工程科学学报
Chinese Journal of Engineering

基于LSTMPPO算法的多机空战智能决策及目标分配

丁云龙 匡敏驰 朱纪洪 祝靖宇 乔直

Intelligent decision making and target assignment of multi-aircraft air combat based on the LSTM–PPO algorithm

DING Yunlong, KUANG Minchi, ZHU Jihong, ZHU Jingyu, QIAO Zhi

引用本文:
丁云龙, 匡敏驰, 朱纪洪, 祝靖宇, 乔直. 基于LSTM－PPO算法的多机空战智能决策及目标分配[J]. 北科大：工程科学学报, 2024, 46(7): 1179–1186. doi: 10.13374/j.issn2095–9389.2023.10.13.003
DING Yunlong, KUANG Minchi, ZHU Jihong, ZHU Jingyu, QIAO Zhi. Intelligent decision making and target assignment of multi–aircraft air combat based on the LSTM－PPO algorithm[J]. *Chinese Journal of Engineering*, 2024, 46(7): 1179–1186. doi: 10.13374/j.issn2095–9389.2023.10.13.003

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2023.10.13.003

工程科学学报, 第 46 卷, 第 7 期: 1179−1186, 2024 年 7 月
Chinese Journal of Engineering, Vol. 46, No. 7: 1179−1186, July 2024
https://doi.org/10.13374/j.issn2095-9389.2023.10.13.003; http://cje.ustb.edu.cn

基于 LSTM–PPO 算法的多机空战智能决策及目标分配

丁云龙[1)], 匡敏驰[1)✉], 朱纪洪[2)], 祝靖宇[2)], 乔　直[2)]

1) 新疆大学计算机科学与技术学院, 乌鲁木齐 830000　2) 清华大学精密仪器系, 北京 100084
✉通信作者, E-mail: kuangmc@tsinghua.edu.cn

摘　要　针对传统多机空战中智能决效率低、难以满足复杂空战环境的需求以及目标分配不合理等问题. 本文提出一种基于强化学习的多机空战的智能决策及目标分配方法. 使用长短期记忆网络(Long short-term memory, LSTM)对状态进行特征提取和态势感知, 将归一化和特征融合后的状态信息训练残差网络和价值网络, 智能体通过近端优化策略(Proximal policy optimization, PPO)针对当前态势选择最优动作. 以威胁评估指标作为分配依据, 计算综合威胁度, 优先将威胁值最大的战机作为攻击目标. 为了验证算法的有效性, 在课题组搭建的数字孪生仿真环境中进行 4v4 多机空战实验. 并在相同的实验环境下与其他强化学习主流算法进行比较. 实验结果表明, 使用 LSTM–PPO 算法在多机空战中的胜率明显优于其他主流强化学习算法, 验证了算法的有效性.

关键词　多机空战; 智能决策; 近端优化策略; 威胁评估; 目标分配
分类号　TP183

Intelligent decision making and target assignment of multi-aircraft air combat based on the LSTM–PPO algorithm

DING Yunlong[1)], *KUANG Minchi*[1)✉], *ZHU Jihong*[2)], *ZHU Jingyu*[2)], *QIAO Zhi*[2)]

1) Xinjiang University, College of Computer Science and Technology, Wulumuqi 830000, China
2) Tsinghua University, Precision Instrument System, Beijing 100084, China
✉Corresponding author, E-mail: kuangmc@tsinghua.edu.cn

ABSTRACT　With the rapid development of intelligent and informationized air battlefields, intelligent air combat has increasingly become key to affecting the outcome of a battlefield. In conventional multi-aircraft air combat, there are issues of low efficiency in intelligent decision-making, difficulty in meeting the needs of complex air combat environments, and unreasonable target allocation. In response to the problems in conventional multi-aircraft air combat, we introduce a long short-term memory–proximal policy optimization algorithm (LSTM–PPO). Using the long short-term memory network to extract features and perceive the situation of the state, an intelligent agent trains the normalized and feature-fused state information residual network and value network, chooses the optimal action through the proximal policy optimization strategy based on the current situation, and embeds a reward function containing expert knowledge during the training process to solve the problem of sparse rewards. Meanwhile, a target allocation algorithm based on threat value calculation is presented. Using angle, speed, and height threat values as the basis for target allocation, the ID of the target aircraft with the highest threat value on the battlefield is calculated in real-time. When the strategy network outputs an action of attack, it conducts target allocation. To confirm the effectiveness of the algorithm, we carried out 4v4 multi-aircraft air combat experiments in a digital twin simulation environment built by our research group. The red team consists of reinforcement learning agents based on LSTM–PPO algorithm, whereas the blue team comprises a finite state machine composed of expert knowledge bases. After more than 1200 rounds of aerial confrontation, the algorithm has been converged, and the win rate of the red team has reached 82%. Furthermore,

