https://doi.org/10.1051/jnwpu/20234110056

深度确定性策略梯度和预测相结合的无人机空战决策研究

李永丰[1], 吕永玺[1,2], 史静平[1,2], 李卫华[1]

(1.西北工业大学 自动化学院, 陕西 西安 710129;
2.陕西省飞行控制与仿真技术重点实验室, 陕西 西安 710129)

摘 要:针对无人机自主空战机动决策过程中遇到的敌方不确定性操纵问题,提出了一种目标机动指令预测和深度确定性策略梯度算法相结合的无人机空战自主机动决策方法。对空战双方的态势数据进行有效的融合和处理,搭建无人机六自由度模型和机动动作库,在空战中目标通过深度 Q 网络算法生成相应机动动作库指令,同时我方无人机通过概率神经网络给出目标机动的预测结果。提出了一种同时考虑了两机态势信息和敌机预测结果的深度确定性策略梯度强化学习方法,使得无人机能够根据当前空战态势选择合适的机动决策。仿真结果表明,该算法可以有效利用空战态势信息和目标机动预测信息,在保证收敛性的前提下提高无人机自主空战决策强化学习算法的有效性。

关 键 词:无人机;空战机动决策;预测;深度确定性策略梯度
中图分类号:V249.1 **文献标志码**:A **文章编号**:1000-2758(2023)01-0056-09

制空权在现代战争中变得越来越重要,在这一领域的最新发展中,无人机(unmanned aerial vehicle,UAV)的研究进展引起了全世界的关注。在攻击方面,新型无人攻击机和多用途无人机应用了精确制导、数据传输和自动控制系统等技术,攻击准确、威力巨大。随着人工智能的迅速发展,无人机在自主空战的应用中具备巨大潜力。无人机将从简单的远程控制演变为智能和自主控制,并配备智能作战决策系统,逐步取代有人驾驶飞机,在提高作战效能的同时降低成本。

在日益复杂的空战环境中,自主机动决策要求无人机在不同的空战情况下自动生成适当的机动控制命令。Ehtamo 等[1]研究了离散化追逃对策的机动决策,通过求解一些反馈解的开环表示来证明离散化的可用性,以终端时间为回报,将这些反馈解应用于一个实际模型飞机和导弹之间复杂的追踪与规避的博弈问题。顾佼佼等[2]采用双矩阵博弈的结构构造空战机动决策模型,并通过改进 Memetic 算法实时求解,以满足实时性的要求。万伟等[3]采用单步预测影响图的方法分析驾驶员的决策过程,通过充分利用信息减小结果的不确定性。之后研究者们将人工智能与空战机动决策联系在一起,利用人工智能系统模拟飞行员空战行为,延伸和扩展飞行员机动决策能力。Kumar 等[4]通过专家系统给出了一种通过空战对抗训练机动决策方法,基于双方几何态势和运动状态预测未来的态势。Smith 等[5]通过遗传学习系统设计了飞机机动决策模型。

与其他的人工智能算法相比,强化学习是一种与环境进行交互的学习方法,它在构建环境和动作映射关系的基础上,通过不断尝试来寻找最优解[6]。Yang 等[7]采用深度 Q 网络(deep Q network,DQN)构建无人机决策模型,使得无人机可以根据敌我两机的态势,从机动动作库中选择相应的机动动作以实现空战,同时该文献采用了基本对抗的方

收稿日期:2022-04-25 **基金项目**:国家自然科学基金(62173277,61573286)、陕西省自然科学基金(2019JM-163,2020JQ-218,2022JM-011)与航空科学基金(20180753006,201905053004)资助
作者简介:李永丰(1995—),西北工业大学博士研究生,主要从事飞行控制与无人机空战方法研究。
通信作者:吕永玺(1990—),西北工业大学助理研究员,主要从事飞行控制与控制方法研究。e-mail:yongxilyu@nwpu.edu.cn

法训练神经网络提高算法的收敛性。然而 DQN 算法生成的主要是机动动作库指令,Bai 等[8]采用深度确定性策略梯度(deep deterministic policy gradient,DDPG)算法输出机动动作指令,使得无人机可以进行连续的机动。Li 等[9]对 DDPG 算法进行了改进,通过添加混合噪声提高收敛速度。同时强化学习算法不只可以应用于单机对抗还可以应用于多机对抗中,但这些所使用的都是三自由度无人机模型,并未考虑无人机自身的姿态特性[10-11]。

态势评估在空战决策中起着举足轻重的作用,通过对信息源的数据进行提取和处理获得精确的位置和身份估计[12]。无人机可以根据战场态势预测目标下一步的行动模式,从而提前做好准备。毛梦月等[13]采用概率神经网络对目标机动单元进行预测,同时结合强化学习算法实现无人机一对一的空中对抗。使用态势评估可以使无人机对战场当前态势、威胁及其重要程度进行实时、完整的评价。

本文首先通过分析现代空战对抗环境下敌我双方的能力参数和空战态势信息,构建空战评估体系,建立了基于两机几何态势、机体稳定性、导弹参数和环境情况的空战优势函数模型;然后在 Matlab/Simulink 中搭建六自由度无人机模型和机动动作库,令敌机通过 DQN 算法生成相应机动动作库指令,同时我方无人机通过概率神经网络(probabilistic neural network,PNN)预测目标机动指令;最后将预测的结果与 DDPG 算法相结合,使得无人机能够自主地进行空战机动决策。仿真结果表明,采用该算法的无人机可以有效地打击目标,实现一对一的空战。

1　空战优势函数

在现代空战环境下,需要根据敌我双方的空战态势和能力参数进行优势函数建模,其中能力参数主要取决于飞机的武器性能、探测能力和电子对抗能力等,下面建立一个基于空战态势威胁和能力参数的空战威胁评估体系。

