基于LSTM-Dueling DQN的无人战斗机机动智能决策

胡东愿，杨任农，左家亮，郑万泽，赵 雨，张 强
（空军工程大学，西安 710051）

摘 要：针对无人作战飞机在一对一自主空战中无法实现智能决策的问题，引入深度强化学习方法，构建无人战斗机战术决策框架，求解智能体对抗的机动指令。首先，建立飞行运动模型和导弹攻击区模型，形成基本的一对一空战对抗环境。其次，利用8个运动变量来构建智能体连续的状态空间，并根据导弹攻击区实时计算结果设计奖惩函数，实现双机对抗决策。最后，使用长短期记忆网络和全连接网络相结合，构建智能体价值网络和目标网络。利用记忆库中的决策样本，对网络进行训练，完成值函数的拟合，实现智能体在任意状态下的决策。仿真试验表明，在典型的案例中，智能体能够有效感知空战场态势，算法给出的决策动作可以积累并保持无人作战飞机的空战优势，完成对目标的打击，决策时间能够满足时效性的要求。

关键词：无人战斗机；空战对抗；机动决策；深度强化学习；值函数搜索；长短期记忆网络
中图分类号：V325 **文献标识码：**A **文章编号：**1009-1300(2021)06-0097-08
DOI：10. 16358/j. issn. 1009-1300. 2021. 6. 005

Intelligent Maneuvering Decision of Unmanned Combat Aircraft Based on LSTM-Dueling DQN

Hu Dongyuan, Yang Rennong, Zuo Jialiang, Zheng Wanze, Zhao Yu, Zhang Qiang
(Air Force Engineering University, Xi'an 710051, China)

Abstract: Aiming at the problem that unmanned combat aircraft cannot make intelligent decisions in one-to-one air combat, a deep reinforcement learning method is introduced to construct the tactical decision framework for unmanned combat aircraft and solve the maneuvering commands of agent confrontation. Firstly, the flight movement model and missile attack zone model are established to form a basic one-to-one confrontation environment. Secondly, eight motion variables are used to construct the continuous state space of the agent, and the reward and punishment functions are designed according to calculation results of the attack area. The decision making is realized under the engagement interaction between agent and combat

收稿日期：2020-11-30；**修回日期：**2021-01-22
作者简介：胡东愿，博士研究生，主要研究方向为智能决策、任务规划。
通讯作者：左家亮，副教授，主要研究方向为智能决策、任务规划。
引用格式：胡东愿，杨任农，左家亮，等. 基于LSTM-Dueling DQN的无人战斗机机动智能决策[J]. 战术导弹技术，2021，(6)：97-104. (Hu Dongyuan, Yang Rennong, Zuo Jialiang, et al. Intelligent Maneuvering Decision of Unmanned Combat Aircraft Based on LSTM-Dueling DQN[J]. Tactical Missile Technology, 2021, (6): 97-104.)

environment. Finally, long-term short-term memory network (LSTM) and fully connected network are combined to build agent value network and target network. By using the decision samples in the memory bank, the network is trained to complete the fitting of the value function and realize the decision making for agents in any state. Simulation tests show that, in typical cases, agents can effectively perceive the situation of the confrontation, and the decision actions given by the algorithm can accumulate and maintain the superiority of unmanned combat aircraft in air combat, and complete the attack on the target. The decision time can meet the requirement of timeliness.

Key words: unmanned combat aircraft; air combat; maneuvering decision; deep reinforcement learning; value function search; LSTM

1 引 言

随着信息化技术和无人作战飞机硬件性能的快速发展，战争向无人化、智能化方向发展的趋势越来越明显[1-3]。未来战场上，利用无人作战飞机进行空中对抗是战争初期获取制空权的首选作战样式。目前无人作战飞机的作战方式主要依靠地面站的人工操作，或是事前设定好飞行程序，根据固定的航线飞行。这种方式可以实现侦察、监视、干扰、或对地突击等任务，但无法实现态势快速变化的空空对抗任务[4-7]。利用智能化的方法，实现无人作战飞机的智能机动决策是目前主攻的研究方向，也是实现智能空空对抗的关键突破口。

智能机动决策是指利用人工智能等方法，根据收集到的战场态势信息，计算出无人作战飞机的最优控制指令，完成相应的战术动作[8]。智能机动决策是实现感知-决策-控制的过程，能够在劣势下规避敌方的跟踪，并逐步积累自身的优势。目前国内外在机动决策方面的研究形成了很多方法，根据算法的设计思想可以将其分为两类：反应式决策和推演式决策。

