文章编号:1674-8190(2024)02-179-09

基于态势评估及DDPG算法的一对一空战格斗控制方法

贺宝记,白林亭,文鹏程
(航空工业西安航空计算技术研究所 人工智能与图形图像研究室,西安 710076)

摘 要: 已有的空中格斗控制方法未综合考虑基于专家知识的态势评估及通过连续性速度变化控制空战格斗的问题。基于深度确定性策略梯度(DDPG)强化学习算法,在态势评估函数作为强化学习奖励函数的基础上,设计综合考虑飞行高度上下限、飞行过载以及飞行速度上下限的强化学习环境;通过全连接的载机速度控制网络与环境奖励网络,实现DDPG算法与学习环境的交互,并根据高度与速度异常、被导弹锁定时间以及格斗时间设计空战格斗结束条件;通过模拟一对一空战格斗,对该格斗控制方法在环境限制学习、态势评估得分以及格斗模式学习进行验证。结果表明:本文提出的空战格斗控制方法有效,能够为自主空战格斗进一步发展提供指导。

关键词: 强化学习;态势评估;深度确定性策略梯度;空战格斗
中图分类号: V323.9 **文献标识码:** A
DOI: 10.16615/j.cnki.1674-8190.2024.02.20

One-on-one air combat control method based on situation assessment and DDPG algorithm

HE Baoji, BAI Linting, WEN Pengcheng
(Laboratory of Artificial Intelligence and Graphic Images, AVIC Xi'an Aeronautics Computing Technique Research Institute, Xi'an 710076, China)

Abstract: The existing aerial combat control methods do not comprehensively consider the situation assessment based on expert knowledge and the control of aerial combat through continuous speed change. Based on the deep deterministic policy gradient (DDPG) reinforcement learning algorithm, a comprehensive reinforcement learning environment is designed that considers flight altitude limits, flight overload and flight speed limits, which is building upon the situation evaluation function as the reward function for reinforcement learning. The interaction between the DDPG algorithm and learning environment is achieved through the fully connected carrier speed control network and the environment reward network. The end condition for air combat is designed based on abnormal height and speed, missile lock time and combat time. By simulating one-on-one air combat, the effectiveness of this combat control method is validated in terms of learning under environmental constraints, situation evaluation scores and combat mode learning. The results show that the air combat control method is effective, and can provide guidance for the further development of autonomous air combat.

Key words: reinforcement learning; situation assessment; deep deterministic policy gradient; air combat

收稿日期: 2023-06-19; **修回日期:** 2023-09-19

通信作者: 贺宝记(1989—),男,博士,工程师。E-mail:hebaoji1989@163.com

引用格式: 贺宝记,白林亭,文鹏程. 基于态势评估及DDPG算法的一对一空战格斗控制方法[J]. 航空工程进展,2024,15(2):179-187.
HE Baoji, BAI Linting, WEN Pengcheng. One-on-one air combat control method based on situation assessment and DDPG algorithm[J]. Advances in Aeronautical Science and Engineering, 2024, 15(2): 179-187. (in Chinese)

0 引 言

随着航空技术的不断发展以及新技术带来战场环境的复杂多变,对空战中操作的实时性有了更高的需求。空战格斗的自主飞行控制方法既可以用于有人战机协助飞行员处理格斗问题,也可以用于成本低、无人员伤亡的无人机飞行控制,因此受到越来越多的关注。

近年来随着深度学习算法的不断发展,其在各个领域也都有较深入的应用[1-3]。针对智能空战引导的方法,国内外研究者进行了研究,钟麟等[4]针对一对一空战决策问题,提出了随机机动决策模型和基于影响图博弈的机动决策模型,并验证了影响图博弈方法可以解决一对一空战机动中的决策问题;姜龙亭等[5]通过动态规划方法改进了一对一空战中的自主攻击方法。随着强化学习算法的发展,Hessel等[6]、Mnih等[7]提出了深度Q网络(DQN)及其变体,该类型强化学习网络的输入是连续值,输出是离散值,不能处理连续控制问题;Lillicrap等[8]提出的深度确定性策略梯度(DDPG)强化学习方法训练参数和输出策略均为连续性数值,满足了连续性控制需求;Schulman等[9-10]提出的置信域策略优化(TRPO)算法及近端策略优化(PPO)算法,解决了先采样、然后更新网络存在训练效率低的问题。通过强化学习方法进行自主飞行控制的研究也有了一定发展,张博超等[11]针对5v5有/无人机协同的空战场景,利用PPO算法,验证了与环境实时交互的空战决策方法;单圣哲等[12]基于Actor-Critic方法提出空战连续决策的统一架构,依据空战训练经验对状态空间、动作空间、奖励及训练科目进行设计,测试多种连续动作空间强化学习算法在高不确定性空战场景下的学习效果并进行了可视化验证;邱妍等[13]使用深度确定性策略梯度(DDPG)实现了向目标点的飞行;张堃等[14]提出的改进近端策略优化算法能够更有效地处理与实践序列相关的无人机自主引导。在一对一格斗控制类场景,针对基于专家知识态势评估的格斗控制相关研究尚不充分。

