航空科学技术
Aeronautical Science & Technology
Jun. 25 2024 Vol. 35 No.06 96-103

基于强化学习的飞行器自主规避决策方法

窦立谦，任梦圆，张秀云，宗群
天津大学，天津 300072

摘　要：考虑飞行器在执行任务过程中存在诸多不可预知的威胁或障碍，为保障飞行器的安全性，本文进行飞行器面向威胁目标的自主规避决策方法研究。首先综合考虑飞行器与威胁目标行为之间的相互影响，提出了基于深度长短期记忆（LSTM）神经网络的轨迹预测算法，实现对威胁目标未来轨迹的预测；然后结合预测信息构建拦截场景下规避机动的马尔可夫决策过程，设计了基于改进双延迟深度确定性策略梯度（P-TD3）的飞行器规避决策方法，以最大化规避过程的总收益为优化目标，实现飞行器自主规避决策。最后通过在虚拟仿真交互平台的试验验证，本文的决策方法提升了网络的收敛速度，具有84%的规避成功率，提高了飞行器对潜在威胁的成功规避概率，有利于增强飞行器的自主性与安全性。

关键词：高超声速飞行器；强化学习；双延迟深度确定性策略梯度；自主规避；机动决策

中图分类号：V249　　**文献标识码：**A　　DOI：10.19452/j.issn1007-5453.2024.06.012

高超声速飞行器通常具有经济性、高效性、安全性、强机动性等特点，已逐渐成为未来空间攻防对抗、应对潜在空间冲突、维护国家安全等方面不可或缺的战略装备，是世界各国航空航天系统的重要研究方向[1-4]。然而，随着飞行器任务与飞行环境的日益复杂，飞行器在执行任务过程中存在诸多不可预知的威胁或障碍，如雷达探测系统及其他飞行器的跟踪、拦截等。因此，研究飞行器自主规避决策方法，对保障飞行器的高效安全飞行、增强飞行器自主能力具有十分重要的意义[5]。

目前飞行器自主机动决策的方法主要分为基于数学模型的传统方法和基于强化学习的人工智能方法。基于数学模型的传统方法包含微分对策法、影响图法、矩阵对策法等[6-10]。杨涛等[11]基于微分对策理论，以飞行器能量为指标，以初始时刻的机动状态、初始位置和速度作为参数建立解析表达式，仿真验证了飞行器的规避效果。Bardhan等[12]根据飞行器与威胁目标攻防模型设计了基于状态方程方法的微分对策制导律，得到了优于经典微分对策理论的规避效能。上述研究均建立在离线规划数学模型的基础上，在实际复杂的飞行环境中，由于无法获得威胁目标的参数信息，飞行器无法在短时间内推导出威胁目标的弹道和制导方式，因此无法自主应对威胁目标的实时跟踪和拦截。

随着人工智能的发展，基于强化学习的人工智能方法可用于求解无模型非线性规划问题，具有求解速度比传统数学算法快的优势，逐渐成为飞行器自主决策领域的研究重点[13-17]。蒋亮等[18]考虑二维平面内向上和向下的推进点火决策，提出了一种基于深度神经网络架构竞争双深度Q网络的飞行器中段突防决策模型，通过引入竞争架构和目标网络架构加快了深度神经网络的收敛速度、增强训练过程中的稳定性。孔维仁等[19]采用状态对抗深度确定性策略梯度算法（SA-DDPG）和逆强化学习算法设计了飞行器自主机动策略生成算法，该算法基于最大熵逆强化学习算法生成奖励，提高了飞行器自主机动策略生成算法的效率。赵宇等[20]将飞行器和多个威胁目标作为多智能体系统，以相对距离和总机动时间为变量设计评价函数，提出了基于多智能体深度确定性策略梯度算法的自主智能决策方法，该方法通过训练实现了飞行器的自主规避脉冲机动。目前的决策理论研究大多集中在无人机等无人系统上，针对飞行器自主规避决策技术的研究还较少。

因此，本文考虑飞行器面临的飞行安全问题，给出了飞行器规避机动场景的任务描述，构建了拦截场景下规避机

收稿日期：2023-10-23；退修日期：2024-01-26；录用日期：2024-02-23
基金项目：国家自然科学基金（62373268，61903349，62073234）；航空科学基金（20170748003）
引用格式：Dou Liqian，Ren Mengyuan，Zhang Xiuyun，et al.Autonomous avoidance decision method for aircraft using reinforcement learning [J].Aeronautical Science & Technology，2024，35(06)：96-103.窦立谦，任梦圆，张秀云，等.基于强化学习的飞行器自主规避决策方法[J].航空科学技术，2024，35(06)：96-103.

