基于深度强化学习与自学习的多无人机近距空战机动策略生成算法

孔维仁[1†], 周德云[1], 赵艺阳[1], 杨婉莎[2]

(1. 西北工业大学 电子信息学院, 陕西 西安 710129; 2. 悉尼大学 计算机学院, 悉尼 2006)

摘要: 为解决多无人机近距空战机动决策问题, 提出一种基于参数共享Q网络与虚拟自我对局的多无人机近距空战机动策略生成算法. 首先, 设计一种适用于不同无人机编队规模的混合马尔可夫博弈模型与多无人机机动决策策略生成强化学习框架—参数共享Q网络, 并通过自编码器对状态空间进行压缩以提高策略学习效率. 然后, 使用虚拟自我对局方法使机动策略收敛至纳什均衡策略. 最后对自编码器的参数选择、策略生成算法的训练过程与机动策略的合理性与迁移性进行了仿真实验. 通过仿真结果表明, 引入自编码器可以有效地提高策略学习效率, 并且使用该算法生成的多无人机近距空战机动策略具有合理性与良好的迁移性.

关键词: 空战决策; 多无人机协同; 强化学习; 虚拟自我对局

引用格式: 孔维仁, 周德云, 赵艺阳, 等. 基于深度强化学习与自学习的多无人机近距空战机动策略生成算法. 控制理论与应用, 2022, 39(2): 352 – 362

DOI: 10.7641/CTA.2021.10120

Maneuvering strategy generation algorithm for multi-UAV in close-range air combat based on deep reinforcement learning and self-play

KONG Wei-ren[1†], ZHOU De-yun[1], ZHAO Yi-yang[1], YANG Wan-sha[2]

(1. School of Electronics and Information, Northwestern Polytechnical University, Xi'an Shaanxi 710129, China;
2. School of Computer Science, The University of Sydney, Sydney 2006, Australia)

Abstract: In order to solve the problem of multi-UAV close-range air combat maneuvering decision-making, a multi-UAV close-range air combat maneuvering strategy generation algorithm based on parameter sharing Q network and neural fictitious self-play is proposed. Firstly, a hybrid Markov game model suitable for different UAV formation sizes and a reinforcement learning framework for generating maneuvering decision strategies of multi-UAV are designed—parameter sharing Q network, and the state space is compressed through the autoencoder to improve the efficiency of strategy learning. Then, using the neural fictitious self-play makes the maneuver strategy converge to the Nash equilibrium strategy. Finally, simulation experiments are carried out on the parameter selection of the autoencoder, the training process of the strategy generation algorithm, and the rationality and portability of the maneuver strategy. The simulation results show that the autoencoder is introduced can effectively improve the efficiency of strategy learning, and the multi-UAV short-range air combat maneuver strategy generated by this algorithm is reasonable and good portability.

Key words: air combat decision-making; multi-UAV cooperation; reinforcement learning; fictitious self-play

Citation: KONG Weiren, ZHOU Deyun, ZHAO Yiyan, et al. Maneuvering strategy generation algorithm for multi-UAV in close-range air combat based on deep reinforcement learning and self-play. *Control Theory & Applications*, 2022, 39(2): 352 – 362

1 引言

随着无人机技术在军事领域的发展, 无人战斗机在战场上的作用越来越重要[1]. 然而, 单架无人机所能完成的作战任务受到了很大约束. 为了适应更加复杂军事任务的需要, 多无人机智能化空战机动决策逐渐成为军事领域的研究热点. 鉴于空战机动在战争发

收稿日期: 2021−02−03; 录用日期: 2021−07−08.

†通信作者. E-mail: k@mail.nwpu.edu.cn; Tel.: +86 18149419636.

本文责任编委: 王龙.

国家自然科学基金项目(61603299, 61612385), 中央高校基本科研业务费专项资金项目(3102019ZX016)资助.

Supported by the National Natural Science Foundation of China (61603299, 61612385) and the Fundamental Research Funds for the Central Universities (3102019ZX016).

展进程中的重要地位, 2020年8月, DARPA公布了名为AlphaDogfight Trail计划的最后一场比赛, 且由苍鹭系统公司大比分获胜[2]. 该项目主要研究空战机动智能算法, 以及如何将研究成果扩展至未来空战. 该项目对于实现智能化作战和人机混合的智能作战系统具有重要意义.

自19世纪60年代以来, 学者对无人机自主空战机动决策进行了大量的研究, 并取得了一些显著的研究成果[3]. 本文把这些成果大致分为两类: 基于对策论的方法与基于人工智能的方法.

在基于对策论的方法中, 包括矩阵对策法[4–6]、影响图法[7–9]、微分对策法[10–12]等. 这些方法在一定程度上为空战决策提供了有效的解决方案, 但也有较大的局限. 例如矩阵对策法得到的策略偏保守, 且随着模型的精度增高, 计算量会急剧上升; 影响图法将空战建立为一个影响图, 影响图可以反应出空战双方状态与决策的影响关系, 如文献[8]给出了一对一空战机动决策问题的影响图, 该图可以将空战这个动态连续的序贯决策问题转换为一个多阶段决策问题, 并使用滚动时域控制等方法进行求解, 然而对于可变规模多对多空战各无人机的状态变化对局部态势和全局态势的影响是很难评价的, 所以影响图的建模是非常困难的; 对于微分对策法, 首先该算法需要精确的数学模型, 计算量大, 且该算法只能解决单纯的追逃问题, 然而空战态势瞬息万变, 随时可能改变攻防关系, 需要多个模型进行交替切换, 若将该方法运用到多对多空战机动决策问题上, 所需的模型的种类和切换次数将会成指数型增加.

在基于人工智能的方法中, 包括专家系统法[13–15]、人工神经网络法[16–18]、深度强化学习法[19–24]等. 这些方法可以一定程度解决基于对策论方法中依赖精确模型、实时性等问题, 但也有较大的局限. 例如专家系统法依赖专家构建知识库, 知识的更新难以满足实时性要求. 人工神经网络法需要大量的空战样本, 且需要大量的人工标注. 深度强化学习方法解决无人机机动决策问题是当前研究的热点, 文献[19]将空战机动决策问题转化为马尔可夫决策过程,并使用深度Q网络算法求解出了单机空战机动策略; 文献[20]在文献[19]的基础上引入了逆强化学习算法来估计更准确的回报函数; 文献[21–23]将空战机动决策问题转化为马尔可夫博弈, 使用多智能强化学习结合最大最小博弈、自学习、机动预测等方法来获得均衡的机动策略. 深度强化学习法虽一定程度上满足智能化空战机动决策的要求, 但与理想效果还相差较远, 主要体现在:

1) 使用强化学习框架主要解决一对一空战机动决策[19–24].

2) 给定了对手的运动规律或机动策略, 使学习得到的空战策略只针对固定机动策略[19–20,24].

3) 假定敌我双方的运动状态是完全已知且准确的[19,21–24].

针对上述强化学习方法的技术难点, 本文以多无人机近距空战为研究对象, 探讨多无人机在近距空战中的智能化空战机动决策生成方法, 基于强化学习框架设计一种适用于无人机规模可变的多无人机机动决策策略生成算法—参数分享的深度Q网络算法(parameter sharing–deep Q network, PS–DQN). 参数分享是指各个Agent共用一个Q值网络参数, 通过设计Agent的状态空间来使Agent协作完成任务.

本文选择使用参数分享的深度Q网络算法的主要原因有两点: 1) 对于多无人机近距空战机动决策问题, 认为各无人机是同构的, 即各无人机的性能参数均相同, 故具备多个Agent公用一个Q值网络的条件; 2) 由于多无人机空战的特点, 无人机初始数量不固定, 而且在空战过程中也会出现损失, Agent的数量是动态变化的, 由于Agent个数动态的不确定性, 导致分布式多智能体强化学习不适合此场景, 参数分享深度Q网络算法由于只存在一个Q值网络, 故天然适用于此场景.

