指挥与控制学报
Journal of Command and Control
ISSN 2096-0204,CN 14-1379/TP

《指挥与控制学报》网络首发论文

题目：基于兵棋推演的空战编组对抗智能决策方法
作者：陈晓轩，冯旸赫，黄金才，刘忠，徐越
收稿日期：2021-05-22
网络首发日期：2021-11-26
引用格式：陈晓轩，冯旸赫，黄金才，刘忠，徐越．基于兵棋推演的空战编组对抗智能决策方法[J/OL]．指挥与控制学报．
https://kns.cnki.net/kcms/detail/14.1379.TP.20211125.1943.002.html

网络首发时间：2021-11-26 10:40:33
网络首发地址：https://kns.cnki.net/kcms/detail/14.1379.TP.20211125.1943.002.html

基于兵棋推演的空战编组对抗智能决策方法

陈晓轩[1] 冯旸赫[1] 黄金才[1] 刘忠[1] 徐越[2]

摘 要: 在空战中，编组对抗战术能够系统性调配己方单元，整合个体能力优势，提高编组对抗能力。兵棋推演技术具有低成本和仿真性，是研究先进空战对抗战术的重要手段。当前基于兵棋研究的空战编组对抗方法主要使用规则或运筹等手段，存在假设不够合理、建模不准确、应变性差等缺陷。强化学习算法可以根据作战数据自主学习编组对抗策略以应对复杂的战场情况，但现有强化学习对作战数据要求高，当动作空间过大时，算法收敛慢，且对仿真平台有较高的要求。针对上述问题，提出了一种融合知识数据和强化学习的空战编组对抗智能决策方法，该决策方法的输入是战场融合态势，使用分层决策框架控制算子选择并执行任务，上层包含使用专家知识驱动的动作选择器，下层包含使用专家知识和作战规则细化的避弹动作执行器、侦察动作执行器和使用强化学习算法控制的打击动作执行器。最后基于典型作战场景进行实验，验证了该方法的可行性和实用性，且具有建模准确、训练高效的优点。

关键词: 空战编组对抗；多算子的协作与控制；多智能体深度强化学习算法；分层决策模型

引用格式 陈晓轩，冯旸赫，黄金才，刘忠，徐越. 基于兵棋推演的空战编组对抗智能决策方法[J]. 指挥与控制学报

Intelligent Decision Method of Air Combat Formation Confrontation Based on War Game

CHEN Xiao-xuan[1] FENG Yang-he[1] HUANG Jin-cai[1] LIU Zhong[1] XU Yue[2]

Abstract: In air combat, formation countermeasure tactics can systematically allocate own units, integrate individual ability advantages and improve formation countermeasure ability. Military chess deduction technology has low cost and simulation. It is an important means to study advanced air combat countermeasures. At present, the air combat formation confrontation method based on military chess research mainly uses rules or operation research, which has some defects, such as unreasonable hypothesis, inaccurate modeling, poor adaptability and so on. Reinforcement learning algorithm can independently learn and organize countermeasure strategies according to combat data to deal with complex battlefield conditions, but the existing reinforcement learning has high requirements for combat data. When the action space is too large, the convergence of the algorithm is slow, and has high requirements for the simulation platform. In view of the above problems, an intelligent decision-making method for air combat formation confrontation integrating knowledge data and reinforcement learning is proposed. The input of the decision-making method is the battlefield fusion situation. The hierarchical decision-making framework is used to control the operator to select and execute the task, and the upper layer includes an action selector driven by expert knowledge, The lower layer includes the bullet avoidance action actuator, reconnaissance action actuator and strike action actuator controlled by reinforcement learning algorithm. Finally, experiments based on typical combat scenarios verify the feasibility and practicability of this method, and has the advantages of accurate modeling and efficient training.

Key words: war-game deduction; multi-operator collaboration and control; multi-agent deep reinforcement learning algorithm; hierarchical decision model

Citation CHEN Xiao-xuan, FENG Yang-He, HUANG Jin-Cai, LIU Zhong, XU Yue. Intelligent decision method of air combat formation confrontation based on war game[J]. Journal of Command and Control

空战编组对抗战术是衡量各国空中作战力量的重要指标。相比于单机作战，空战编组可以共享态势信息，合理分配对空搜索任务并执行协同作战任务[1]。现代战斗机的传感器和武器更加先进，空战竞争日趋激烈，对空战编组对抗战术提出了更高的要求。兵棋推演，特别是实时策略类兵棋推演，借助计算机仿真技术，对现实世界军事问题的模拟水平日益增长，能够支撑不对称不完全信息下的动态博弈[2]，辅

收稿日期 2021-05-22
Manuscript received May 22, 2021

1. 国防科技大学系统工程学院，湖南 长沙 410072 2. 解放军 31102 部队，江苏 南京 210016
1. College of Systems Engineering, National University of Defense Technology, Changsha Hunan 410072, China 2. Unit 31102 of PLA , Nanjing Jiangsu 210016, China

