DOI:10.11918/202005108

深度强化学习的无人作战飞机空战机动决策

李永丰[1],史静平[1,2],章卫国[1,2],蒋 维[1]

(1. 西北工业大学 自动化学院,西安 710029;2. 陕西省飞行控制与仿真技术重点实验室(西北工业大学),西安 710029)

摘 要:无人作战飞机(unmanned combat aerial vehicle,UCAV)在进行空战自主机动决策时,面临大规模计算,易受敌方不确定性操纵的影响。针对这一问题,提出了一种基于深度强化学习算法的无人作战飞机空战自主机动决策模型。利用该算法,无人作战飞机可以在空战中自主地进行机动决策以获得优势地位。首先,基于飞机控制系统,利用 MATLAB/Simulink 仿真平台搭建了六自由度无人作战飞机模型,选取适当的空战动作作为机动输出。在此基础上,设计了无人作战飞机空战自主机动的决策模型,通过敌我双方的相对运动构建作战评估模型,分析了导弹攻击区的范围,将相应的优势函数作为深度强化学习的评判依据。之后,对无人作战飞机进行了由易到难的分阶段训练,并通过对深度 Q 网络的研究分析了最优机动控制指令。从而无人作战飞机可以在不同的态势情况下选择相应的机动动作,独立评估战场态势,做出战术决策,以达到提高作战效能的目的。仿真结果表明,该方法能使无人作战飞机在空战中自主的选择战术动作,快速达到优势地位,极大地提高了无人作战飞机的作战效率。

关键词:无人作战飞机;深度强化学习;空战自主机动决策;六自由度;优势函数;深度 Q 网络

中图分类号:V279 **文献标志码**:A **文章编号**:0367-6234(2021)12-0033-09

Maneuver decision of UCAV in air combat based on deep reinforcement learning

LI Yongfeng[1], SHI Jingping[1,2], ZHANG Weiguo[1,2], JIANG Wei[1]

(1. School of Automation, Northwestern Polytechnical University, Xi'an 710029, China; 2. Shaanxi Provincial Key Laboratory of Flight Control and Simulation Technology (Northwestern Polytechnical University), Xi'an 710029, China)

Abstract: When an unmanned combat aerial vehicle (UCAV) is making the decision of autonomous maneuver in air combat, it faces large-scale calculation and is susceptible to the uncertain manipulation of the enemy. To tackle such problems, a decision-making model for autonomous maneuver of UCAV in air combat was proposed based on deep reinforcement learning algorithm in this study. With this algorithm, the UCAV can autonomously make maneuver decisions during air combat to achieve dominant position. First, based on the aircraft control system, a six-degree-of-freedom UCAV model was built using MATLAB/Simulink simulation platform, and the appropriate air combat action was selected as the maneuver output. On this basis, the decision-making model for the autonomous maneuver of UCAV in air combat was designed. Through the relative movement of both sides, the operational evaluation model was constructed. The range of the missile attack area was analyzed, and the corresponding advantage function was taken as the evaluation basis of the deep reinforcement learning. Then, the UCAV was trained by stages from the easy to the difficult, and the optimal maneuver control command was analyzed by investigating the deep Q network. Thereby, the UCAV could select corresponding maneuver actions in different situations and evaluate the battlefield situation independently, making tactical decisions and achieving the purpose of improving combat effectiveness. Simulation results suggest that the proposed method can make UCAV choose the tactical action independently in air combat and reach the dominant position quickly, which greatly improves the combat efficiency of the UCAV.

Keywords: unmanned combat aerial vehicle (UCAV); deep reinforcement learning; autonomous maneuver decision in air combat; six-degree-of-freedom; advantage function; deep Q network

收稿日期:2020-05-22

基金项目:国家自然科学基金(62173277,62073266,61573286);
陕西省自然科学基金(2019JM-163,2020JQ-218)

作者简介:李永丰(1995—),男,博士研究生;
史静平(1980—),男,教授,博士生导师;
章卫国(1956—),男,教授,博士生导师

通信作者:史静平,2017100622@mail.nwpu.edu.cn

目前无人作战飞机(unmanned combat aerial vehicle, UCAV)被广泛应用于军事领域[1],UCAV在过去主要从事战场监视、吸引火力和通信中继等任务,随着武器装备的传感器、计算机及通信等技术的发展,性能不断提升,未来的 UCAV 将逐步升级成为可以执行空中对抗、对地火力压制和参与制空

权的夺取等作战任务的主要作战装备之一。尽管 UCAV 的性能提升很大,但大多数的任务都离不开人工干预,控制人员通过基站在地面对 UCAV 进行控制,这种控制方法有延迟且易受到电磁干扰。因此研究 UCAV 的自主作战能力已经成为空军发展的必然趋势,装备了无人作战决策系统的 UCAV 将逐步取代飞行员的位置,以达到减少成本,提高战斗力的作用。在近距离格斗的阶段,UCAV 应根据当前的空战态势及时选取合适的飞行控制指令,抢占有利的位置,寻找击落敌机的机会并保护自己[2]。

在空战条件下,飞机模型本身为非线性同时目标的飞行轨迹是不确定的,这些都将给 UCAV 的机动决策带来许多不便,因此良好的机动决策是 UCAV 自主空战的一个重要环节,自动机动决策要求 UCAV 能在不同的空战环境下自动生成飞行控制指令。常规的机动决策控制方法包括最优化方法、博弈论法、矩阵对策法、影响图法、遗传算法、专家系统、神经网络方法以及强化学习方法等。文献[3]将空战视为一个马尔可夫过程,通过贝叶斯推理理论计算空战情况,并自适应调整机动决策因素的权重,使目标函数更加合理,保证了无人战斗机的优越性。文献[4]设计了一个基于遗传学习系统的飞机机动决策模型,通过对机动的过程加以优化来解决空战环境未知情况下的空战决策问题,可以在不同的空战环境中产生相应的战术动作,但该方法的参数设计存在主观性,不能灵活应用。文献[5]利用统计学原理研究 UCAV 的空战机动决策问题,具有一定的鲁棒性,但该算法实时性能较差无法应用于在线决策。文献[6]将可微态势函数应用于 UCAV 微分对策中,可以快速反应空战环境,但由于实时计算的局限性很难解决复杂的模型。文献[7]采用博弈论对 UCAV 空战决策进行建模,对不同的空战环境具有通用性。虽然这些决策算法可以在一定程度上提高决策的效率、鲁棒性和寻优率,但由于这些决策模型存在推理过程较为频繁,会浪费大量时间寻优等问题,导致 UCAV 的响应变慢,并不适用于当今的战场环境。