空战的主要形式为双方参战飞机迎头飞行,互相发射空空导弹,因此在空战态势中主要考虑敌我两机的相对角度、距离、速度和高度等因素。在能力参数方面对于空战双方而言,机载雷达和空空导弹的性能对于战场态势的影响较大,雷达决定飞机能否有效地跟踪目标,而导弹性能决定了其载机能否有效地攻击目标机,因此本文的优势建模需要同时兼顾双方几何态势、机载雷达和导弹的性能。

1.1　几何态势建模

双方几何态势定义如图 1 所示:

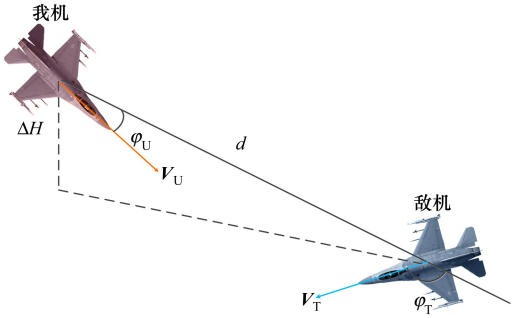

图 1　敌我双方空战态势图

图中,红机为我方无人机,蓝机为敌方无人机。$\boldsymbol{V}_{\mathrm{U}}$ 为我方无人机速度矢量;$\boldsymbol{V}_{\mathrm{T}}$ 为敌方无人机速度矢量;d 为我方无人机到敌方无人机的距离;φ_{U} 为我方机头指向与目标线的夹角,称为目标方位角;φ_{T} 为敌方机头指向与目标线的夹角,称为目标航向角;ΔH 为我机相对于敌机的高度差。

除此之外,两机机载雷达的主要性能指标包括搜索方位角 $\varphi_{R\max}$、搜索距离 $d_{R\max}$。而两机空空导弹的主要性能指标包括最大离轴发射角 $\varphi_{M\max}$、最大攻击距离 $d_{M\max}$、最小攻击距离 $d_{M\min}$、最大不可逃逸区距离 $d_{MK\max}$、最小不可逃逸区距离 $d_{MK\min}$ 等。

1.1.1　角度优势函数

当我机相对于敌机的目标方位角越小时,我方的攻击优势越大;而敌机相对于我机的目标方位角越小时,敌方所处的攻击劣势越大,这时我方对敌方形成追击态势。根据角度优势对空战态势的影响构建相应的角度优势函数如(1)式所示。

$$f_{\varphi}=\frac{360^{\circ}-|\varphi_{\mathrm{U}}|-|\varphi_{\mathrm{T}}|}{360^{\circ}} \tag{1}$$

1.1.2　距离优势函数

在空战过程中,我机通过雷达搜索目标,并在锁定目标后对敌机进行追击,在到达导弹的有效射程后对敌方进行火力打击。随着敌我双方距离的增加,机载雷达搜索目标的有效性和空空导弹对于敌方的打击效果也会大大降低,根据机载雷达和导弹的性能构建相应的距离优势函数如(2)式所示。

$$f_d=\begin{cases}0, & d\geqslant d_{R\max}\\ 0.5\mathrm{e}^{\left(-\frac{d-d_{M\max}}{d_{R\max}-d_{MK\max}}\right)}, & d_{M\max}<d\leqslant d_{R\max}\\ 2^{\left(-\frac{d-d_{MK\max}}{d_{M\max}-d_{MK\max}}\right)}, & d_{MK\max}<d\leqslant d_{M\max}\\ 1, & d_{MK\min}<d\leqslant d_{MK\max}\\ 0, & d\leqslant d_{MK\min}\end{cases}\tag{2}$$

1.1.3 速度优势函数

进行空战时,我方无人机的机动速度越大,则在攻击过程中可以更快地进入攻击范围,同时在受到敌机攻击时也能进行规避或逃离战场。如果速度过快,则会超过其最佳的作战速度范围,在近距空战时无人机的灵活性也将大大降低。因此需要综合考虑速度对于空战的影响,设置角速度优势函数如(3)式所示。

$$f_v=\begin{cases}\mathrm{e}^{\left(-\frac{V_{\mathrm{U}}-V_{\mathrm{Ubest}}}{V_{\mathrm{Ubest}}}\right)}, & 1.5V_{\mathrm{T}}<V_{\mathrm{U}}\\ 1, & V_{\mathrm{T}}<V_{\mathrm{U}}\leqslant 1.5V_{\mathrm{T}}\\ \dfrac{2V_{\mathrm{U}}}{V_{\mathrm{T}}}-1, & 0.5V_{\mathrm{T}}<V_{\mathrm{U}}\leqslant V_{\mathrm{T}}\\ 0, & V_{\mathrm{U}}\leqslant 0.5V_{\mathrm{T}}\end{cases}\tag{3}$$

式中:V_{U} 为我方无人机速度;V_{T} 为敌机速度。

1.1.4 高度优势函数

在空战过程中无人机一直处于高空环境中,当我方无人机相对于敌方较高且高度差处于最佳攻击高度范围($\Delta h_{M\min},\Delta h_{M\max}$)内时,其空战优势最大。但如果高度差过大,也会不利于无人机发动攻击。此时无人机的高度优势函数如(4)式所示。

$$f_h=\begin{cases}1, & \Delta h_{M\min}<\Delta H\leqslant\Delta h_{M\max}\\ \mathrm{e}^{-\frac{(\Delta H-\Delta h_{M\min})^2}{2(\Delta h_{M\max}-\Delta h_{M\min})^2}}, & \Delta H<\Delta h_{M\min}\\ \mathrm{e}^{-\frac{(\Delta H-\Delta h_{M\max})^2}{2(\Delta h_{M\max}-\Delta h_{M\min})^2}}, & \Delta H>\Delta h_{M\max}\end{cases}\tag{4}$$