助训练指挥员的宏观战略决策和局部战术决策的作战能力，是研究空战编组对抗战术的重要工具。

传统的兵棋推演技术主要采用规划和运筹学知识，夏阳升等[3]提出了一种结合车机载体协同完成多区域覆盖侦察任务的新模式，使用 0-1 整数规划建模技术进行建模，应用于小型无人机在战场区域侦察中。张可等[4]设计了关键点推理遗传模糊系统，结合遗传算法和模糊系统理论构成智能算法推理得到了陆战兵棋推演的行军安全点。刘满等[5]设计了一款引擎，通过挖掘兵棋历史推演数据，提取棋子历史位置概率、夺控热度、观察度等评价属性，利用多属性综合评价软优选算法和兵棋基本规则决策出棋子下步行动。邹烨翰等[6]对作战推演的相关理论基础进行归纳研究，认为在理论上重视兵棋推演中的随机性和复杂性，运用军事运筹学，对问题进行求解研究，能够改善作战模拟系统性能，如果将其应用到各级决策中去，就有可能起到倍增作战能力的作用[7]。但上述方法存在假设不够合理、建模不够准确、应变性差等缺陷。

棋类智能体阿尔法狗战胜围棋世界冠军李世石事件[9]，展现了强化学习技术在智能决策领域的优势，将强化学习技术运用于兵棋推演中是当前军事智能研究的重要方向。Ciancarini P 等[10]在军棋中采用蒙特卡洛树搜索的智能体架构，以较少的领域特定知识获得显著更好的实验结果。Sun Y 等[11]建立了一个基于先验知识的 DQN 智能决策模型用于兵棋推演中的坦克动作控制。针对多智能体控制规划问题，Tan M[12]提出了离散化策略的方法，即对每一个算子，都根据它的观测历史学习训练一个决策网络，但是单个算子常常存在局部观测的状态，即它只能观测到战场的部分态势环境[13]，导致单个算子只能学习得到基于局部观测的最优动作，而不能获得对于全局而言最优的动作。为了研究多智能体强化学习问题，Foerster 等[14]提出了分布决策，集中训练的学习范式，并在星际争霸平台取得了较好的成绩。强化学习需要智能体与环境交互并得到大量高质量的数据用于训练智能体，但是对于空战编组对抗任务而言，动作空间和状态空间随着算子数目的增加而快速增加，会导致单纯的强化学习训练、收敛困难，国内鲜有使用强化学习和知识数据融合控制的空战编组对抗研究。

基于 2020 年的智能博弈挑战赛兵棋推演平台，本文针对兵棋推演中空战编组对抗这一核心问题，提出了一种融合知识数据和强化学习的空战编组对抗智能决策方法，从构建决策方法使用的分层决策框架开始研究，首先确定分层框架输入的静态数据和动态数据的融合方式，接着设计该决策方法的核心分层决策框架执行和训练架构。最后，构建了典型空战对抗环境，设计了强化学习算法的状态空间和动作空间，通过设计推进函数推进该智能决策方法不断与环境交互获得样本数据，用积累的样本数据进行训练。实验结果表明，融合知识数据和强化学习的空战编组对抗智能决策方法控制的红方空战编组与纯规则控制的蓝方空战编组对抗，对抗平均得分可以达到 28 分，在作战中基本可以获得制空权，验证了该方法的可行性和实用性。