反应式决策是通过实现设立好的规则，利用当前的状态进行规则匹配，利用规则中的优化模型求解出相应的机动动作。例如专家系统法、贝叶斯网络法和模糊推理法等[9-15]。或是通过设立“优势函数”，将其转化为当前状态下的最优化问题，例如优化理论法。反应式决策方法能够保证问题的可解性，而且计算时间较快，能够满足实时性的需求，但决策效果往往容易陷入局部的最优解，或是无法满足规则之外的态势判断。

推演式决策分析双方从当前状态往后的n步决策及其可能出现的态势情况，从而给出当前的机动决策解。该方法能够从全局或部分全局出发，避免了局部最优解。例如微分对策法、动态规划法、蒙特卡洛搜索法[16-19]。微分对策法从控制角度出发，为整个飞行过程进行数学公式建模，建模过程复杂，求解难度较大。动态规划法通过将双方对抗过程按照时域或空域划分为各个相互关联的子阶段，通过求解每一个子问题的最优解来达到全局最优，计算的时间复杂度较大，且必须提前知道双方的机动策略。蒙特卡洛方法使用仿真搜索的方式来弥补双方策略的不确定性问题，但求解依然无法解决时效性问题。

近年来随着深度强化学习在AlphaGo、AlphaZero、AlphaStar等复杂问题中的成功运用[20]，使得利用强化学习方法来解决空战机动决策问题成为了一种可能，人工智能技术的运用成为无人作战飞机对抗领域的一个热点。强化学习是一种无监督的试错学习方法，无人作战飞机通过与环境的交互得到奖赏回报，根据回报值最大化的原则不断学习并适应环境，能够通过线下学习和线上决策的方式满足决策的实时性要求。

目前强化学习在空战对抗中的运用主要有基于值函数搜索的Q学习方法和基于策略搜索的Actor-Critic方法。文献［21］中使用强化学习深度梯度策略的方法，探索连续动作空间的机动决策问题，通过优化算法生成空战机动动作值，保证了动作值的正确性，但网络训练消耗资源较大。

文献［22］中使用DQN构建智能体决策，利用深度网络拟合Q值，但算法收敛效果较差，智能性不强。文献［23］中使用AC框架解决机动决策问题并改进回报函数，但仅在二维空间中有效。

本文利用强化学习框架，基于竞争网络的深度强化学习（Dueling DQN）算法，研究连续状态空间的机动决策问题。使用LSTM网络增强智能体的记忆功能，并对状态-动作对的值函数进行拟合，形成状态-记忆-决策三层机制，加快智能体的收敛时间。利用记忆库分别训练结构相同但参数不同的价值网络和目标网络，解决传统强化学习过拟合的问题，改善算法收敛效果。

2　对抗问题描述

2.1　飞行模型构建

以地面坐标系为基础，建立无人机飞行模型，攻击机的质心运动学方程为：

$$\begin{cases}\dot{x}_T = v_T\cos\chi_T\cos\gamma_T\\ \dot{y}_T = v_T\sin\chi_T\\ \dot{z}_T = v_T\cos\chi_T\sin\gamma_T\end{cases}\tag{1}$$

其中，$\dot{x}_T$、$\dot{y}_T$、$\dot{z}_T$表示飞机在惯性坐标系中x、y和z三个方向上的位置变化率，χ_T和γ_T表示目标机的航迹倾角和航迹偏航角，v_T表示攻击机速度。

攻击机的质心动力学方程为：

$$\begin{cases}\dot{v}_T = g(n_{Tt}-\sin\chi_T)\\ \dot{\chi}_T = (n_{Tf}\cos\mu_T-\cos\chi_T)g/v_T\\ \dot{\gamma}_T = n_{Tf}g\sin\mu_T/(v_T\cos\theta_T)\end{cases}\tag{2}$$

其中，θ_T、φ_T和μ_T分别表示飞机的航迹俯仰角、航迹偏航角和速度滚转角。n_{Tt}和n_{Tf}表示飞机的切向过载和法向过载。

2.2　双机对抗运动模型

在分析空战态势及空战决策过程中，需要知道两机之间位置关系和运动参数的关系，需要构建双机相对运动模型，双机相对运动示意图如1所示。

图1中$o_Rx_Ry_Rz_R$和$o_Bx_By_Bz_B$分别表示红方飞机和蓝方飞机的机体坐标系，d表示双方飞机的距离，v_B和v_R分别表示蓝方飞机和红方飞机的速度，χ_{Rr}和γ_{Rr}表示蓝方在红方机体坐标系上相对于红方的相对倾斜角和相对偏航角；χ_{Br}和γ_{Br}表示红方在

