓝方无人机根据目标分配原则，红方 1 号与蓝方 1 号互为目标，红方 2 号和蓝方 2 号互为目标。红方 1 号和蓝方 1 号相互发现后，双方的雷达均转为跟踪模态。红方 1 号与蓝方 1 号受到对方威胁选择了脱离机动，红方 1 号选择掉头躲避威胁，向红方 2 号方向靠拢，而蓝方 1 号在空中完成大转弯后，立刻回转，对红方 1 号发动二次攻击。红方 2 号和蓝方 2 号发现对方后，执行脱离战术。红方 2 号在完成躲避蓝方 2 号攻击后，立刻向蓝方 1 号快速突进，迅速跟踪目标。蓝方 1 号正在向红方 1 号发起第 2 次进攻，当蓝方 1 号受到警告来不及做出脱离机动就被红方 2 号纳入导弹攻击区内，经过红方 2 号的导弹攻击区解算，判定蓝方 1 号被击毁。与此同时蓝方 2 号机成功脱离红方 2 号机的跟踪，立刻回转，向红方 2 号机发起第 2 次进攻，由于自身规避威胁时转弯半径过大，无法有效牵制红方 2 号机。

如图 7(b)所示，蓝方 2 号机抓住红方 2 号机攻击蓝方 1 号机的时机，迅速调整攻击角度，进行攻击占位，迅速发现红方 2 号机，转入跟踪模态。而红方 2 号机立刻做出脱离机动，但是由于进攻蓝方 1 号机时太过前突，无法有效躲过蓝方的攻击，根据蓝方 2 号机的导弹攻击区解算，判定红方



(a) Combat trajectories at 350 s　(b) Combat trajectories at 450 s　(c) Combat trajectories at 530 s

图 7　进攻脱离战术

Fig.7　Offense disengagement tactics

2号机被蓝方2号机击落。与此同时,红方1号机发现友机受到威胁后,加速向前前突,调整好自身的攻击角度进行攻击占位,对蓝方2号机发起进攻。

如图7(c)所示,红方1号机在蓝方2号机进攻红方2号机时,对蓝方2号机进行攻击,随着两机的距离越来越近,红方1号机的雷达已经跟踪上蓝方2号机,蓝方2号机接收到雷达和导弹告警后,立刻做出脱离机动,掉头向后跑去。而红方调整好攻击角度后,加速向前将蓝机纳入导弹攻击区后,红方1号机根据双方态势和蓝机规避优势导弹攻击区解算杀伤概率,最后判定蓝方2号机被红方1号机击落。经过一系列对抗,最终红方取得了本轮2V2超视距空战的胜利。图8表示双机空战脱离中各架飞机的存活状态、雷达模态和导弹攻击区状态。