1 空战编组对抗智能决策方法

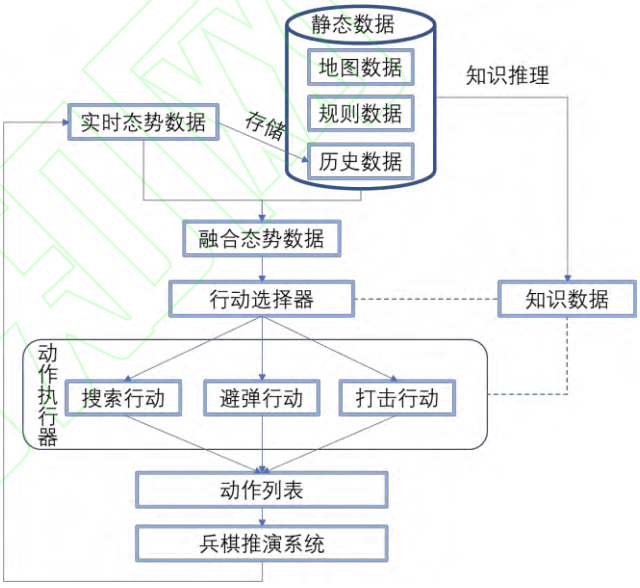

图 1 智能决策方法

Fig. 1 Intelligent decision model

本文提出了一种融合知识数据和强化学习的空战编组对抗智能决策方法，如图 1 所示。方法的输入是战场实时态势信息和静态数据，首先对输入数据进行态势处理，数据经过清洗、筛选、提取、打包、归一化以及格式化表示后，形成融合态势数据，输入分层决策框架的上层行动选择器，行动选择器根据内置逻辑确定搜索动作、避弹动作、打击动作的可行性和优先级，选择可以执行的行动，接着调用下层对应行动的执行器，执行器细化行动细节，形成动作列表，输入兵棋推演系统，系统执行对应动作后可以生成新的战场实时态势。与此同时，根据静态数据进行知识推理，可以获取敌方的装备数据、常见编组模式、常见作战模式、巡逻热点区域。获得上述知识后，可用于行动选择器的逻辑设计，可以用于动作执行器中搜索动作和躲避动作的逻辑设计和参数设置，也可以用

于设计打击行动用的 QMIX 算法的奖励函数设计。图 2 展示了一种根据专家经验设置的行动选择器。在该模型中，智能体首先判断是否执行侦察动作，接着判断是否需要躲避敌方导弹攻击，最后使用弹目匹配模块执行打击行为。

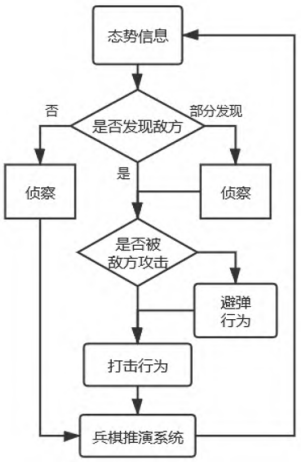

图 2 一种根据专家经验设置的行动选择器

Fig. 2 An action selector based on expert experience

1.1 动作执行器

1.1.1 侦察行动

侦察行动根据专家经验设置，当敌方未出现在我方的探测区域时，我方战斗机组成编组向前推进，同时尽可能增加融合侦察覆盖范围，以减小侦察盲区。当敌方全部目标出现在我方探测区域时，我方战斗机迅速机动至有利位置。当敌方部分目标出现在我方探测区域时，我方战斗机根据历史经验选择继续侦察或是打击敌方目标。通过挖掘分析地图数据，兵棋规则，装备属性，历史数据等，记录敌方经常出现的区域位置，确定其侦察范围，寻找我方被发现概率较小并且能够侦察到敌方概率较大的侦察航线，确定侦察动作的细节参数，有利于侦察行动的展开。

1.1.2 避弹行动

分析态势信息，当我方战斗机相较于敌方战斗机，处于不利位置，或者发现敌方发射导弹向我方攻击时，我方采取避弹动作。该动作根据专家经验设置，首先解析态势信息，建立列表，记录侦测到的敌方导弹，和敌方有威胁的战斗机，接着根据航向及位置信息研判其攻击对象（我方战机），可能有多个单位对我方战机造成威胁，战机需要综合考虑敌方的威胁程度和攻击方式，选择合理的躲避方式，比如飞离敌方威胁区、施放红外干扰弹、低空突防等。

1.1.3 打击行动

根据战机的局部观测，判断当前战机是否可以执行打击动作，以及可以打击的目标，接着使用 QMIX 算法实现我方飞机对敌方目标的打击分配关系（弹目匹配），算法细节将在下一节具体解释。

2 基于 QMIX 算法的弹目匹配实现

2.1 DRQN 算法

深度强化学习主要研究解决的问题是序贯决策问题，为了该问题进行有效的分析，学者们提出了马尔可夫决策过程(MDP)理论对决策过程进行建模，如图3所示。

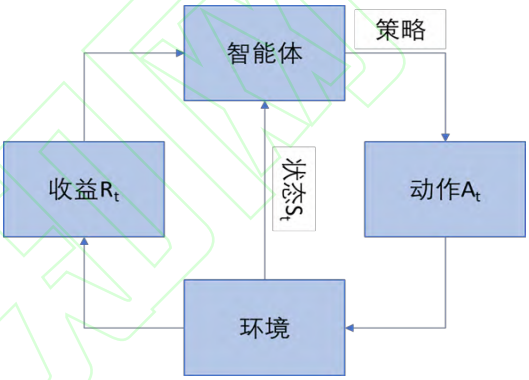

图 3 MDP 示意图

Fig. 3 MDP diagram

MDP由一个五元组定义组$\langle S, A, P, R, \gamma \rangle$定义，在每一个时刻$t$，智能体观察环境并将观察$S_t$以某种格式进行表征，在此基础上根据策略$\pi(a \mid s_t)$选择动作$A_t$。动作$A_t$作用于环境，智能体会观察到新的状态$S_{t+1}$并获得收益$R_t$。$P$表示状态转移概率，在状态$S_t$时，采取动作$a$，智能体以概率$p$获得后继状态$s'$和奖励值$r$。$\gamma$表示折扣因子，$\gamma \in [0,1]$，是优化中长期奖赏与立即奖赏之间的权衡。