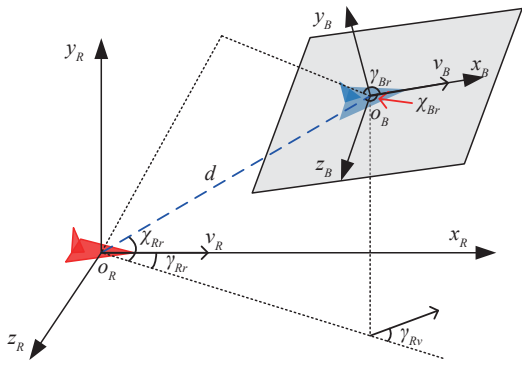


图1　空战双方相对方位关系

蓝方机体坐标系上相对于蓝方的相对倾斜角和相对偏航角；χ_R和γ_R表示红方飞机的航迹倾斜角和航迹偏航角；χ_B和γ_B表示蓝方飞机的航迹倾斜角和航迹偏航角；γ_{Rv}表示蓝方飞机的相对进入角（蓝方飞机的速度在红方飞机机体坐标系中$o_Rx_Ry_R$平面内的投影与两机连线的投影延长线的夹角）。在红方机体坐标系中，红方与蓝方的距离公式为：

$$d = \sqrt{(x_B-x_R)^2+(y_B-y_R)^2+(z_B-z_R)^2}\tag{3}$$

蓝方相对于红方的相对倾斜角为：

$$\sin\chi_{Rr} = \frac{z_B-z_R}{d}\tag{4}$$

蓝方相对于红方的相对偏航角为：

$$\cos\gamma_{Rr} = \frac{x_B-x_R}{d\cos\chi_{Rr}}\tag{5}$$

距离、相对倾斜角和相对偏航角对时间的导数为：

$$\begin{aligned}\dot{d} &= v_Be_B\cdot e_r - v_Re_R\cdot e_r\\ &= v_B\cos\chi_B\cos\chi_{Br}\cos(\chi_t-\chi_{Br}) + v_B\sin\chi_B\sin\chi_{Br} -\\ &\quad v_R\cos\chi_R\cos\chi_{Rr}\cos(\chi_R-\chi_{Rr}) + v_R\sin\chi_R\sin\chi_{Rr}\end{aligned}\tag{6}$$

$$\begin{aligned}\dot{\chi}_{Rr} = [&-v_B\cos\chi_B\sin\chi_{Rr}\cos(\chi_B-\chi_{Rr}) + v_B\sin\chi_B\sin\chi_{Rr} +\\ &v_R\cos\chi_R\cos\chi_{Rr}\cos(\chi_R-\chi_{Rr}) - v_R\sin\chi_R\cos\chi_{Rr}]/d\end{aligned}\tag{7}$$

$$\dot{\gamma}_{Rr} = \frac{v_B\cos\chi_B\sin(\gamma_B-\gamma_{Rr}) - v_R\cos\chi_R\sin(\gamma_R-\gamma_{Rr})}{d\cos\chi_{Rr}}\tag{8}$$

2.3　导弹攻击区模型

攻击区是衡量空空导弹作战能力的重要指标，在进行攻击区仿真前，必须确定好限制条件，不同的限制条件计算出来的导弹攻击区相差很大。导弹攻击区常见的表示方式是以目标机为中心，

考察载机从不同的方位攻击目标时的最大、最小发射距离。然而，对于作战仿真，攻击区是发射导弹的攻击机飞行状态和发射角以及目标机飞行状态、机动能力、目标进入角的函数。结合本文考察的问题，导弹攻击区的限制条件主要如下：

从导弹的角度出发：

（1）导弹的最大、最小飞行高度；

（2）击中目标前导弹飞行的最小速度；

（3）安全限制；

（4）导弹最大飞行时间。

其中，安全限制主要指导弹离开载机的最小安全距离限制以及导弹解除保险装置的最短时间。导弹最大的飞行时间主要受到导弹系统中工作时间较短部件的约束，如弹上电源和弹上制冷系统。