收稿日期: 2023−10−13

we assessed the performance of four other mainstream reinforcement learning algorithms in 4v4 air combat experiments under the same experimental conditions. It is shown that the deep Q-network (DQN) and soft actor-critic (SAC) algorithms have difficulties in dealing with high-dimensional continuous action spaces and multiagent collaboration. The multi-agent deep deterministic policy gradient algorithm (MADDPG) employs a multi-agent strategy and cooperative training, so it exhibits a significantly higher win rate than the DQN and SAC algorithms. The multi-agent proximal policy optimization (MAPPO) algorithm has a relatively high failure rate and is not stable enough to deal with enemy aircraft's strategies in some cases. The LSTM–PPO algorithm shows a significantly higher win rate than other mainstream reinforcement learning algorithms in multi-aircraft collaborative air combat, which confirms the effectiveness of the LSTM–PPO algorithm in dealing with high-dimensional continuous action spaces and multi-aircraft collaborative operations.

KEY WORDS　multi-aircraft air combat; intelligent decision; proximal policy optimization; threat assessment; dynamic target assignment

随着科技的不断进步和军事领域的发展, 智能空战作为现代战争的重要组成部分, 备受各国关注. 在空战中, 多架战机的协同作战被认为是提升整体作战水平和执行任务能力的关键因素. 因此, 研究者们的研究逐步由单机向“集群”方向发展[1].

自 20 世纪 60 年代以来, 学者们对多机空战智能决策问题进行了广泛而深入的研究, 取得了一些重要的成果. 这些研究主要可以分为传统的决策方法和基于人工智能的决策方法两大类. 传统的决策方法主要包括矩阵决策法[2−4]、影响图法[5−7]和动态博弈法[8−9]. 此类方法虽然给空战决策提供了解决方案, 但随着空战环境的复杂化, 此类方法在建模和计算精度方面都存在困难. 基于人工智能的方法主要包括专家系统法[10−11]、遗传算法[12−14]、人工神经网络法[15−17] 等. 此类方法可以通过感知态势环境、接收历史状态信息等, 自主高效的输出战机机动决策. 但此类方法也面临一定的局限性, 专家系统需要根据规则构建知识库, 它难以满足空战环境的需要, 从而导致决策的效果和实时性受到限制. 遗传算法通常需要进行大量的迭代和评估个体的适应度, 导致计算量急剧上升. 人工神经网络法需要大量带有人工标注的样本, 但复杂的空战环境使样本标注十分困难. 近年, 也有许多学者致力于强化学习的研究, 孔维仁等[18] 提出利用参数共享 Q 网络来解决多机空战决策问题. 此方法仅限于近距空战, 未对超视距空战做出考量.

针对多智能体空战强化学习算法的挑战, 本文以 4v4 模拟空战环境为基础, 探索了多机空战中的智能决策和目标分配问题. 首先, 从强化学习建模的角度出发, 对战机的机理、状态空间、动作空间以及奖励函数进行建模. 利用 LSTM–PPO (Long short-term memory–proximal policy optimization algorithm) 算法进行信息提取和智能决策, 以实现更精准的决策策略. 同时, 在决策过程中嵌入威胁指标为依据的目标分配算法, 实现实时目标分配. 最后, 通过模拟对抗实验, 对算法的整体性能进行评估.

1　强化学习建模

1.1　空战建模

4v4 多机空战属于部分可观测马尔可夫决策过程 (Partially observable markov decision process, POMDP), 战机在每一个时间步都会面临决策, 但它无法直接观测到系统的完全状态, 而是通过部分信息来推断状态. 4 架战机可共享当前获得的状态信息. 空战基本模型如图 1 所示.

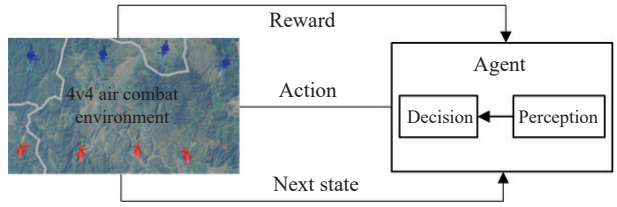

图 1　空战基本模型

Fig.1　Basic model of air combat

1.2　状态空间建模

本文所提及的状态空间主要由导弹信息和飞机信息组成. 导弹信息包括导弹类型 M_{Type}、导弹发射信息 M_{Castable} 以及剩余导弹数量信息 M_{Left} 等. 飞机信息由敌我飞机的高度 (H_{self}、H_{enemy})、位置 (P_{self}、P_{enemy})、速度 (V_{self}、V_{enemy})、俯仰角 φ、滚转角 θ 及偏航角 ϕ 组成. 同时, 飞机信息还包括两架飞机之间的方位信息, 采取经典的 McGrew 法[19] 对战机方位信息进行表示. 其中视线角 (Aspect angle, AA) 表示我机速度方向与双方飞机连线的夹角;天线角 (Antenna train angle, ATA) 表示敌机速度方向与双方飞机连线的夹角;水平交叉角 (Horizonal crossing angle, HCA) 是双方飞机速度方向的夹角. R 表示敌我飞机之间的距离信息. 战机方位信息如图 2 所示.