动的马尔可夫决策过程，提出基于改进双延迟深度确定性策略梯度（P-TD3）的飞行器规避决策方法。通过考虑威胁目标的行为对飞行器决策的影响，在自主规避决策方法中加入了轨迹预测网络，依据获得的预测信息进行规避决策。通过仿真试验，本文的决策方法实现了飞行器的主动规避，有效提高了飞行器对潜在威胁的成功规避概率，对飞行器自主规避技术研究具有一定的参考价值。

1 飞行器模型

1.1 飞行器动力学模型

本文将飞行器面向威胁目标的规避任务转化为马尔可夫决策过程，其中假设飞行器为T，威胁目标为M。在双方交互过程中，假设飞行器为质量无变化的刚体，只考虑飞行器在三维空间中的质心运动，则其运动模型为

$$\begin{cases} \dot{x}_{\mathrm{T}}=v_{\mathrm{T}}\cos\theta_{\mathrm{T}}\cos\varphi_{\mathrm{T}} \\ \dot{y}_{\mathrm{T}}=v_{\mathrm{T}}\sin\theta_{\mathrm{T}} \\ \dot{z}_{\mathrm{T}}=-v_{\mathrm{T}}\cos\theta_{\mathrm{T}}\sin\varphi_{\mathrm{T}} \\ m\dot{v}_{\mathrm{T}}=F\cos\alpha\cos\sigma-X-mg\sin\theta_{\mathrm{T}} \\ mv_{\mathrm{T}}\dot{\theta}_{\mathrm{T}}=F\left(\sin\alpha\sin\beta-\cos\alpha\sin\sigma\cos\beta\right)+Y\sin\beta+Z\cos\beta \end{cases} \tag{1}$$

式中，$x_{\mathrm{T}},y_{\mathrm{T}},z_{\mathrm{T}}$为地面坐标系下飞行器位置信息；$v_{\mathrm{T}},\theta_{\mathrm{T}},\varphi_{\mathrm{T}}$分别为飞行器的速度、航迹角、航向角；$\alpha,\beta,F$分别为飞行器迎角、倾侧角、推力；$\sigma$为飞行器侧滑角，设置为0；$X,Y,Z$分别为飞行器所受阻力、升力、侧向力。

1.2 威胁目标制导律

在飞行器规避威胁目标的场景中，假设威胁目标以比例导引制导律拦截飞行器，其控制形式为

$$\begin{cases} n_1=k_1\left|\dot{r}_{\mathrm{los}}\right|\dot{q}_{\varepsilon} \\ n_2=k_2\left|\dot{r}_{\mathrm{los}}\right|\dot{q}_{\delta} \end{cases} \tag{2}$$

式中，k_1,k_2为比例导引系数；r_{los}为威胁目标与飞行器的视距；$\dot{r}_{\mathrm{los}}$为视距变化率；$\dot{q}_{\varepsilon}$为视线高低角变化率；$\dot{q}_{\delta}$为视线方位角变化率；n_1为威胁目标的垂直面控制量；n_2为威胁目标的水平面控制量，式中涉及物理量可参考图1。

由此，威胁目标的运动学方程可表示为式（3）

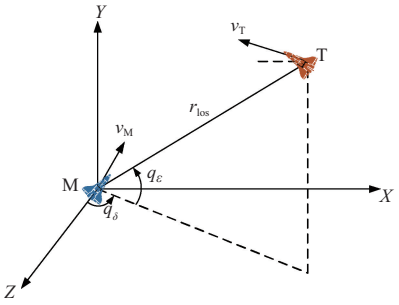


图1 三维比例导引律的几何关系

Fig.1 Geometric relation of three dimensional proportional guidance law

$$\begin{cases} \dot{x}_{\mathrm{M}}=v_{\mathrm{M}}\cos\theta_{\mathrm{M}}\cos\varphi_{\mathrm{M}},\ \dot{y}_{\mathrm{M}}=v_{\mathrm{M}}\sin\theta_{\mathrm{M}} \\ \dot{z}_{\mathrm{M}}=-v_{\mathrm{M}}\cos\theta_{\mathrm{M}}\sin\varphi_{\mathrm{M}},\ \dot{v}_{\mathrm{M}}=0 \\ \dot{\theta}_{\mathrm{M}}=g(n_y-\cos\theta_{\mathrm{M}})/v_{\mathrm{M}} \\ \dot{\varphi}_{\mathrm{M}}=-gn_z/(v_{\mathrm{M}}\cos\theta_{\mathrm{M}}) \end{cases} \tag{3}$$

式中，$x_{\mathrm{M}},y_{\mathrm{M}},z_{\mathrm{M}}$为威胁目标位置信息；$v_{\mathrm{M}},\theta_{\mathrm{M}},\varphi_{\mathrm{M}}$分别为威胁目标的速度、航迹角、航向角；$g$为重力加速度。

2 威胁目标轨迹预测

在飞行器规避过程中，飞行器与威胁目标之间的行为耦合相关，提前获取威胁目标的未来轨迹可以为飞行器的机动决策过程提供依据，使其尽早规避威胁。由此，本文基于深度长短期记忆（LSTM）网络设计如图2所示的威胁目标预测网络，该网络以飞行器与威胁目标的历史状态信息为输入，通过数据处理、特征提取以及双层LSTM网络的时序分析，最终在网络输出层输出预测的威胁目标未来轨迹。

（1）数据处理层

根据飞行器与威胁目标的运动模型，飞行器与威胁目标的历史信息包含过去每一时刻各自位置、速度、航迹角、航向角信息，所预测的未来轨迹是威胁目标在下一时刻的位置信息，在地面坐标系下表示为$x'_{\mathrm{M}},y'_{\mathrm{M}},z'_{\mathrm{M}}$。考虑该预测过程是有监督的学习问题，预测网络首先在数据处理层将飞行器与威胁目标的历史数据分解为训练样本和标签，其