此外，从载机的角度出发，载机的姿态与机动过载大小也限制着导弹的发射。并且载机和目标机的运动应该符合各自的飞行剖面。

根据以上限制条件，给出典型导弹攻击区的定义如下：

（1）最大攻击包线

最大攻击包线又称可能攻击区，即满足上述限制条件的空空导弹有效攻击的最大区域或者边界，反应导弹的极限性能。

（2）不可逃逸攻击区

在满足约束条件下，载机发射导弹后，目标不管作何种机动都不能摆脱导弹攻击的有效边界或者区域。超过该边界，目标可能会摆脱导弹的跟踪。

（3）最小攻击包线

最小攻击包线是在保证载机安全的基础上，能对目标机形成一定攻击的最小距离。当超过该距离时，导弹击中目标机或引信引爆时产生的爆破碎片对攻击机不产生影响。

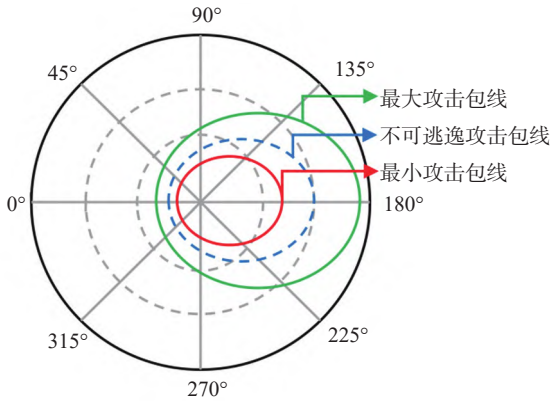


图2 攻击区示意图

3 机动决策模型

3.1 强化学习决策框架

根据强化学习中智能体与环境互动的过程，建立无人作战飞机空战的强化学习框架。将红蓝双方无人作战飞机看成红方智能体和蓝方智能体，利用第二部分中的运动模型和攻击区模型构建空战环境。利用智能体对抗来实现决策的优化，对抗的双方可以同时进行学习。

如图3所示，智能体中的红方和蓝方相互独立，其结构相同、功能相同，但参数不同，主要实现从状态到动作指令的映射过程。环境中主要包含状态转移模型，攻击区模型和奖励函数。智能体根据当前状态决策出战术动作，环境中的状态转移模型根据此刻的状态和动作计算出动作完成后的下一个状态，在状态转移过程中同时利用攻击区模型计算每一时刻的三种包线，判断是否满足攻击条件，奖惩函数利用攻击区的计算结果给出智能体的回报值。环境将回报值和下一时刻的状态传递到智能体中，至此，完成了一次对抗决策循环。

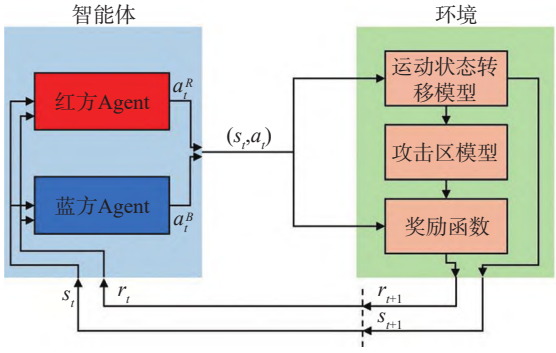


图3 强化学习机动决策框架

t时刻，智能体从环境中获得的状态信息s_t与回报信息r_t，并执行动作a_t作用到环境中，环境的状态发生改变，转移到下一个状态s_{t+1}，并反馈回报r_{t+1}。

强化学习的目标就是找到一个策略，使得智能体的长期积累的回报最大，即：

$$\pi^* = \arg\max_{\pi} E_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t] \tag{9}$$

其中，γ为折扣因子，用来计算累计回报，实际上是对不同时刻奖赏的信度分配，一般地，随时间的增长，奖赏对于最优策略的影响也就越低；$E_{\pi}[.]$为策略π^*下的期望。

强化学习中的值函数包含状态值函数$V(s)$和行为值函数$Q(s,a)$。状态值函数为：

$$V^{\pi}(s) = E_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t | s_0 = s] \tag{10}$$

行为值函数为：

$$Q^{\pi}(s) = E_{\pi}[\sum_{t=0}^{\infty}\gamma^t r_t | s_0 = s, a_0 = a] \tag{11}$$

求解最优策略π^*的过程即为最优化上述值函数的过程。

3.2　状态空间与动作空间

分析红蓝双方的对抗相对运动模型，强化学习中的状态空间可以通过若干个相对运动特征来表示。三维对抗中，可以直接获取双机的空间相对位置以及速度，并据此计算出相关的角度，便可描述出整个对抗过程。同时考虑到导弹攻击区与双机对抗的高度联系较为紧密，选取双机对抗模型中8个状态特征表征状态空间：

$$s = \{\Delta x, y_R, y_B, \Delta z, \gamma_{Rr}, \chi_{Rr}, \gamma_R, \chi_R\} \tag{12}$$

其中，Δx、Δz表示红蓝双方战机在正北方向、正东方向的位置差，y_R、y_B表示红蓝双方战机的高度，γ_{Rr}、χ_{Rr}, γ_R、χ_R分别表示红蓝双方战机相对角度和红方自身角度。