1.1.5 几何态势函数

几何态势函数需要综合考虑两机间相对角度、距离、速度和高度对于空战态势的影响,构建几何空战态势优势函数如(5)式所示。

$$\begin{cases}f_A=2(w_1f_\varphi+w_2f_d+w_3f_v+w_4f_h)-1\\ w_1+w_2+w_3+w_4=1\end{cases}\tag{5}$$

式中,w_1,w_2,w_3 和 w_4 为各项优势函数占几何态势函数的权重,在空战中角度优势函数所占的权重比应大于其他优势函数占总函数的权重比。

1.2 稳定性优势函数

无人机机身的稳定性在空战中尤其重要,当迎角 α 和侧滑角 β 过大时会导致无人机失速并且降低操纵性。同时无人机在飞行过程中俯仰角和滚转角的急剧变化会导致震荡,影响稳定性,基于此构建无人机机身稳定性优势函数如(6)式所示。

$$f_B=\begin{cases}-0.1p-0.1q, & |\alpha|\leqslant 20^\circ\ \text{且}\ |\beta|\leqslant 30^\circ\\ -5, & \text{其他}\end{cases}\tag{6}$$

式中,p 和 q 分别表示我方无人机的滚转角速率和俯仰角速率。

1.3 作战优势函数

当导弹攻击目标时,能否命中目标受到多方面因素的限制。当无人机同目标的距离保持在最小攻击距离 $d_{M\min}$ 和最佳攻击距离 $d_{M\mathrm{best}}$ 之间,同时无人机方位角 φ_{U} 小于搜索方位角 $\varphi_{R\max}$,目标方位角 φ_{T} 小于 90°,我方无人机可以通过发射空空导弹命中目标。此时导弹优势函数如(7)式所示。

$$f_C=\begin{cases}10 & d_{M\min}<d\leqslant d_{M\mathrm{best}}\ \text{且}\ \varphi_{\mathrm{U}}<\varphi_{R\max}\ \text{且}\ \varphi_{\mathrm{T}}<90^\circ\\ 0 & \text{其他}\end{cases}\tag{7}$$

同时为了避免无人机在仿真学习过程中,与目标间的距离超过最大搜索距离 $d_{R\max}$ 导致丢失目标,或者飞行高度过低导致触地,构建环境优势函数如(8)式所示。

$$f_D=\begin{cases}-5 & d\geqslant d_{R\max}\ \text{或}\ H_{\mathrm{U}}\leqslant h_{\min}\\ 0 & \text{其他}\end{cases}\tag{8}$$

式中,H_{U} 为我方无人机的飞行高度。

1.4 综合优势函数

无人机优势函数建模由敌我双方空战几何态势函数、稳定性优势函数和作战优势函数组成。其中几何态势函数基于敌我双方空间占位的动态因素进行综合分析,是整个综合优势函数的基础。稳定性函数通过设置负的奖励值减少指令变化,避免飞行震荡。作战函数体现了最终空战的结果,即我方无人机在避免失速和坠毁的基础上,通过导弹攻击敌机并取得最终的胜利。将这几部分优势函数相加得到综合优势函数,将其作为奖励值应用于深度强化学习算法的训练中,如(9)式所示。

$$f=f_A+f_B+f_C+f_D\tag{9}$$

各优势函数的参数值如表 1 所示。

表 1　优势函数中各参数值

参数名	数值	参数名	数值
$d_{R\max}$/km	50	$d_{M\text{best}}$/km	3
$d_{M\max}$/km	10	$d_{M\min}$/km	1
$d_{MK\max}$/km	5	$d_{MK\min}$/km	1
$\Delta h_{M\max}$/km	1	$\Delta h_{M\min}$/km	0.5
$\varphi_{M\max}$/(°)	90	$h_{\min}$/km	0.5
w_1	0.4	w_2	0.25
w_3	0.1	w_4	0.25

2　无人机和机动预测模型

2.1　无人机六自由度方程

本文采用 F16 飞机模型的数据设计无人机,通过升降舵偏角、方向舵偏角、副翼偏角和油门(δ_e,δ_a,δ_r,δ_T)控制该模型,在惯性坐标系下,无人机的六自由度方程通常由动力学方程和运动学方程组成,可以得到无人机的非线性六自由度方程如(10)~(13)式所示。

力方程组

$$\begin{cases}\dot{V}=(u\dot{u}+v\dot{v}+w\dot{w})/V\\ \dot{\alpha}=\dfrac{u\dot{w}-w\dot{u}}{u^2+w^2}\\ \dot{\beta}=(\dot{v}V-v\dot{V})/(V^2\cos\beta)\end{cases}\tag{10}$$

式中:V,α 和 β 分别表示无人机的速度、迎角和侧滑角;$[u,v,w]^{\mathrm{T}}$ 为无人机的三轴速度分量。

力矩方程组

$$\begin{cases}\dot{p}=(c_1r+c_2p)q+c_3\bar{L}+c_4N\\ \dot{q}=c_5pr-c_6(p^2-r^2)+c_7M\\ \dot{r}=(c_8p-c_2r)q+c_4\bar{L}+c_9N\end{cases}\tag{11}$$

式中:p,q 和 r 分别表示无人机的滚转角速率、俯仰角速率和偏航角速率;$[\bar{L},\ M,\ N]^{\mathrm{T}}$ 为无人机在机体坐标系 3 个轴上合成力矩的分量。