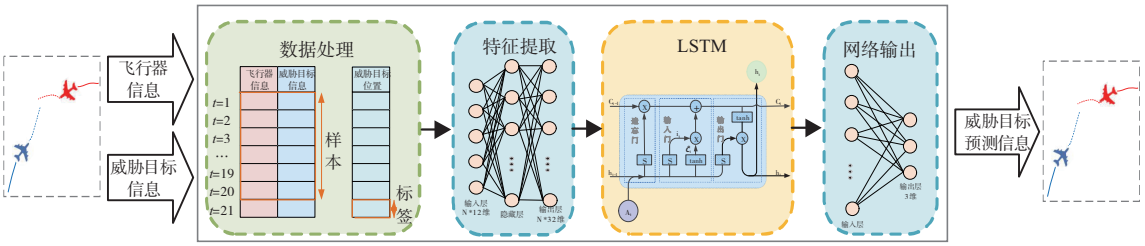


图2 威胁目标轨迹预测网络图

Fig.2 Trajectory prediction network of threat target

中训练样本为前n个时刻内的双方历史信息，标签为威胁目标在第n时刻的位置信息。预测网络通过该过程构造出规避过程的数据集，为进一步训练提供依据。

(2)特征提取层

考虑神经网络在拟合非线性关系中的优势，预测网络利用一个全连接网络作为特征提取层，将输入数据映射到新的空间，以便于预测网络在拟合过程中探索状态信息间隐含的相关性。

(3)LSTM网络层

预测网络利用LSTM网络的特殊结构学习基于双方历史信息的时间序列数据之间的关系，通过其中的遗忘门输出对前一时刻信息的取舍概率以控制是否丢弃，通过输入门确定要添加到当前时刻的新信息，通过输出门计算当前时刻的隐藏状态，并依据时间顺序循环计算，最终获得与前序数据相关的当前时刻的隐藏状态。

(4)网络输出层

网络输出层为一个线性变换网络，预测网络最终通过网络输出层将LSTM输出的当前时刻的隐藏状态映射为三维数据，即输出对威胁目标未来位置的预测值，表示为$\hat{x}'_{\mathrm{M}},\hat{y}'_{\mathrm{M}},\hat{z}'_{\mathrm{M}}$。

在预测网络的训练过程中，预测网络采用均方差函数作为损失函数，如式(4)所示

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(O_i-P_i\right)^2 \tag{4}$$

式中，n表示每一回合中训练过程批量样本的个数，$i\in[1,n]$代表该批量样本中的第i个样本，O_i表示第i个样本中威胁目标真实的未来位置，即$x'_{\mathrm{M}},y'_{\mathrm{M}},z'_{\mathrm{M}}$；$P_i$表示该样本下神经网络预测的威胁目标未来位置，即$\hat{x}'_{\mathrm{M}},\hat{y}'_{\mathrm{M}},\hat{z}'_{\mathrm{M}}$。网络以最小化威胁目标未来位置的真实值与预测值的均方差损失为网络参数的优化目标，利用反向传播算法逐步更新网络参数，从而实现对威胁目标未来信息的预测。

通过威胁目标预测网络输出的预测信息，可以辅助飞行器感知威胁目标的未来运动趋势，为飞行器的规避决策奠定基础。

3 飞行器自主规避决策方法

针对飞行器自主规避决策问题，本文首先基于飞行器运动模型，考虑威胁目标的行为对飞行器决策的影响，综合飞行器的机动能力、双方的状态信息以及第2节中的预测信息，建立了面向飞行器规避任务的马尔可夫决策过程；然后设计了基于改进双延迟深度确定性策略梯度的飞行器自主规避决策方法(P-TD3)，其结构如图3所示，利用该算法求解最优策略；最终通过迭代不断更新决策网络与评价网络的权值，实现飞行器智能自主规避。

3.1 飞行器自主规避马尔可夫决策过程

面向飞行器自主规避决策任务，本文参考飞行器运动模型，综合考虑飞行器的机动能力和双方的状态信息，建立了面向飞行器规避任务的马尔可夫决策过程，其各个要素空间的定义如下。

(1)状态空间S：考虑规避任务需求，将飞行器的状态信息、威胁目标的状态信息以及对威胁目标的预测信息作为飞行器面向规避任务的状态s，即式(5)。其中，根据上述飞行器运动模型，考虑飞行器与威胁目标的相对运动，飞行器与威胁目标的状态信息包含各自的位置、速度、航迹角和航向角。