根据作战飞机动力学方程和控制量可以计算出相应状态变量的值，计算公式如下：

$$\begin{aligned}\dot{\gamma} &= \frac{g}{v}\left(n\cos\phi - \cos\gamma\right)\\ \dot{\psi} &= \frac{gn\sin\phi}{v\cos\gamma}\end{aligned} \tag{13}$$

其中，ϕ表示滚转角，γ为航迹角，n表示法向过载，g为重力加速度，v为速度。

为简化动作空间，选取空战中5个基本机动动作，即：向左转弯、向右转弯、跃升、俯冲以及保持当前运动状态。即动作空间定义为：$\{tl, tr, up, dn, st\}$。为不同动作设置相应的控制量如下：

（1）向左转弯tl：

$$\begin{cases}\phi_R = 60°, n_R = 1\\ \phi_B = 60°, n_B = 1\end{cases} \tag{14}$$

（2）向右转弯tr：

$$\begin{cases}\phi_R = -60°, n_R = 1\\ \phi_B = -60°, n_B = 1\end{cases} \tag{15}$$

（3）跃升up：

$$\begin{cases}\phi_R = 0°, n_R = 6\\ \phi_B = 0°, n_B = 5\end{cases} \tag{16}$$

（4）俯冲dn：

$$\begin{cases}\phi_R = 0°, n_R = -3\\ \phi_B = 0°, n_B = -3\end{cases} \tag{17}$$

（5）保持当前运动状态st：

$$\dot{\gamma} = 0, \dot{\psi} = 0 \tag{18}$$

3.3　奖励函数设计

当智能体从当前状态s_t执行动作指令，环境模型将智能体转移到下一个状态s_{t+1}，并返回回报值r_t。利用导弹攻击区可以准确地反应智能体在当前态势下的奖惩情况。在智能体从s_t转换到s_{t+1}过程中，利用导弹攻击区模型求解每一时刻双方的攻击区边界值，从而计算出奖惩值。

以红方智能体为例，考虑红方导弹对蓝方进行攻击的奖惩情况：

$$r_{\text{red\_attack}} = \begin{cases}-50, & D < D_{\min}\\ 100, & D_{\min} < D < D_{\text{non-escape}}\\ 50, & D_{\text{non-escape}} < D < D_{\max}\\ 0, & D > D_{\max}\end{cases} \tag{19}$$

其中，D为红蓝双方的距离，$D_{\min}$为当前态势下的红方导弹最小攻击包线距离，$D_{\text{non-escape}}$为当前态势下的红方导弹不可逃逸距离，$D_{\max}$为当前态势下的红方导弹最大攻击包线距离。

同时，智能体必须考虑对方导弹对自身的安全威胁情况，同样以红方智能体为例，若红方处于蓝方的不可逃逸区内，将获得最大的惩罚项。考虑红方受到蓝方导弹威胁设计的奖惩函数如下：

$$r_{\text{red\_damage}} = \begin{cases}0, & D < D'_{\min}\\ -100, & D'_{\min} < D < D'_{\text{non-escape}}\\ -50, & D'_{\text{non-escape}} < D < D'_{\max}\\ 0, & D > D'_{\max}\end{cases} \tag{20}$$

其中，$D'_{\min}$为当前态势下蓝方导弹对红方的最小攻

击包线距离，$D'_{\text{non-escape}}$ 为当前态势下蓝方导弹的不可逃逸距离，$D'_{\max}$ 为当前态势下蓝方导弹的最大攻击包线距离。

综合红方智能体导弹对蓝方的攻击情况以及红方智能体受蓝方导弹的威胁情况，针对红方智能体设计的总体奖惩函数如下：

$$r_{\text{red}} = \alpha \times r_{\text{red\_attack}} + \beta \times r_{\text{red\_damage}} \tag{21}$$

其中，α、β 为系数，当 $\alpha > \beta$ 时，红方智能体的战术策略更倾向于先保全自身，以安全为主。当 $\alpha < \beta$ 时，红方智能体的战术策略以冒险为主。

4 LSTM-Dueling DQN 算法模型

4.1 价值网络构建

由于空战战场较大，智能体的状态特征量是连续的、多维的，离散化后的状态空间较大，增加了决策的搜索难度。利用深度网络的非线性拟合能力来逼近 $Q(s,a)$，可以解决连续状态空间中的“维度灾难”问题。