基于人工智能的方法包括神经网络法、专家系统法以及强化学习算法。文献[8]采用了专家系统法,通过预测双方的态势和运动状态生成相应的机动指令控制 UCAV 飞行,但不足之处在于规则库的构建较为复杂,通用性差。文献[9]采用了自适应神经网络技术设计 PID 控制器,对高机动目标具有较强的跟踪精度,但神经网络方法需要大量的空战样本,存在学习样本不足的问题。与以上两种方法相比,强化学习算法是一种智能体与环境之间不断试错交互从而进行学习的行为,智能体根据环境得到的反馈优化自己的策略,再根据策略行动,最终达到最优策略。由于强化学习的过程通常不考虑训练样本,仅通过环境反馈得到的奖励对动作进行优化,可以提高了学习的效率,是一种可行的方法[10]。文献[11]将空战时的状态空间模糊化、归一化作为强化学习算法的输入,并将基本的空战动作作为强化学习的输出,使得 UCAV 不断与环境交互从而实现空战的优势地位。在此基础上,文献[12 – 13]将神经网络与强化学习相结合,提高了算法的运算效率,但这些文章都没有考虑飞机的姿态变化。

本文提出了一种深度强化学习(deep reinforcement learning, DRL)算法来解决 UCAV 自主机动决策作战的问题,并在 MATLAB/Simulink 环境中搭建了某种六自由度 UCAV 模型,充分考虑了其非线性。同时选取适当的空战动作作为 UCAV 的机动输出,建立空战优势函数并设计 UCAV 空战机动决策模型。通过强化学习方法可以减少人为操纵的复杂性,保证计算结果的优越性,提高 UCAV 的作战能力,而神经网络可以提升实时决策能力。最后通过仿真将该方法应用于 UCAV 机动作战决策中,证明了其有效性和可行性。

1 UCAV 运动学建模

1.1 UCAV 运动模型

本文所研究的 UCAV 运动模型如图 1 所示,在研究 UCAV 运动时,把 UCAV 视为左右对称的理想刚体,其运动主要表现为速度及 3 个姿态角的变化情况,对 UCAV 的操纵主要依赖于发动机推力以及 UCAV 的气动舵面。采用六自由度方程描述 UCAV 在机动决策和仿真时的运动状态,具体参数如下:质量为 3.93 kg,机长为 1.47 m,机翼面积为 0.264 5 m^2,翼展长为 0.89 m,平均气动弦长为 0.336 m。

图 1 UCAV 模型图

Fig. 1 UCAV model diagram

1.2　运动学方程

在惯性坐标系当中，无人机六自由度方程通常可以描述为机体坐标系下的力方程组、力矩方程组、运动方程组与导航方程组，无人机六自由度方程的通常状态变量是：$[V,\alpha,\beta,p,q,r,\phi,\theta,\psi,x,y,z]$。对于无人机 12 个状态量的非线性六自由度方程如下（欧式坐标系）：

$$\dot{V}=(u\dot{u}+v\dot{v}+w\dot{w})/V \tag{1}$$

$$\dot{\alpha}=(u\dot{w}-w\dot{u})/(u^2+w^2) \tag{2}$$

$$\dot{\beta}=(\dot{v}V-v\dot{V})/(V^2\cos\beta) \tag{3}$$

$$\dot{\phi}=p+\tan\theta(r\cos\phi+q\sin\phi) \tag{4}$$

$$\dot{\theta}=q\cos\phi-r\sin\phi \tag{5}$$

$$\dot{\psi}=(r\cos\phi+q\sin\phi)/\cos\theta \tag{6}$$

$$\dot{p}=(c_1r+c_2p)q+c_3L+c_4N \tag{7}$$

$$\dot{q}=c_5pr-c_6(p^2-r^2)+c_7M \tag{8}$$

$$\dot{r}=(c_8p-c_2r)q+c_4L+c_9N \tag{9}$$

$$\dot{x}=u\cos\theta\cos\psi+v(\sin\phi\sin\theta\cos\psi-\cos\phi\sin\psi)+w(\sin\phi\sin\psi+\cos\phi\sin\theta\cos\psi) \tag{10}$$

$$\dot{y}=u\cos\theta\sin\psi+v(\sin\phi\sin\theta\sin\psi+\cos\phi\cos\psi)+w(-\sin\phi\cos\psi+\cos\phi\sin\theta\sin\psi) \tag{11}$$

$$\dot{z}=-u\sin\theta+v\sin\phi\cos\theta+w\cos\phi\cos\theta \tag{12}$$

式中$[u,v,w]$为机体系 3 个轴上的速度分量。

本文根据上述 UCAV 非线性模型，使用 PID 算法设置控制律，同时考虑姿态对 UCAV 空战决策的影响，搭建基本操纵动作库，选取适当的机动动作作为 UCAV 的输出。之后通过深度强化学习算法得到 UCAV 在不同的态势下的机动动作，使得该算法能对 UCAV 进行精准控制。

2　空战机动决策模型

2.1　方案

通过对 UCAV 自主战术决策系统进行研究，可以使 UCAV 具备更高的自主性，能独立应对突发事件以提高任务执行的效率，同时可以提高系统适应环境的能力。图 2 为 UCAV 自主决策模块，将我方无 UCAV 和目标的态势估计进行综合评价，输入机动决策模块中，得到机动库的控制指令，再对我方 UCAV 进行控制。

2.2　强化学习原理

强化学习算法主要由以下 5 个部分组成：智能体、环境、状态 S、动作 A 和观测回报 R。在时间 t 时刻，智能体会产生动作 A_t 并与环境之间进行交互，在动作执行后，智能体的状态由 S_t 转移成 S_{t+1}，并得到环境的回报值 R_t。就这样，智能体在与环境的交互中不断修改自身参数，在经过多次运算后得到最优解，如图 3 所示。