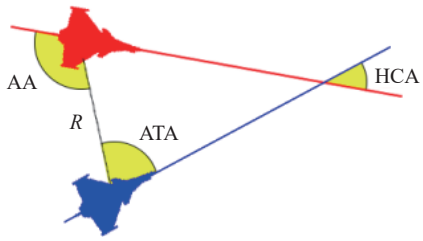

图 2 方位角度示意图

Fig.2 Azimuth angle diagram

综上所述, 总体状态空间定义如下:

$$S=[M_{\text{Type}},M_{\text{Castable}},M_{\text{Left}},H_{\text{self}},H_{\text{enemy}},P_{\text{self}},P_{\text{enemy}},V_{\text{self}},V_{\text{enemy}},\varphi,\theta,\phi,\text{AA},\text{ATA},\text{HCA}] \tag{1}$$

1.3 动作空间建模

为了应对复杂多变的空战环境, 并为智能体在空战过程中开发出更多组合动作, 本文将针对战机设定六种基础动作. 通过合理的灵活运用这些基础动作, 能够做出一些空战中常见的机动, 如殷麦曼机动和眼镜蛇机动[20]. 这六种基础动作包括直飞、追踪、盘旋、筋斗、攻击和躲避. 动作描述如表 1 所示.

1.4 奖励函数建模

面对复杂多变的空战环境, 奖励函数直接影响智能体的训练. 本文针对多机空战引入全局奖励、局部奖励以及节点事件奖励[20]. 全局奖励, 即智能体获得胜利, 获得+100 的奖励, 失败获得–100 的奖励, 平局奖励为–10, 如公式 (2) 所示, $R(s,a)$表示在当前状态s下采取动作a所获得的奖励.

$$R(s,a)=\begin{cases}+100, & \text{combat aircrafts win}\\ -10, & \text{combat aircrafts tie}\\ -100, & \text{combat aircrafts lose}\end{cases} \tag{2}$$

全局奖励函数存在奖励稀疏和奖励延迟的问题. 只有当一场战斗结束, 智能体才会得到奖励的反馈, 不利于智能体的训练. 针对这一问题, 引入局部奖励和节点事件奖励, 局部奖励, 如公式 (3) 所示.

$$R_{\text{A}}=\begin{cases}80\mathrm{e}^{-\frac{\text{Ang}^2}{1300}}\left(\dfrac{D}{1000}\right), & D\leqslant 1000\\ 80\mathrm{e}^{-\frac{\text{Ang}^2}{1300}}\left(\dfrac{39000-D}{38000}\right), & 1000<D\leqslant 20000\\ 80\mathrm{e}^{-\frac{\text{Ang}^2}{1300}}\left(\dfrac{10000}{D}\right), & 2000<D\end{cases}$$

$$R_{\text{T}}=\begin{cases}160\mathrm{e}^{-\frac{\text{ang}^2}{144}}\left(1-\dfrac{t}{20}\right), & 0\leqslant t<20\\ 0, & 20\leqslant t\end{cases} \tag{3}$$

R_{A} 表示我机相对与敌机的优势度, D表示两机之间的距离, Ang 表示两机之间的夹角, 当 Ang 趋向于 0 时, 我方战机相对于敌方战机处于追尾状态, 随着距离的减少, 优势值会有明显的提升. R_{T} 表示敌机导弹的威胁度, ang 表示导弹与我机之间的夹角, t表示导弹预计到达时间. 当 ang 趋向于 0 时, 敌机导弹的威胁呈最大值状态.

局部奖励带有持续性和连贯性, 节点事件奖励则具有瞬时性, 当战机触发某个事件, 会立即获得奖励, 具体奖励值如表 2 所示.

2 多机空战智能决策及目标分配

在多机协同作战中, 目标分配合理可以减少导弹的冗余, 以最少的资源击败敌方战机. 本文提出威胁值优先的目标分配算法, 利用角度、速度和高度威胁指数进行威胁值计算, 选择威胁值最大的敌机作为目标战机.