运动方程组

$$\begin{cases}\dot{\phi}=p+\tan\theta(r\cos\phi+q\sin\phi)\\ \dot{\theta}=q\cos\phi-r\sin\phi\\ \dot{\psi}=(r\cos\phi+q\sin\phi)/\cos\theta\end{cases}\tag{12}$$

式中,ϕ,θ 和 ψ 分别表示滚转角、俯仰角和偏航角。

导航方程组

$$\begin{cases}\dot{x}=V\cos\gamma\cos\chi\\ \dot{y}=\cos\gamma\sin\chi\\ \dot{z}=-V\sin\gamma\end{cases}\tag{13}$$

式中,γ 和 χ 分别表示航迹倾斜角和航迹偏航角。

2.2　目标机动动作库

搭建无人机飞行控制系统如图 2 所示。图中的速度控制器通过动力系统控制无人机的飞行速度,航迹倾斜角控制器和滚转角控制器通过伺服机构控制无人机的姿态,由空战机动决策模块生成速度、航迹倾斜角和滚转角指令控制无人机进行机动。

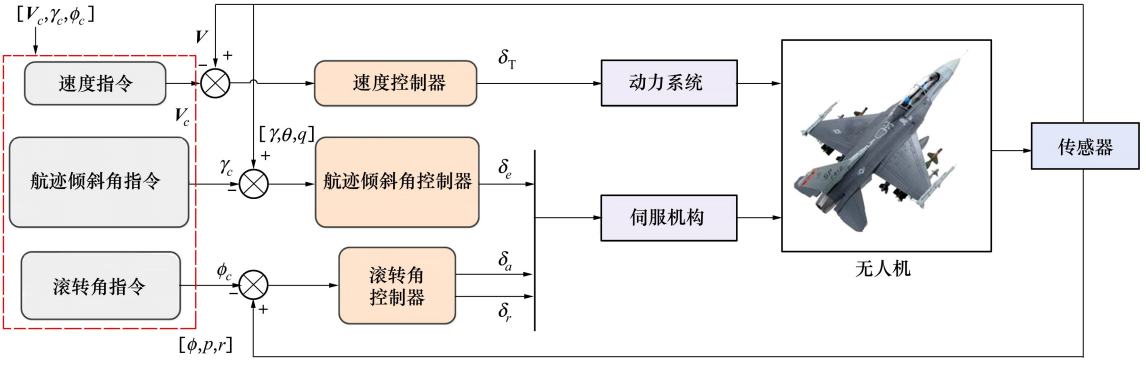

图 2　飞行控制系统

目标在空战中通过 DQN 算法生成相应的机动动作库指令,并通过机动动作库生成不同的动作指令。在高度方面控制航迹倾斜角,使其分别进行爬升、平飞和俯冲的机动;侧向上控制滚转角,使其分别进行左滚转、平飞和右滚转的机动;速度方面使其加速、减速或保持当前的速度,因此总计有 27 种不同的机动动作可供选择。

航迹倾斜角指令

$$\boldsymbol{\gamma}_c=[\gamma_{\min},\gamma_a,\gamma_{\max}]\tag{14}$$

式中,$\boldsymbol{\gamma}_c$,$\gamma_{\min}$,γ_a 和 $\gamma_{\max}$ 分别表示给定航迹倾斜角指令、最小航迹倾斜角、当前航迹倾斜角和最大航迹倾斜角值。

滚转角指令

$$\boldsymbol{\phi}_c = [\phi_{\text{left}}, \phi_a, \phi_{\text{right}}] \tag{15}$$

式中，$\boldsymbol{\phi}_c$，ϕ_{left}，ϕ_a 和 ϕ_{right} 分别表示给定滚转角指令、左滚转、当前滚转角和右滚转值。

速度指令

$$\boldsymbol{V}_c = [V_{\min}, V_a, V_{\max}] \tag{16}$$

式中，$\boldsymbol{V}_c$，$V_{\min}$，V_a 和 $V_{\max}$ 分别表示给定速度指令、最小速度、当前速度和最大速度值。

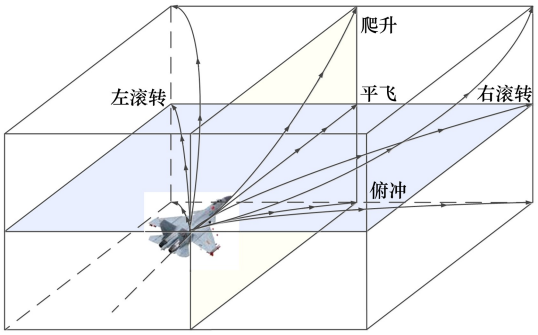

图 3 机动动作库

机动动作库的动作指令参数如表 2 所示。

表 2 机动动作库中各参数值

参数名	数值	参数名	数值
$\gamma_{\min}/(°)$	−30	$\gamma_{\max}/(°)$	30
$\phi_{\text{left}}/(°)$	−75	$\phi_{\text{right}}/(°)$	75
$V_{\min}/(\text{m}\cdot\text{s}^{-1})$	150	$V_{\max}/(\text{m}\cdot\text{s}^{-1})$	350

2.3 机动预测

概率神经网络是径向基函数神经网络的一个分支，由输入层、隐含层、求和层和输出层组成。首先通过输入层将特征向量传入网络，在隐含层中计算欧氏距离，并采用高斯函数作为网络的传递函数。之后在求和层处将隐含层所得到的所有同类神经元的输出累加求和再取平均，最后通过竞争函数取出分值最高的特征作为预测结果。