在一对一空战格斗中,通常用态势评估函数来分析空战中载机相对于敌机的优势[15-17]。目前在一对一空战中的态势评估函数主要包含进入角优势、方位角优势、速度优势以及高度优势,并根据载机相对于敌机的位置,针对上述几种优势做变权重综合来判断相对于敌机的空中优势。态势评估函数除了与格斗双方当前的相对角度、速度、高度有关外,还与飞机自身特性(最优空战高度、最优空战速度、雷达探测距离、最大搜索角度等)、所携带导弹特性(最大射程、最大不可逃逸距离、最小不可逃逸距离以及偏向角度等)相关[17]。态势评估函数解决的是当前空战中优势问题,并不考虑飞机的飞行限制,比如高度、速度、过载等,且不对下一步的行动做出指导。除此之外,态势评估还与强化学习结合用于空战的机动决策,丁林静等[18]设计了一种空战机动决策的动态模糊Q学习模型,并选取典型的空战动作作为强化学习基本行动。

本文设计基于DDPG强化学习算法的一对一空战格斗方法,其中强化学习环境在态势评估函数的基础上增加飞行高度上下限、飞行过载以及飞行速度上下限,强化学习环境与DDPG算法的交互通过全连接的格斗速度控制与环境奖励网络实现,格斗终止条件包含被导弹锁定时间、速度与高度异常以及格斗时间;通过一对一空战格斗模拟,验证该格斗控制方法在环境限制学习、态势评估得分以及格斗模式学习上的有效性,以期为自主空战格斗发展提供指导。

1 一对一空战格斗模型训练流程

本文中强化学习模型是格斗控制模型,模型输出是对载机三维方向速度的增减控制;位置、角度、高度的更新均由速度的更新计算得到。为了简化模型,环境中敌机的性能参数及导弹参数与载机相同。环境部分中的敌机格斗控制是通过正在训练的模型控制。

本文中空战格斗训练如图1所示,流程主要包含两个部分:DDPG算法部分以及环境部分。其中,DDPG算法部分负责格斗模型的不断优化,环境部分主要是通过优化后的模型计算新的敌机与载机状态,并返回给算法部分新状态下的奖励函数。

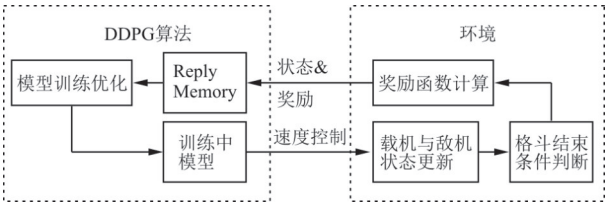


图1　一对一空战格斗模型训练整体流程

Fig. 1　One-on-one air combat model training overall process

DDPG算法部分与环境部分的交互包含两个部分。

1) 环境中生成的载机与敌机的状态信息存储在用于模型训练的Reply Memory库中。本文中假设飞机速度与飞机对称面夹角为0,不考虑侧滑角影响;飞机过体轴的铅垂面与对称面的夹角为0,不考虑滚转角的影响;速度控制模型输入包含11个参数,对应于环境中奖励函数需要的参数,分别为:载机高度、敌机高度、敌机相对载机在x、y、z三个方向的位置,载机x、y、z三个方向的速度,敌机x、y、z三个方向的速度。

在模型训练中,载机高度、敌机高度的单位均为最优空战高度;载机、敌机的三个方向速度单位为最优空战速度;敌机相对载机在三个方向的距离单位设定为50 km。根据上述的设定可以将参数(载机高度、速度、与敌机相对距离等)控制在1.0附近,其中,z对应飞机的飞行高度。

2) 模型推理的结果用于载机速度的更新。推理模型的激活函数为tanh函数,其值域为(−1.0, 1.0),载机的速度更新以及位置更新函数如公式(1)~公式(3)所示。

$$\boldsymbol{A}_{\mathrm{rf}}=\boldsymbol{A}_{\mathrm{f}}\times\boldsymbol{I}\times\left[X_{\mathrm{ol}},Y_{\mathrm{ol}},Z_{\mathrm{ol}}\right]^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{L}_{\mathrm{f}}=\boldsymbol{L}_{\mathrm{f}}+\left(\boldsymbol{V}_{\mathrm{f}}+\frac{\boldsymbol{A}_{\mathrm{rf}}}{2}\right)\times\delta T\times 0.001 \tag{2}$$

$$\boldsymbol{V}_{\mathrm{f}}=\boldsymbol{V}_{\mathrm{f}}+\boldsymbol{A}_{\mathrm{rf}} \tag{3}$$