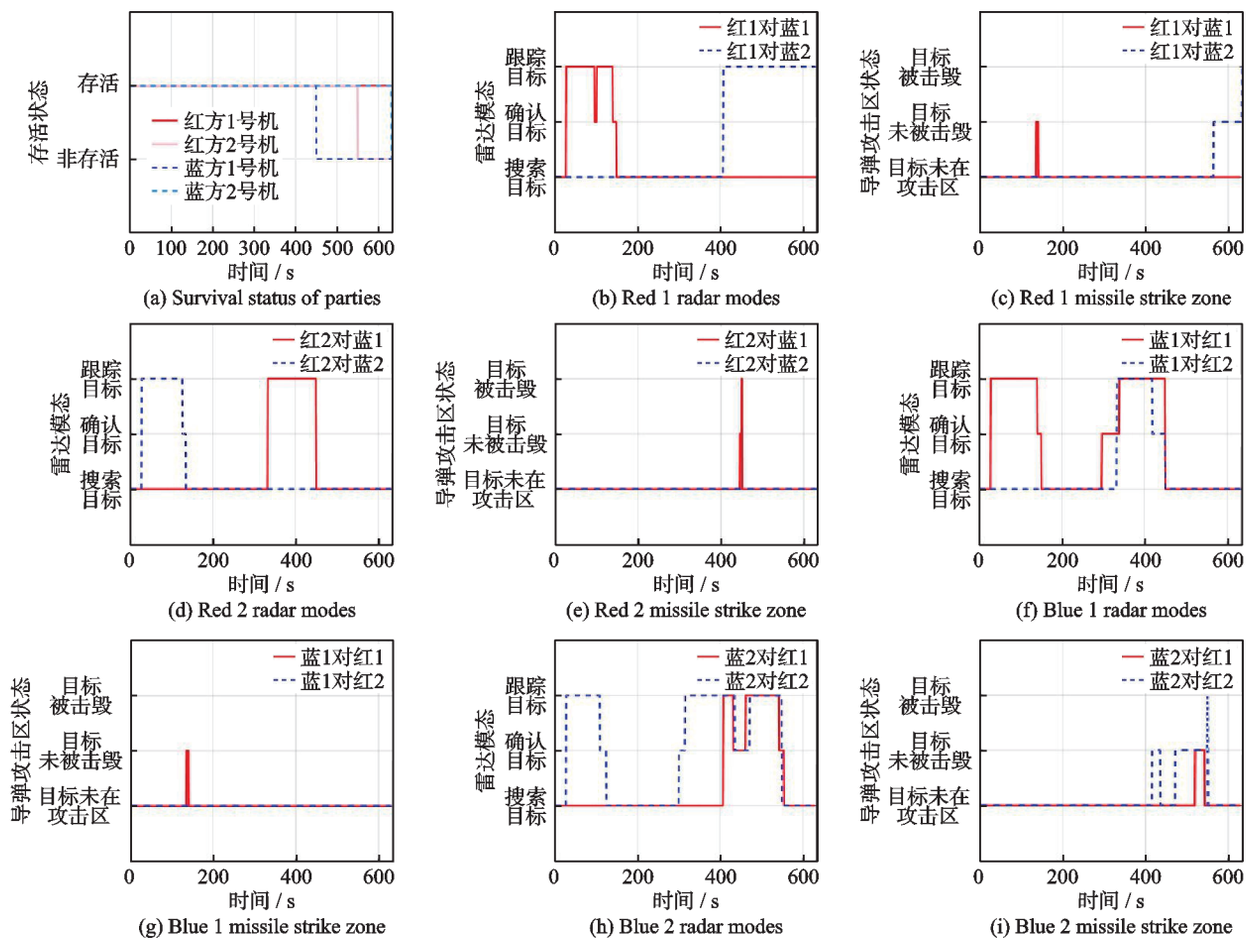


图8 2V2空战

Fig.8 2V2 air combat

4 结 论

本文针对多架飞机,研究了一种超视距协同空战决策算法。

根据超视距空战的特点,给出无人机运动控制模型;确定机载火控雷达探测区模型及其工作阶段;根据目标的规避优势、敌我双方距离和双方高度差建立了导弹攻击区和毁伤概率模型。

采用集中式训练分布式执行架构处理多架无人机同步决策和无人机之间既有协作又有竞争的问题;针对如何设置学习率的问题,设计了学习率衰减机制来提升网络的收敛速度和稳定性;然后,利用LSTM网络改进了网络结构,使网络可以通过敌我历史动作序列提取敌方的战术特征,从而做出更有利的空战决策;提出了基于衰减因子的奖励函数机制和加速网络训练。

仿真结果表明所提出的多机协同超视距空战决策算法满足无人机协同作战的需求,仿真中涌现出了一些专家级的超视距协同空战战术。

参考文献:

[1] 杨伟. 关于未来战斗机发展的若干讨论[J]. 航空学报, 2020, 41(6): 524377.
YANG Wei. Development of future fighters[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(6): 524377.

[2] STILLION J. Trends in air-to-air combat implications for future air superiority[M]. Washington DC, USA: Center for Strategic and Budgetary Assessments, 2015.

[3] 孙智孝，杨晟琦，朴海音，等. 未来智能空战发展综述[J]. 航空学报，2021，42(8)：525799.
SUN Zhixiao, YANG Shengqi, PIAO Haiyin, et al. A survey of air combat artificial intelligence[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(8): 525799.

[4] BURGIN G H, OWENS A J. An adaptive maneuvering logic computer program for the simulation of one-on-one air-to-air combat: NASA-CR-2582[R]. [S.l.]: NASA, 1975.

[5] BURGIN G. Improvements to the adaptive maneuvering logic program: NASA-CR-3985[R]. [S.l.]: NASA, 1986.

[6] GOODRICH K, MCMANUS J. Development of a tactical guidance research and evaluation system (TGRES)[C]//Proceedings of Flight Simulation Technologies Conference and Exhibit. Boston, USA: AIAA, 1989: 3312.

[7] MCMANUS J, GOODRICH K. Application of artificial intelligence (AI) programming techniques to tactical guidance for fighter aircraft[C]//Proceedings of Guidance, Navigation and Control Conference. Boston, USA: AIAA, 1989: 3525.

[8] ERNEST N, CARROLL D, SCHUMACHER C, et al. Genetic fuzzy based artificial intelligence for unmanned combat aerial vehicle control in simulated air combat missions[J]. Journal of Defense Management, 2016, 6(1): 1-7.

[9] Defense Advanced Research Projects Agency. AlphaDog-fight trials go virtual for final event[EB/OL]. [2022-04-31]. https://www.darpa.mil/news-events/2020-08-07.

[10] 张强，杨任农，俞利新，等. 基于Q-network强化学习的超视距空战机动决策[J]. 空军工程大学学报(自然科学版)，2018，19(6)：8-14.
ZHANG Qiang, YANG Rennong, YU Lixin, et al. BVR air combat maneuvering decision by using Q-network reinforcement learning[J]. Journal of Air Force Engineering University (Natural Science Edition), 2018, 19(6): 8-14.

[11] LI Y F, SHI J P, JIANG W, et al. Autonomous maneuver decision-making for a UCAV in short-range aerial combat based on an MS-DDQN algorithm[J]. Defence Technology, 2022, 18(9): 1697-1714.

[12] SUN Z X, PIAO H Y, YANG Z, et al. Multi-agent hierarchical policy gradient for air combat tactics emergence via self-play[J]. Engineering Applications of Artificial Intelligence, 2021, 98: 104112.

[13] ISCI H, KOYUNCU E. Reinforcement learning based autonomous air combat with energy budgets [C]//Proceedings of AIAA SCITECH 2022 Forum. San Diego, USA: AIAA, 2022: 0786.

[14] PAN Q, ZHOU D Y, HUANG J C, et al. Maneuver decision for cooperative close-range air combat based on state predicted influence diagram[C]//Proceedings of 2017 IEEE International Conference on Information and Automation (ICIA). Macao, China: IEEE, 2017: 726-731.

[15] WANG L H, HU J W, XU Z, et al. Autonomous maneuver strategy of swarm air combat based on DDPG[J]. Autonomous Intelligent Systems, 2021, 1(1): 1-12.

（编辑：张蓓）