采用概率神经网格构建预测模块，首先需要收集学习过程中的敌我两机飞行状态和空战态势，以及该态势下敌方根据机动动作库所做出的机动动作指令；在采集了一定数量的样本后建立预测模块；最后在实际应用中通过该预测模块以及两机空战态势对敌机的机动指令选择进行预测。输入该神经网络的特征向量包含 10 个变量如(17)式所示

$$\boldsymbol{s} = [\varphi_{\text{U}}, \varphi_{\text{T}}, \varphi_{\text{UT}}, \theta_{\text{U}}, \theta_{\text{T}}, V_{\text{U}}, V_{\text{T}}, d, H_{\text{U}}, \Delta H] \tag{17}$$

式中：φ_{UT} 为我机速度矢量和敌机速度矢量之间的夹角；θ_{U} 和 θ_{T} 分别为我机和敌机的俯仰角；H_{U} 为我方无人机当前的飞行高度。

将上述 10 个变量做归一化处理后作为特征向量输入概率神经网络模型。而神经网络的输出值为航迹倾斜角、滚转角和速度的指令类别所组成的 27 种机动动作指令值，即$\{1,2,\cdots,27\}$组成。

3 基于 DDPG 的空战机动决策设计

本文采用 DDPG 算法对无人机空战机动决策进行研究，因此首先介绍 DDPG 算法原理，然后基于 DDPG 算法，设计结合目标预测结果的无人机自主空战机动决策的相应过程。

3.1 DDPG 算法

DDPG 算法有 4 个网络分别为：动作(Actor)在线网络 Q、Actor 目标网络 Q'、判断(Critic)在线网络 μ 和 Critic 目标网络 μ'。令 θ^Q 和 $\theta^{Q'}$ 分别为 Critic 在线和目标网络的参数，θ^μ 和 $\theta^{\mu'}$ 分别为 Actor 在线和目标网络的参数。在 t 时刻，Actor 在线网络主要根据当前状态 S_t 选择相应的动作 a_t，同时环境进行更新从而形成新的状态 S_{t+1}；Critic 在线网络主要根据奖赏值 r_t 计算 Q 值，从而对动作的选择进行评估。

DDPG 算法采用了经验回放训练强化学习过程，通过一定量的数据样本(s_t, a_t, r_t, s_{t+1}) 打破了数据间的关系使得神经网络的训练收敛，同时采用了目标网络打破时分偏差。

对 Critic 在线网络的损失函数进行计算如(18)式所示。

$$L(\theta^Q) = \frac{1}{N}\sum_i^N (y_i - Q(s_i, a_i|\theta^Q))^2 \tag{18}$$

式中，$y_i = r_i + \sigma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$ 表示目标函数，σ 表示折扣系数。

更新 Actor 在线网络如(19)式所示。

$$\nabla_{\theta^\mu}\mu|_{s_i} \approx \frac{1}{N}\sum_i \nabla_{a_i}Q(s_i, a_i|\theta^Q)\,\nabla_{\theta^\mu}\mu(s_i|\theta^\mu) \tag{19}$$

目标函数的参数值更新如(20)式所示

$$\begin{cases}\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'} \\ \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}\end{cases} \tag{20}$$

3.2 空战机动决策

对于敌方无人机而言，需要根据两机态势信息从 DQN 自主空战机动决策模块中选择机动动作库中的机动指令并进行空战。而对于我方无人机而言，除了要考虑两机的空战态势信息外，还需要对敌

机的机动指令选择进行一定的预测,将预测结果与两机态势同时输入 DDPG 自主空战机动决策模块中如图 4 所示。

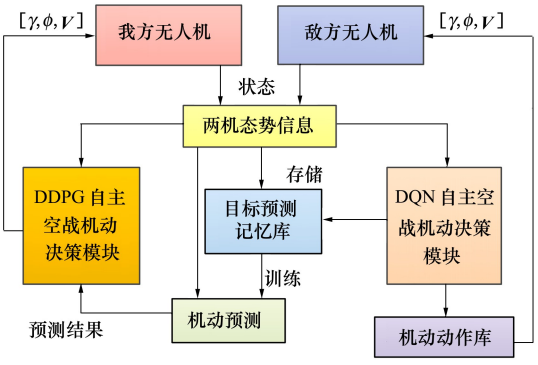

图 4　空战自主机动决策

图中,DDPG 自主空战机动决策模块生成的控制指令是连续的,分别为航迹倾斜角控制指令、滚转角控制指令和速度控制指令。由于 DDPG 算法能够解决连续动作空间问题,因此可以根据无人机当前飞行状态输出连续的航迹倾斜角、滚转角和速度的控制指令如(21)式所示。

$$[\boldsymbol{\gamma}_c,\boldsymbol{\phi}_c,\boldsymbol{V}_c]=[\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{V}]+[\Delta\boldsymbol{\gamma}_c,\Delta\boldsymbol{\phi}_c,\Delta\boldsymbol{V}_c]\cdot T_s \tag{21}$$

式中,T_s 为每次决策间的时间间隔,$[\Delta\boldsymbol{\gamma}_c,\Delta\boldsymbol{\phi}_c,\Delta\boldsymbol{V}_c]$ 分别为航迹倾斜角、滚转角和速度的变化率。其中,航迹倾斜角变化率范围为(-30°/s,30°/s),滚转角变化率范围为(-30°/s,30°/s),速度变化率范围为(-10 m/s,10 m/s)。通过对这 3 个机动指令进行控制实现我方无人机的自主空战机动决策。

4　仿真分析

4.1　仿真步骤

空战决策训练时,首先在 Matlab/Simulink 中搭建敌我两机非线性模型、机动动作库、两机空战态势信息收集和空战自主机动决策等模块。之后设置每次训练的轮数和每轮的最大持续时间,每当我机判定其导弹击中敌机、被敌机导弹所击中、到达持续回合时间、丢失目标或飞行高度过低时,结束该轮训练重新进入下一轮并重置仿真环境。