式中:$\boldsymbol{A}_{\mathrm{f}}$为模型在$x$、$y$、$z$方向的输出,范围是(−1.0, 1.0);$\boldsymbol{A}_{\mathrm{rf}}$为载机实际的速度变化;$\left[X_{\mathrm{ol}},Y_{\mathrm{ol}},Z_{\mathrm{ol}}\right]^{\mathrm{T}}$分别为$x$、$y$、$z$三个方向的过载映射,模型训练中设定为3.0、3.0、0.9,其中3.0对应x、y方向过载,上限约为$1g$,0.9对应于高度方向过载,最大约为$0.3g$。

过载上限属于飞机自身性能,其限制了速度的变化率,公式(1)通过对速度变化的映射设置飞机过载。公式(2)表示载机的位置更新,此处假设在δT的时间段内,飞机的速度变化是匀变速,0.001表示从米到位置单位千米的转换。公式(3)表示速度的变化。在环境中敌机的位置以及速度更新与上述公式相同。

2　DDPG算法描述以及网络设定

2.1　DDPG算法描述

在DDPG算法中,本文采用Actor网络拟合策略函数,在此处对应载机的速度操控;使用Critic网络拟合状态值,此处对应环境的奖励网络,使用Reply Memory作为样本存储仓库。在训练开始时,Critic网络打分结果差,Actor网络也是近似随机选择动作,随着网络不断训练,Critic网络评判越来越准确,Actor网络也会逐渐选择高回报的动作来迎合Critic网络。

2.2　DDPG中网络结构设定

设定Actor网络以及Critic网络均为6层全连接神经网络,网络参数分别如图2~图3所示。

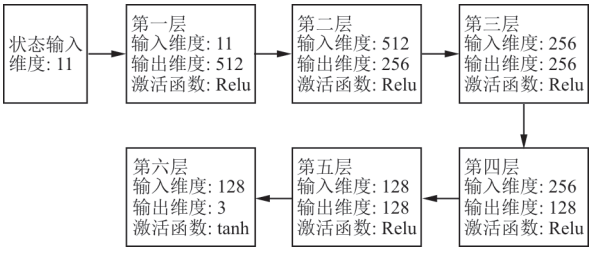


图2　Actor网络结构示意图

Fig. 2　Actor network structure

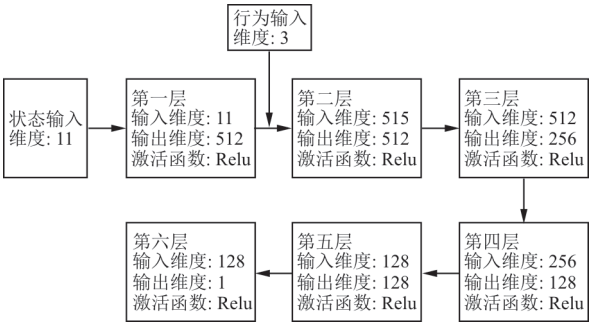


图3　Critic网络结构示意图

Fig. 3　Critic network structure

Actor网络为全连接网络,对应速度控制部分,输入维度为11项,对应11种状态参数;中间隐藏

层维度分别为(512,256,256,128,128),激活函数均为Relu;输出层对应为x、y、z方向的速度控制,激活函数为tanh,维度为3。Critic网络也为全连接网络,对应奖励函数拟合部分,其在第二层中输入网络中加入了行为输入,因此输入维度为515维;中间层维度分别为(512,256,256,128,128),激活函数均为Relu,最后输出维度对应奖励,因此维度为1。

3 强化学习中环境设定

3.1 环境中态势评估函数以及奖励函数设定

一对一空战中空战态势如图4所示,其中载机位置为P,目标机位置为T;载机与目标机的距离为D,目标机方位角为θ,进入角为q。目标线是指载机P到敌机T的连线;目标方位角θ是指载机航向与目标线的夹角,目标航向角φ是指敌机航向与目标线的夹角;目标进入角q是指敌机航向与目标线延长线的夹角;方位角与进入角的方向一致,规定右偏为正,左偏为负。

环境中的优势函数包含角度优势函数、速度优势函数、距离优势函数以及高度优势函数。优势函数的选取参考文献[17]。

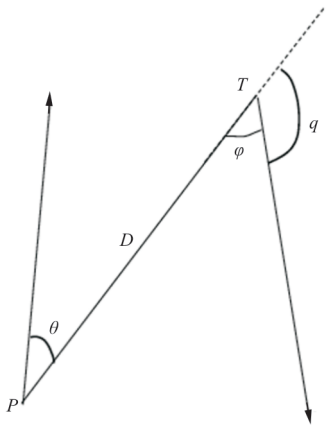


图4 一对一态势示意图

Fig. 4 One-on-one situation diagram

1) 角度优势函数

角度优势函数主要包含方位角优势和进入角优势,目标机方位角$|\theta|$越小,进入角$|q|$越大,空空导弹的攻击区域范围越广,角度态势越好,一对一格斗中角度态势函数如公式(4)~公式(5)所示。