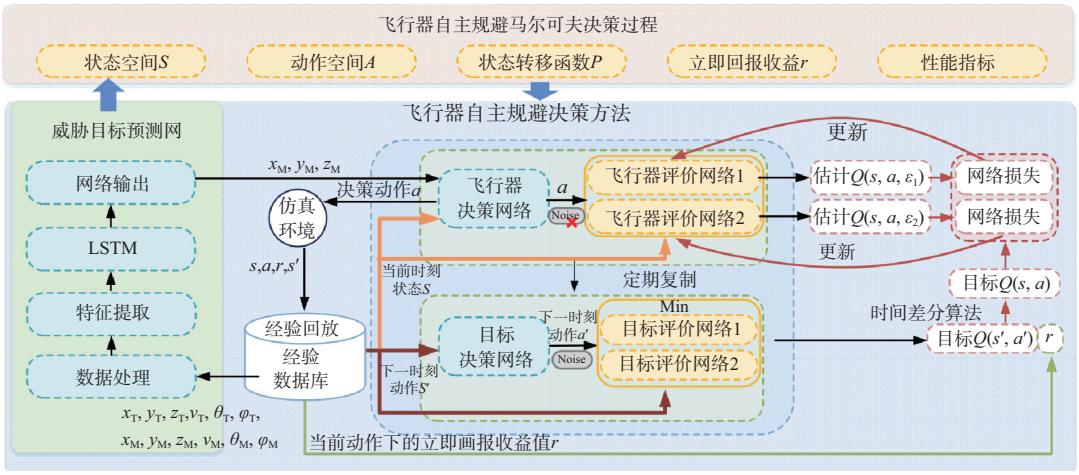


图3 飞行器自主规避决策方法结构图

Fig.3 The structure of aircraft autonomous avoidance decision method

$$s=\left(x_{\mathrm{T}},y_{\mathrm{T}},z_{\mathrm{T}},v_{\mathrm{T}},\theta_{\mathrm{T}},\varphi_{\mathrm{T}},x_{\mathrm{M}},y_{\mathrm{M}},z_{\mathrm{M}},v_{\mathrm{M}},\theta_{\mathrm{M}},\varphi_{\mathrm{M}},\hat{x}'_{\mathrm{M}},\hat{y}'_{\mathrm{M}},\hat{z}'_{\mathrm{M}}\right)\in S \tag{5}$$

式中，$\hat{x}'_{\mathrm{M}},\hat{y}'_{\mathrm{M}},\hat{z}'_{\mathrm{M}}$是预测网络对威胁目标的预测信息，即威胁目标未来位置的预测值。

（2）动作空间A：本文考虑飞行器常用机动方式，飞行器的控制量一般为迎角、倾侧角以及推力。本文为减少飞行器燃料消耗，将推力F设置为0，并将飞行器迎角、倾侧角作为动作空间，即$a=[\alpha,\beta]\in A$。

（3）状态转移函数P：将飞行器的运动学方程式（1）作为飞行器的状态转移函数。

（4）立即回报收益值r：为了减小飞行器自主规避决策行为对后续任务的影响，设置了任务目标点以限制飞行器采取不合理的规避决策，因此考虑双方相对位置、与目标点的相对位置建立奖惩机制，确定单步决策立即回报收益值，如式(6)所示

$$r=\begin{cases}-c_1*d_{\mathrm{end}}^2-c_2*\left(\dfrac{1}{d_{\mathrm{T}}}-\dfrac{1}{\rho_{\mathrm{A}}}\right)^2, d_{\mathrm{T}}\leqslant\rho_{\mathrm{A}}\\ -c_1*d_{\mathrm{end}}^2, \text{ other}\end{cases} \tag{6}$$

式中，d_{T}为飞行器与威胁目标的距离；ρ_{A}为威胁目标的威胁半径；d_{end}为与任务目标点的距离；c_1,c_2为常数系数。从式(6)可以看出，飞行器距离威胁目标越近，则奖励值越小；飞行器距离目标点越近，奖励值越大。

根据上述建立的面向飞行器规避任务的马尔可夫决策过程，设计其性能指标为最大化决策过程飞行器的总收益，即式（7）

$$\max_{\pi} Q=\max_{\pi}\sum_{t=1}^{\infty}\gamma^{t-1}r_t \tag{7}$$

式中，Q表示飞行器在完整决策π过程中获得的总收益；r_t表示t时刻飞行器获得的立即回报收益值；γ为折扣因子，表示未来累积回报收益值相对于当前决策的重要程度；为了实现飞行器自主规避过程，设置决策过程的优化目标是使Q最大。

3.2 飞行器自主规避决策求解方法

为了求解3.1节中面向规避决策任务的策略，实现任务收益的最大化，本文提出基于预测信息的改进双延迟深度确定性策略梯度算法，其网络结构如图3所示，包含威胁目标预测网络、飞行器决策网络、目标决策网络、飞行器评价网络1、飞行器评价网络2、目标评价网络1和目标评价网络2。其中威胁目标预测网络的结构如图2所示，其余网络均由三层全连接网络组成。威胁目标预测网络通过历史状态数据获取威胁目标的预测信息，飞行器决策网络输入飞行器的状态信息、威胁目标的状态信息以及对威胁目标的预测信息，依据确定性策略输出飞行器机动动作，即迎角和倾侧角。飞行器评价网络接收动态环境的状态信息和飞行器的机动动作信息，输出飞行器在该状态下采取此机动动作可能获得的总收益值，用来评估该动作的好坏，从而指导决策网络的改进。

根据3.1节中决策过程的性能指标，P-TD3方法中飞行器决策网络的目标就是最大化决策过程的总收益。由于强化学习通过改变网络参数实现优化，因此可将式（7）描述为式(8)

$$\max_{\pi(\vartheta)} J(\vartheta)=\max_{\pi(\vartheta)}\sum_{t=1}^{\infty}\gamma^{t-1}r_t\left(s_t,\pi\left(a_t|s_t\right)\right) \tag{8}$$