DDPG 和目标机动指令预测相结合的无人机空战机动决策算法的具体步骤如算法 1 所示。

算法 1　DDPG 和预测相结合的无人机空战机动决策

1. 初始化记忆回放单元 D 和目标预测记忆库 D',单次学习的样本数 m 和随机噪声 N_t
2. 初始化 Critic 在线网络 $Q(\boldsymbol{s},a|\theta^Q)$ 和目标网络 Q',随机生成参数 θ^Q 和 $\theta^{Q'}=\theta^Q$;初始化 Actor 在线网络 $\mu(s|\theta^\mu)$ 和目标网络 μ',随机生成参数 θ^μ 和 $\theta^{\mu'}=\theta^\mu$
3. 生成目标机动动作库,训练目标机动决策模块,将训练过程中的两机态势和敌机机动指令存入目标预测记忆库 D',从中选取一定数据作为训练集搭建预测模块
4. for episode = 1,2, …,M do
5. 　初始化敌我双方无人机的状态,敌机获取当前两机态势 s_t' 从机动动作库中选择机动动作指令 a_t'
6. 　for step = 1,2,…,T do
7. 　　我机根据两机态势 $\boldsymbol{s}_t'$ 预测敌机的机动动作指令为 a_t'',令 $\boldsymbol{s}_t=[s_t',a_t'']$,在 Actor 在线网络中生成随机动作策略 $a_t=\mu(\boldsymbol{s}_t|\theta^\mu)+N_t$
8. 　　分别执行敌我双方的动作 a_t 和 a_t',得到奖励值 r_t 及新的两机态势 s_{t+1}',并预测敌机的下一步机动动作为 a_{t+1}'',令 $\boldsymbol{s}_{t+1}=[s_{t+1}',a_{t+1}'']$
9. 　　将数据样本 $(\boldsymbol{s}_t,a_t,r_t,\boldsymbol{s}_{t+1})$ 存入 D 中
10. 　　从 D 中随机抽取一批样本 $(\boldsymbol{s}_i,a_i,r_i,\boldsymbol{s}_{i+1})$
11. 　　令 $y_t=r_t+\sigma Q'(\boldsymbol{s}_{i+1},\mu'(\boldsymbol{s}_{i+1}|\theta^{\mu'})|\theta^{Q'})$
12. 　　根据目标函数 $(y_i-Q(\boldsymbol{s}_i,a_i|\theta^Q))^2$ 对 Critic 在线网络使用梯度下降法进行更新
13. 　　使用随机策略梯度对 Actor 在线网络进行更新:

$$\nabla_{\theta^\mu}\boldsymbol{\mu}|_{s_i}\approx\frac{1}{N}\sum_i\nabla_{a_i}Q(\boldsymbol{s}_i,a_i|\theta^Q)\nabla_{\theta^\mu}\mu(\boldsymbol{s}_i|\theta^\mu)$$

14. 　　更新 Critic 目标网络 Q' 和 Actor 目标网络 μ':

$$\theta^{Q'}\leftarrow\tau\theta^Q+(1-\tau)\theta^{Q'}$$
$$\theta^{\mu'}\leftarrow\tau\theta^\mu+(1-\tau)\theta^{\mu'}$$

15. 　end for
16. end for

4.2　仿真环境设置

设置空战决策训练中单次训练轮数为 M = 5 000,每轮最大的持续时间为 T=200 s,DDPG 自主空战机动决策模块的决策时间间隔为 0.2 s,DQN 自

主空战机动决策模块的决策时间间隔为 1 s。

在 DDPG 自主空战机动决策模块中，Critic 网络和 Actor 网络都由双层全连接神经网络组成，隐含层大小分别为 1 024 和 512，Critic 网络和 Actor 网络的学习率分别为 0.01 和 0.005，隐藏层激活函数为 ReLU。令单次学习的样本数 $m=128$，随机噪声方差为 0.4，每步方差衰减率为 10^{-5}，平滑因子 $\tau=0.001$，记忆回放区的大小为 10^6。

DQN 决策模块的在线 Q 网络也由双层全连接神经网络组成，隐含层大小分别为 1 024 和 512，使用 Tanh 函数作为激活函数，输出层为 Purelin 函数。

4.3 仿真训练

每次训练时我方无人机处于固定的位置，同时初始速度和初始方向也固定，而敌方无人机的位置随机，两机的初始状态如表 3 所示。

表 3 训练时敌我双方无人机初始状态

初始状态	我方无人机	敌方无人机
x/m	0	(−15 000,15 000)
y/m	0	(−15 000,15 000)
h/m	3 000	(2 000,4 000)
v/(m·s^{-1})	200	(150,350)
ψ/(°)	0	(0,360)

设我方飞机为红机，敌方飞机为蓝机。首先令红机做匀速直线运动，蓝机随机生成初始状态，并对蓝机的 DQN 自主空战机动决策模块进行训练，一共训练 5 000 轮，之后从记忆库中随机选取 5 000 组数据作为训练集生成预测模块。令蓝机根据之前训练所得到的 DQN 自主空战机动决策模块从机动动作库中生成机动指令，而红机分别对含目标机动指令预测和不含预测的 DDPG 算法进行训练，分别训练 5 000 轮，耗时 29.5 h 和 27.2 h。两者每轮训练生成的回报值随训练轮数的变化如图 5 所示。