$$T_\theta=\begin{cases}0.1-\dfrac{|\theta|-\theta_{\mathrm{rmax}}}{10(\pi-\theta_{\mathrm{rmax}})} & (\theta_{\mathrm{rmax}}<|\theta|\leqslant\pi)\\ 0.3-\dfrac{|\theta|-\theta_{\mathrm{mmax}}}{10(\theta_{\mathrm{rmax}}-\theta_{\mathrm{mmax}})} & (\theta_{\mathrm{mmax}}<|\theta|\leqslant\theta_{\mathrm{rmax}})\\ 0.8-\dfrac{|\theta|-\theta_{\mathrm{mkmax}}}{2(\theta_{\mathrm{mmax}}-\theta_{\mathrm{mkmax}})} & (\theta_{\mathrm{mkmax}}<|\theta|\leqslant\theta_{\mathrm{mmax}})\\ 1-\dfrac{|\theta|}{5\theta_{\mathrm{mkmax}}} & (0\leqslant|\theta|\leqslant\theta_{\mathrm{mkmax}})\end{cases} \tag{4}$$

$$T_q=\begin{cases}e^{-\frac{|q|-\pi}{2\pi}} & (\dfrac{\pi}{3}<|q|\leqslant\pi)\\ e^{-\frac{\frac{\pi}{2}-|q|}{\frac{\pi}{3}}} & (0\leqslant|q|\leqslant\dfrac{\pi}{3})\end{cases} \tag{5}$$

式中:T_θ为方位角优势;T_q为进入角优势;θ_{rmax}为最大搜索方位角;θ_{mmax}为最大离轴发射角;θ_{mkmax}为不可逃逸圆锥角,其中$0<\theta_{\mathrm{mkmax}}<\theta_{\mathrm{mmax}}<\theta_{\mathrm{rmax}}<\pi$。

方位角和进入角之间存在耦合,如公式(6)所示。

$$T_a=T_\theta^{\gamma\theta}T_q^{\gamma q} \tag{6}$$

式中:$\gamma\theta$,γq分别为方位角和进入角的权重,其权重满足$\gamma\theta+\gamma q=1$,随着方位角的不同,其权重也在变化,如公式(7)所示。

$$(\gamma\theta,\gamma q)=\begin{cases}(0.4,0.6) & (\theta_{\mathrm{rmax}}<|\theta|\leqslant\pi)\\ (0.5,0.5) & (\theta_{\mathrm{mmax}}<|\theta|\leqslant\theta_{\mathrm{rmax}})\\ (0.66,0.34) & (\theta_{\mathrm{mkmax}}<|\theta|\leqslant\theta_{\mathrm{mmax}})\\ (0.75,0.25) & (0\leqslant|\theta|\leqslant\theta_{\mathrm{mkmax}})\end{cases} \tag{7}$$

2) 速度优势函数

在空战中,较快的速度能帮助战机占领优势位置,而过快的速度对应转弯半径大,油耗高,本文采用基于最佳空战速度建立空战速度优势函数,速度优势函数如公式(8)~公式(9)所示。

当$v_0>1.5v_{\mathrm{t}}$时:

$$T_v=\begin{cases}e^{-\frac{v_{\mathrm{m}}-v_0}{v_0}} & (v_0<v_{\mathrm{m}})\\ 1 & (1.5v_{\mathrm{t}}<v_{\mathrm{m}}\leqslant v_0)\\ \dfrac{v_{\mathrm{m}}}{v_{\mathrm{t}}}-0.5 & (0.6v_{\mathrm{t}}<v_{\mathrm{m}}\leqslant 1.5v_{\mathrm{t}})\\ 0.1 & (v_{\mathrm{m}}\leqslant 0.6v_{\mathrm{t}})\end{cases} \tag{8}$$

当$v_0 \leqslant 1.5v_t$时:

$$T_v=\begin{cases} e^{-\frac{v_m-v_0}{v_0}} & (v_0<v_m) \\ \dfrac{2}{5}\left(\dfrac{v_m}{v_t}+\dfrac{v_m}{v_0}\right) & (0.6v_t<v_m\leqslant v_0) \\ 0.1 & (v_m\leqslant 0.6v_t) \end{cases} \tag{9}$$