在实际决策情况下，智能体常常无法观测到完整的状态信息S_t，这时的观测值就从S_t变成了O_t，智能体的O_t序列不一定能满足马尔可夫性质，因此,需要使用时序网络辅助深度强化学习来辅助智能体记忆并补充状态信息[15]，算法1展示了基于LSTM的DRQN算法。

算法1：基于LSTM的DRQN算法

输入：S_t

输出：A_t

Initialize replay memory D to capacity E

Initialize action-value function Q with random weigths

For each episode do:

Initialize sequence $S_1=\{x_1\}$ and preprocessed sequenced $\phi_1=\phi(s_1,h_1)$

For t in range(1, T) do:

With probability ε select a random action a_t

Otherwise select $a_t=\max_a Q^*\left(\phi\left(s_t,h_t\right),a;\theta\right)$

Execute action a_t in emulator and observe reward r_t and image x_{t+1}

Set $S_{t+1}=S_t,a_t,x_{t+1}$ and preprocess $\phi_{t+1}=\phi(s_{t+1},h_{t+1})$

Store transition $(\phi_t,a_t,r_t,\phi_{t+1})$ in D

End for

Sample randomly an episode E from D

Sample random mini batch of transition $(\phi_j,a_j,r_j,\phi_{j+1})$ from E

Set $y_j=\begin{cases} r_j & \text{if episode terminates at step j+1} \\ r_j+\gamma\max_{a'}\hat{Q}\left(\phi_{j+1},a';\theta^-\right) & \text{otherwise} \end{cases}$

Perform a gradient descent step on $\left(y_j-Q\left(\phi_j,a_j;\theta\right)\right)^2$ according to equation

End for

2.2 QMIX 算法

在实际世界和实时策略游戏中，单个智能体存在局部观测的约束，再加上智能体间的通信受限，对于每一个智能体而言，需要根据它所处的具体环境建立对应的离散式策略。当智能体数量增加时，离散化策略所构成的联合动作决策空间过大会导致传统的单智能体算法失效。因此,研究人员提出了集中训练，分布执行的学习范式。

一方面，集中训练要求智能体能够获得一个基于全局状态和联合动作的Q_{total}。另一方面，当算子数量过多时，Q_{total}难以直接习得，即使可以习得，也没有直接的方法可以转化成每个算子可以根据单独的观测，选取单独动作的分布式策略[15]。

因此，Tabish Rashid等提出了QMIX算法[16]，包含一组DRQN网络一个混合网络，组中的每个DRQN网络对应一个分布式执行的策略，混合网络把一组DRQN网络组合输出的一组Q_a，以一种复杂的非线性模式加权组合，从而输出Q_{tot}，同时能够保持一致性。因此，QMIX算法可以以一种因子化的表示方法，来表示复杂的中心化动作值函数。这样的表示方法根据智能体的数量变化，可以有很好的伸缩性，并且允许分散化的策略在线性时间内，可以通过单独的argmax操作容易获得结果。

$$\frac{\partial Q_{total}}{\partial Q_a}>=0,\forall a. \tag{1}$$

为了保证一致性，只需要确保全局最优是由所有算子的局部最优所组成的就可以了，算法通过约束混合网络的参数为正数，使得式(1)所示的约束满足。对于每一个智能体a，都有一个DRQN网络输出它单独的值函数$Q_a\left(s_a,u_a\right)$，在每一个时间步把当前局部观测值s_a和上一步的动作u_a作为输入，如图3所示。

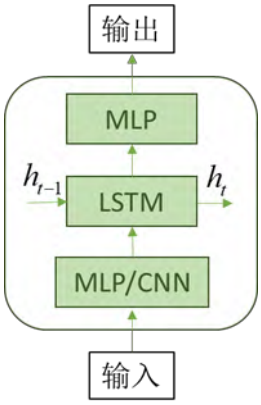

图 4 深度循环 Q 网络图

Fig. 4 Deep recurrent Q network diagram

2.3 基于 QMIX 算法的弹目匹配技术

在空战编组对抗中，合理且有效的武器分配方案是提升我方飞机的导弹打击成功率，保证对抗胜利的前提条件之一，针对当前空战形势下敌我交战出现目标不明确、分配不均衡、打击拦截效果不佳、统筹协调不高和资源浪费的情况，本研究希望使用QMIX算法，通过不断地与环境交互，学习对战经验，帮助空战编组的武器目标分配策略收敛到最优，提升智能决策框架的总体作战性能。