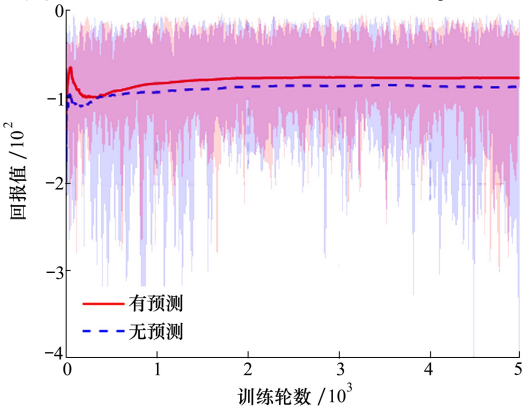

图 5 回报值

图 5 中阴影部分为每轮训练生成的回报值，红线为有预测的平均回报值，蓝线为无预测的平均回报值，两者最终都趋于收敛。

随机生成 1 000 个测试集进行测试，两机的初始状态如表 3 所示，对比是否采用机动指令预测对于无人机胜率的影响。测试结果显示，有预测的无人机胜率为 94.0%，无预测的无人机胜率为91.3%，可以看出通过对目标机动指令进行预测可以提升 DDPG 算法训练过程中的平均回报值和最终的胜率。

通过设计 3 个案例来观察含预测的 DDPG 自主空战机动决策模块的训练结果。

1）算例一中令敌机做匀速直线运动，敌机初始位置 $(x_T,y_T,h_T)=(5\,000,5\,000,3\,000)$ m，速度为 200 m/s，航向角为 0°，两机空战轨迹如图 6 所示。

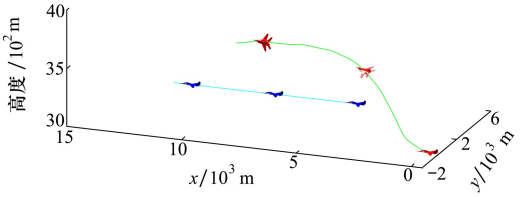

图 6 算例一中两机空战轨迹

我机首先向右飞行并爬升，在接近了敌机后向左滚转以最大程度拉进敌方与我方的距离，此时无人机同敌机的距离保持在最佳攻击距离内，同时位于敌机的后方，这时可以认为我机能成功击落敌机。

2）算例二中令敌机做右盘旋运动时，航迹倾斜角为 20°，滚转角为 30°。设敌机初始位置为 $(x_T,y_T,h_T)=(0,5\,000,3\,000)$ m，速度为 200 m/s，航向角为 180°，两机空战轨迹如图 7 所示。

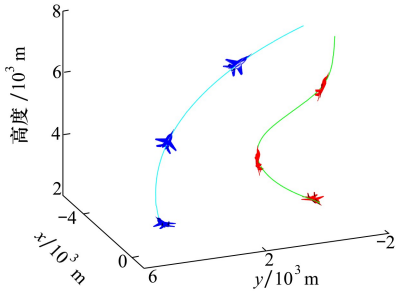

图 7 算例二中两机空战轨迹

我方无人机首先向右飞行并爬升，在达到角度的优势后保持爬升姿态，向左滚转令机头朝向敌机，一定时间后达到攻击距离判定此时我方无人机的导

弹可以击中敌机。

3) 算例三中令敌机通过 DQN 算法生成机动动作库指令从而进行空战,设敌机初始位置为 $(x_T, y_T, h_T) = (0, 7\,000, 3\,000)$ m,初始速度为 200 m/s,航向角为 180°,两机空战轨迹如图 8 所示。

图 9 为空战中敌我两机的滚转角和速度的值随时间的变化,结合空战轨迹可以看出我机一开始以速度为代价,向右滚转并爬升,在一开始就达到了相对优势的位置,之后一直保持这种优势直到导弹锁定目标并判定击中。

从上述几个算例可以看出该算法可以有效应用于空战决策中,同时我方无人机在飞行过程中姿态角不会发生急剧变化,能够平稳持续飞行。

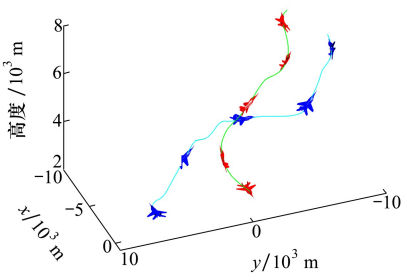

图 8　算例三中两机空战轨迹

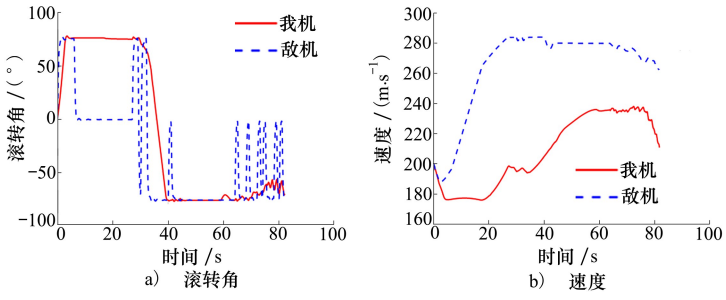

图 9　算例三中两机态势图

5　结　论

本文研究了一种预测和 DDPG 结合的无人机空战机动决策过程。

1) 除了分析角度、速度、高度以及敌我双方距离对空战因素的影响,还综合考虑了非线性无人机机身的稳定性、导弹性能以及其在环境中所处的方位对空战综合态势的影响。这一方面加强了综合态势评估结果的可靠性,另一方面也方便无人机在训练过程中保持稳定防止失速,有利于工程实践。

2) 采用概率神经网络搭建目标预测模块,通过研究两机空战态势和敌机的机动动作库,使得我方无人机对敌机的机动进行一定的预测并应用于空战决策中,可以提高无人机的空战能力。