式中:T_v为速度优势函数;v_0为最优空战速度;v_m为载机速度;v_t为敌机速度。

3) 距离优势函数

距离优势函数与所携带导弹的相关参数以及雷达相关参数有关,如公式(10)所示。

$$T_d=\begin{cases} 0.18e^{-\frac{D-D_{rmax}}{D_{rmax}}} & (D_{rmax}<D) \\ 0.5e^{-\frac{D-D_{mmax}}{D_{rmax}-D_{mmax}}} & (D_{mmax}<D\leqslant D_{rmax}) \\ 2^{-\frac{D-D_{mkmax}}{D_{mmax}-D_{mkmax}}} & (D_{mkmax}<D\leqslant D_{mmax}) \\ 1 & (D_{mkmin}<D\leqslant D_{mkmax}) \\ 2^{-\frac{D-D_{mkmin}}{D_{mmin}-D_{mkmin}}} & (D_{mmin}\leqslant D\leqslant D_{mkmax}) \\ 0 & (D<D_{mmin}) \end{cases} \tag{10}$$

式中:T_d为距离优势;D为敌机与载机之间的距离;D_{rmax}为雷达最大探测距离;D_{mmax}为导弹最大攻击距离;D_{mkmax}为导弹不可逃逸最大距离;D_{mkmin}为不可逃逸最小距离;D_{mmin}为导弹最小攻击距离,其中$D_{mmin}<D_{mkmin}<D_{mmax}<D_{mmax}<D_{rmax}$。

4) 高度优势函数

本文采用基于最优作战高度的高度优势函数,如公式(11)所示。

$$T_h=\begin{cases} e^{-\frac{h_m-h_{bh}}{h_{bh}}} & (h_{bh}\leqslant h_m) \\ e^{\frac{h_m-h_{bh}}{h_t}} & (h_t\leqslant h_m<h_{bh}) \\ \dfrac{h_m}{h_t}-0.5 & (0.6h_t\leqslant h_m<h_t) \\ 0.1 & (h_m<0.6h_t) \end{cases} \tag{11}$$

式中:T_h为高度优势;h_m为载机高度;h_t为敌机高度;h_{bh}为空战最优高度。

5) 态势评估函数以及环境的奖励函数

战机的态势中距离与角度耦合,因此评估函数如公式(12)所示。

$$T_m=T_h\times w_h+T_v\times w_v+T_a^{wa}\times T_d^{wd}\times w_c \tag{12}$$

式中:w为速度、高度、距离和角度耦合后的权重。

奖励函数定义为在当前状态下,载机与敌机的态势评估函数差,如公式(13)所示。

$$R=T_m-T_t \tag{13}$$

式中:T_t为敌机的态势评估函数。

3.2　环境中速度以及高度限制

在空战中,战机的速度、高度并不是任意的,其受自身参数以及自然条件限制,比如战机高度过低会撞到地面。因此对态势评估函数进行矫正,以满足上述的限制。

1) 高度限制

高度限制主要设置在高度过低时对评估函数的矫正,如公式(14)所示。

$$T_m=\begin{cases} T_m & (h_m>0.4h_{bh}) \\ \dfrac{10h_m-h_{bh}}{9h_{bh}}T_m & (h_m\leqslant 0.4h_{bh}) \end{cases} \tag{14}$$

在$0.4h_{bh}$开始对态势评估函数做惩罚,在$0.1h_{bh}$时,态势评估函数降为0,即对高度优势函数的矫正。高度优势函数中通过比较载机相对于敌机最优空战高度,计算其高度优势;但并未对高度过低做限制,加入高度限制函数后能提升强化学习效率,使得载机飞行高度不会过低导致撞向地面。此处仅对高度过低进行了限制,这是因为在高度优势函数中已经对过高的高度进行了指数级的惩罚。

2) 速度限制

速度限制主要设置在速度过高时对评估函数的矫正,如公式(15)所示。

$$T_m=\begin{cases} T_m & (v_m<2v_0) \\ \left(3-\dfrac{v_m}{v_0}\right)T_m & (v_m\geqslant 2v_0) \end{cases} \tag{15}$$

在$2v_0$开始对态势评估函数进行惩罚,在$3v_0$时,态势评估函数降为0,即对速度过高时的限制,速度过低时会终止格斗。速度优势函数中通过比较载机相对于敌机、最优空战速度,计算载机的速度优势;其对于速度过低的状态设置了最低的速度优势为0.1,而对于速度过高的情况下虽然有指数衰减的函数,但并未限制其飞行速度上限,通过加入速度限制能完成飞机速度上限设置,提升强化学习效率。

3.3 格斗结束条件

格斗中结束的判定条件包含3类。

1）速度、高度异常类，载机或者敌机速度过大以及z方向的速度过大均会触发格斗结束。

2）格斗时间过长类，每一局的格斗事件均设定为固定时间，过长的时间对训练过程的采样不利。

3）被导弹锁定时间足够长事件，那么认为一方被击落，不再训练。

4 深度学习训练环境设定

4.1 DDPG训练参数设定

DDPG中Reply Memory样本数设置为500万；模型训练批大小设置为64；Actor网络以及Critic网络的学习率均为10^{-4}；DDPG中噪声设置为0.2；模拟步数上限设置为500万步。