随后, 通过设计Agent的状态空间, 使Agent提取附近友方与敌方无人机的态势特征, 从而使多无人机拥有进行合作空战所需的必要信息; 该状态空间可以部分解决难点3, 只需在一定范围内给出双方准确的运动状态; 同时, 为解决难点2, 本文使用虚拟自我对局(fictitious self-play)使在不给定敌方机动决策的情况下迭代的增强空战机动策略的智能水平, 并收敛到纳什均衡策略.

2　多无人机近距空战问题描述与建模

2.1　多无人机近距空战描述

多无人机近距空战的目的是通过编队内无人机的协作在保证己方编队损失最小的情况下, 尽快歼灭敌方无人机编队. 本文将多无人机近距空战建模为一个混合马尔可夫博弈(mixed Markov game), 将每一个无人机作为博弈中的一个Agent. 将空战编队双方分为红方(己方)和蓝方(敌方)分别采用集合$\mathcal{R}$和$\mathcal{B}$表示. 为了方便对多无人机近距空战建模与生成机动决策, 本文提出以下假设.

假设 1　在编队内部, 每架无人机可以与一定范围内的无人机进行无延时通信.

假设 2　每架无人机可以探测到一定范围内的敌方无人机的准确位置.

假设 3　所有无人机均在同一高度机动飞行[23–25].

本文红蓝双方战机采用文献[25]中开发的一种模拟无人机的动力学方程, 在惯性系下构建无人机的运

动模型. 每架无人机的运动状态由位置(x, y)、速度v、航迹偏角ψ、滚转角φ与滚转角变化率$\dot{\varphi}$定义为

$$\begin{cases} \dot{x} = v\cos\psi, \\ \dot{y} = v\sin\psi, \\ \dot{v} = u_{\rm t}, \\ \dot{\varphi} = u_{\dot{\varphi}}, \\ \dot{\psi} = \dfrac{g\tan\varphi}{v}, \end{cases} \tag{1}$$

其中$(u_{\rm t}, u_{\dot{\varphi}})$为无人机的切向与法向控制量.

2.2 多无人机近距空战建模

2.2.1 状态空间设计

本文提出的PS–DQN算法使Agent共用Q值网络, 则应保证每个Agent的状态向量维度相同; 同时Agent的状态不仅需要反映出自身无人机的运动状态, 而且还需反映出友方无人机与敌方无人机的信息. 这样才可以保证多个Agent共用Q值网络且可以学习出协同策略. 按照上述设计思路, 将Agent的状态空间S分为3个部分: 当前空战态势信息$S_{\rm c}$, 上一步空战态势信息$S_{\rm p}$与上一步无人机动作$a_{\rm p}$

$$S \triangleq S(t) = S_{\rm c}^{\frown} S_{\rm p}^{\frown}[a_{\rm p}], \tag{2}$$

其中$^{\frown}$为向量连接符. 当前空战态势信息$S_{\rm c}$又分为本机运动状态信息$S_{\rm c}^{\rm o}$、双方战损信息$S_{\rm c}^{\rm a}$、友方态势信息$S_{\rm c}^{\rm f}$与敌方态势信息$S_{\rm c}^{\rm e}$. 其中, 本机运动状态信息$S_{\rm c}^{\rm o}$为

$$[x\ \ y\ \ v\ \ \psi\ \ \varphi], \tag{3}$$

双方战损信息$S_{\rm c}^{\rm a}$为

$$[|\mathcal{R}|\ \ |\mathcal{B}|], \tag{4}$$

其中: $|\mathcal{R}|, |\mathcal{B}|$分别表示现存的我方与敌方无人机数量. 由于$S_{\rm c}^{\rm f}$与$S_{\rm c}^{\rm e}$的设计思路相同, 现以$S_{\rm c}^{\rm f}$为例描述其设计思路. 为可以使用固定维度的向量来描述友方无人机信息, 首先, 本文根据无人机的速度方向, 将无人机所处的平面平均分为6个区域, 如图1所示. 本机相对于任意无人机的方位关系均可以归纳为双方划分的6个区域内, 即共有36种方位关系. 然后, 为每一种方位关系设计一个4维向量用于表示友方无人机的位置信息

$$[c\ \ d_{\rm sum}\ \ d_{\rm max}\ \ d_{\rm min}], \tag{5}$$

其中: c表示在区域内的友方无人机个数, $d_{\rm sum}$表示归一化距离总和, $d_{\rm max}$表示归一化距离最大值, $d_{\rm min}$表示归一化距离最小值. 归一化距离$d_{\rm norm}$为

$$d_{\rm norm} = \begin{cases} 0, & d > D_{\rm max}, \\ 1 - \dfrac{d}{D_{\rm max}}, & d \leqslant D_{\rm max}, \end{cases} \tag{6}$$

其中: d为双机距离, $D_{\rm max}$为最大通信(攻击)距离. 友方态势信息$S_{\rm c}^{\rm f}$共有144维, 同理敌方态势信息$S_{\rm c}^{\rm e}$共有144维. 所以, 当前状态信息$S_{\rm c}$共有295维. 上一步态势信息$S_{\rm p}$储存上一时刻的空战态势信息, 故与$S_{\rm c}$结构相同; 上一步无人机动作$a_{\rm p}$为独热编码(one-hot), 储存该无人机上一时刻的动作, 无人机的动作空间在第2.2.2节中定义.

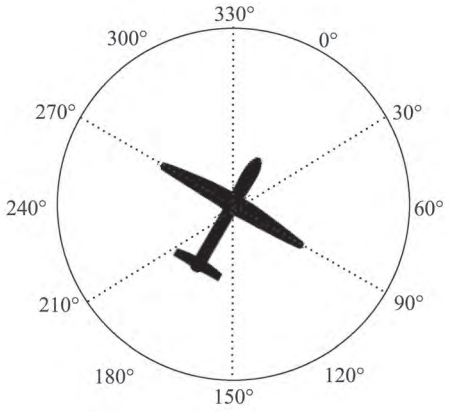


图 1 无人机空间划分

Fig. 1 UAV space division

本文考虑到, S维度过高且$S_{\rm c}^{\rm f}$与$S_{\rm c}^{\rm e}$稀疏程度较高, 故采用自编码器(autoencoder)对S进行降维, 具体方法将在第3.1节中阐述.

2.2.2 动作空间设计

在多无人机近距空战中, 无人机动作空间是连续的, 分别为无人机的切向控制量$u_{\rm t}$和法向控制量$u_{\dot{\varphi}}$. 为了满足DQN算法框架, 本文将无人机动作空间离散化, 按照美国国家航空航天局(NASA)学者设计的基本机动动作库[26]并结合假设3设计了5种机动动作, 分别为匀速直飞、最大加速直飞、最大减速直飞、最大过载左转与最大过载右转. 5个机动动作与控制量的映射关系如表1所示. 其中: $u_{\rm t}^{\rm max}$, $u_{\rm t}^{\rm min}$为$u_{\rm t}$的最大值与最小值, $u_{\dot{\varphi}}^{\rm max}$为$u_{\dot{\varphi}}$的最大值.