2.1 多机目标分配

2.1.1 威胁值计算

本文提引入三个关键威胁评估参数[21], 角度、速度和高度. 根据公式分别计算敌机 P_j对友机 P_i的威胁值, 利用公式计算出对当前友机P_i威胁值最大的敌机 ID, 并将其作为打击目标. 用公式 (4) 计算角度威胁值.

$$\alpha_{ij}=(\alpha_i+\alpha_j)/360^\circ \tag{4}$$

表 1 动作空间设计

Table 1 Action design

Type	Parameter	Description
Straight	Target pitch angle	A straight flight is a steady, horizontal, and continuous straight flight.
Track	Target location and bearing	Tracking the behavior of observing a target and approaching its location to achieve continuous observation and contact with the target.
Hover	Roll and pitch	The act of continuously rotating and circulating in a confined space.
Loop	Pitch	A somersault is the act of turning and tumbling rapidly in the air.
Attack	Target prediction position	An attack is the act of violating, attacking, and causing harm to a target by means of action, force, or weapons.
Escape	Alarm information	Avoidance is the taking of measures and actions to avoid or circumvent potential threats, dangers, or attacks.

表 2　节点事件奖励设计

Table 2　Node event reward design

Name	Reward Value	Description
Hit	+100	The missile hit the enemy target.
Lose	–100	Hit by an enemy missile.
Tie	–10	The remaining fighters on both sides are 0.
crash	–100	Their plane hit the ground and crashed.
Scanning	+10	Radar picked up enemy aircraft.
Escape	+10	Successfully evaded the missile at close range.
Past	+10	Close pass and contact with enemy aircraft.

其中, α_i表示我机的视线角(AA), α_j表示敌机的天线角(ATA). 当α_i趋向于 0 或α_j趋向于 180°时, 我机可以更快调整机头方向, 对目标发射导弹.

根据公式 (5) 计算速度威胁值.

$$V_{ij}=\begin{cases}1, & 1.5\,v_i \leqslant v_j \\ v_j/v_i-0.5, & 0.6v_i < v_j < 1.5v_i \\ 0.1, & v_j \leqslant 0.6v_i\end{cases} \tag{5}$$

其中, V_{ij}表示敌机对我机的速度威胁值, v_i表示我机速度, v_j表示敌机速度.

根据公式 (6) 计算高度威胁值.

$$H_{ij}=\begin{cases}1, & h_{ij} \geqslant 5 \\ 0.5+0.1h_{ij}, & -4 < h_{ij} < 5 \\ 0.1, & h_{ij} \leqslant -4\end{cases} \tag{6}$$

其中, H_{ij}表示敌机对我机的高度威胁值, h_{ij}表示敌我之间战机的高度距离差.

2.1.2　目标分配

通过公式计算出敌机对我机的威胁值, 利用公式 (7) 计算综合威胁值.

$$T_j=\sum_{i=1}^{N}\left(\sum_{j=1}^{M}(V_{ij}+H_{ij})\frac{\ln{\alpha_{ij}}^2}{360^\circ}\right) \tag{7}$$

其中, T_j表示第 j架敌机对我机的综合威胁值, N表示我机的数量, M表示敌机的数量, V_{ij},H_{ij},α_{ij}分别代表速度、高度和角度威胁值. 选取综合威胁值最大的战机作为目标.

2.2　决策网络设计

空战决策的目标是追求最优解, 即在保证友方飞机最大生存的同时, 尽可能多地击败敌机. 这是一个复杂而具有挑战性的任务, 需要充分利用飞机获取的时序性状态信息, 并采用先进的技术和算法来实现智能决策. 本文旨在扩展这一领域的研究, 使用一种基于深层 LSTM 模型和全局观测量与部分观测量相分离的 Actor-critic 网络结构[22].

LSTM 是循环神经网络的一种变体, 用于序列数据建模. 它通过引入门控机制和记忆单元解决了传统 RNN 中的长期依赖问题. Actor 网络和 Critic 网络分别使用部分观测空间和全局观测空间, 通过使用全连接神经网络对观测的状态信息做特征提取, 进而使用 LSTM 进一步时序信息特征提取. Actor 网络将提取的时序特征作为残差网络的输入, 最终输出动作的概率分布, Critic 网络利用全连接神经网络输出状态价值, 决策网络流程如图 3 所示.

2.3　决策流程

在 4v4 的空战环境中, 先由战机获取战场信息, 将战场信息转化成包括状态信息和奖励的态势信息, 并作为经验数据发送至数据库. 对态势信息进行预处理, 提取威胁评估指数(角度、速度、高度), 并根据威胁值计算出目标机 ID, 只有当决策动作为 Attack 时, 智能体才会得到目标机 ID. 用归一化后的数据作为决策网络的输入, 输出战机动作, 将动作放入数据库, 与状态信息和奖励组成经验数据, 用来对决策网络做参数更新. 同时, 智能体执行动作并判断是否取得胜利, 当成功消灭所有敌机时被视为胜利. 如果未达成胜利条件, 则继续获取战场信息, 决策流程如图 4 所示.