4.2 态势评估相关参数设定

1）态势评估函数参数设定

方位角态势函数中，雷达最大搜索角65°，导弹最大离轴发射角35°，不可逃逸角20°。

速度态势函数中，最优空战速度设置为320 m/s。

距离态势函数中，雷达最大探测距离140 km，导弹最大攻击距离80 km，最大不可逃逸距离60 km，最小不可逃逸距40 km，导弹最小攻击距离10 km，最优空战高度设置为5 km。

态势优势函数中各个态势函数的权重分别为：角度权重0.6，距离权重0.4，速度权重0.3，高度权重0.2，距离与角度耦合后权重为0.5。

2）环境初始化参数设定

环境初始化敌机与载机距离为30～70 km；高度为3.0～6.5 km；初始化速度大小为$0.5v_0$～$0.8v_0$，z方向速度大小为$-0.3v_0$～$0.3v_0$，每一步训练时间间隔δT为0.3 s。

3）格斗终止条件参数设定

速度、高度类限制：载机或敌机速度$>2.2v_0$或$<0.2v_0$；z方向速度$>v_0$；高度低于$0.1h_{bh}$或者高于$2h_{bh}$，格斗结束。

格斗时间长度限制：格斗时间超过20 min，格斗结束。

格斗被锁定事件限制：被锁定1 min，格斗结束。

5 训练效果验证

模型训练完成后通过最终训练模型与中间训练模型的一对一格斗效果，判断模型训练效果。效果评测包含以下3个方面：模型能不能学习到格斗中的环境限制，包含速度限制、高度限制等；随着模型的训练，能不能获取到更优的行动策略，使得格斗中态势评估函数得分更优；模型能否学习到格斗模式。通过这三个方面评测确定DDPG算法结合态势评估在一对一格斗中的应用效果。

5.1 模型训练可以学习到环境限制信息

模型训练能否学习到速度、高度等限制可以通过计算不同训练步数下的敌机能否完成一定步数的格斗训练来评测。此处设置载机与敌机的格斗上限为1.5万次，载机配置第500万次训练的模型；敌机配置不同训练次数的模型，每一个模型与载机对抗50次。格斗中敌机或载机速度、高度触发格斗终止条件判定为格斗结束，被锁定不判定为格斗结束。格斗结果如图5所示。

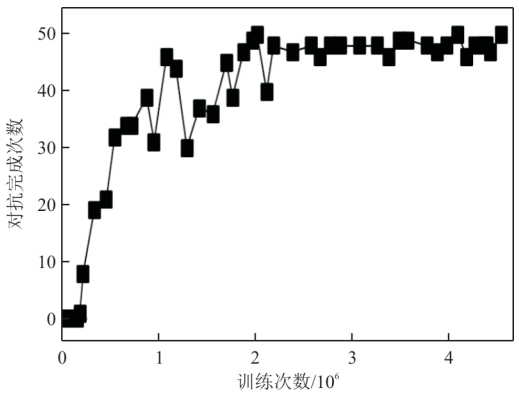


图5 不同训练次数下学习到环境限制信息

Fig. 5 The environmental limitation information was learned under different training times

从图5可以看出：在训练次数小于18万次时，模型尚未能充分学习到环境信息，因此敌机速度

或者高度总是超出限制,导致格斗结束;在18万到约200万次训练中,敌机学习到了速度以及高度限制信息,完成格斗的次数不断上升;在约240万次训练后,能有效格斗完成的次数均大于45次。这说明随着不断的训练,模型训练可以学习到速度以及高度的限制。

5.2　模型训练可以获取到更高的态势评估得分

为评测空战格斗模型的训练效果,载机配置第500万次训练的模型,敌机配置不同训练次数的模型。通过模拟不同训练次数下,载机与敌机格斗优势,可以判断不同的训练次数下模型是否学习到了更优的格斗控制策略。载机对不同敌机训练次数下的格斗优势如图6所示。