式中，ϑ表示飞行器决策网络的权值；$J(\vartheta)$表示该决策算法中基于神经网络拟合的决策过程的总收益。s_t表示t时刻双方的状态信息；a_t表示t时刻飞行器采取的决策动作，$\pi(a_t|s_t)$表示在状态s_t下依据当前网络参数ϑ输出动作值为a_t的概率，r_t表示t时刻飞行器获得的立即回报收益值。因此，飞行器决策网络通过最小化梯度$\nabla_{\vartheta}J(\vartheta)$优化该决策网络的权值，基于贝尔曼方程与梯度下降方法可将$\nabla_{\vartheta}J(\vartheta)$描述为如下形式

$$\nabla_{\vartheta}J(\vartheta)=E\left[\nabla_{\vartheta}\pi\left(a_t|s_t\right)\bullet\nabla_{\pi}Q^{\pi}(s_t,a_t,\varepsilon)\right] \tag{9}$$

式中，∇为梯度计算符号；$Q^{\pi}(s_t,a_t,\varepsilon)$为飞行器评价网络输出的估计$Q$值；$\varepsilon$表示飞行器评价网络的权值。

为使飞行器评价网络输出的估计$Q^{\pi}(s_t,a_t,\varepsilon)$值近似真实值$Q^{\pi}(s_t,a_t)$，采用均方差函数作为损失函数优化其权值参数，即使式(10)最小

$$L(\varepsilon)=E[(Q^{\pi}(s_t,a_t,\varepsilon)-Q^{\pi}(s_t,a_t))^2] \tag{10}$$

式中，$Q^{\pi}(s_t,a_t)$表示真实值，可以根据时间差分算法近似表示。为了防止过估计，P-TD3方法采用两套具有不同网络参数的飞行器评价网络，选择两个评价网络的输出中较小的估计值进行计算，如式(11)所示

$$Q(s_t,a_t)=r_t+\gamma\min\left\{Q_1(s_t',a_t'),Q_2(s_t',a_t')\right\} \tag{11}$$

式中，s_t',a_t'表示下一时刻飞行器的状态与动作，$Q_1(s_t',a_t'),Q_2(s_t',a_t')$分别是飞行器评价网络1、飞行器评价网络2针对下一时刻状态估计的累积回报收益值。其思想是在训练过程中两个网络的更新目标均为拟合实际Q值，但由于网络参数初始值不同，计算的Q值也不同，此时采用其中最小的Q值将有利于避免过估计。最终通过梯度下降方法最小化梯度$\nabla_{\varepsilon}L(\varepsilon)$更新飞行器评价网络权值，即

$$\nabla_{\varepsilon}L(\varepsilon)=E[(Q^{\pi}(s_t,a_t,\varepsilon)-Q^{\pi}(s_t,a_t))\cdot\nabla_{\varepsilon}Q^{\pi}(s_t,a_t,\varepsilon)] \tag{12}$$

此外，为保障网络更新过程的稳定性，提升训练的收敛速度，P-TD3方法将飞行器决策网络、飞行器评价网络1、飞行器评价网络2的参数采用软更新方法定期复制到目标决策网络、目标评价网络1、目标评价网络2中，如式(13)所示

$$\begin{aligned}\vartheta' &\leftarrow \tau\vartheta+(1-\tau)\vartheta'\\ \varepsilon_1' &\leftarrow \tau\varepsilon_1+(1-\tau)\varepsilon_1'\\ \varepsilon_2' &\leftarrow \tau\varepsilon_2+(1-\tau)\varepsilon_2'\end{aligned} \tag{13}$$

式中，ϑ'为目标决策网络的参数；ϑ为飞行器决策网络的参数；$\varepsilon_1',\varepsilon_2'$分别为目标评价网络1和目标评价网络2的权重；$\varepsilon_1,\varepsilon_2$分别为飞行器评价网络1和飞行器评价网络2的权重；$\tau$为软更新系数。

考虑强化学习训练阶段需要从经验数据库中提取数据，P-TD3方法在飞行器每次与智能仿真交互平台的交互过程中，将当前时刻的状态信息s_t、决策动作a_t、立即回报收益值r_t及下一时刻的状态信息s_t'以集合的方式(s_t,a_t,r_t,s_t')存入经验数据库中；然后采用随机经验回放机制训练威胁目标预测网络、飞行器决策网络与评价网络，以保障威胁目标预测网络的适应性和飞行器决策网络的有效性。

4 仿真结果与分析

由于强化学习的训练过程需要与环境不断交互，本文基于Python算法和Unity3D游戏引擎搭建了可视化飞机虚拟仿真平台，其效果如图4所示。该平台根据飞行器模型与威胁目标运动规律在Unity中搭建物理模型，根据所提出的算法在Python端搭建神经网络模型，通过ML-agent安装包实现Python端和Unity端之间的数据传输。在仿真试验中，部分场景的飞行器初始条件设置见表1。飞行器决策算法中各部分网络结构参数设置见表2，在训练过程中，设置$\gamma=0.9$，预测网络、决策网络、评价网络的学习率为1×10^{-3}，目标网络的软更新率为$\tau=5\times10^{-3}$，用于预测的历史状态时间步长为$n=20$；经验数据库的大小为1×10^{6}，同时每500时间步从经验数据库中提取批量大小为512的数据样本用于网络训练。