3　仿真实验与结果

3.1　实验想定

如图 5 所示, 红蓝双方均为 4 架飞机, 红方是基于强化学习的空战智能体, 蓝方为专家系统-有

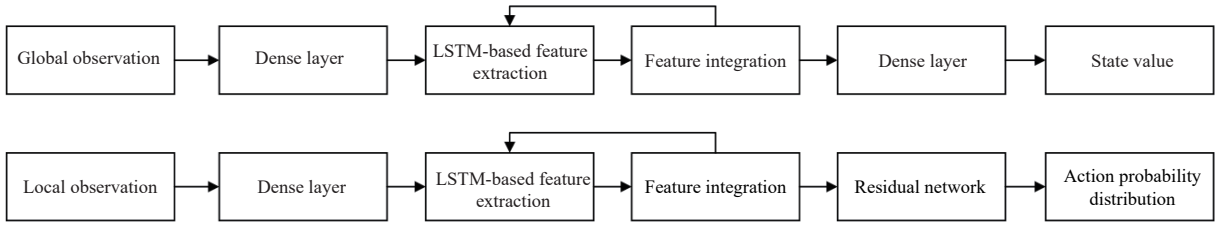

图 3　决策网络流程图

Fig.3　Decision network flow diagram

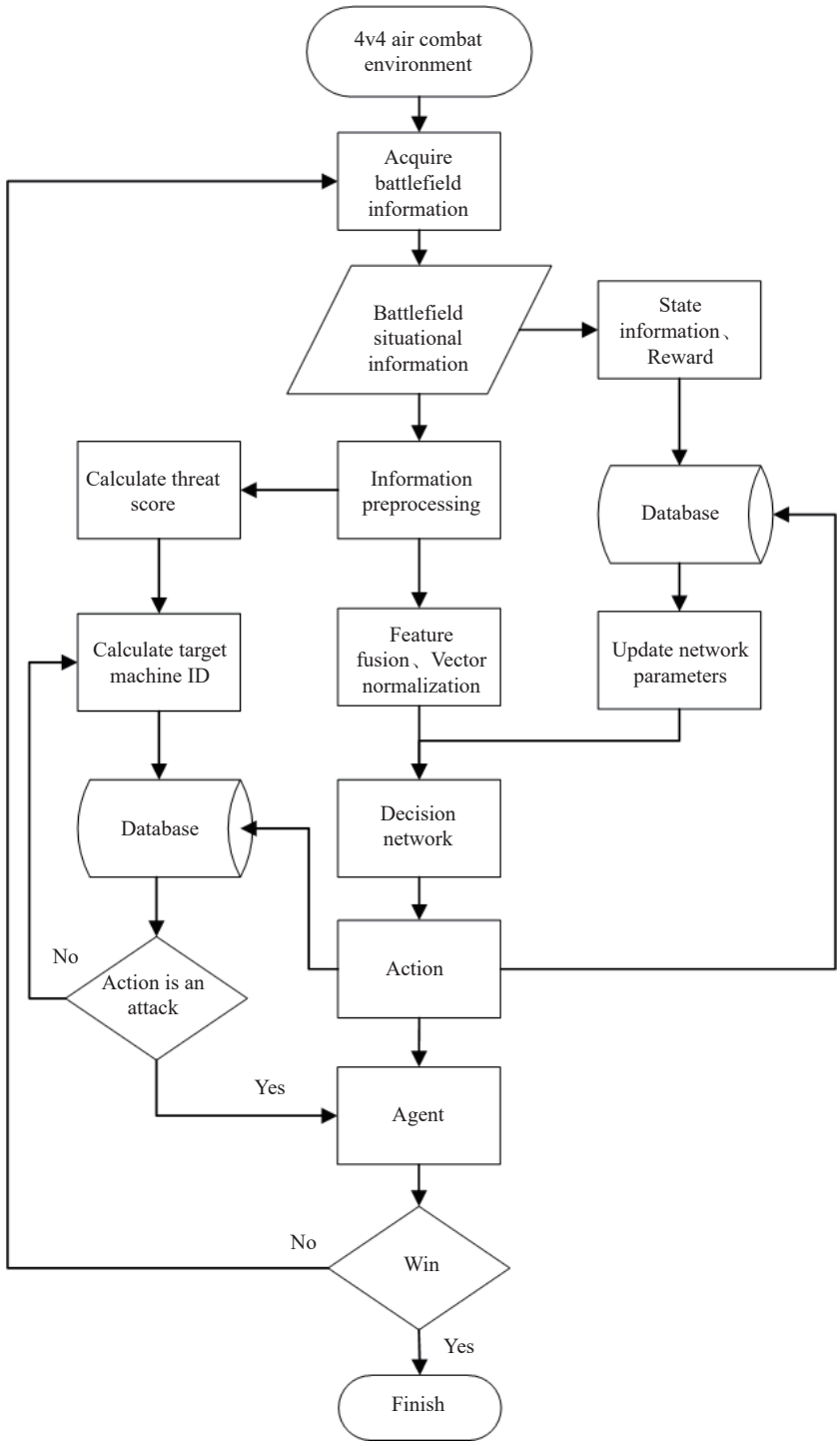

图 4 决策过程流程图

Fig.4 Decision process flow chart

限状态机 (Finite state machine, FSM), 每架飞机均携带六枚导弹.

3.2 实验环境

基于面向强化学习的空战仿真平台[23], 该环境可以模拟空战过程, 并显示飞机、导弹轨迹等功能, 并且 1∶1 还原战机的状态信息, 仿真实验环境如图 6 所示.

3.3 实验仿真与分析

3.3.1 4v4 多机空战

由于强化学习在训练过程中出现的抖动较大, 描绘曲线不易于观察, 为了能够有效可视化曲线的大致趋势, 所以本文对原始数据进行指数滑动平均(Exponential moving average)处理. 在图 7、图 8 和图 9 中, 浅色曲线代表原始数据, 而深色曲线则表示经过滑动平均处理后的数据.

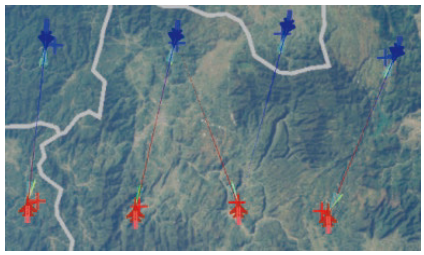

图 5　4v4 实验对抗

Fig.5　4v4 experimental confrontation