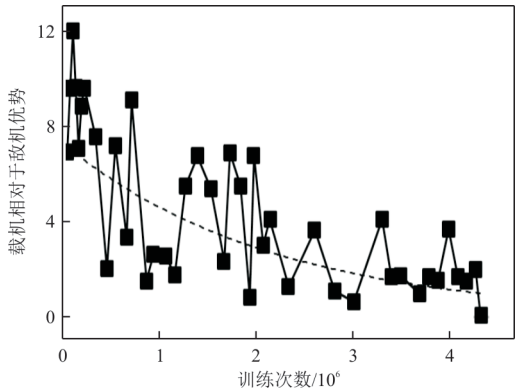


图6　不同训练次数下载机相比于敌机优势

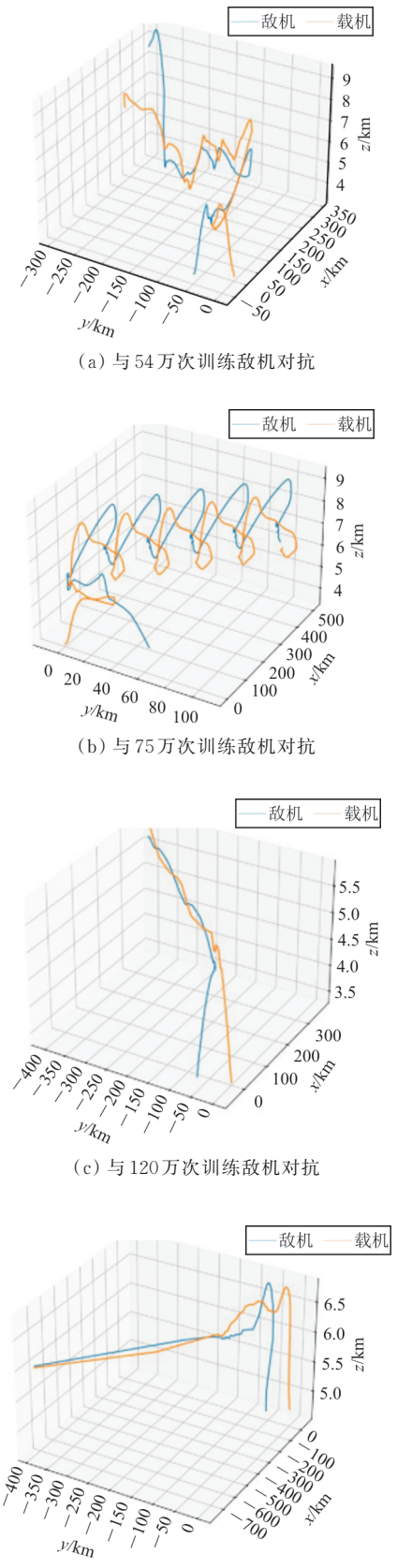
Fig. 6　Different training times download machine advantages over enemy aircraft

图6中纵坐标表示载机相对于敌机优势是在整个空战过程中载机的态势评估函数与敌机态势评估函数的差值的总和,此数值越小表示载机的空战格斗优势越小,如果此数值为负数,表示敌机优势高于载机。从图6可以看出:随着训练次数的不断上升,载机获得的优势不断下降,这说明随着模型训练得越来越成熟,敌机学习到了更优的格斗控制策略,从而使得载机的优势不断减少。

5.3　模型训练可以学习到格斗模式

随着模型训练,模型是否学习到一些格斗模式,可以通过载机与敌机格斗的轨迹分析。

载机与敌机的格斗轨迹如图7所示,载机选用500万次训练模型,敌机选用不同的训练次数。



(a) 与54万次训练敌机对抗

(b) 与75万次训练敌机对抗

(c) 与120万次训练敌机对抗

(d) 与450万次训练敌机对抗

图7　不同训练次数下格斗轨迹

Fig. 7　Combat trajectory under different training times

从图7可以看出:在模型未充分训练时(54万次),敌机并未学习到一些行为模式,行为尚处于混乱状态;在75万次模拟时,敌机已学习到一些行为模式,格斗模式为不规则的螺旋追逐线;在120万次的训练后,格斗已变为较规则的螺旋追逐线;在450万次训练后,螺旋追逐线更加平滑,此时格斗优势约为1:1;与120万次相比,450万次训练的敌机模型在固定对抗次数下完成螺旋追逐圈数目减少,且螺旋追逐线高度上下限逐渐变小。通过与75、120及450万次训练敌机对抗比较可知:其格斗轨迹均为螺旋追逐线,说明该方法学习到了类似螺旋追逐的格斗模式并收敛于该模式;随着训练次数增加,螺旋追逐线逐渐平滑,螺旋追逐线的高度上下限逐渐变小,完成单个螺旋追逐圈需要对抗次数逐渐增大,由此可知随着敌机训练次数增加,其螺旋追逐线会更加平滑、高度上下限更小且完成单个螺旋追逐圈需要的对抗次数更大。综上所述,训练中模型学会了类似螺旋追逐线的格斗模式。

6 结 论

1) 通过计算不同训练次数下的敌机能否完成一定次数的格斗,证明本文提出的方法能够学习到高度、速度等环境限制。

2) 通过计算不同训练次数下,载机与敌机格斗中环境奖励,证明充分的训练可以获取更高的态势评估得分。

3) 通过对不同训练阶段格斗行为、轨迹的分析,证明本文提出的飞行控制方法能够学习到一定的格斗模式。

本文通过将态势评估及DDPG强化学习应用于空战格斗控制,说明可以通过在强化学习格斗环境中将专家知识引入格斗控制模型,可以作为专家知识在格斗控制中应用的一种参考方案。

参 考 文 献

[1] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis, 2017, 42: 60-88.

[2] MOODY J, SAFFEL M. Deep learning for financial forecasting[J]. Journal of Computational Finance, 2018, 22(1): 1-21.