研究根据上述仿真设置进行网络训练，其中轨迹预测网络的预测结果如图5所示，图中以时间为横坐标，分别以威胁目标位置信息x_M,y_M,z_M为纵坐标，其中黄线为真实轨迹，蓝线为预测轨迹，可以看出，预测网络对未来时刻的预测轨迹与真实轨迹的趋势一致且偏差较小，其预测误差平均为67m。同时，本文对比了考虑预测信息的P-TD3算法与普通TD3算法的网络训练过程的奖励值变化，如图6所示，可以看出，网络在150回合后学会规避决策，而考虑预测信息的P-TD3算法收敛速度更快。这说明提前感知对方态势对飞行器实现规避决策具有指导作用，这也与一般战场经验相符。

图4 仿真平台效果图

Fig.4 The rendering of simulation platform

表1 飞行器的初始参数设置

Table 1 Initial parameters of the aircraft

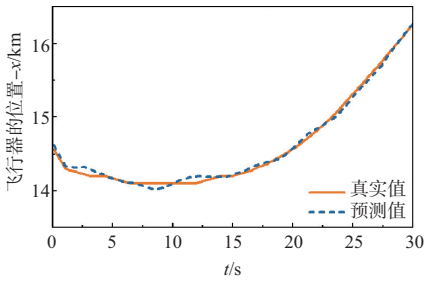
参数	飞行器	威胁目标
制导律	—	比例导引
初始位置/km	(0,14,0)	(2,0,12)
初始速度/(m/s)	1500	1500
航迹角/(°)	0	0
航向角/(°)	0	0

表2 网络参数设置

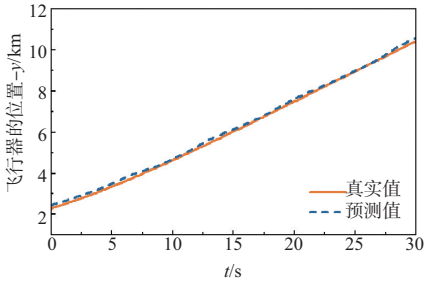
Table 2 Network parameters setting

网络	名称	参数
预测网络	LSTM层节点数	64
	LSTM层节点数	256
飞行器决策网络	全连接神经网络隐藏节点数	256
	全连接神经网络输出节点数	3
	全连接神经网络层数	3
	网络激活函数(全连接层)	Tanh
飞行器评价网络	全连接神经网络隐藏节点数	256
	全连接神经网络输出节点数	1
	全连接神经网络层数	3
	网络激活函数(全连接层)	ReLU

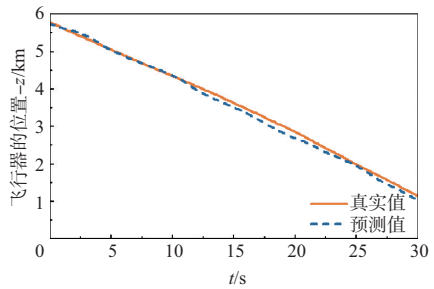
图7显示了飞行器决策过程，红线分别是飞行器在随机机动、无机动或P-TD3策略时的轨迹，蓝线表示在飞行器的不同策略下威胁目标根据其制导律产生的轨迹变化。图8为飞行器在不同策略下的控制量输出，图9是飞行器在随机机动、无机动或P-TD3策略时，威胁目标根据其制导律产生的过载量变化。可以看出，通过训练，飞行器在接近威胁目标时通过拉大过载自主规避威胁，并且在规避过程中有效消耗了威胁目标的过载量。仿真在不同测试环境下统计了算法的规避脱靶量见表3，与随机机动策略相比，所提出的算法规避脱靶量平均增加了41.4m，成功率提升了22%，与普通TD3算法相比，本文算法的规避性能也有所提