表 1 机动动作库

Table 1 Maneuver library

序号	动作名称	独热编码	$u_{\rm t}$	$u_{\dot{\varphi}}$
1	匀速直飞	[0 0 0 0 1]	0	0
2	最大加速直飞	[0 0 0 1 0]	$u_{\rm t}^{\rm max}$	0
3	最大减速直飞	[0 0 1 0 0]	$u_{\rm t}^{\rm min}$	0
4	最大过载左转	[0 1 0 0 0]	0	$u_{\dot{\varphi}}^{\rm max}$
5	最大过载右转	[1 0 0 0 0]	0	$-u_{\dot{\varphi}}^{\rm max}$

2.2.3 奖励函数设计

强化学习算法框架利用Agent与环境交互获得奖赏信息, 根据最大奖赏原则选择动作, 得到最优策略. 奖励函数为强化学习Agent提供了有用的反馈, 对策略学习结果有显著影响[27]. 本文将奖励函数定义为

$$R \triangleq R(S, n_{\rm e}) = R^{\rm r} + \lambda(n_{\rm e})(R^{\rm g} + R^{\rm l}), \tag{7}$$

其中: R^{r}为真实奖励函数, R^{g}为全局奖励函数, R^{l}为局部奖励函数, n_{e}为第n_{e}个训练周期(episode). 真实奖励函数R^{r}是描述空战的最终结果, 它真正表明了多无人机近距空战的目标. 然而R^{r}是一个非常稀疏的奖励函数, 这样的奖励函数学习到有效的空战机动策略是非常困难的[28], 因此, 本文通过设计R^{g}与R^{l}对真实奖励函数进行了奖励塑造(reward shaping). $\lambda(t)$ 是一个随着训练周期的增加而逐渐减小的因子, 它的目的在于在强化学习训练初期按照R^{g}与R^{l}的指导快速的学到有效的空战机动策略, 同时在强化学习训练后期按照R^{r}的指导使无人机完成真正的空战目标. $\lambda(n_{\mathrm{e}})$的表达式为

$$\lambda(n_{\mathrm{e}}) = \mathrm{e}^{-kn_{\mathrm{e}}}. \tag{8}$$

1) 真实奖励函数设计.

根据多无人机近距空战的目的, 设计真实奖励函数R^{r}:

$$R^{\mathrm{r}} \triangleq R^{\mathrm{r}}(S) = \begin{cases} \dfrac{1}{m-|\mathcal{R}|+1}, & |\mathcal{B}|=0, \\ -\dfrac{1}{n-|\mathcal{B}|+1}, & |\mathcal{R}|=0, \\ 0, & \text{其他}, \end{cases} \tag{9}$$

其中R^{r}表示当多无人机空战结束后, 若我方无人机获胜, 则根据我方的损失程度得到一个正的奖励值; 反之, 根据我方的损失程度得到一个负的奖励值. 当多无人机空战正在进行, 则奖励值一直为0.

2) 全局奖励函数设计.

全局奖励函数R^{g}反映了多无人机空战全局的战损信息, 对于整个编队无人机接收到的奖励值都是相同的, 其形式为

$$R^{\mathrm{g}} \triangleq R^{\mathrm{g}}(S) = \frac{1}{m-|\mathcal{R}|+1} - \frac{1}{n-|\mathcal{B}|+1}, \tag{10}$$

从R^{g}中可以看出, 在多无人机空战时, 我方无人机损失越少, 敌方无人机损失越大, 则奖励值越大, 反之越小. 并且当多无人机空战结束后, R^{g}等于R^{r}.

3) 局部奖励函数设计.

局部奖励函数R^{l}反映了多无人机空战中各无人机的局部战场信息, 本文参考文献[25]中定义的局部态势, 且不考虑格斗导弹的前向攻击能力. 我方无人机取得优势需要满足3个条件:

1) 双机距离d小于等于$D_{\max}$.

2) 我方无人机视界角在指定视界范围内, 本文设置为30°.

3) 我方无人机的天线偏转角在指定天线偏转角范围内, 本文设置为30°.

根据上述空战局部态势优势满足条件, 设计局部奖励函数R^{l}:

$$R^{\mathrm{l}} \triangleq R^{l}(S) = \frac{|\mathcal{B}|_{\mathrm{head}} - |\mathcal{B}|_{\mathrm{tail}}}{n}, \tag{11}$$

其中: $|\mathcal{B}|_{\mathrm{head}}$为图1中无人机头部区域中敌方无人机个数, $|\mathcal{B}|_{\mathrm{tail}}$为图1中无人机尾部区域中敌方无人机个数. 从$R^{\mathrm{l}}$中可以看出, 当无人机在局部态势中处于优势, R^{l}越大, 反之, R^{l}越小.

3　多无人机近距空战机动决策生成算法设计

3.1　状态空间自编码器设计

多无人机近距空战机动决策生成算法的第1步是收集数据来训练自动编码器(autocoder). 在第2.2.1节中, 作者设计了多无人机空战的状态空间. 然而, 状态空间的维度过高, 这将会使训练强化学习Agent昂贵且耗时[29].

自动编码器是一种基于人工神经网络(artificial neural network, ANN)的特征表达网络, 用于将高维数据压缩成小的潜在表示(latent)[30]. 本文设计为3层神经网络结构, 包括编码器和解码器两部分, 其中, 编码器完成从输入信号到输出表征的映射转换, 解码器实现输出表征逆向映射回输入空间, 获取重构输入. 根据第2.2.1节可知, 编码器的输入输出维度为144; 本文选取Sigmoid函数作为激活函数.自动编码器的训练过程即为最小化重构误差函数J_{AE}的过程, 其表达式为

$$J_{\mathrm{AE}} = \sum_{i=1}^{p} \|y_i - x_i\|^2 + \lambda \sum \boldsymbol{W}^2, \tag{12}$$

其中: p为输入样本的个数; λ为L_2正则化系数, 用于减少权重的大小来防止过拟合; J_{AE}中的第1项为误差项平方和均值, 第2项为正则化项.

3.2　参数分享深度Q网络算法(PS–DQN)

3.2.1　算法描述

为使无人机编队生成合理的近距空战机动策略, 本节提出参数分享的深度Q网络算法(PS–DQN)用于求解混合马尔可夫博弈模型, 得到合理的多无人机近距空战机动策略.

PS–DQN算法整体框架图如图2所示, 该算法整体框架共包含2个部分: Q网络训练部分与多无人机空战仿真部分. 多无人机空战仿真部分用于仿真多无人机空战场景; Q网络训练部分是PS–DQN算法的核心部分, 它由两个Q值网络和一个经验回放记忆池(experience replay memory)组成. 从图2可以看出, 我方无人机均使用同一个Q值网络进行机动策略的获取与更新, 我方所有无人机共享Q值网络的参数.

3.2.2　PS–DQN原理与算法步骤

强化学习是一个反复迭代的过程, 每一次迭代要解决两个问题: 给定一个策略求取Q值函数, 根据Q

值函数来更新策略. 在环境中的Agent的目的是最大化长期未来奖励, 为得到每个状态下执行每个动作后所得到长期未来奖励, 定义状态–动作值函数$Q_\pi(s,a)$

$$Q_\pi(s,a)=\mathrm{E}[\sum_{t=0}^{\infty}\gamma^t r_t|s_0=s,a_0=a], \quad (13)$$