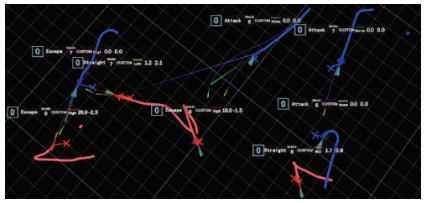

图 6　4v4 实验仿真

Fig.6　4v4 experimental simulation

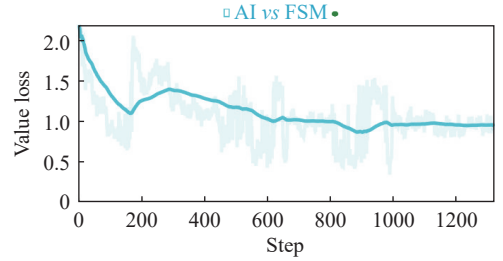

图 7　4v4 空战价值损失函数曲线

Fig.7　4v4 air combat value loss function curve

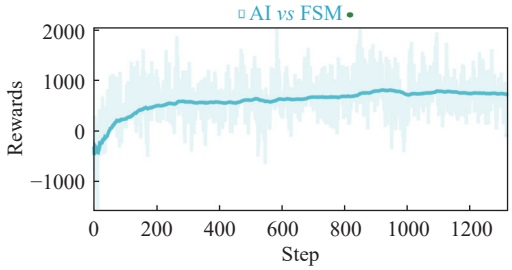

图 8　4v4 空战奖励函数曲线

Fig.8　4v4 air combat reward function curve

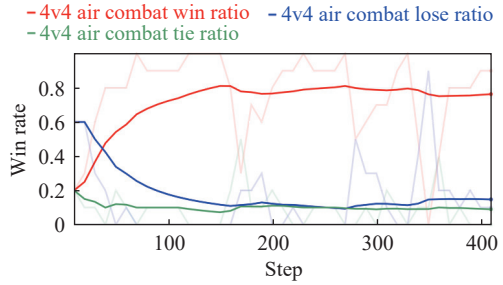

图 9　4v4 多机空战胜负率统计图

Fig.9　Statistical chart of the victory rate in multi-aircraft air combat

经过 1200 多轮的迭代, 智能体不断更新策略参数和价值参数, AI *vs* FSM 的价值损失函数趋于稳定, 说明 LSTM–PPO 算法已经收敛, 如图 7 所示.

多机空战获胜的关键在战机共享状态信息的情况下, 友机之间协同决策, 从而获得更高的奖励. 战机获得的奖励由负到正, 表现出智能体逐步学习的过程, 并且最终获得较高的奖励, 如图 8 所示.

如图 9 所示, 统计近 400 场空战数据, 由 LSTM–PPO 算法做决策的智能体胜率稳定在 80%, 以明显的优势击败专家系统.