(a)威胁目标的预测结果– x_M



(b)威胁目标的预测结果–y_M



(c)威胁目标的预测结果–z_M

图5　威胁目标轨迹预测仿真结果

Fig.5　Result of trajectory prediction for threat targets

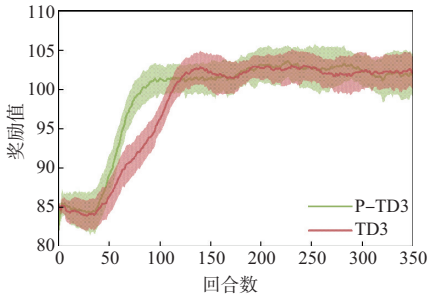


图6　奖励值变化曲线

Fig.6　The change curve of reward value

升，验证了本文算法的有效性与优势。

5 结束语

本文针对飞行器面临的飞行安全问题，首先考虑到威胁目标的行为对飞行器决策的影响，设计了基于LSTM神经网络的轨迹预测算法，预测威胁目标未来轨迹；然后综合

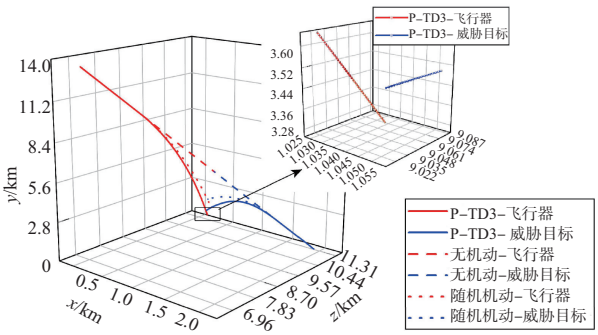
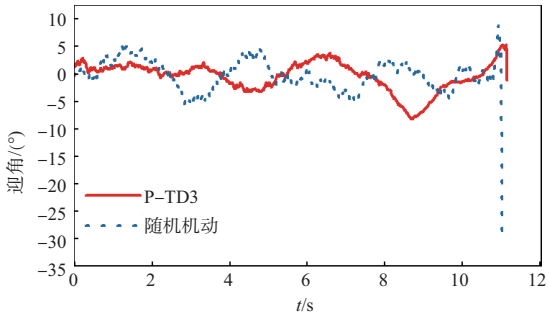
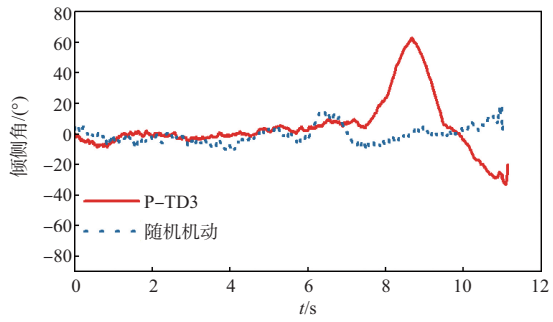


图7　飞行器规避决策轨迹

Fig.7　Avoidance decision trajectory of the aircraft



(a) 飞行器的迎角控制量

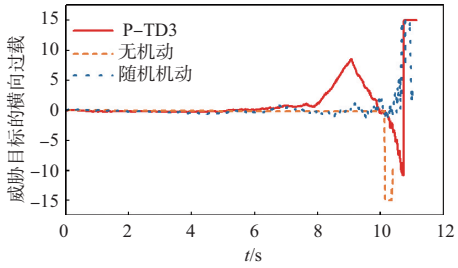


(b) 飞行器的倾侧角控制量

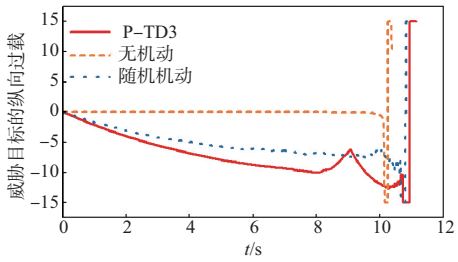
图8　飞行器控制量变化

Fig.8　The change curve of control value

预测信息与马尔可夫决策过程理论将飞行器面向威胁目标的规避任务转化为马尔可夫决策过程，依据飞行器运动模型，建立面向飞行器规避任务的马尔可夫决策过程；最终设计了基于改进双延迟深度确定性策略梯度的飞行器自主规避决策方法求解最优策略，通过迭代更新决策网络与评价网络的权值，实现飞行器自主规避决策。试验表明，考虑预测信息的飞行器决策方法有利于网络训练的收敛，可以实现飞行器的智能自主规避，并有效提升了飞行器规避威胁目标的成功率，可以为保障飞行器安全自主飞行提供支撑。



(a) 威胁目标的横向过载



(b) 威胁目标的纵向过载

图9 威胁目标的过载量

Fig.9 Overload of the threat target

表3 算法成功率

Table 3 Success rate of algorithms

机动方法	平均脱靶量/m	成功率/%
无机动	5.57	0
随机机动	66.31	62
TD3	98.87	80
P–TD3	107.17	84

参考文献

[1] 张秀云,李智禹,宗群,等. 复杂环境影响下空天飞行器智能决策与控制方法发展分析[J]. 空天技术,2022, 1(1):39-53+82.
Zhang Xiuyun, Li Zhiyu, Zong Qun, et al. Analysis of the development of intelligent flight control methods for aerospace vehicle under the influence of complex environment[J]. Aerospace Technology, 2022, 1 (1):39-53+82.(in Chinese)

[2] 王长青. 空天飞行技术创新与发展展望[J]. 宇航学报,2021, 42(7): 807-819.
Wang Changqing. Technological innovation and development prospect of aerospace vehicle[J]. Journal of Astronautics, 2021, 42(7): 807-819. (in Chinese)

[3] 窦立谦,唐艺璠,张秀云. 执行器故障下临近空间飞行器容错控制重构[J]. 天津大学学报(自然科学与工程技术版),2023, 56(2):160-168.
Dou Liqian, Tang Yifan, Zhang Xiuyun. Fault-tolerant control reconstruction of near space vehicle under actuator faults[J]. Journal of Tianjin University (Science and Technology), 2023, 56(2):160-168.(in Chinese)

[4] 宋庆国. 百年未有之大变局下的航空科技发展[J]. 航空科学技术,2021, 32(3):1-5.
Song Qingguo. The development of aviation science and technology under changes unseen in a century[J]. Aeronautical Science & Technology, 2021, 32(3):1-5.(in Chinese)

[5] 符小卫,吴迪,支辰元. 基于改进向量场直方图算法的无人机动态避障策略[J]. 航空科学技术,2023, 34(9):100-109.
Fu Xiaowei, Wu Di, Zhi Chenyuan. Dynamic obstacle avoidance of UAV based on improved vector field histogram algorithm[J]. Aeronautical Science & Technology, 2023, 34(9):100-109.(in Chinese)

[6] Shen Zhipeng, Yu Jianglong, Dong Xiwang, et al. Deep neural network-based penetration trajectory generation for hypersonic gliding vehicles encountering two interceptors[C]. 2022 41st Chinese Control Conference(CCC), 2022:3392-3397.