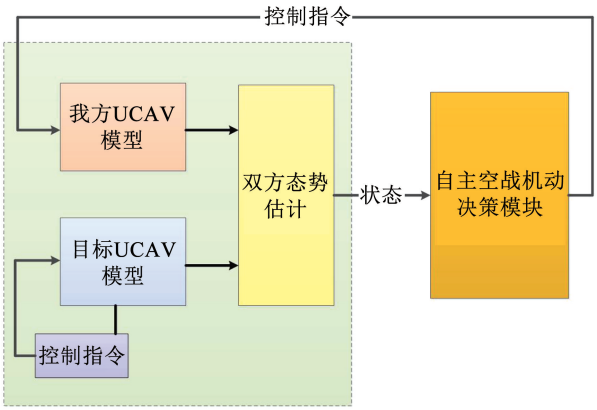

图 2　UCAV 自主决策模块

Fig. 2　UCAV autonomous decision module

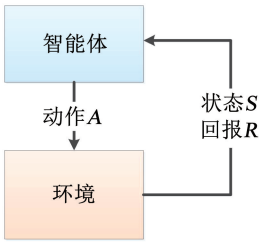

图 3　强化学习基本框架图

Fig. 3　Basic framework of reinforcement learning

强化学习的计算过程就是一个不断探索最优策略的过程，策略指的是状态到动作的映射，通过符号 π' 表示，下式为状态 S 下所对应的每个动作的概率，即

$$\pi'(a|s)=p[A_t=a|S_t=s] \tag{13}$$

对于强化学习算法而言，希望每一个状态所对应的动作都能使其价值最大化，需要找到策略：

$$\pi'_*(a|s)=\underset{a\in A}{\operatorname{argmax}}\,Q_*(a|s) \tag{14}$$

Q 强化学习算法是对状态 - 动作对的值 $Q(s,a)$ 进行迭代，在学习过程中选择动作 a 时，即

$$Q(s_t,a_t)=r_t+\gamma\sum_{s_{t+1}}P_{a_t}(s_t,s_{t+1})\max_a Q(s_{t+1},a) \tag{15}$$

$Q(s_t,a_t)$ 的更新公式为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\delta[r_t+\gamma\max_{a\in A(s)}Q(s_{t+1},a)-Q(s_t,a_t)] \tag{16}$$

式中：δ 为学习率，γ 为折扣率，r_t 为 t 时刻的综合优势函数。

可以看出该 Q 强化学习算法包含了综合优势函数和所选动作后的状态值，具有远视性，长期看来无限趋于稳定。

2.3　优势函数

对于空中格斗决策来说，将我方 UCAV 和目标之间的瞬时空中态势作为一个奖惩信号，构建相应

的空战优势函数,可以使得决策系统选择合适的机动动作,提高我机对敌机的空战优势。通常来说,传统的环境奖赏包括方位角奖赏、速度奖赏、距离奖赏和高度奖赏,并由这几部分加权得到综合空战态势评估值,但这种态势评估的加权值都是主观值,无法准确适应不同的空战武器。为解决该问题,本文针对 UCAV 空对空导弹的攻击方式设计了相应的优势函数[14]。典型的空空导弹攻击区间是攻击机的前方一定距离和角度的锥形范围,如图 4 所示。

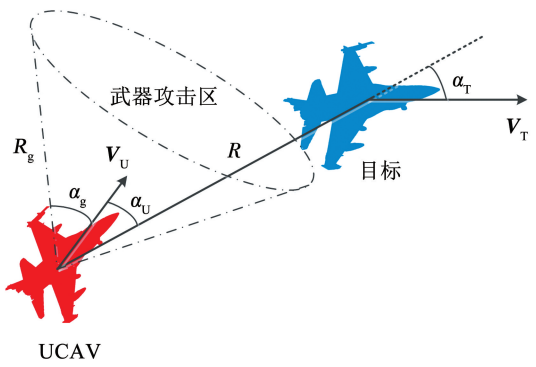

图 4 空战态势

Fig. 4 Air combat situation

图 4 中:$\boldsymbol{V}_U$、$\boldsymbol{V}_T$ 分别为 UCAV 和目标的速度向量,令 V_U、V_T 分别为向量 $\boldsymbol{V}_U$、$\boldsymbol{V}_T$ 的速度大小;R 为 UCAV 和目标之间的距离;α_U、α_T 分别为 UCAV 和目标之间的连线和各自速度向量的夹角;R_g 为 UCAV 导弹的最大攻击距离;α_g 为 UCAV 导弹的最大攻击角度。

在空战环境中,追尾飞机处于优势状态,被追飞机处于劣势状态,两机相向相背或同向飞行时则处于均势状态,本文通过两机方位角计算角度优势:

$$f_\alpha(\alpha)=\frac{2\pi-\alpha_U-\alpha_T}{2\pi} \tag{17}$$

对空空导弹而言,命中率主要与距离有关,为了使距离参数函数对于距离的变化不敏感,从而使无人机决策具有鲁棒性,计算方位角、距离优势的函数为

$$\eta_A=\begin{cases}\dfrac{2\pi-\alpha_U-\alpha_T}{2\pi}, R\leqslant R_g\\ \dfrac{2\pi-\alpha_U-\alpha_T}{2\pi}e^{-\frac{(R-R_g)^2}{2\sigma_R^2}}, R>R_g\end{cases} \tag{18}$$

式中 σ_R 为距离标准偏差。

如果 UCAV 与目标之间的距离小于导弹攻击距离,UCAV 速度矢量与两者间距离矢量的夹角小于 UCAV 导弹的攻击角度,同时目标的速度矢量与两者间距离矢量的夹角小于 90°。则说明目标处于 UCAV 武器的攻击范围内,可以发射导弹并拦截,并结束这个仿真回合进入下一个回合。此时 UCAV 的奖赏值为

$$\eta_U=\begin{cases}5, R\leqslant R_g, \alpha_U<\alpha_g, \alpha_T<\pi/2\\ 0, \text{其他}\end{cases} \tag{19}$$

当满足式(19)中的条件时,UCAV 得到奖赏值,同时为了训练 UCAV 规避敌机的攻击,目标也存在攻击武器,当目标满足相同条件时说明我方处于劣势,得到负的奖赏值。

$$\eta_B=\eta_U-\eta_T \tag{20}$$

其中:

$$\eta_T=\begin{cases}5, R\leqslant R_g, \alpha_T<\alpha_g, \alpha_U<\pi/2\\ 0, \text{其他}\end{cases}$$