3.3.2　算法对比分析

为了充分验证算法的有效性和可行性. 在相同实验环境进行 4v4 空战实验, 使各类主流强化学习算法与状态机对战. 等待算法收敛完成, 主流强化学习算法性能如表 3 所示. 我们评估了五种不同的强化学习算法在 4v4 空空中对抗中的表现. 这些算法包括 DQN(Deep Q-network)[24]、SAC(Soft actor-critic)[25]、MADDPG(Multi-agent deep deterministic policy gradient)[26]、MAPPO(Multi-agent proximal policy optimization)[27] 和 LSTM–PPO, 对比试验的指标包括 4v4 空战的胜率、败率和平局率. DQN 算法在 4v4 空战胜率方面表现最低, 仅为 36%, SAC 算法在胜率方面相对提升, 达到 41%, 这表明 DQN 和 SAC 算法在处理高维连续动作空间和多智能体协作方面依然面临困难. MADDPG 算法使用了多智能体策略和协同训练, 能够更好地应对多智能体环境中的博弈和合作, 所以该算法在 4v4 空战中表现出更高的胜率. MAPPO 算法在胜率方面与 MADDPG 算法相近, 但 MAPPO 算法的败率为 44%, 这表明 MAPPO 算法在一些情况下不够稳定且难以适应敌机的策略. LSTM–PPO 算法在 4v4 空战中表现出色, 胜率高达 82%. 通过引入 LSTM 网络结构做时序处理并结合 PPO 算法做智能决策, 算法显著地提升了 4v4 空战中的性能.

表 3　算法性能基准对比

Table 3　Algorithm performance benchmark comparison　%

Algorithm name	4v4 air combat win rate	4v4 air combat loss rate	4v4 air combat tie rate
DQN	36	57	7
SAC	41	54	5
MADDPG	53	40	7
MAPPO	52	44	4
LSTM–PPO (ours)	82	12	6

4　结论

本文旨在研究 4v4 多机空战的决策和目标分

配问题, 并提出一种利用 LSTM–PPO 算法做多机空战决策. 使用威胁指数作为目标分配依据, 以解决传统多机空战中难以面对复杂状态空间和目标分配不合理等问题. 同时利用多 GPU 并行处理环境数据, 从而加速智能体的收敛.

为了评估算法的有效性, 本文设计 4v4 空战环境, 与基于专家系统设计的有限状态机进行对抗, 同时, 使用主流的强化学习算法进行对比. 实验结果表明, 智能体以 82% 的胜率战胜有限状态机, 并以较大的优势击败其他主流的强化学习算法.

虽然基于 LSTM–PPO 的算法在 4v4 多机空战中取得一定的效果. 但是在奖励函数的设计、动作空间的连续化等方面还需要考虑做出进一步改进. 让多机智能体具有更强大的协同作战能力, 是本文今后进一步的研究方向.

参 考 文 献

[1] Duan H B, Li P. Autonomous control for unmanned aerial vehicle swarms based on biological collective behaviors. *Sci Technol Rev*, 2017, 35(7): 17
(段海滨, 李沛. 基于生物群集行为的无人机集群控制. 科技导报, 2017, 35(7): 17)

[2] Deng K, Peng X Q, Zhou D Y. Study on air combat decision method of UAV based on matrix game and genetic algorithm. *Fire Contr Command Contr*, 2019, 44(12): 61
(邓可, 彭宣淇, 周德云. 基于矩阵对策与遗传算法的无人机空战决策. 火力与指挥控制, 2019, 44(12): 61)

[3] Xu G D, Lv C, Wang G H, et al. Research on UCAV autonomous air combat maneuvering decision-making based on bi-matrix game. *Ship Electron Eng*, 2017, 37(11): 24
(徐光达, 吕超, 王光辉, 等. 基于双矩阵对策的 UCAV 空战自主机动决策研究. 舰船电子工程, 2017, 37(11): 24)

[4] Su M C, Lai S C, Lin S C, et al. A new approach to multi-aircraft air combat assignments. *Swarm Evol Comput*, 2012, 6: 39

[5] Wan W, Jiang C S, Wu Q X. Application of one-step prediction influence diagram in air combat maneuvering decision. *Electron Opt Contr*, 2009, 16(7): 13
(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. 电光与控制, 2009, 16(7): 13)

[6] Virtanen K, Raivio T, Hamalainen R P. Modeling pilot's sequential maneuvering decisions by a multistage influence diagram. *J Guid Contr Dyn*, 2004, 27(4): 665

[7] Pan Q, Zhou D Y, Huang J C, et al. Maneuver decision for cooperative close-range air combat based on state predicted influence diagram // 2017 *IEEE International Conference on Information and Automation* (*ICIA*). Macao, 2017: 726