与简单的 DQN 网络不同，Dueling DQN 网络将 Q 网络分成了两部分，价值函数 $V(s,\omega,\alpha)$ 和优势函数 $A(s,a,\omega,\beta)$，网络的最终结果包含两部分：

$$Q(s,a,\omega,\alpha,\beta) = V(s,\omega,\alpha) + A(s,a,\omega,\beta) \tag{22}$$

其中，ω 为价值函数和优势函数共同的网络参数，α 为价值函数独立部分的网络参数，β 为优势函数独立部分的网络参数。

以红方智能体为例，智能体的状态由 8 个变量组成，因此网络的输入包含 8 个节点，网络的输出为各动作的价值函数，中间层为隐含层，具体结构如图 4 所示。

为增强智能体从状态变量中的感知能力，使决策动作具有一定的连贯性，引入长短期网络单元作为隐含层的基本神经元，并对网络的输入层进行改进。

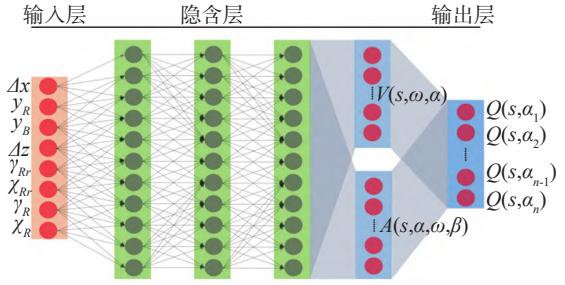


图 4 价值网络结构图

在 DQN 网络中，输入层是每一时刻的智能体状态变量，以红方智能体为例，在 t 时刻，网络的输入为 $\text{input} = s_t = \{\Delta x, y_R, y_B, \Delta z, \gamma_{Rr}, \chi_{Rr}, \gamma_R, \chi_R\}$，使用 LSTM 网络作为基本单元时，利用环境状态转移模型进行采样，使用连续 4 个时刻的状态序列作为网络的输入 (s_1, s_2, s_3, s_4)，如图 5 所示。

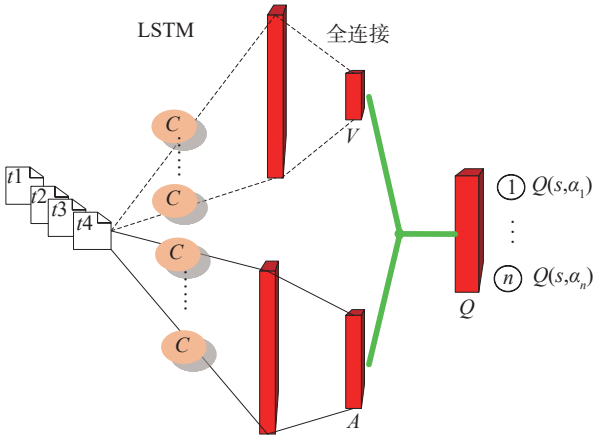


图 5 LSTM-Dueling DQN 价值网络结构图

4.2 记忆库构建

记忆库是“价值网络”的训练样本库，通过智能体不断与环境交互试错学习得到。记忆库中每一组数据记录的策略执行情况为 $< s, a_R, a_B, r, s' >$。其中，s 为当前状态，a_R、a_B 为红蓝双方采取的动作，r 为对应的奖赏函数值，s' 为执行完动作后双方进入的下一时刻状态。

在学习过程中，不断有新的策略执行数据加入到记忆库，为保证最新的数据能够被用来训练网络，需设置记忆库的更新规则。设置记忆库的最大容量为 $mem_{\max}$，新进入的第 n 组数据在记忆库的位置记为 mem_{index}，利用取模运算计算该位置，有：$mem_{\text{index}} = n \,(\text{mod}\, mem_{\max})$，并取代之前在该位置的数据。

4.3 价值网络训练

价值网络训练的目标是对 Q 值尽可能准确地做出估计。利用以往执行策略的数据，利用梯度下降法更新网络参数，使得“价值网络”尽可能逼近准确的 Q 值。定义损失函数 $loss$：

$$loss = \frac{1}{2}\left[r(s) + \gamma \max_{a'} Q(s', a';\ \theta^-) - Q(s, a;\ \theta)\right]^2 \tag{23}$$

其中，$r(s)$表示奖赏函数，γ表示折扣率，$Q(s',a';\theta^-)$、$Q(s,a;\theta)$分别为利用目标网络估计的动作值函数和利用价值网络估计的动作值函数。利用随机梯度下降（SGD）算法，更新网络参数，设置小样本的数量为100组。整个训练示意图如图6所示。