为了避免 UCAV 在飞行过程中失速、飞行过低或过高、远离目标或与目标发生碰撞,应限制 UCAV 的速度不小于 20 m/s,高度不小于 200 m,距离限制在[100 m,3 000 m]之间。

$$\eta_C=\begin{cases}-10, V_U<20\ \text{m/s}, H<200\ \text{m},\\ \qquad R<100\ \text{m}, R>3\ 000\ \text{m}\\ 0, \text{其他}\end{cases} \tag{21}$$

同时由于该 UCAV 为六自由度非线性模型,机动动作的选择不仅要考虑敌我态势,还需要考虑 UCAV 选择机动动作时的状态,使得 UCAV 基于当前态势所选择的机动动作可以完整的执行下去,避免 UCAV 的失控。对于固定翼飞机而言,三轴力和三轴力矩的大小与迎角和侧滑角相关,因此决定其失控与否和飞行品质的关键是气流角。在飞机做机动动作时,要避免其因惯性或扰动超出飞行包线从而导致飞机的失控,需要对气流角加以保护,可以将 UCAV 的迎角限制在[-20°, 20°]之间,侧滑角限制在[-30°, 30°]之间,当超出限制时给予负的奖励值,使得该决策机制可以避免选择造成 UCAV 失控的机动指令。

$$\eta_D=\begin{cases}-10, -\pi/9\leqslant\alpha\leqslant\pi/9, -\pi/6\leqslant\beta\leqslant\pi/6\\ 0, \text{其他}\end{cases} \tag{22}$$

由于单次空战为一个作战回合,最终结果会影响之前的空战动作,需要根据时间差给之前的步骤添加奖赏值,则综合优势函数为

$$\eta=\eta_A+0.95^{\Delta t}(\eta_B+\eta_C+\eta_D) \tag{23}$$

式中 Δt 为 t 时刻到这一作战回合结束的剩余时间。

则 t 时刻的综合优势函数为

$$r_t=\eta(t) \tag{24}$$

2.4 状态空间

由于该空战环境为三维空间,为了充分展现两机的飞行状态和空战态势,图 2 中输入自主空战机动决策模块的状态空间包含 10 个变量:

$$S=[\alpha_U,\alpha_T,\alpha_{UT},\theta_U,\theta_T,V_U,V_T,R,H_U,\Delta H] \tag{25}$$

式中：α_{UT}为 UCAV 速度矢量和目标速度矢量之间的夹角；θ_U、θ_T 分别为 UCAV 和目标的俯仰角；H_U 为 UCAV 的当前飞行高度；$\Delta H = H_U - H_T$ 为 UCAV 相对于目标的高度差。需要将上述状态空间做归一化处理后再输入神经网络模型。

2.5　基本机动动作库

空战机动动作库分为两类，一类是典型的战术动作库，另一类是基本的机动动作库。典型战术动作库包括眼镜蛇机动、榔头机动、螺旋爬升等，但这些战术动作本质上是各个基础动作组合而成的，同时特殊的机动动作目前必须依靠人机紧密配合，协调完成，否则 UCAV 的状态可能会超过正常的包线范围，导致 UCAV 有失控风险。因此本文采用美国国家航空航天局提出的基本机动动作库[15] 作为 UCAV 机动动作库的选择范围，如图 5 所示。

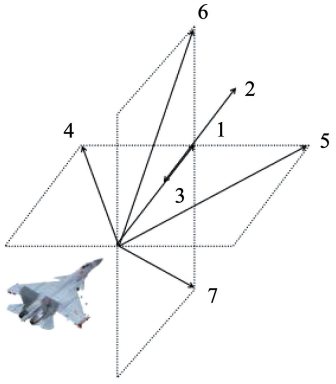

图 5　基本机动动作库

Fig. 5　Basic maneuver library

包括以下 7 种机动动作：1）定常平飞；2）最大加力加速；3）最大加力减速；4）最大加力左转；5）最大加力右转；6）最大加力爬升；7）最大加力俯冲。对于基本操纵动作库的实现，采用欧式坐标系下的机动动作指令$[V_C, H_C, \phi_C]$实现各种机动，建立自主作战决策的候选动作库。

匀速直线运动：

$$[V_C, H_C, \phi_C] = [V_A, H_A, 0] \tag{26}$$

最大加力加速飞行：

$$[V_C, H_C, \phi_C] = [V_{max}, H_A, 0] \tag{27}$$

最大加力减速飞行：

$$[V_C, H_C, \phi_C] = [V_{min}, H_A, 0] \tag{28}$$

最大加力左转：

$$[V_C, H_C, \phi_C] = [V_{max}, H_A, \phi_{left}] \tag{29}$$

最大加力右转：

$$[V_C, H_C, \phi_C] = [V_{max}, H_A, \phi_{right}] \tag{30}$$

最大加力爬升：

$$[V_C, H_C, \phi_C] = [V_{max}, H_{max}, 0] \tag{31}$$

最大加力俯冲：

$$[V_C, H_C, \phi_C] = [V_{max}, H_{min}, 0] \tag{32}$$

式中：V_C 为 UCAV 的速度指令，V_A、V_{max}、V_{min}分别为 UCAV 当前的速度、最大速度和最小速度；H_C 为 UCAV 的高度指令，H_A、H_{max}、H_{min}分别为 UCAV 的当前高度、最大高度和最小高度；ϕ_C 为 UCAV 的滚转角指令，ϕ_{left}、ϕ_{right}分别为 UCAV 的最大向左滚转角和最大向右滚转角。

将这 7 种机动动作作为 UCAV 机动决策的输出，控制 UCAV 的飞行。同时由于 UCAV 缺少人类感知飞机状态的能力，需要对上述机动动作做出限制，通过对俯仰角、滚转角和推力指令的大小进行限制，从而对控制输出端做必要的约束，以防 UCAV 的迎角、侧滑角和速度的值过大或过小从而导致失控。令控制输出端的俯仰角指令范围在[-20°，20°]之间，滚转角指令范围在[-60°，60°]之间，推力指令范围在[-10 N，30 N]之间。