[8] Ma Y Y, Wang G Q, Hu X X, et al. Cooperative occupancy decision making of multi-UAV in beyond-visual-range air combat: A game theory approach. *IEEE Access*, 2019, 8: 11624

[9] Li S Y, Chen M, Wang Y H, et al. Air combat decision-making of multiple UCAVs based on constraint strategy games. *Def Technol*, 2022, 18(3): 368

[10] Wang X, Wang W J, Song K P, et al. UAV air combat decision based on evolutionary expert system tree. *Ordnance Ind Autom*, 2019, 38(1): 42
(王炫, 王维嘉, 宋科璞, 等. 基于进化式专家系统树的无人机空战决策技术. 兵工自动化, 2019, 38(1): 42)

[11] Fu L, Xie F H, Meng G L, et al. An UAV air-combat decision expert system based on receding horizon control. *J Beijing Univ Aeronaut Astronaut*, 2015, 41(11): 1994
(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. 北京航空航天大学学报, 2015, 41(11): 1994)

[12] Jiang Y, Wang D B, Bai T T, et al. Multi-UAV objective assignment using Hungarian fusion genetic algorithm. *IEEE Access*, 2022, 10: 43013

[13] Xie J F, Yang Q M, Dai S L, et al. Air combat maneuver decision based on reinforcement genetic algorithm. *J Northwest Polytech Univ*, 2020, 38(6): 1330

[14] Li G L, Wang Y X, Lu C, et al. Multi-UAV air combat weapon-target assignment based on genetic algorithm and deep learning // 2020 *Chinese Automation Congress* (*CAC*). Shanghai, 2020: 3418

[15] Xue J J, Zhu J, Xiao J Y, et al. Panoramic convolutional long short-term memory networks for combat intension recognition of aerial targets. *IEEE Access*, 2020, 8: 183312

[16] Teng T H, Tan A H, Tan Y S, et al. Self-organizing neural networks for learning air combat maneuvers // *The* 2012 *International Joint Conference on Neural Networks* (*IJCNN*). Brisbane, 2012: 1

[17] Zhang H P, Huang C Q. Maneuver decision-making of deep learning for UCAV thorough azimuth angles. *IEEE Access*, 2020, 8: 12976

[18] Kong W R, Zhou D Y, Zhao Y Y, et al. Maneuvering strategy generation algorithm for multi-UAV in close-range air combat based on deep reinforcement learning and self-play. *Contr Theory Appl*, 2022, 39(2): 352
(孔维仁, 周德云, 赵艺阳, 等. 基于深度强化学习与自学习的多无人机近距空战机动策略生成算法. 控制理论与应用, 2022, 39(2): 352)

[19] McGrew J S, How J P, Williams B, et al. Air-combat strategy using approximate dynamic programming. *J Guid Contr Dyn*, 2010, 33(5): 1641

[20] Zhu J Y, Kuang M C, Zhou W Q, et al. Mastering air combat game with deep reinforcement learning. *Def Technol*, 2023

[21] Zhang J H, Wang Y W, Meng F J. Target threat sequencing and allocation for multi-aircraft air combat. *Fire Contr Command Contr*, 2013, 38(12): 96
(张基晗, 王玉文, 孟凡计. 多机空战目标威胁排序及打击分配. 火力与指挥控制, 2013, 38(12): 96)

[22] Zhu J Y, Zhang H L, Kuang M C, et al. Under the sparse reward

based on the study of unmanned aerial vehicle (UAV) air combat simulation. *J Syst Simul*, https://doi.org/10.16182/j.issn1004731x.joss.23-0349
(祝靖宇, 张宏立, 匡敏驰, 等. 稀疏奖励下基于课程学习的无人机空战仿真. 系统仿真学报, https://doi.org/10.16182/j.issn1004731x.joss.23-0349)

[23] Zhou W Q, Zhu J H, Kuang M C. An unmanned air combat system based on swarm intelligence. *Sci China Inform Sci*, 2020, 50(3): 363
(周文卿, 朱纪洪, 匡敏驰. 一种基于群体智能的无人空战系统. 中国科学:信息科学, 2020, 50(3): 363)

[24] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning [J/OL]. *arXiv preprint* (2013-12-19) [2023-10-13].https://arxiv.org/abs/1312.5602

[25] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor // *International Conference on Machine Learning*. Vienna, 2018: 1861

[26] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. *Adv Neural Inf Proc Syst*, 2017, 30

[27] Liu X X, Yin Y, Su Y Z, et al. A multi-UCAV cooperative decision-making method based on an MAPPO algorithm for beyond-visual-range air combat. *Aerospace*, 2022, 9(10): 563