图 5 弹目匹配技术示意图

Fig. 5 Diagram of missile target matching technology

基于QMIX算法的弹目匹配技术如图5所示，*n*架战斗机将局部观测分别输入DRQN网络，输出*Q*值最高的动作，即攻击敌方第*j*到*k*架飞机。

3 实验设计与结果分析

3.1 实验平台简介

本文使用的兵棋推演平台是 2020 年的智能博弈挑战赛平台，基于该平台，设计了红蓝方多机空战对抗想定。考虑到海空作战力量速度快，活动空间大，同时，结合战局紧凑性和 AI 计算复杂度，想定问题设计如下：红方有一战斗机编队前往目标空域执行空中巡逻任务，在空中遭遇执行拦截任务的蓝方编队，希望红方战斗机能够选择最优的决策动作序列，以最少的损失歼灭敌蓝方编队，想定示意如图 6 所示。

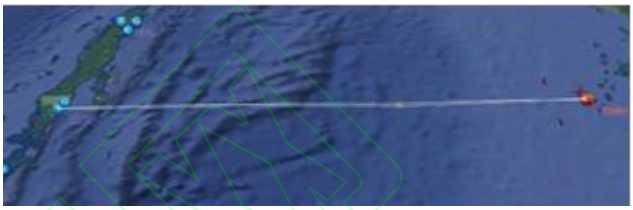

图 6 想定示意图

Fig. 6 Scenario diagram

战场中心为坐标原点，向右为 x 轴正轴，向上为 y 轴正轴，红方岛屿机场坐标为(146 700, -300 0)，蓝方南部岛屿指挥所坐标为(-131 154, -878 88)。红方战斗机在目标空域巡逻，占据制空权。蓝方战斗机从机场出发飞往目标空域，企图歼灭红方战斗机群并夺取制空权。红蓝双方战斗机雷达均处于开机状态，战斗机群按 5 千米的纵坐标差依次排开。在该想定中红蓝双方飞机算子配置相等，各自包含 4 架战斗机，想定区域为直径 250 km 的无人区域，飞机实体的各项性能如表 1 所示。战斗机的雷达探测角度为 120 度，探测距离为 100 km，识别距离为 60 km，战斗机在空中机动时，雷达自动开机，同时双方都具有地面雷达，雷达探测角度为 360 度，探测距离为 250 km，识别距离为 150 km。

表 1 飞机实体性能表

Table 1 Aircraft performance table

类型	速度范围	高度范围	探测角度	探测距离	打击距离	导弹数量
战斗机	900~1 000 km/h	[0,180 00]米	120 °	100 km	80 km	6 枚中程导弹

3.2 神经网络整体架构设计

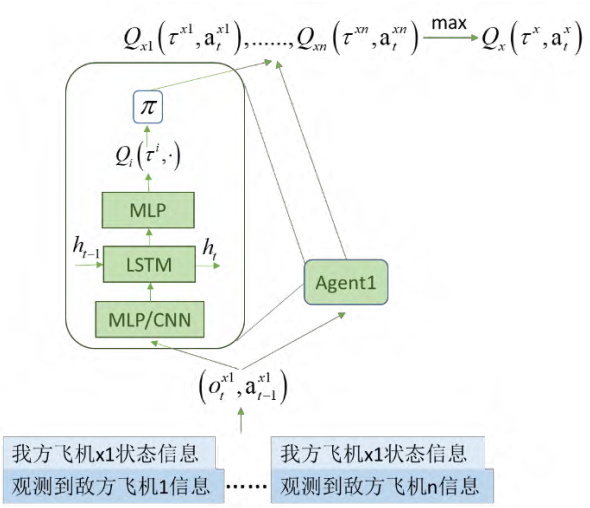

图 7 网络架构

Fig. 7 DRQN network

神经网络架构设计如图5所示，在*T*时刻，需要进行决策的第x_1架飞机的探测距离是100 km，其视野范围内共有*N*架敌方飞机。为确定需要打击敌方的哪一架飞机。首先，将需要决策的第x_1架飞机，观测到的敌方的第1架飞机的状态信息拼接以后输入DRQN网络，如图6所示。

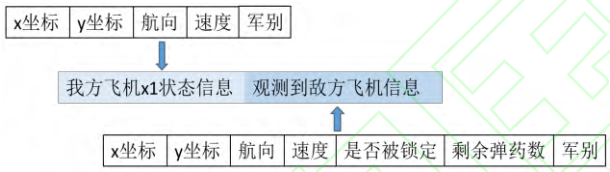

图 8 状态空间

Fig. 8 State space

将输入信息经过一层MLP网络，处理态势信息，并输出一组特征向量，将特征向量和隐藏层信息输入到LSTM层中，根据时序训练的要求，将智能体当前时间步的特征向量和上一步的隐藏层信息融合生成新的特征向量，最后将特征向量输入MLP以获得*Q1*值。