[7] Mishley A, Shaferman V. Linear quadratic guidance laws with intercept angle constraints and varying speed adversaries[J]. Journal of Guidance Control and Dynamics, 2022, 45(11): 2091-2106.

[8] Shen Zhipeng, Yu Jianglong, Dong Xiwang, et al. Penetration trajectory optimization for the hypersonic gliding vehicle encountering two interceptors[J]. Aerospace Science and Technology, 2022, 121(2): 107363.

[9] Turetsky V, Weiss M, Shima T. A combined Linear Quadratic/Bounded control differential game guidance law[J]. IEEE Transactions on Aerospace and Electronic Systems, 2021, 57 (5): 3452-3462.

[10] Wang Yaokun, Zhao Kun, Guirao J L G, et al. Online intelligent maneuvering penetration methods of missile with respect to unknown intercepting strategies based on reinforcement learning [J]. Electronic Research Archive, 2022, 30(12):4366-4381.

[11] Yang Tao, Geng Lina, Duan Mingkuan, et al. Research on the evasive strategy of missile based on the theory of differential game[C].34th Chinese Control Conference (CCC) , 2015: 5182-5187.

[12] Bardhan R, Ghose D. Nonlinear differential games-based impact-angle-constrained guidance law[J]. Journal of Guidance, Control, and Dynamics, 2015, 38(3):384-402.

[13] 崔雅萌,王会霞,郑春胜,等. 高速飞行器追逃博弈决策技术

[J].指挥与控制学报,2021, 7(4):403-414.
Cui Yameng, Wang Huixia, Zheng Chunsheng, et al. Pursuit-evasion game decision technology of high speed vehicles[J]. Joural of Command and Control, 2021, 7(4):403-414.(in Chinese)

[14] 朱雅萌,张海瑞,周国峰,等. 一种基于深度强化学习的机动博弈制导律设计方法[J]. 航天控制,2022, 40(3): 28-36.
Zhu Yameng, Zhang Hairui, Zhou Guofeng, et al. A design method of maneuvering game guidance law based on deep reinforcement learning[J]. Aerospace Control, 2022, 40(3):28-36.(in Chinese)

[15] Huang Hongji, Yang Yuchun, Wang Hong, et al. Deep reinforcement learning for UAV navigation through massive MIMO technique[J]. IEEE Transactions on Vehicular Technology, 2020, 69(1):1117-1121.

[16] Ouahouah S, Bagaa M, Prados-Garzon J, et al. Deep-reinforcement-learning-based collision avoidance in UAV environment [J]. IEEE Internet of Things Journal, 2022, 9(6):4015-4030.

[17] Kong Xue, Ning Guodong, Yang Ming, et al. A maneuvering penetration strategy via integrated flight/propulsion guidance and control method for air-breathing hypersonic vehicle[C]. 2018 IEEE CSAA Guidance, Navigation and Control Conference (CGNCC), 2018:1-6.

[18] Jiang Liang, Nan Ying, Li Zhihan. Realizing midcourse penetration with deep reinforcement learning[J]. IEEE Access, 2021, 9: 89812-89822.

[19] Kong Weiren, Zhou Deyun, Zhen Yang, et al. UAV autonomous aerial combat maneuver strategy generation with observation error based on state-adversarial deep deterministic policy gradient and inverse reinforcement learning[J]. Electronics, 2020, 9(7):1121.

[20] Zhao Yu, Zhou Ding, Bai Chengchao, et al. Reinforcement learning based spacecraft autonomous evasive maneuvers method against multi-interceptors[C]. 2020 3rd International Conference on Unmanned Systems (ICUS) , 2020:1108-1113.

Autonomous Avoidance Decision Method for Aircraft Using Reinforcement Learning

Dou Liqian, Ren Mengyuan, Zhang Xiuyun, Zong Qun

Tianjin University, Tianjin 300072, China

Abstract: There are many unpredictable threats or obstacles in the course of the mission of the aircraft. In order to solve the problem of autonomous avoidance decision of aircraft facing threat targets, firstly, a trajectory prediction algorithm based on deep Long Short-Term Memory (LSTM) neural network is proposed to predict the future trajectory of threat targets by considering the interaction between aircraft and threat targets. Secondly, the Markov decision process of evasive maneuver in the interception scenario was constructed combined with the prediction information . Then, the avoidance decision method based on progressed double delay depth deterministic strategy gradient (P-TD3) was proposed to maximize the benefits of the circumvention process to achieve intelligent autonomous avoidance decisions for the aircraft. Finally, the simulation experiments verify that the decision-making method improves the convergence speed of the network and has an 84% success rate of avoidance, which improves the probability of successful avoidance of potential threats and enhances the autonomy and safety of the aircraft.

Key Words: hypersonic aircraft; reinforcement learning; double delay depth deterministic strategy gradient; autonomous avoidance; maneuver decision

Received: 2023-10-23; **Revised:** 2024-01-26; **Accepted:** 2024-02-23

Foundation item: National Natural Science Foundation of China (62373268,61903349,62073234); Aeronautical Science Foundation of China (20170748003)