其中: π为Agent的策略函数, 它可以为随机性或者确定性函数. γ为折扣系数, 取值区间为$[0,1]$. 对于多无人机近距空战场景, 在式(14)中, s和a的取值空间已分别在第2.2.1节与第2.2.2节给出. 得到状态–动作值函数$Q_\pi(s,a)$后, 得到相应的贪婪策略$\pi(s)$

$$\pi(s)=\arg\max_a Q(s,a). \quad (14)$$

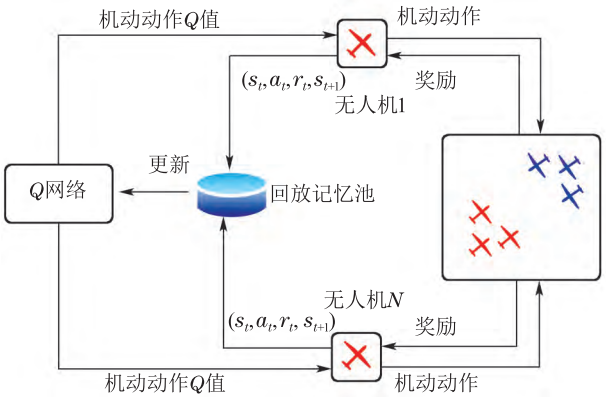


图 2 PS–DQN算法整体框架图

Fig. 2 PS–DQN algorithm diagram

深度Q网络是基于深度学习与强化学习思想而提出的机器学习方法[31–32]. 使用神经网络(Q网络)对Q值函数进行近似表达. 本文采用全连接神经网络对Q值函数进行拟合. Q网络的输入层接受由第2.2.1节设计出的特征状态或自动编码器压缩后的特征状态(在第4.1节进行对比实验), 与当前时刻无人机的动作, 输出层输出对应的Q值. PS–DQN算法的损失函数可表示为

$$L(\theta)=\mathrm{E}[(y-Q(s_t,a_t;\theta))^2], \quad (15)$$

$$y=r_t+\gamma\max_{a_{t+1}}Q'(s_{t+1},a_{t+1};\theta'), \quad (16)$$

其中: $Q(s_t,a_t;\theta)$为估值Q网络, θ为估值Q网络的参数; $Q'(s_t,a_t;\theta')$为目标Q网络, θ'为目标Q网络的参数. PS–DQN算法步骤在算法1中给出.

算法 1 PS–DQN算法.

Input: Q网络$Q(S,a;\theta^{\mathrm{Q}})$,目标$Q$网络参数$Q'(S,a;\theta'^{\mathrm{Q}})$学习速率$\alpha$, 强化学习记忆池$D_{\mathrm{RL}}$, 探索系数$\epsilon$, 仿真回合数$M$, 编队无人机个数$N_{\mathrm{UCAV}}$, 最大单局运行步长$T$, batch规模$N_{\mathrm{batch}}$, 目标$Q$网络更新间隔$C$.

for episode $=1,\cdots,M$ do

初始化空战仿真环境, 并对每一架无人机得到初始状态S_0^i;

while $t<T$并且该场空战未分出胜负 do

for $i=1,\cdots,N_{\mathrm{UCAV}}$ do

在S_t^i状态下, 使用ϵ-greedy策略选取执行动作a_t^i;

end

将所有a_t^i组合为动作向量a_t; 输入a_t至空战仿真环境得到回报R_t与下一步状态S_{t+1};

for $i=1,\cdots,N_{\mathrm{UCAV}}$ do

将$<S_t^i,a_t^i,R_t^i,S_{t+1}^i>$保存至$D_{\mathrm{RL}}$;

end

在D_{RL}中随机采样N_{batch}大小的batch: $<S_j,a_j,R_j,S_j'>$;

$$y_j=\begin{cases}R_j, & \text{回合结束},\\ R_j+\gamma\max_{a'}Q(S_j',a';\theta'^{\mathrm{Q}}), & \text{其他},\end{cases}$$

构造误差函数$(y_j-Q(S_j,a_j;\theta^{\mathrm{Q}}))^2$, 对$\theta^{\mathrm{Q}}$使用mini-batch梯度下降法进行更新, 学习速率为α; 每C步将θ^{Q}赋值给θ'^{Q};

end

end

3.3 神经网络虚拟自我对局

3.3.1 算法描述

由于将多无人机近距空战机动决策问题建模为马尔可夫博弈问题, 若将敌方无人机编队建模为交互环境的一部分, 会造成第3.2节提出的强化学习算法训练过程不稳定[33]; 若设定敌方无人机编队为基于规则的机动策略, 则无法保证得到的空战策略收敛到纳什均衡策略[34].

为获得多无人机近距空战机动决策问题的纳什均衡策略, 本文引入神经网络虚拟自我对局(neural fictitious self-play, NFSP)方法作为策略生成的主框架[35]. 虚拟自我对局(fictitious self-play, FSP)是一种被证明在二人零和博弈中可以收敛到纳什均衡的机器学习算法. 在FSP的学习过程中, 两个玩家通过相互博弈不断优化各自的博弈策略, 最终得到纳什均衡策略. 对于多无人机近距空战机动决策问题, 敌我双方使用相同的策略更新方法, 故以我方无人机编队为例, 给出NFSP算法框架图如图3所示.

3.3.2 NFSP算法原理与算法步骤

从图3中可以看出, 敌我无人机编队均拥有两套空战策略: 对于对手的最优反应策略B以及自己的平均策略Π, 且使用神经网络来拟合最优策略以及平均策略. 无人机编队的行动策略σ是通过对最优反应策略与平均策略按照η的概率选取, 即

$$\sigma=\eta B+(1-\eta)\Pi, \quad (17)$$

$\eta\in(0,1)$称为预测参数(anticipatory parameter). 每当无人机编队依据其行动策略采取动作并获得空战

环境的反馈后, 将记忆片段$< S, a, R, S' >$, 存入强化学习记忆池和模仿学习记忆池(当行动策略为最优反应策略时, 且只储存$< S, a >$片段)中. 随后, 使用PS–DQN算法更新最优反应策略网络$Q(s, a)$参数, 使用模仿学习算法更新平均策略网络$\Pi(s, a)$的参数.

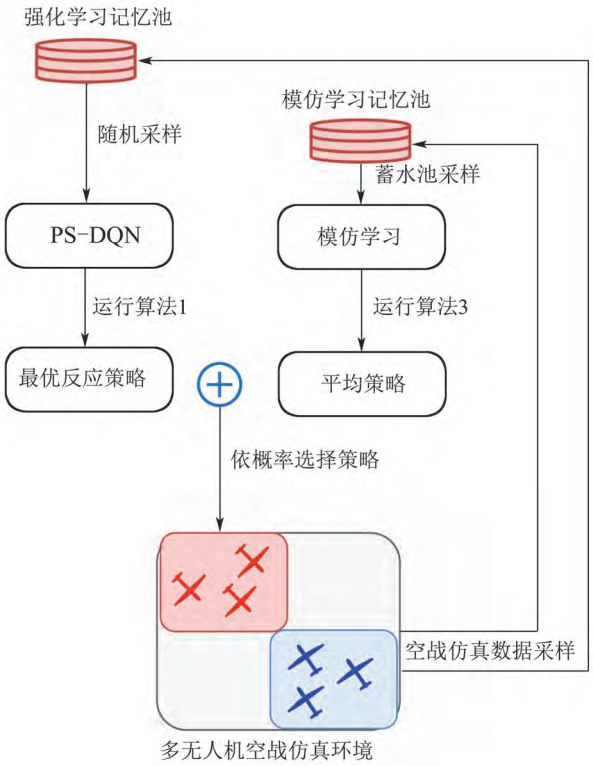


图 3　PS–NFSP算法整体框架图

Fig. 3　PS–NFSP algorithm diagram

对于通过模仿学习算法得到平均策略的目标是保持一个过去迭代轮的最优反应策略的平均混合策略组:

$$\Pi_{k+1} = \frac{1}{k+1}\sum_{i=1}^{k+1} B_i = \frac{k}{k+1}\Pi_k + \frac{1}{k+1}B_{k+1}, \tag{18}$$

Π_k和B_{k+1}分别表示上一轮(第k轮)迭代的平均策略与当前轮(第$k+1$轮)迭代的最优反应策略. 通过上式看出当前轮的平均策略可以通过所有轮的最优反应策略进行均匀采样进行获得. 这也是模仿学习记忆池只储存最优反应策略采样记忆片段的原因.