同理，将第x_1架飞机与观测到的其他N-1架敌方飞机的状态信息分别组合并获得*Q2*……*Qn*值，最后将得到的*N*个*Q*值组成一个列表并取argmax，*Q*值最大的第*n*架敌方飞机即为我方所选的攻击目标，记录下当前时间步打击的敌方目标和Q_{x1}值。动作空间集为：[无打击动作，打击敌方飞机1，打击敌方飞机2……打击敌方飞机*n*]。

假设我方共有*m*架飞机，依据此方法，获得我方第x_1架飞机到第x_m架飞机的Q值列表，将该Q值列表作为混合网络的输入，经过网络计算后，输出一个$Q_{tot}(\tau,a)$值。

混合网络的权重由独立的超网络产生。每一个超网络把全局状态*s*作为输入并生成混合网络的一层参数。图7解释了混合网络和它的超网络。

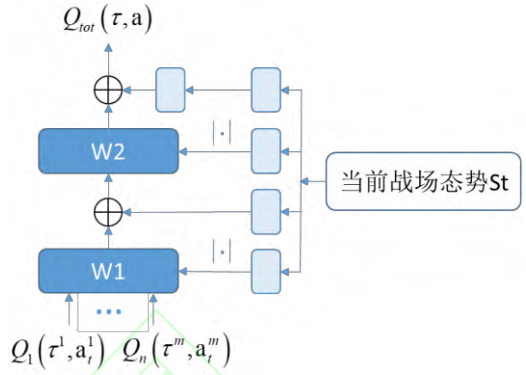

图 9 混合网络

Fig. 9 Mixing network

*T*时刻的全局状态*s*由我方所有飞机的信息，敌方所有飞机的信息，所有的导弹信息组成，如图8所示。

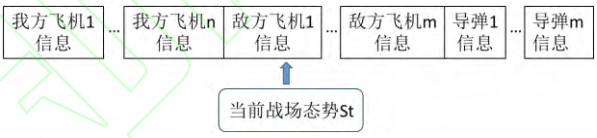

图 10 战场态势

Fig. 10 Battlefield situation

全局状态*s*先经过一层单独的线形层，再经过一层绝对值激活函数，为了确保混合网络的权重是非负的，超网络的输出结果是一个向量，这个向量根据*w1*的要求被重整成为了一个有适当尺寸的矩阵。

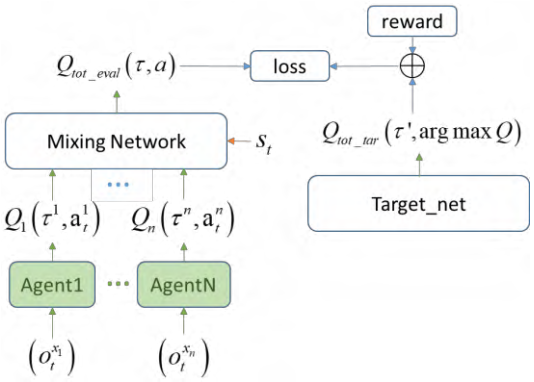

图 11 神经网络更新过程

Fig. 11 Neural network update process

输入经过混合网络后生成Q_{tot}，同时使用目标网络和存储的下一帧观测动作奖励信息，根据式（2）：

$$y^{tot}=r+\gamma\max_{\mathbf{u}'}Q_{tot}\left(\boldsymbol{\tau}',\mathbf{u}',s';\theta^{-}\right) \tag{2}$$

计算出目标网络对当前状态动作对的估值，计算y^{tot}和Q_{tot}的差值从而生成损失函数并反向传播，其中θ^-是target网络的参数。损失函数的定义如式（3）所示，b是从记忆池里采样得到的状态迁移四元组(s,a,s',r)的数量大小，训练过程如图9所示。

$$\mathcal{L}(\theta)=\sum_{i=1}^{b}\left[\left(y_i^{tot}-Q_{tot}(\tau,\mathbf{u},s;\theta)\right)^2\right] \tag{3}$$