3　深度强化学习自主作战决策

3.1　深度强化学习

对于传统的强化学习而言，通常采用表格的形式记录值函数模型，这种方法可以稳定得出不同状态和动作下函数的值。但在面对复杂问题时，状态和行动的空间较大，需要花费很多时间检索表格中相应状态的值，难以求解。由于深度学习将特征学习融入模型中，具有自学习性和鲁棒性，适用于不同的非线性模型。但深度学习不能对数据规律进行无偏差估计，需要大量的数据反复计算才能达到较高的精度。因此，本文将深度学习和强化学习算法相结合，得到深度强化学习算法，并使用深度 Q 网络（Deep Q network，DQN）作为优化算法，将态势信息输入神经网络并输出机动动作值，同时不断与环境进行交互得到最优机动动作，使得 UCAV 能自主的进行作战决策，提高其智能性[16]。

在 UCAV 空战决策过程中，需要对我方 UCAV 和敌机的飞行状态和空战态势进行分析，采用卷积神经网络（convolutional neural network，CNN）计算每一个状态动作对的长期折扣期望，并将 Q 函数网络作为评判依据，遍历不同状态下的所有机动动作。同时为了让学习的数据更接近独立分布的数据，需要建立一个数据库，将一段时间内的状态、动作、奖励和该动作下一步的状态存储起来，每次学习时使用存储区内的小部分样本，与 2.2 节的 Q 强化学习算法相比可以打乱原始数据的相关性，减小发散。

为了解决算法的不确定性，DQN 算法还建立了一个结构相同的目标网络用于更新 Q 值，该目标网

络具有和 Q 函数网络一样的初始结构,但参数固定不动,每隔一段时间将 Q 函数网络的参数赋给该目标网络,使其一定时间内的 Q 值保持不变。可以通过梯度下降法最小化损失函数 $L(\theta^{\mu})$ 来得到最优解:

$$L(\theta^{\mu}) = E[y_t - Q(s_t, a_t) \mid \theta^{\mu}] \tag{33}$$

其中 y_t 为目标参数,即

$$y_t = r_t + \gamma \max_a Q(s_{t+1}, a \mid \theta^{\mu'})$$

式中:θ^{μ} 为 Q 函数网络参数,$\theta^{\mu'}$ 为目标网络参数。

则 DQN 模型如图 6 所示。

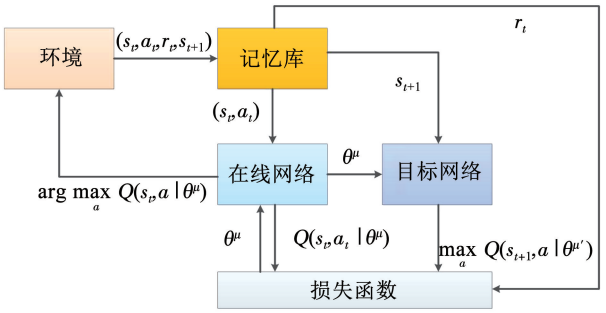

图 6 DQN 模型

Fig. 6 DQN model

3.2 训练步骤

在空战决策的训练中,UCAV 根据上述深度强化学习算法进行机动决策。整个训练过程由多个空战回合组成,每当 UCAV 判定击中敌机、被敌机击中、到达最大回合时间或处于式(21)、(22)所描述的错误态势时,结束该空战回合重新进入一个新的空战回合,并重置仿真环境。在训练过程中采用 ε-greedy 策略,一开始以 100% 的概率随机产生行动,随着仿真的进行,不断减小概率直至 10%,使得该策略不断向最优靠拢。同时为了反应学习的效果,需要在训练过程中定期判断其决策能力,在进行判断时令随机概率降为 0,使得决策模型直接输出最大的 Q 值动作,并统计其在结束时的优势函数值,与不同时期相对以此判断其学习效率。

深度强化学习算法的具体步骤如算法 1 所示。

算法 1 基于深度强化学习的 UCAV 空战自主机动决策过程。

1. 初始化记忆回放单元 D,其容量为 R
2. 初始化在线 Q 网络,随机生成参数 θ^{μ}
3. 初始化目标 Q'网络,随机生成参数 $\theta^{\mu'} = \theta^{\mu}$
4. 初始化 ε 的值为 1
5. **for** episode = 1, 2, … , M **do**
6. 初始化双方 UCAV 模型的状态,获取当前态势
7. **if** episode 为 N 的倍数 **then**
8. 执行评估,评估时 $\varepsilon = 0$
9. **end if**
10. **for** step = 1, 2, … , T **do**
11. 以 ε 的概率从基本机动动作库中的 7 种机动动作中随机选择一个动作,否则选择动作 $a_t = \arg\max_a Q(s_t, a \mid \theta^{\mu})$
12. 执行动作 a_t,得到奖励 r_t 及新的状态 s_{t+1}
13. 将数据样本(s_t, a_t, r_t, s_{t+1})存入 D 中
14. 判断该空战回合是否结束
15. **end for**
16. 从 D 中随机抽取一批样本$(s_{t'}, a_{t'}, r_{t'}, s_{t'+1})$
17. 令 $y_{t'} = r_{t'} + \gamma \max_a Q'(s_{t'+1}, a \mid \theta^{\mu'})$
18. 根据目标函数$(y_{t'} - Q(s_{t'}, a_{t'} \mid \theta^{\mu}))^2$ 使用梯度下降法进行更新
19. 每隔 C 回合更新目标 Q'网络,令 $\theta^{\mu'} = \theta^{\mu}$
20. 逐步减小 ε 的值直至 0.1
21. **end for**

4 仿真实验

4.1 参数设置

为验证本文提出的深度强化学习算法,设我方 UCAV 为红机,敌机为蓝机。将红蓝两机的仿真环境限制在同一空域范围内,x 坐标轴范围为 $x \in [-5\ \text{km}, 5\ \text{km}]$,$y$ 坐标轴范围为 $y \in [-5\ \text{km}, 5\ \text{km}]$,$z$ 坐标轴范围为 $z \in [0\ \text{km}, 2\ \text{km}]$。深度强化学习算法所涉及的各个参数取值见表 1。

表 1 参数取值情况

Tab. 1 Parameter values

R_g/m	α_g	V_{max}/(m·s^{-1})	V_{min}/(m·s^{-1})
300	π/4	90	20
H_{max}/m	H_{min}/m	ϕ_{left}	ϕ_{right}
2 000	200	π/3	−π/3