由于模仿学习记忆池是通过所有轮的最优反应策略采样进行增量更新的, 为了构建一个最优反应的平均策略的无偏估计, 需要从每一轮的最优反应策略中抽样相同数量的片段. 故使用蓄水池抽样[36](reservoir sampling)的方式随机的从模仿学习记忆池中采样训练数据进行模仿学习. 模仿学习算法本质为监督学习算法, 是使用神经网络对平均策略函数进行拟合的过程. 模仿学习算法步骤在算法2中给出.

算法 2　模仿学习算法.

Input: 平均策略网络$\Pi(S, a; \theta^{\Pi})$, 学习速率α, 模仿学习记忆池$D_{\rm SL}$, batch规模$N_{\rm batch}$.

1) 在$D_{\rm SL}$中进行蓄水池采样$N_{\rm batch}$大小的batch: $< S_j, a_j >$;

2) 构造误差函数$-\log \Pi(S, a; \theta^{\Pi})$, 并对$\theta^{\Pi}$使用mini-batch梯度下降法进行更新,学习速率为α.

根据上述NFSP算法原理, 在算法3给出NFSP算法步骤.

算法 3　NFSP算法.

Input: 多无人机空战对抗环境Γ,预测参数η, 算法2–3中提到的超参数.

Output: 平均策略$\Pi(S, a; \theta^{\Pi})$.

初始化: $\Pi(S, a; \theta^{\Pi})$, $Q(S, a; \theta^{\rm Q})$, $Q'(S, a; \theta'^{\rm Q})$, $D_{\rm RL}$, $D_{\rm SL}$;

for　episode $= 1, \cdots, M$　do

初始化空战仿真环境;

$$\sigma = \begin{cases} B, & \text{with prob. } \eta, \\ \Pi, & \text{with prob. } 1-\eta, \end{cases}$$

while　$t < T$并且该场空战未分出胜负　do

for　$i = 1, \cdots, N_{\rm UCAV}$　do

依据行为策略σ抽取动作a_t^i;

end

将所有a_t^i组合为动作向量a_t; 输入a_t至空战仿真环境得到回报R_t与下一步状态S_{t+1};

for　$i = 1, \cdots, N_{\rm UCAV}$　do

将$< S_t^i, a_t^i, R_t^i, S_{t+1}^i >$保存至$D_{\rm RL}$;

if σ为最优反应策略　then

将$< S_t^i, a_t^i >$保存至$D_{\rm SL}$;

end

end

使用算法1更新$Q(S, a; \theta^{\rm Q})$和$Q'(S, a; \theta'^{\rm Q})$的参数; 使用算法2更新$\Pi(S, a; \theta^{\Pi})$的参数;

end

end

4　仿真实验与分析

4.1　空战仿真平台搭建

本文空战仿真平台采用Python 3编程语言, 基于OpenAI团队开发的gym强化学习环境开发包进行开发[37], 该仿真平台能够在一定的空域内仿真多无人机近距空战. 对于每架无人机, 平台采用式(1)的微分方程组来解算每架无人机的运动状态, 同时仿真平台收集所有无人机的运动状态计算在第2.2.1节中提出的状态向量, 并分发给每架无人机, 作为该无人机当前的空战状态.

本仿真平台中, 假设红蓝编队处于同一高度水平, 无人机的空域范围设置在5 km以内, 假定各架无人机

机动能力相同, 且最小速度为100 m/s; 最大速度为300 m/s; 最大通信距离为5 km; 最大探测距离为2 km, 滚转角变化率为40(°)/s. 每个仿真步长代表0.15 s且最大仿真时间为20 s. 由于本空战仿真平台没有对近距格斗导弹进行建模与仿真, 故本平台对于击落的判别方式为满足第2.2.3节给出的稳定跟踪条件且持续3个仿真步长.

4.2 自编码器latent维度选择实验

4.2.1 实验设计

在第3.1节中, 提出了使用状态空间自编码器来降低状态空间维度, 以加快空战策略学习速率. 本节将选取不同latent维度: 8, 16, 32和64维, 用来考察不同latent维度对原始状态的表示精度, 并选取合理的latent维度.

在本实验中, 双方无人机使用随机策略, 通过对不同规模空战仿真进行随机采样, 得到40000组空战状态. 其中, 30000组空战状态作为训练集, 剩余10000组空战状态作为测试集. 空战规模有: 1v1, 2v2, 4v4, 4v2这4种. 使用pytorch对自编码器进行实现, 采用Adam优化算法[38]进行训练, 学习速率α为0.01, β_1为0.9, β_2为0.999. 迭代次数选为30000, batch大小为64, 在采样的时候对数据进行打乱, 用于消除batch内数据之间的相关性.

4.2.2 实验结果与讨论

算法训练过程中的损失变化图如图4所示, 横坐标为训练步数, 纵坐标为自编码器每次训练的损失, 可以看出随着训练步数的增加, 训练损失逐渐下降最终收敛趋近于稳定, 说明自编码器训练收敛. 通过观察不同latent维度收敛到的训练误差可以发现, 当latent维度为32与64时, 收敛后的训练误差在误差范围内是相同的. 说明latent为32维时, latent就足以表示144维的原始空战状态.

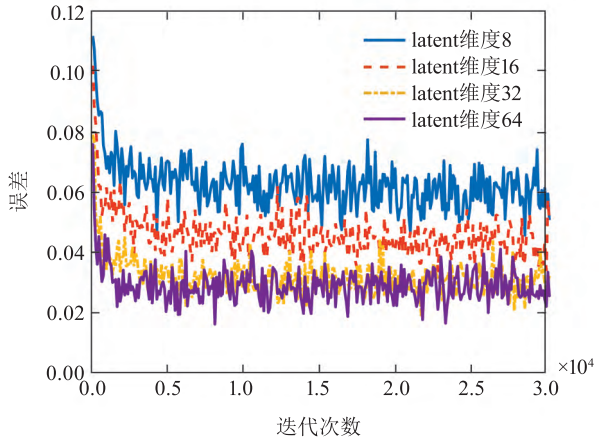


图 4 自编码器训练损失变化图

Fig. 4 Autoencoder training loss diagram

现使用测试集中的任意空战状态输入到训练完成的latent维度为32的自编码器中, 得到的编码平均误差为0.031, 与训练集的误差相似, 说明未出现过拟合现象. 以某一空战状态为例, 原始空战状态与通过自编码器解码的空战状态对比图如图5所示.

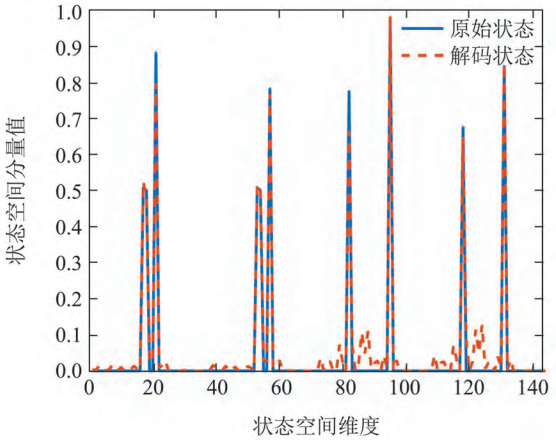


图 5 原始状态与自编码器解码状态对比图

Fig. 5 The diagram comparising original state and autoencoder decoding state

从图5中发现, 原始空战状态有稀疏的特性, 且自编码器解码的空战状态可以很好的的反映出原始空战状态不为0的分量, 平均误差在0.1以内, 足以满足空战机动策略生成算法所需要的精度.

4.3 策略生成算法训练实验

4.3.1 实验设计

本节对提出的空战机动策略生成算法的训练过程进行实验, 研究引入自编码器与神经网络虚拟自我对局对PS–DQN算法训练效果的影响. 故本节设计3个训练算法进行对比:

算法 1 假定敌方无人机编队为贪婪策略, 引入自编码器的PS–DQN算法.

算法 2 引入神经网络虚拟自我对局, 不引入自编码器的PS–DQN算法.