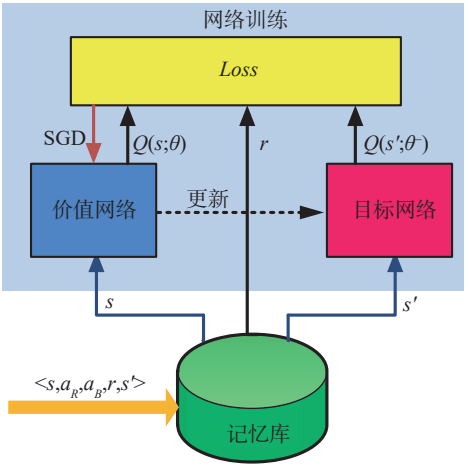


图6　价值网络训练过程示意图

5　案例仿真分析

在本次试验中，采用的硬件为Intel（R）Xeon（R）CPU ES-2643，内存16 GB，64位操作系统。在双方对抗过程中，设定空战决策周期为0.5 s，每次训练前进行初始化，固定双方的初始位置和相对态势，红方无人作战飞机的三维坐标$Position_R$:(3000, 4000, 8000)，红方的航迹倾角和航迹偏角分别为$\gamma_R = 0°, \psi_R = 0°$，速度大小$v_R$ = 350 m/s。蓝方战机的空间三维坐标$Position_B$:(30000, 4000, 9000)，蓝方战机航迹倾角和航迹偏角分别为$\gamma_B = 0°, \psi_B = 0°$，速度大小为$v_B$ = 300 m/s。

红蓝双方对抗10000回合进行网络学习，在每个回合中最大的仿真时间为10 min，若对抗时间超出10 min且仍未达到终止条件，则结束本轮对抗。训练完成后，终止蓝方价值网络参数的学习，红方智能体以此为基础，继续进行5000次对抗训练，并进行参数更新。训练过程中记录红方每一局的平均单步回报值，在5000次的红方网络训练中，每5次再取一次智能体平均回报值，得到的奖励函数曲线如图7所示。

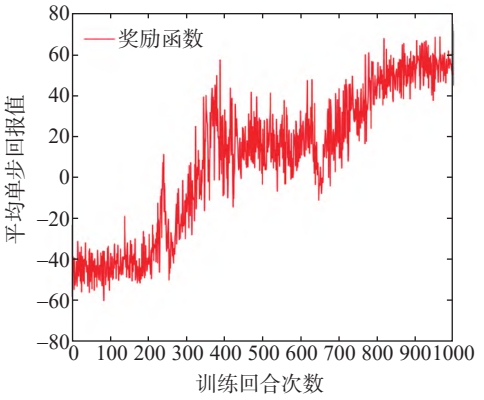


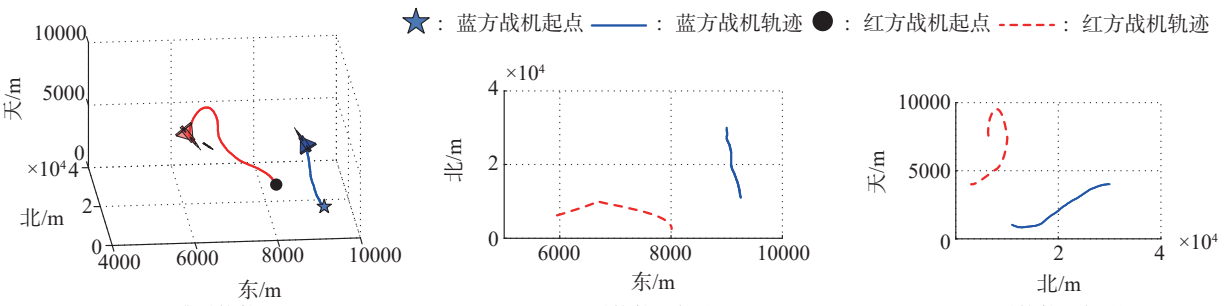
图7　强化学习对抗训练奖励函数曲线

在训练初期，双方使用相同的网络参数，对态势的感知和决策能力相差不大。但初始状态下，红方相较于蓝方处于弱势状态，红方的回报值为负数。经过一定次数的训练之后，红方智能体价值网络参数不断更新优化，此时蓝方的网络模型和参数保持不变。从奖励函数曲线图中可以看出，智能体的决策能力在逐渐增强，能够改变自身的劣势，从环境中获取正回报。经过800次对抗训练之后，智能体在该场景下开始收敛，最终奖励函数值能够稳定在50左右。

红方完成训练后，测试双方的智能体决策模型，测试结果如图8所示。通过双方对抗轨迹发现，停止参数更新的蓝方智能体决策能力弱于红方。在对抗开始，蓝方具备一定的角度优势，通过拉升高度，进一步保存优势。但经过多次对抗并进行参数优化的红方智能体能够快速爬升，积累高度能量，避开蓝方的攻击，随后俯冲转身，营造攻击条件，最终获取绝对的优势战胜蓝方。