DQN 算法的参数设置如下:使用一个两层全连接前馈神经网络作为在线 Q 网络,有 10 个输入状态和 7 个输出值,其中网络有两个隐藏层,单位大小分别为 1 000 和 500,使用 TANH 函数作为激活函数,在最后的输出层采用 PURELIN 函数进行激活。设置学习率 $\delta = 0.01$,折扣系数 $\gamma = 0.9$,记忆回放单元 D 的缓冲区大小为 10^6,在存储了 10 000 个样本之后神经网络开始训练,每次抽取的训练样本数量为 1 000,目标网络每 4 000 步更新一次。

在仿真的过程中每一步的决策时间 $t = 1$ s,每一次作战的最大回合时间为 40 s,每进行 500 次作战回合对神经网络的学习能力进行一次评估,查看其停止作战时的奖赏值。

4.2 强化学习与深度强化学习仿真时间对比

强化学习的计算过程是一个迭代寻找最优策略的过程,需要耗费一定的时间。传统的强化学习相对于深度强化学习而言状态空间较大,遇到复杂问题时需要花费很多时间检索表格中相应状态的值,但是 UCAV 在现实中执行机动决策时要求的决策

时间非常短，否则无法进行有效的决策。同时不同大小的状态空间也会对结果造成影响，对状态空间的设定具有主观性。

根据输入状态空间的 10 个变量建立不同复杂程度的强化学习 Q 值表格，Q 值表 2 的大小是 Q 值表 1 的两倍，Q 值表 3 的大小是 Q 值表 2 的两倍，分别仿真 1 000 个作战回合，单次作战的最大回合时间为 40 s，基本采样时间为 0.02 s。对比每仿真 1 s 深度强化学习和不同复杂程度的强化学习在决策中所花费的时间，可以验证算法的时效性。

如图 7 所示，仿真使用的计算机为 AMD Ryzen 7 3700X 8-Core Processor CPU 和 NVIDIA GeForce GTX 1660 SUPER 显卡。

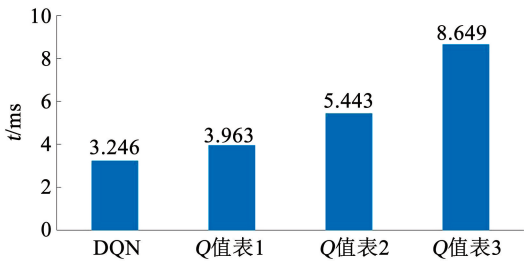

图 7　每仿真 1 s 的决策时间

Fig. 7　Decision time per second of simulation

从图 7 中可以看出，传统强化学习决策所花费的时间与其 Q 值表的大小有关，而深度强化学习在决策时花费的时间低于强化学习，在进行空战时能更快的做出有效的决策。

表 2　第 1 次和第 2 次训练的初始位置

Tab. 2　Initial positions of the first and the second training

训练方式	无人机	x/m	y/m	z/m	V/(m·s^{-1})	θ/(°)	ϕ/(°)	ψ/(°)
基础训练	红机	0	0	500	50	2.4	0	0
	蓝机	[400,1 000]	[−1 000,1 000]	[400,600]	[20,40]	1.6	0	[−90,90]
特定场景训练	红机	0	0	500	50	2.4	0	0
	蓝机	600	600	500	30	1.6	0	30

4.3　空战训练

由于 UCAV 空战环境复杂，直接训练会产生大量无效样本，致使学习算法的效率降低，需要先让目标在不同的初始状态下进行训练，之后再实现不同环境下 UCAV 的自主机动作战。

第 1 次训练以目标做匀速直线飞行运动的场景对 UCAV 依次进行基础训练和特定的空战场景训练，一开始红方战机处于优势地位，红方战机和蓝方战机的初始位置见表 2，蓝方战机做匀速直线运动。根据上文给出的 DQN 算法进行学习，首先对神经网络进行 20 000 个回合的基础训练，之后对具体的情况进行训练，经过 250 000 个作战回合后敌我双方的 UCAV 轨迹仿真如图 8 所示。

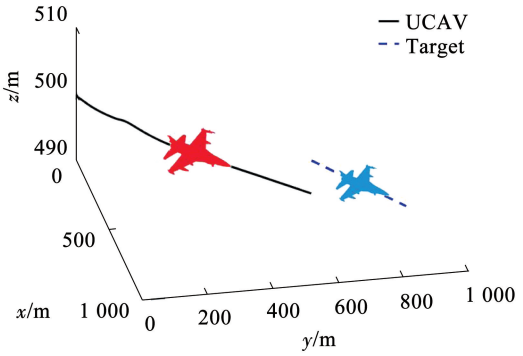

图 8　第 1 次训练时双方的立体轨迹

Fig. 8　Stereo trajectory of both sides at the first training

从图 8 中可以看出蓝方想要远离红机，但红方首先向左偏转，保持与蓝方相近的方位角和高度，接着加速追向目标，最终使蓝方处于红机武器攻击范围内，达到优势地位，说明该 DQN 算法的确可以快速有效的提高 UCAV 的自主作战能力。

通过对比图 9、10 中经过了基础训练和未经训练的评估奖赏值可以看出，经过了一定基础训练的 DQN 算法学习效率明显提高，能较快的使我方 UCAV 处于优势位置。

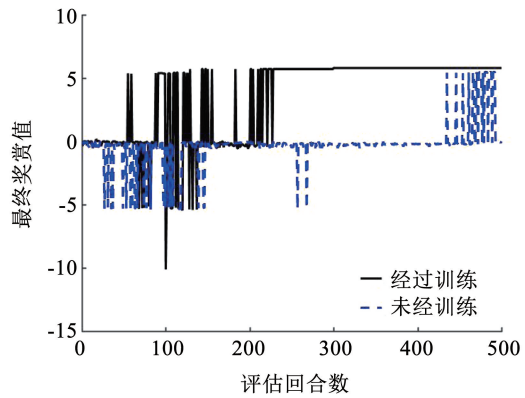

图 9　第 1 次训练时每次评估的最终奖赏值

Fig. 9　Final reward values of each assessment at the first training

第 2 次训练时以目标做匀速盘旋飞行的场景对 UCAV 依次进行基础训练和特定的空战场景训练，红方战机和蓝方战机的初始位置不变，蓝方战机做俯仰角为 10°，滚转角为 −20°的匀速盘旋飞行。重复相同的训练方法，经过 25 000 个作战回合后敌我