算法 3 同时引入神经网络虚拟自我对局与自编码器的PS–DQN算法.

本节将从两个方面分析PS–DQN算法的训练效果,首先通过对比3个算法策略的可利用性(exploitability)[39]来研究策略是否收敛到纳什均衡策略, 用以判断引入NFSP是否可以改善策略的可利用性, 然后对比算法2–3的平均策略网络训练的收敛情况, 用以判断引入自编码器是否可以加快策略收敛的速度. 策略可利用性是用来衡量当前双方策略是否达到纳什均衡的指标, 计算方法为

$$\epsilon(\sigma)=\frac{u_1(\sigma_1,B_2(\sigma_1))+u_2(B_1(\sigma_2),\sigma_2)}{2},\quad(19)$$

$u_1(\sigma_1,B_2(\sigma_1))$为我方采用当前策略$\sigma_1$, 对方使用针对我方当前策略的最优反应策略$B_2(\sigma_1)$所得到的真

实奖励(式(9)所示). 从上式可以看出, 若双方策略达到纳什均衡, $\epsilon(\sigma)$值为0; 相反, 若双方背离纳什均衡策略, 则$\epsilon(\sigma)$越大. 故策略的可利用性是衡量双方策略是否达到纳什均衡的指标.

整体仿真实验参数设置: 采用4v2的空战规模, 初始无人机运动状态随机; 仿真步数设置为200000; 强化学习与模仿学习记忆池规模为100000; 预测参数η为0.1; 网络训练开始时的步数为1000; 单次episode最大步数为150. PS–DQN算法参数设置: 采用4层全连接网络作为Q网络, 使用Dueling网络结构[40], 2层隐藏层节点个数分别为128, 64; Q网络学习速率为0.001; 折扣系数γ为0.95; 探索概率ϵ为动态衰减因子, 初始为1.0, 终止为0.001; 目标网络更新周期为1000. 模仿学习算法参数设置: 模仿网络采用4层全连接网络, 两层隐藏层节点个数分别为60, 30, 输出层为softmax层; 网络学习速率为0.001.

4.3.2　实验结果与讨论

平均策略网络训练过程中的损失变化图如图6所示, 由于算法1只有Q网络, 故图中只有算法2–3的平均策略网络损失变化图. 在图中, 横坐标为训练步数, 纵坐标为平均策略网络每次训练的损失, 可以看出随着训练步数的增加, 训练损失逐渐下降最终收敛趋近于稳定, 说明平均策略网络收敛. 对比算法2与算法3的收敛曲线, 可以发现算法3相比于算法2有更快的收敛速率, 说明引入自编码器可以改善空战机动策略的学习效率.

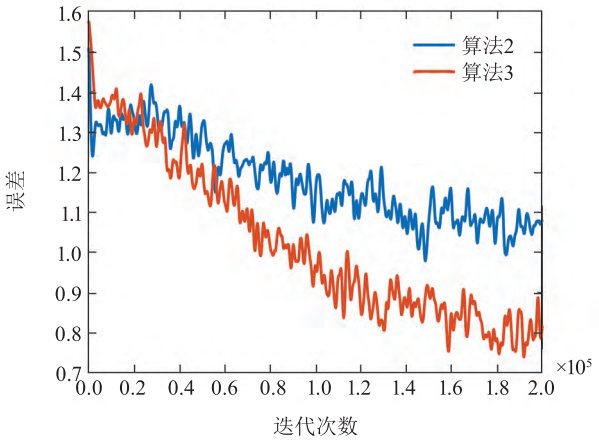


图 6　平均策略网络训练损失变化图

Fig. 6　Average policy network training loss diagram

3个算法平均策略可利用性的变化过程如图7所示. 训练初期, 3个算法的平均策略可利用性在0附近震荡, 这是由于在训练初期最优反应策略网络与平均策略网络无法提供准确的策略, 使双方均无法进行有效的攻防, 双方均无战损; 训练中期, 3个算法的平均策略可利用性均上升, 这是由于平均策略网络平均的最优反应策略过少, 而导致平均策略网络无法反应出真实的平均策略, 然而这时最优反应策略对平均策略有较好的针对性, 故此时的策略可利用性逐渐增加; 训练后期, 算法2–3的策略可利用性逐渐减小, 且收敛到0附近, 这说明双方的平均策略逐步趋近于纳什均衡策略, 算法1的策略可利用性依然增大, 这是由于敌方策略为贪婪策略, 导致我方策略收敛至针对贪婪策略的最优策略, 然而该策略是与纳什均衡策略相背离的, 故策略可利用性依然增大. 同时通过对比算法2–3的策略可利用性变化曲线可以发现, 算法3的策略可利用性峰值与收敛速度均好于算法2, 这说明引入自编码器确实有益于策略的学习效率.

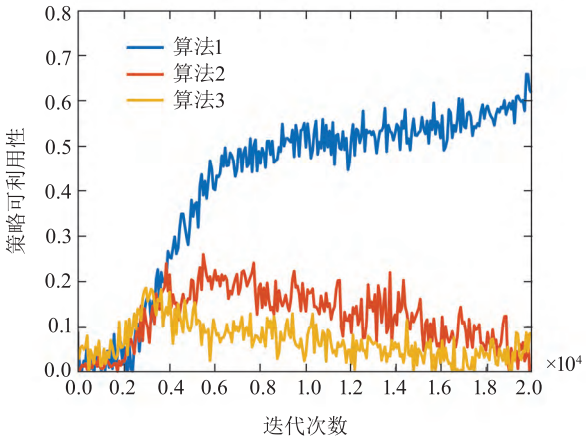


图 7　平均策略可利用性变化图

Fig. 7　Exploitability of average policy diagram

4.4　机动策略合理性与迁移性实验

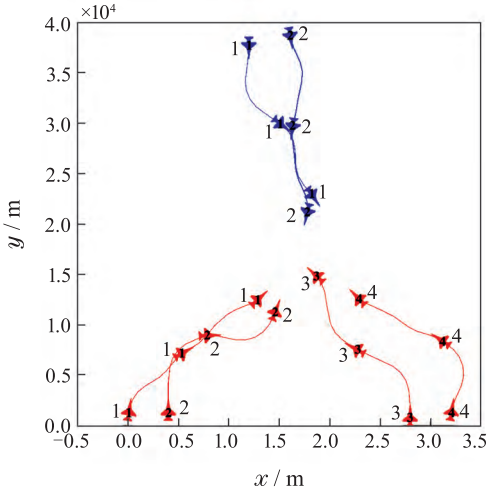
4.4.1　实验设计

本节对使用算法3得到的多无人机近距空战机动策略的合理性与迁移性进行实验, 对于策略有效性, 本节首先使用两个双机编队在近距空战空域内对抗一个双机编队, 且双方为迎头的均衡态势, 如图8(a). 该空战规模与策略生成算法训练实验相同. 该实验将给出该空战想定下的整体无人机机动过程, 用于分析空战机动策略是否合理, 是否符合基本的空战原则[41]. 随后, 验证空战规模为1v1, 2v2, 4v2, 4v4下的多无人机近距空战机动策略的合理性, 即机动策略的迁移性实验. 此外, 由于1v1空战的基础性与特殊性, 本实验也将给出1v1空战的整体无人机机动过程, 用以直观的分析空战机动策略的迁移性.