3.3 实验结果

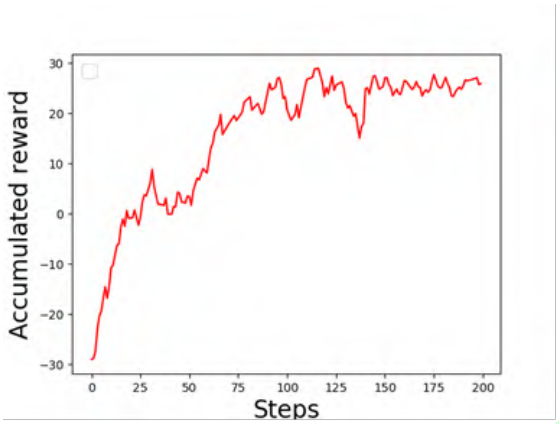

图 12 累积奖赏值随时间变化曲线

Fig. 12 Curve of cumulative reward value over time

本次实验对QMIX算法中的神经网络进行了4 000回合的训练。先分析兵棋推演决策序列 τ 的累积奖赏值随训练时间的变化，累积奖赏值R是每个序列 τ 下单步奖赏值的叠加，可以反映QMIX学习到策略的好坏，在本次实验中，单步奖赏值得定义如下，每击落一架敌方飞机奖励值加20，我方损失一架飞机奖励值减20，我方飞机每躲避敌方一枚导弹奖励值加1。为了评估算法效果，我们的空战对抗每进行20次重复实验，就将其20个累积奖赏值求和取均值并记录，如图10所示，累积奖赏在[0,20]迅速增大，但在[20,50]迅速下降，且出现在第100回合和第250回合时出现剧烈下降的情况，这是因为训练初期学习样本不够，且初期智能体探索具有一定盲目性。训练中期（[50，100]）间累积奖赏值整体呈上升趋势,期间曲线也存在中小幅度的挂起与坠落。训练末期（[100，200]）间的累积奖赏值基本收敛，达到将近20，相比训练初期的-28，训练结果较好。

图 13 损失函数训练曲线

Fig. 13 Loss function training curve

再分析公式3所示的强化学习损失函数，其测算了目标网络和评估网络的逼近程度。由图11可见，损失函数基本在3 000回合的时候得到较小值且相对稳定。虽然训练过程中，由于智能体决定动作时具有一定的探索性、智能体训练过程中参数有较大的波动以及达到局部最优解，使得图11中神经网络的损失函数下降过程中不断存在微弱噪音。曲线在大致[500，700]，[1 300，1 500]存在相对明显的噪音，并在[1 300，1 450]存在较大抖动。但是从整个训练过程看，损失函数呈下降趋势，从2 000最终下降至250左右，训练结果较理想。

3.4 行为分析

对实验数据进行复盘，经过训练的红方战斗机编组已经展现出了一定的个体战术与协同配合，对训练后的红方进行编组对抗战术进行分析。红方六架战机两两组成编队，按照巡航速度，编队 1 向东北方向飞行，编队 2 往正东方向飞行，编队 3 向东南方向飞行，到达预定阵位后 3 组编队改变航向往正东方向飞行，此时我方战机的侦察雷达全开，基本上可以覆盖南北直径为 250km 的作战区域，如图 14 所示，当侦察到敌方战机时，我方战机调用打击行动模块，充分发挥武器射程优势，使用远程空空导弹打击敌方战机。在遇到敌方战机袭击或是我方战机处于不利位置时，战机调用避弹模块，执行大角度转弯机动，以尽快拉开与敌方攻击算子的位置，并向有利位置机动。

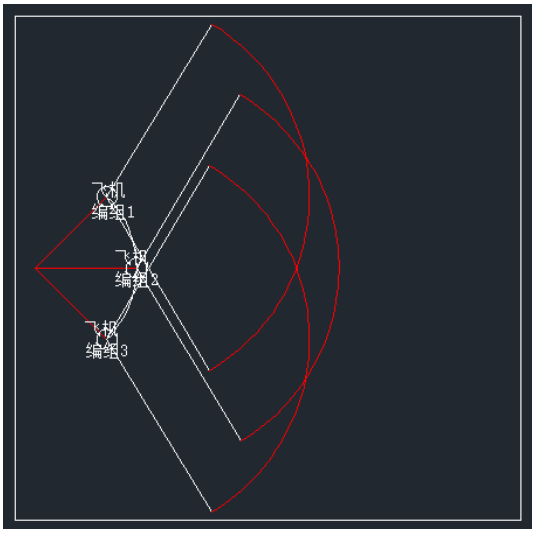

图 14 3 组编队侦察范围示意图

Fig. 14 Diagram of detect range of three formation

4 结论

针对复杂的空战编组对抗问题，在兵棋推演平台上建立了一个典型空战想定用于空战对抗研究，提出了一种基于知识数据和强化学习融合的空战编组智能决策方法，设计了空战编组对抗的分层逻辑框架，上层包含使用专家知识驱动的动作选择器，下层包含避弹动作执行器、侦察动作执行器和打击动作执行器，针对打击动作中的战斗机的弹目匹配这一关键问题，基于马尔可夫决策理论建模，确定空战智能体的输入状态空间和空战动作集，使用QMIX算法对智能体进行策略训练，通过与传统规划控制的对手进行对抗实验，验证了该智能决策方法的优势。下一步研究方向是将该分层决策模型与不同算法控制的不同风格的对手进行对抗实验，研究算法的泛化性。

References

1 余敏建, 嵇慧明, 韩其松等. 基于合作协同进化的多机空战目标分配[J]. 系统工程与电子技术, 2020, 42(6): 1290-1300.