双方的 UCAV 轨迹仿真如图 11 所示。

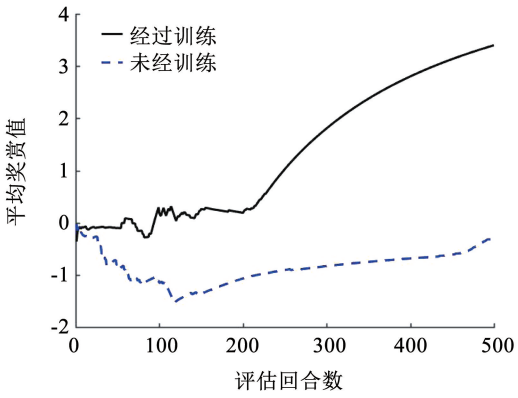

图 10　第 1 次训练时的平均奖赏值

Fig. 10　Average reward values at the first training

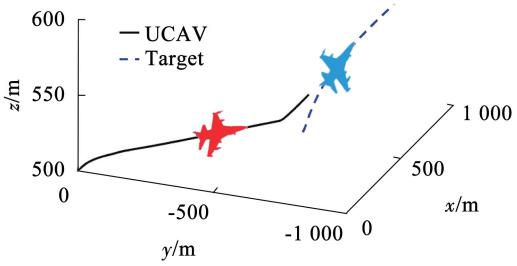

图 11　第 2 次训练时双方的立体轨迹

Fig. 11　Stereo trajectory of both sides at the second training

在图 11 中，红方首先向右偏转，保持与蓝方相近的方位角和高度，接着加速追向目标，由于蓝方处于相对较高的位置，红方为了追击蓝方迅速爬升，最终使得蓝方处于红方武器攻击范围内，达到优势地位，第 2 次训练时的平均奖赏值如图 12 所示。

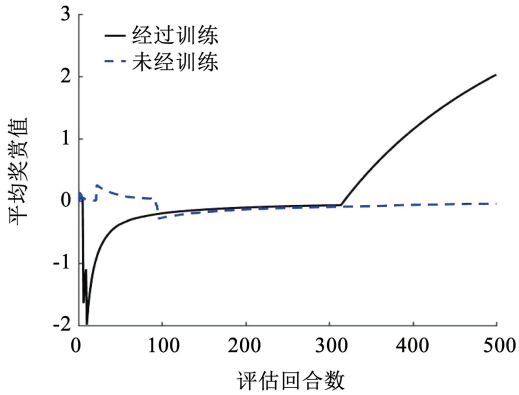

图 12　第 2 次训练时的平均奖赏值

Fig. 12　Average reward values at the second training

第 3 次训练时同样以目标做匀速直线飞行运动的场景对 UCAV 依次进行基础训练和特定的空战场景训练，但一开始红方战机处于劣势地位，红方战机和蓝方战机的初始位置见表 3，蓝机做匀速直线运动。同样对神经网络进行 20 000 个回合的基础训练，之后对具体的情况进行训练，经过 250 000 个作战回合后敌我双方的 UCAV 轨迹仿真如图 13、14 所示。

表 3　第 3 次训练的初始位置

Tab. 3　Initial positions of the third training

训练方式	无人机	x/m	y/m	z/m	V/(m·s^{-1})	θ/(°)	ϕ/(°)	ψ/(°)
基础训练	红机	0	0	500	50	2.4	0	0
	蓝机	[-1 000, -400]	[-1 000, 1 000]	[400, 600]	[20, 60]	1.6	0	[-90, 90]
特定场景训练	红机	0	0	500	50	2.4	0	0
	蓝机	-600	-600	500	40	1.6	0	30

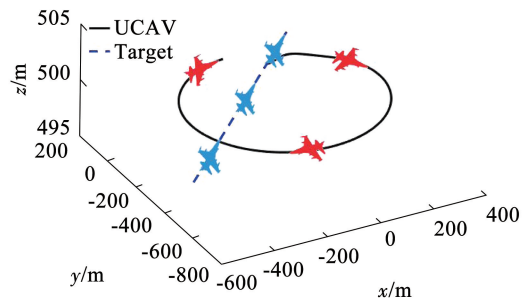

图 13　第 3 次训练时双方的立体轨迹

Fig. 13　Stereo trajectory of both sides at the third training

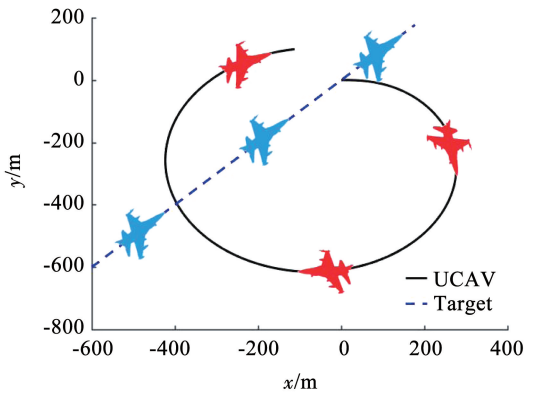

图 14　第 3 次训练时双方的平面轨迹

Fig. 14　Plane trajectory of both sides at the third training

从图 13、14 中可以看出，面对蓝机的追击，红机向右偏转，绕到了蓝机的身后，由劣势转化为优势，最终扭转局面取得胜利。第 3 次训练时的平均奖赏值如图 15 所示。

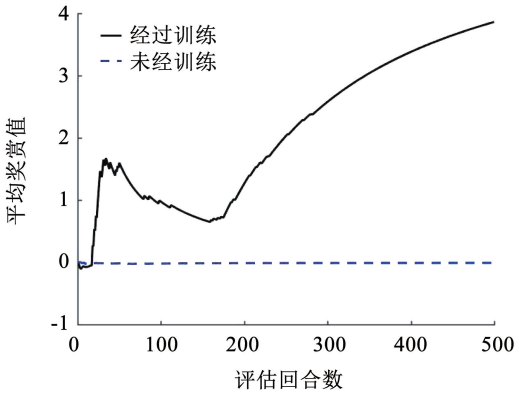

图 15　第 3 次训练时的平均奖赏值

Fig. 15　Average reward values at the third training

由上述 3 组仿真实验可以看出红方在不同的初始态势下都可以通过一定的机动决策占据有利态势，证明了 DQN 算法的有效性。同时，经过了一定基础训练的 DQN 算法学习效率明显提高，具有较高的智能性，可以有效地引导 UCAV 进行空战。