6　结　论

本文设计了基于LSTM-Dueling DQN的强化学习方法，同时为对抗双方决策机动动作。试验表明，在未知对方的机动时，智能体仍能作出合理的机动决策。且决策一次动作的时间约为0.02 s，达到实时性的要求。该研究对于智能自主空战的机动决策具有重要的理论价值和现实意义，不足之处在于动作的选取上与实际空战仍有一定的差



(a) 三维对抗轨迹图 (b) 对抗轨迹投影图 (c) 对抗轨迹投影图

图8 红蓝对抗结果示意图

距，下一步的工作将针对机动控制量连续取值的问题进行研究。

[参 考 文 献]

[1] 厉博. 国外无人作战飞机发展回顾与趋向分析[J]. 飞航导弹，2019，(10)：43-48.

[2] 姜进晶，汪民乐，姜斌. 无人机作战运用研究[J]. 飞航导弹，2019，(1)：41-44.

[3] 赵振平，路瑞敏，王锦程，等. 智能无人飞行器技术发展与展望[J]. 战术导弹技术，2017，(3)：1-7.

[4] 贾治辉，薛楠. 侦察活动中无人机的应用探讨[J]. 公安学刊（浙江警察学院学报），2019，(3)：32-39.

[5] 朱瑞，施麟，陈宁. 基于移动平台的海域无人机监视监测系统研究[J]. 江苏科技信息，2020，37 (10)：40-43.

[6] 陈圣哉，彭章友. 基于无人机的低复杂度干扰源定位算法研究[J]. 工业控制计算机，2020，33 (5)：82-84.

[7] 丰雨轩，刘树光，解武杰，等. 基于改进Hopfield网络的对地攻击型无人机自主能力评价[J]. 北京航空航天大学学报，2020，47 (4)：835-843.

[8] 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.

[9] Han S. Analysis of relative combat power with expert system [J]. Journal of Digital Convergence，2016，(14)：143-150.

[10] Gacovski Z，Deskovski S. Modelling of combat actions via fuzzy expert system[C]. RTO NMSG Conference on “Future Modeling and Simulation Challenges”，Breda，Netherlands，November 12-14，2001.

[11] 罗元墙. 基于动态贝叶斯网络的空战决策方法研究[D]. 沈阳：沈阳航空航天大学，2018.

[12] 孟光磊，罗元强，梁宵，等. 基于动态贝叶斯网络的空战决策方法[J]. 指挥控制与仿真，2017，39 (3)：49-54.

[13] Sathyan A，Ernest N D，Cohen K. An efficient genetic fuzzy approach to UAV swarm routing [J]. Unmanned Systems，2016，4 (2)：117-127.

[14] Ernest N，Cohen K，Kivelevitch E，et al. Genetic fuzzy trees and their application towards autonomous training and control of a squadron of unmanned combat aerial vehicles[J]. Unmanned Systems，2015，3 (3)：185-204.

[15] 王兴虎，程家林，郭强，等. 防空压制任务中智能协同作战体系研究[J]. 无人系统技术，2020，3 (4)：10-21.

[16] 傅莉，王晓光. 无人战机近距空战微分对策建模研究[J]. 兵工学报，2012，33 (10)：1210-1216.

[17] Ma Y，Ma X，Xiao S. A case study on air combat decision using approximated dynamic programming [J]. Mathematical Problems in Engineering，2014 (4)：38-42.

[18] 黄长强，赵克，韩邦杰，等. 一种近似动态规划的无人机机动决策方法[J]. 电子与信息学报，2018，40 (10)：2447-2452.

[19] 何旭，景小宁，冯超. 基于蒙特卡洛树搜索方法的空战机动决策[J]. 火力与指挥控制，2018，43 (3)：34-39.

[20] Mnih V，Kavukcuoglu K，Silver D，et al. Human-level control through deep reinforcement learning [J]. Nature，2015，(518)：529-533.

[21] 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.

[22] Liu P，Ma Y. A deep reinforcement learning based intelligent decision method for UCAV air combat [C]. Asian Simulation Conference，Springer Singapore，August 27-29，2017.

[23] Kurniawan B，Vamplew P，Papasimeon M，et al. An empirical study of reward structures for actor-critic reinforcement learning in air combat maneuvering simulation [C]. Advances in Artificial Intelligence，32nd Australasian Joint Conference，Adelaide，SA，Australia，December 2-5，2019.