3) 由于 DDPG 算法可以生成连续的动作指令,与 DQN 算法相比在空战中无人机的机动性更强。实验证明,该算法具有收敛性和空战可靠性,这对于无人机在空战中的实际应用具备一定的参考意义。

参考文献:

[1] EHTAMO H, RAIVIO T. On Applied nonlinear and bilevel programming for some pursuit-evasion games[J]. Journal of Optimization Theory and Applications, 2001, 108(1): 65-96

[2] 顾佼佼, 赵建军, 刘卫华. 基于博弈论及 Memetic 算法求解的空战机动决策框架[J]. 电光与控制, 2015, 22(1): 20-23
GU Jiaojiao, ZHAO Jianjun, LIU Weihua. Air combat maneuvering decision framework based on game theory and memetic algorithm[J]. Electronics Optics & Control, 2015, 22(1): 20-23 (in Chinese)

[3] 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[J]. 电光与控制, 2009, 16(7): 13-17
WAN Wei, JIANG Changsheng, WU Qingxian. Application of one-step prediction influence diagram in air combat maneuvering decision[J]. Electronics Optics & Control, 2009, 16(7): 13-17 (in Chinese)

[4] KUMAR S, JAIN S, KUMAR H. Prediction of jatropha-algae biodiesel blend oil yield with the application of artificial neural networks technique[J]. Energy Sources, 2018, 41(7/8/9/10/11/12): 1285-1295

[5] SMITH R E, DIKE B A, MEHRA R K. Classifier systems in combat: two-sided learning of maneuvers for advanced fighter aircraft[J]. Computer Methods in Applied Mechanics and Engineering, 2000, 186(2/3/4): 421-437

[6] 丁林静, 杨啟明. 基于强化学习的无人机空战机动决策[J]. 火力与指挥控制, 2018,49(2): 29-35
DING Linjing, YANG Qiming. Research on air combat maneuver decision of UAVs based on reinforcement learning[J]. Avionics Technology, 2018, 49(2): 29-35 (in Chinese)

[7] YANG Q, ZHANG J, SHI G, et al. Maneuver decision of UAV in short-range air combat based on deep reinforcement learning

[J]. IEEE Access, 2020, 8: 363-378
[8] BAI S, SONG S, LIANG S, et al. UAV maneuvering decision-making algorithm based on twin delayed deep deterministic policy gradient algorithm[J]. Journal of Artificial Intelligence and Technology, 2022, 2(1): 16-22
[9] LI B, YANG Z P, CHEN D Q, et al. Maneuvering target tracking of UAV based on MN-DDPG and transfer learning[J]. Defence Technology, 2021, 17(2): 457-466
[10] WANG L, HU J, XU Z, et al. Autonomous maneuver strategy of swarm air combat based on DDPG[J]. Journal of Artificial Intelligence and Technology, 2021, 1(1): 232-243
[11] ZHANG J, YANG Q, SHI G, et al. UAV cooperative air combat maneuver decision based on multi-agent reinforcement learning [J]. Journal of Systems Engineering & Electronics, 2021, 32(6): 1421-1438
[12] 韩占朋, 王玉惠, 程聪. 态势估计方法研究综述[J]. 航空兵器, 2013(1): 14-19
HAN Zhanpeng, WANG Yuhui, CHENG Cong. Summary on situation assessment method research[J]. Aero Weaponry, 2013 (1): 14-19 (in Chinese)
[13] 毛梦月, 张安, 周鼎, 等. 基于机动预测的强化学习无人机空中格斗研究[J]. 电光与控制, 2019, 26(2): 5-10
MAO Mengyue, ZHANG An, ZHOU Ding, et al. Reinforcement learning of UCAV air combat based on maneuver prediction[J]. Electronics Optics and Control, 2019, 26(2): 5-10 (in Chinese)

UAV′s air combat decision-making based on deep deterministic policy gradient and prediction

LI Yongfeng[1], LYU Yongxi[1,2], SHI Jingping[1,2], LI Weihua[1]

(1.School of Automation, Northwestern Polytechnical University, Xi′an 710129, China;
2.Shaanxi Provincial Key Laboratory of Flight Control and Simulation Technology, Xi′an 710129, China)

Abstract: To solve the enemy uncertain manipulation problem during a UAV′s autonomous air combat maneuver decision-making, this paper proposes an autonomous air combat maneuver decision-making method that combines target maneuver command prediction with the deep deterministic policy algorithm. The situation data of both sides of air combat are effectively fused and processed, the UAV′s six-degree-of-freedom model and maneuver library are built. In air combat, the target generates its corresponding maneuver library instructions through the deep Q network algorithm; at the same time, the UAV on our side gives the target maneuver prediction results through the probabilistic neural network. A deep deterministic policy gradient reinforcement learning method that considers both the situation information of two aircraft and the prediction results of enemy aircraft is proposed, so that the UAV can choose the appropriate maneuver decision according to the current air combat situation. The simulation results show that the method can effectively use the air combat situation information and target maneuver prediction information so that it can improve the effectiveness of the reinforcement learning method for UAV′s autonomous air combat decision-making on the premise of ensuring convergence.

Keywords: UAV; air combat maneuver decision-making; prediction; deep deterministic policy gradient

引用格式: *李永丰, 吕永玺, 史静平, 等. 深度确定性策略梯度和预测相结合的无人机空战决策研究[J]. 西北工业大学学报, 2023,41(1): 56-64*

LI Yongfeng, LYU Yongxi, SHI Jingping, et al. UAV′s air combat decision-making based on deep deterministic policy gradient and prediction[J]. Journal of Northwestern Polytechnical University, 2023, 41(1): 56-64 (in Chinese)