5　结　论

1）本文在 MATLAB/Simulink 环境中搭建了 UCAV 六自由度模型，设计了一个 UCAV 空战自主机动决策的仿真平台，实现了空战实验的闭环仿真。该仿真平台成本低，易于实现，并且各个函数都采用了模块化设计，易于更新和替换。

2）将机动动作库和基于导弹武器攻击区建立的优势函数应用于该仿真平台上，同时对模型采用由易到难的训练方法，可以使 UCAV 与不同运动状态下的目标进行空战，并且保证 UCAV 最终能够到达优势地位。

3）结果显示，深度强化学习算法可以通过不断与环境之间试错交互从而进行学习，能有效提高 UCAV 的自主作战能力，得到的仿真结果具有较高的工程参考价值。根据深度强化学习算法所得到的机动控制指令具有鲁棒性、远视性和时效性。

参考文献

[1] DUAN Haibin, SHAO Shan, SU Bingwei, et al. New development thoughts on the bio-inspired intelligence based control for unmanned combat aerial vehicle[J]. Science China (Technological Sciences), 2010, 53(8): 2025. DOI: 10.1007/s11431-010-3160-z

[2] 郭昊, 周德云, 张堃. 无人作战飞机空战自主机动决策研究[J]. 电光与控制, 2010, 17(8): 28
GUO Hao, ZHOU Deyun, ZHANG Kun. Study on UCAV autonomous air combat maneuvering decision-making[J]. Electronics Optics and Control, 2010, 17(8): 28. DOI: 10.3969/j.issn.1671-637X.2010.08.007

[3] HUANG Changqiang, DONG Kangsheng, HUANG Hanqiao, et al. Autonomous air combat maneuver decision using Bayesian inference and moving horizon optimization[J]. Journal of Systems Engineering and Electronics, 2018, 29(1): 86. DOI: 10.21629/JSEE.2018.01.09

[4] SMITH R E, DIKE B A, MEHRA R K, et al. 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. DOI: 10.1016/S0045-7825(99)00395-3

[5] 国海峰, 侯满义, 张庆杰, 等. 基于统计学原理的无人作战飞机鲁棒机动决策[J]. 兵工学报, 2017, 38(1): 160
GUO Haifeng, HOU Manyi, ZHANG Qingjie, et al. UCAV robust maneuver decision based on statistics principle [J]. Acta Armamentarii, 2017, 38(1): 160. DOI: 10.3969/j.issn.1000-1093.2017.01.021

[6] XU Guangyan, WEI Shenna, ZHANG Hongmei. Application of situation function in air combat differential games[C]//Proceedings of the 36th Chinese Control Conference (CCC). Dalian: IEEE, 2017: 5865. DOI: 10.23919/ChiCC.2017.8028286

[7] 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
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. DOI: 10.3969/j.issn.1671-637X.2015.01.005

[8] 傅莉, 谢福怀, 孟光磊, 等. 基于滚动时域的无人机空战决策专家系统[J]. 北京航空航天大学学报, 2015, 41(11): 1994
FU Li, XIE Fuhuai, MENG Guanglei, et al. An UAV air-combat decision expert system based on receding horizon control[J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41(11): 1994. DOI: 10.13700/j.bh.1001-5965.2014.0726

[9] ROSALES C, MIGUEL S C, ROSSOMANDO F G. Identification and adaptive PID control of ahexacopter UAV based on neural networks[J]. International Journal of Adaptive Control and Signal Processing, 2019, 33(1): 74. DOI: 10.1002/acs.2955

[10] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[J]. IEEE Transactions on Neural Networks, 1998, 9(5): 1054. DOI: 10.1109/tnn.1998.712192

[11] 丁林静, 杨啟明. 基于强化学习的无人机空战机动决策[J]. 航空电子技术, 2018, 49(2): 29
DING Linjing, YANG Qiming. Research on air combat maneuver decision of UAVs based on reinforcement learning[J]. Avionics Technology, 2018, 49(2): 29. DOI: 10.3969/j.issn.1006-141X.2018.02.06

[12] 孙楚, 赵辉, 王渊, 等. 基于强化学习的无人机自主机动决策方法[J]. 火力与指挥控制, 2019, 44(4): 142
SUN Chu, ZHAO Hui, WANG Yuan, et al. UCAVautonomic maneuver decision-making method based on reinforcement learning [J]. Fire Control & Command Control, 2019, 44(4):142. DOI: 10.3969/j.issn.1002-0640.2019.04.029

[13] YANG Qiming, ZHANG Jiandong, SHI Guoqing, et al. Maneuver decision of UAV in short-range air combat based on deep reinforcement learning[J]. IEEE Access, 2019, 8: 363. DOI: 10.1109/ACCESS.2019.2961426

[14] YANG Qiming, ZHU Yan, ZHANG Jiandong, et al. UAV air combat autonomous maneuver decision based on DDPG algorithm [C]//Proceedings of the 15th International Conference on Control and Automation. [S.l.]: IEEE, 2019: 39. DOI: 10.1109/ICCA.2019.8899703

[15] FRED A, GIRO C, MICHAEL F, et al. Automated maneuvering decisions for air-to-air combat: AIAA-87-2393 [R]. [S.l.]: AIAA, 1987: 2393. DOI: 10.2514/6.1987-2393

[16] 周毅, 马晓勇, 郜富晓, 等. 基于深度强化学习的无人机自主部署及能效优化策略[J]. 物联网学报, 2019, 3(2): 47
ZHOU Yi, MA Xiaoyong, GAO Fuxiao, et al. Autonomous deployment and energy efficiency optimization strategy of UAV based on deep reinforcement learning[J]. Chinese Journal on Internet of Things, 2019, 3(2): 47. DOI: 10.11959/j.issn.2096-3750.2019.00106

（编辑　张　红）