YU M J, JI H M, HAN Q S, et al. Multi-aircraft air combat target allocation based on cooperative co-evolutionary[J]. Systems Engineering and Electronics, 2020, 42(6): 1290-1300.

2 胡晓峰, 贺筱媛, 陶九阳. AlphaGo 的突破与兵棋推演的挑战[J]. 科技导报, 2017, 35(21): 49-60.

HU X F, HE X Y, TAO J Y. AlphaGo's breakthrough and challenges of wargaming[J]. Science & Technology Review, 2017, 35(21): 49-60.

3 夏阳升,石建迈,陈超,等. 车机协同多区域覆盖侦察路径规划方法[J]. 指挥与控制学报, 2020, 6(4): 372-380.

XIA Y S, SHI J M, CHEN C, et al. Path planning method for multi-area reconnaissance by cooperated ground vehicle and drone[J]. Journal of Command and Control, 2020, 6(4): 372-380.

4 张可,郝文宁,余晓晗,靳大尉,邵天浩.基于遗传模糊系统的兵棋推演关键点推理方法[J].系统工程与电子技术,2020,42(10): 2303-2311.

ZHANG K, HAO W N, YU XH, et al. Wargame key point reasoning method based on genetic fuzzy system[J]. Systems Engineering and Electronics, 2020 , 42(10): 2303-2311.

5 刘满,张宏军,郝文宁,等.战术级兵棋实体作战行动智能决策方法[J].控制与决策,2020,35(12):2977-2985.

LIU M, ZHANG H J, HAO W N, et al. Research on intelligent decision-making method of tactical-level wargames[J]. Control and Decision, 2020,35(12):2977-2985

6 邹烨翰,冯旸赫,程光权, 等.面向军事条令条例的本体构建技术[J].指挥与控制学报,2019,5(1):47-54.

ZOU Y H, FENG Y H, CHENG G Q, et al. Ontology developing technology for military regulations[J]. Journal of Command and Control, 2019, 5(1): 47-54

7 张明星,程光权,刘忠, 等.多武器协同作战发射时序规划方法[J].指挥与控制学报,2017,3(1):10-18.

ZHANG M X, CHENG G Q, LIU Z, et al. Schedule of launch sequential timing in multiple weapons cooperative engagement[J]. Journal of Command and Control, 2017, 3(1): 10-18

8 Volodymyr Mnih,Koray Kavukcuoglu, et al. Human-level control through deep reinforcement learning[J]. Nature: International weekly journal of science,2015,518(7540).

9 David Silver,Julian Schrittwieser, et al. Mastering the game of Go without human knowledge[J]. Nature: International Weekly Journal of Science,2017,550(7676).

10 Ciancarini P, Favini G P , et al. Monte Carlo tree search in Kriegspiel[J].Artificial Intelligence, 2010, 174(11):670-684.

11 SUN Y, YUAN B, ZHANG T, et al. Research and implementation of intelligent decision based on a priori knowledge and DQN algorithms in wargame environment[J]. Electronics, 2020, 9(10):1668.

12 TAN M. Multi-agent reinforcement learning: Independent vs. cooperative agents[J]. Machine Learning Proceedings, 1993: 330-337.

13 梁星星，冯旸赫，马扬，等. 多 Agent 深度强化学习综述[J]. 自动化学报,2020,46(12):2537-2557.

LIANG X X, FENG Y H, MA Y, et al. Deep multi-agent reinforcement learning: a survey[J]. Acta Automatica Sinica, 2020, 46(12): 2537-2557

14 FOERSTER, J., FARQUHAR, G., AFOURAS, T, Nardelli, N , et al. Counterfactual multi-agent policy gradients. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence,

2018.
15 刘朝阳, 穆朝絮, 孙长银.深度强化学习算法与应用研究现状综述[J].智能科学与技术学报,2020,2(4):314-326.
LIU Z Y, MU C X , SUN C Y. An overview on algorithms and applications of deep reinforcement learning[J]. Chinese Journal of Intelligent Science and Technology, 2020,2(4):314-326.
16 Tabish Rashid, Mikayel Samvelyan, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. arXiv preprint arXiv:1803.11485.

陈晓轩（1998—），男，硕士，学生，主要研究方向为强化学习、深度学习。
冯旸赫（1985—），男，博士，副教授，主要研究方向为因果发现与推理、主动学习及强化学习。
黄金才（1973–），男，博士，研究员，主要研究方向人工智能和任务规划。
刘 忠（1968—），男，博士，教授，主要研究方向为通用人工智能、多智能体系统及强化学习。本文通信作者。
E-mail: Liuzhong@nudt.edu.cn
徐 越（1968—），男，硕士，高工，主要研究方向为任务规划和模拟仿真。