[3] LIANG X, WANG Z, DONG, et al. Deep reinforcement learning for autonomous driving: a review[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(11): 3835-3853.

[4] 钟麟, 佟明安, 钟卫, 等. 基于影响图的空战激动决策模型[J]. 系统仿真学报, 2007, 19(2): 410-412.
ZHONG Lin, TONG Ming'an, ZHONG Wei, et al. Multistage influence diagram for sequential maneuvering decisions used in air combat[J]. Journal of System Simulation, 2007, 19(2): 410-412. (in Chinese)

[5] 姜龙亭, 寇亚楠, 王栋, 等. 改进近似动态规划法的攻击占位决策[J]. 火力与控制, 2019, 44(7): 135-141.
JIANG Longting, KOU Yanan, WANG Dong, et al. Attack placeholder decision based on improved approximate dynamix programming[J]. Fire Control and Command Control, 2019, 44(7): 135-141. (in Chinese)

[6] HESSEL M, MODAYLI J, HASSELT H V, et al. Rainbow: combining improvements in deep reinforcement learning[C]// 2018 AIAA Computer Science Conference. US: AIAA, 2018: 1-8.

[7] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[C]// 2013 NIPS Deep Learning Workshop. US: NIPS, 2013: 1-12.

[8] LILLICRAP T P, HUNT J J, ALEXANDER P, et al. Continuous control with deep reinforcement learning[C]// 2016 AIAA Computer Science Conference. US: AIAA, 2016: 1-8.

[9] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization[J]. Computer Science, 2015, 25: 1889-1897.

[10] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithm[C]// 2017 AIAA Computer Science Conference. US: AIAA, 2017: 1-9.

[11] 张博超, 温晓玲, 刘璐, 等. 基于近端策略优化的空战决策算法研究[J]. 航空工程进展, 2023, 14(2): 145-151.
ZHANG Bochao, WEN Xiaoling, LIU Lu, et al. Research on air combat decision algorithm based on proximal policy optimization[J]. Advances in Aeronautical Science and Engineering, 2023, 14(2): 145-151. (in Chinese)

[12] 单圣哲, 杨孟超, 张伟伟, 等. 自主空战连续决策方法[J]. 航空工程进展, 2022, 13(5): 47-58.
SHAN Shengzhe, YANG Mengchao, ZHANG Weiwei,

et al. Continuous decision-making method for autonomous air combat[J]. Advances in Aeronautical Science and Engineering, 2022, 13(5): 47-58. (in Chinese)

[13] 邱妍, 赵宝奇, 邹杰, 等. 基于PPO算法的无人机近距离空战自主引导方法[J]. 电光与控制, 2023, 30(1): 9-14.
QIU Yan, ZHAO Baoqi, ZOU Jie, et al. An autonomous guidance method of UAV in close air combat based on PPO algorithm[J]. Electronics Optics & Control, 2023, 30(1): 9-14. (in Chinese)

[14] 张堃, 李珂, 时昊天, 等. 基于深度强化学习的UAV航路自主引导机动控制决策算法[J]. 系统工程与电子技术, 2020, 42(7): 1567-1574.
ZHANG Kun, LI Ke, SHI Haotian, et al. Autonomous guidance maneuver control and decision making algorithm based on deep reinforcement learning UAV route[J]. System Engineering and Electronics, 2020, 42(7): 1567-1574. (in Chinese)

[15] 肖冰松, 方洋旺, 胡诗国, 等. 一种新的超视距空战威胁评估方法[J]. 系统工程与电子技术, 2009, 31(9): 2163-2166.
XIAO Bingsong, FANG Yangwang, HU Shiguo, et al. New threat assessment method in beyond-the-horizon range air combat[J]. System Engineering and Electronics, 2009, 31(9): 2163-2166. (in Chinese)

[16] 张洪波, 李国英, 丁全心, 等. 超视距空战下的态势评估技术研究[J]. 电光与控制, 2010, 17(4): 9-13.
ZHANG Hongbo, LI Guoying, DING Quanxin, et al. Research on situation assessment in BVR air combat[J]. Electronics Optics & Control, 2010, 17(4): 9-13. (in Chinese)

[17] 吴文海, 周思羽, 高丽, 等. 基于导弹攻击区的超视距空战态势评估改进[J]. 系统工程与电子技术, 2011, 33(12): 2679-2686.
WU Wenhai, ZHOU Siyu, GAO Li, et al. Improvements of situation assessment for beyond visual range air combat based on missile launching envelope analysis[J]. System Engineering and Electronics, 2011, 33(12): 2679-2686. (in Chinese)

[18] 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, 2022, 49(2): 29-35.
DING Linjing, YANG Qiming. Reinforcement learning based unmanned aerial vehicle maneuver decision-making [J]. Avionics Technology, 2022, 49(2): 29-35. (in Chinese)

(编辑:丛艳娟)