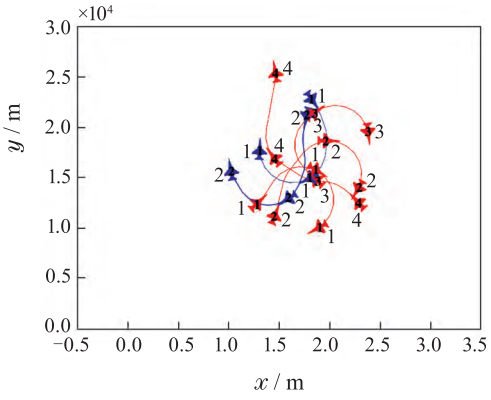
4.4.2　实验结果与讨论

4v2空战机动过程如图8所示, 空战初始红方有两个双机编队呈平行态势, 蓝方有一个双机编队, 双方为迎头的均衡态势. 在空战进行0 s至4 s之间, 双方战机均选择正面接敌, 进入1 km^{2}内的空域内进行缠斗(图8(a)). 空战进行4 s至8 s之间, 红方1, 2, 3号无人机进行大过载右转试图稳定跟踪蓝方编队的尾部, 以锁定并构成攻击条件, 同时红方4号无人机进行加速直飞用于诱敌与后续包夹, 此时蓝方编队合力追击红方

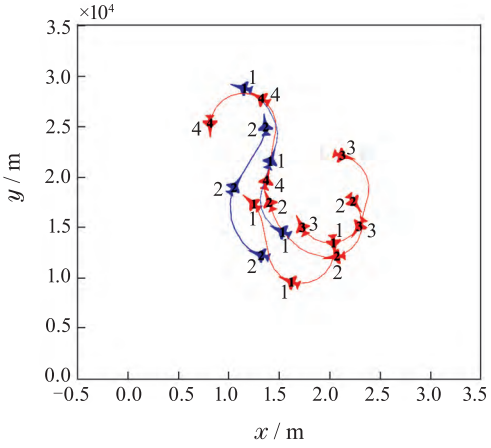
4号无人机, 意在锁定红方4号无人机(图8(b)). 空战进行8 s至12 s之间, 红方红方1, 2, 3号无人机成功锁定蓝方2号无人机并满足攻击条件, 进而歼灭蓝色2号无人机, 红方4号无人机进行大机动左转试图摆脱蓝方1号无人机的追踪, 蓝方1号无人机由于目标的大机动导致脱锁(图8(c)). 空战进行12 s至20 s之间, 由于红方态势占绝对优势, 故蓝方1号无人机实施连续的S形机动试图脱离红方无人机编队的追击, 红方编队各无人机从多个角度追击蓝方无人机并最终成功击落(图8(d)).



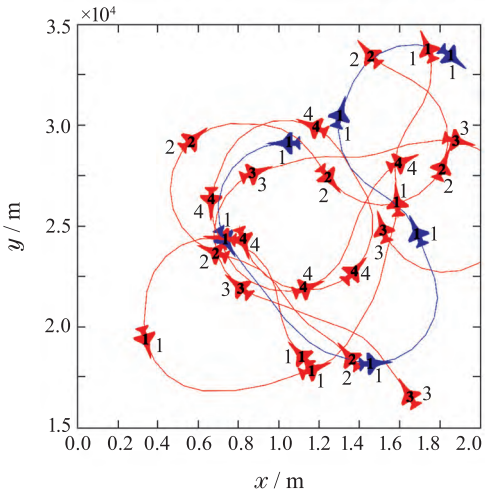
(a) 0～ 4 s



(b) 4～ 8 s



(c) 8～ 12 s



(d) 12～ 20 s

Fig. 8 4v2空战机动过程图

Fig. 8 Maneuver process diagram of 4v2 air combat

从4v2空战整体过程可以看出, 双方的机动动作是可以被直观解读的, 这说明得到的空战机动策略是合理的.

1v1空战机动过程如图9所示, 空战初始红蓝双方为迎头的均衡态势, 双方无人机在4 s后形成互相“咬尾”的缠斗模式, 一直往复下去. 这在1v1空战实战中是非常常见的, 只是由于模型的简化, 没有势能的损耗而导致的高度下降. 从1v1空战整体过程可以看出, 双方的机动动作是可以被直观解读的, 说明空战机动策略迁移至1v1空战场景也是有效的.

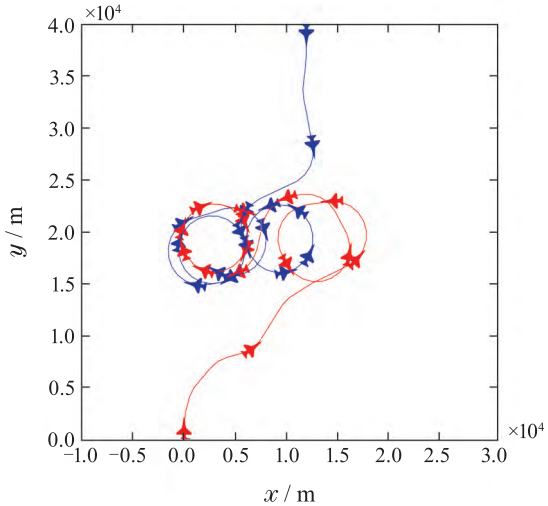


图 9 1v1空战机动过程图

Fig. 9 Maneuver process diagram of 1v1 air combat

随后对空战规模为1v1, 2v2, 4v2, 4v4下的多无人机近距空战分别进行了1000次的对抗仿真, 初始空战态势随机, 得到以红方视角的空战结果如表2所示.

从表2中可以看出, 对于双方无人机个数相同的空战仿真结果是对称的, 说明双方的平均策略拥有达到了均衡的必要条件, 且均可以有效的歼灭对手. 这说

明得到的策略迁移到不同空战规模均是有效且合理的. 对于4v2的空战规模, 红方4机无人机编队明显优势于蓝方的双机编队, 这与实际空战相符.

表 2 不同空战规模下的多无人机近距空战结果

Table 2 Air combat results under different air combat scales

空战规模	完胜	优势	均势	劣势	完败
1v1	253	—	523	—	224
2v2	131	216	260	243	150
4v2	410	373	100	85	32
4v4	40	356	241	329	34

5 结论

本文采用深度强化学习与自学习相结合的技术, 提出了一种解决多无人机近距空战机动决策问题的算法. 该算法允许所有编队中的无人机共用一个Q网络来使算法满足在不同无人机编队规模下具有良好的迁移性. 同时, 引入了自编码器与神经网络虚拟自我对局机制, 使算法可以高效率的学习到达纳什均衡的无人机空战机动策略.

目前本文通过空战仿真实验验证了算法的可行性, 下一步的工作将加入高度影响将仿真环境从二维拓展到三维, 并考虑雷达和武器情况做现代空战的研究, 使算法适应更加复杂的战场环境.

参考文献:

[1] LEE D, KIM S, SUK J. Formation flight of unmanned aerial vehicles using track guidance. *Aerospace Science and Technology*, 2018, 76(5): 412 – 420.

[2] HAMBLING D. AI outguns a human fighter pilot. *New Scientist*, 2020, 247(3297): 12.

[3] SHIN H, LEE J, KIM H , et al. An autonomous aerial combat framework for two-on-two engagements based on basic fighter maneuvers. *Aerospace Science and Technology*, 2017, 72: 305 – 315.

[4] DENG Ke, PENG Xuanqi, ZHOU Deyun. UAV air combat decision based on matrix game and genetic algorithm. *Firepower and Command Control Control*, 2019, 44(12): 61 – 66.
(邓可, 彭宣淇, 周德云. 基于矩阵对策与遗传算法的无人机空战决策. 火力与指挥控制, 2019, 44(12): 61 – 66.)

[5] XU Guangda, LU Chao, WANG Guanghui, et al. Research on autonomous mobility decision of UCAV air combat based on binary matrix game. *Ship Electronic Engineering*, 2017, 37(11): 24 – 28, 39.
(徐光达, 吕超, 王光辉, 等. 基于双矩阵对策的UCAV空战自主机动决策研究. 舰船电子工程, 2017, 37(11): 24 – 28, 39.)

[6] SU M C, LAI S C, LIN S C, et al. A new approach to multi-aircraft air combat assignments. *Swarm and Evolutionary Computation*, 2012, 6: 39 – 46.

[7] WAN Wei, JIANG Changsheng, WU Qingxian. Application of single-step prediction influence diagram method in air combat manipulation decision-making . *Electronic Optics & Control*, 2009, 16(7): 13 – 16, 28.
(万伟, 姜长生, 吴庆宪. 单步预测影响图法在空战机动决策中的应用. 电光与控制, 2009, 16(7): 13 – 16, 28.)

[8] VIRTANEN K, RAIVIO T, HAMALAINEN R P. Modeling pilot's sequential maneuvering decisions by a multistage influence diagram. *Journal of Guidance, Control, and Dynamics*, 2004, 27(4): 665 – 677.

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

[10] WANG Yining, JIANG Yuxian. Intelligent differential game method in air combat decision making. *Journal of Flight Mechanics*, 2003, 21(1): 66 – 70.
(王义宁, 姜玉宪. 空战决策中的智能微分对策法. 飞行力学, 2003, 21(1): 66 – 70.)

[11] WANG Yu, ZHANG Weiguo, FU Li, et al. Particle swarm optimization for air combat nash equilibrium. *Control Theory & Applications*, 2015, 32(7): 857 – 865.
(王昱, 章卫国, 傅莉, 等. 基于精英改选机制的粒子群算法的空战纳什均衡策略逼近. 控制理论与应用, 2015, 32(7): 857 – 865.)

[12] PARK H, LEE B Y, TAHK M J, et al. Differential game based air combat maneuver generation using scoring function matrix. *International Journal of Aeronautical and Space Sciences*, 2016, 17(2): 204 – 213.

[13] ZHAO W, ZHOU D. Application of expert system in sequencing of air combat multi-target attacking. *Electronics Optics & Control*, 2008, 15(2): 23 – 26.

[14] SHENYU G A O. Research on expert system and decision support system for multiple air combat tactical maneuvering. *Systems Engineering—Theory & Practice*, 1999, (8): 76 – 79, 126.

[15] PARK S J, PARK S S, CHOI H L, et al. An expert data-driven air combat maneuver model learning approach. *AIAA Scitech 2021 Forum*. San Diego, CA: AIAA, 2021: 0526.

[16] XUE J, ZHU J, XIAO J, et al. Panoramic convolutional long short-term memory networks for combat intension recognition of aerial targets. *IEEE Access*, 2020, 8: 183312 – 183323.

[17] TENG T H, TAN A H, TAN Y S, et al. Self-organizing neural networks for learning air combat maneuvers. *The 2012 International Joint Conference on Neural Networks (IJCNN)*. Brisbane, QLD, Australia: IEEE, 2012: 1 – 8.

[18] ZHANG H, HUANG C. Maneuver decision-making of deep learning for UCAV thorough azimuth angles. *IEEE Access*, 2020, 8: 12976 – 12987.

[19] YANG Q, ZHANG J, SHI G, et al. Maneuver decision of UAV in short-range air combat based on deep reinforcement learning. *IEEE Access*, 2019, 8: 363 – 378.

[20] KONG W, ZHOU D, YANG Z, et al. UAV autonomous aerial combat maneuver strategy generation with observation error based on state-adversarial deep deterministic policy gradient and inverse reinforcement learning. *Electronics*, 2020, 9(7): 1121.

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

[22] KONG W, ZHOU D, YANG Z, et al. Maneuver strategy generation of UCAV for within visual range air combat based on multi-agent reinforcement learning and target position prediction. *Applied Sciences*, 2020, 10(15): 5198.

[23] MA Wen, LI Hui, WANG Zhuang, et al. Manipulation decision making in close air combat based on deep random game. *Systems Engineering and Electronics*, 2021, 43(2): 443 – 451.
(马文, 李辉, 王壮, 等. 基于深度随机博弈的近距空战机动决策. 系统工程与电子技术, 2021, 43(2): 443 – 451.)

[24] YANG Q, ZHU Y, ZHANG J, et al. UAV air combat autonomous maneuver decision based on DDPG algorithm. *2019 IEEE 15th International Conference on Control and Automation (ICCA)*. Edinburgh UK: IEEE, 2019: 37 – 42.

[25] MCGREW J S, HOW J P, WILLIAMS B, et al. Air-combat strategy using approximate dynamic programming. *Journal of Guidance, Control, and Dynamics*, 2010, 33(5): 1641 – 1654.

[26] AUSTIN F, CARBONE G, FALCO M, et al. Automated maneuvering decisions for air-to-air combat. *Guidance, Navigation and Control Conference*. Monterey, USA: AIAA, 1987: 2393.

[27] SUTTON R S, BARTO A G. *Introduction to Reinforcement Learning*. Cambridge: MIT Press, 1998.

[28] DING N, SORICUT R. *Cold-start reinforcement learning with softmax policy gradient*. arXiv preprint arXiv: 1709.09346, 2017.

[29] LANGE S, RIEDMILLER M. Deep auto-encoder neural networks in reinforcement learning. *The 2010 International Joint Conference on Neural Networks (IJCNN)*. Barcelona, Spain: IEEE, 2010: 1 – 8.

[30] JIANG F, WANG K, DONG L, et al. Stacked autoencoder-based deep reinforcement learning for online resource scheduling in large-scale MEC networks. *IEEE Internet of Things Journal*, 2020, 7(10): 9278 – 9290.

[31] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, 518(7540): 529 – 533.

[32] GU S, LILLICRAP T, SUTSKEVER I, et al. Continuous deep q-learning with model-based acceleration. *International Conference on Machine Learning*. New York, USA: PMLR, 2016: 2829 – 2838.

[33] LI S, WU Y, CUI X, et al. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019, 33(1): 4213 – 4220.

[34] SILVER D, HUBERT T, SCHRITTWIESER J, et al. *Mastering chess and shogi by self-play with a general reinforcement learning algorithm*. arXiv preprint arXiv: 1712.01815, 2017.

[35] HEINRICH J, LANCTOT M, SILVER D. Fictitious self-play in extensive-form games. *International Conference on Machine Learning*. Lille, France: PMLR, 2015: 805 – 813.

[36] AGGARWAL C C. On biased reservoir sampling in the presence of stream evolution. *Proceedings of the 32nd International Conference on Very Large Data Bases*. Seoul, Korea: DBLP, 2006: 607 – 618.

[37] BROCKMAN G, CHEUNG V, PETTERSSON L, et al. *Openai Gym*. arXiv preprint arXiv: 1606.01540, 2016.

[38] KINGMA D P, BA J. *Adam: A method for stochastic optimization*. arXiv preprint arXiv: 1412.6980, 2014.

[39] LOCKHART E, LANCTOT M, PEROLAT J, et al. *Computing approximate equilibria in sequential adversarial games by exploitability descent*. arXiv preprint arXiv: 1903.05614, 2019.

[40] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning. *International Conference on Machine Learning*. New York, USA: PMLR, 2016: 1995 – 2003.

[41] HERBST W B. Dynamics of air combat. *Journal of Aircraft*, 1983, 20(7): 594 – 598.

作者简介:

孔维仁 博士研究生, 目前研究方向为多智能体强化学习在军事方面的应用, E-mail: k@mail.nwpu.edu.cn;

周德云 博士, 教授, 目前研究方向为新一代智能火控系统建模与仿真, E-mail: dyzhou@nwpu.edu.cn;

赵艺阳 博士研究生, 目前研究方向为复杂作战环境下的作战任务规划与智能优化算法, E-mail: zhaoyiyang@mail.edu.cn;

杨婉莎 硕士研究生, 目前研究方向为多智能体强化学习算法与人工智能在游戏下的应用, E-mail: yangwansha1997@gmail.com.