Journal of System Simulation

Volume 36 | Issue 7 Article 2

7-15-2024

Research on Learnable Wargame Agent Driven by Battle Scheme

Yifeng Sun
Strategic Support Force Information Engineering University, Zhengzhou 450001, China,
yfsun001@163.com

Zhi Li
Strategic Support Force Information Engineering University, Zhengzhou 450001, China;

Jiang Wu
Strategic Support Force Information Engineering University, Zhengzhou 450001, China;

Yubin Wang
PLA 66389 Troops, Zhengzhou 450000, China

Research on Learnable Wargame Agent Driven by Battle Scheme

Abstract

Abstract: To enable the agent to cope with complex battle scenarios and objectives in wargame, a learnable wargame agent architecture driven by a battle scheme is proposed. By analyzing the "attachment characteristics" and "loose coupling characteristics" of the agent to wargame system, the learnable requirements of the agent are obtained. In the design of the agent framework, battle schemes are used to reduce the learning range of the agent. The finite state machine corresponds to the knowledge of the operational phase in the battle scheme, and the decision-making space of the agent is determined according to the framework of the battle scheme. A learnable deep neural network is designed to explore key decision space. The neural network uses prior knowledge imitation learning mode and deep reinforcement learning mode. This architecture can iteratively explore optimal deployment and collaboration issues for multiple chessmen that are difficult for humans to fully tease out.

Keywords

wargame, agent, battle scheme, deep neural network, reinforcement learning, imitation learning

第 36 卷第 7 期 系统仿真学报© Vol. 36 No. 7
2024 年 7 月 Journal of System Simulation Jul. 2024

作战方案驱动的可学习兵棋推演智能体研究

孙怡峰[1]，李智[1]，吴疆[1]，王玉宾[2]
(1. 战略支援部队信息工程大学，河南 郑州 450001；2. 中国人民解放军66389部队，河南 郑州 450000)

摘要： 为了使智能体能够应对兵棋推演中的复杂作战场景和作战目的，提出作战方案驱动的可学习兵棋推演智能体架构。剖析智能体对兵棋系统的“依附特性”和“松耦合特性”，得到智能体的可学习要求；在智能体框架设计中，使用作战方案压减智能体学习范围。通过有限状态机对应作战方案中的作战阶段知识，依据作战方案框架确定智能体决策空间，设计可学习的深层神经网络实施关键决策空间探索，神经网络采用先验知识模仿学习模式和深度强化学习模式。该架构能迭代探索人类难以充分梳理清楚的多棋子最优部署和协作问题。

关键词： 兵棋推演；智能体；作战方案；深层神经网络；强化学习；模仿学习
中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X(2024)07-1525-11
DOI: 10.16182/j.issn1004731x.joss.23-0477
引用格式: 孙怡峰, 李智, 吴疆, 等. 作战方案驱动的可学习兵棋推演智能体研究[J]. 系统仿真学报, 2024, 36(7): 1525-1535.
Reference format: Sun Yifeng, Li Zhi, Wu Jiang, et al. Research on Learnable Wargame Agent Driven by Battle Scheme [J]. Journal of System Simulation, 2024, 36(7): 1525-1535.

Research on Learnable Wargame Agent Driven by Battle Scheme

Sun Yifeng[1], *Li Zhi*[1], *Wu Jiang*[1], *Wang Yubin*[2]
(1. Strategic Support Force Information Engineering University, Zhengzhou 450001, China;
2. PLA 66389 Troops, Zhengzhou 450000, China)

Abstract: To enable the agent to cope with complex battle scenarios and objectives in wargame, a learnable wargame agent architecture driven by a battle scheme is proposed. By a*nalyzing the "attachment characteristics" and "loose coupling characteristics" of the agent to wargame system, the learnable requirements of the agent are obtained. In the design of the agent framework, battle schemes are used to reduce the learning range of the agent. The finite state machine corresponds to the knowledge of the operational phase in the battle scheme, and the decision-making space of the agent is determined according to the framework of the battle scheme. A learnable deep neural network is designed to explore key decision space. The neural network uses prior knowledge imitation learning mode and deep reinforcement learning mode.* This architecture can iteratively explore optimal deployment and collaboration issues for multiple chessmen that are difficult for humans to fully tease out.

Keywords: wargame; agent; battle scheme; deep neural network; reinforcement learning; imitation learning

0 引言

以AlphaGo的成功研发为起点，对智能决策的研究取得突飞猛进的进展。DeepMind推出的MuZero智能体在没有传授棋类运行规则的情况下，通过自我观察掌握了围棋、国际象棋、将棋

收稿日期：2023-04-21 修回日期：2023-06-12
第一作者：孙怡峰(1976-)，男，副教授，博士，研究方向为人工智能与信息安全等。E-mail：yfsun001@163.com

和雅达利(Atari)游戏[1]。而针对即时战略游戏，DeepMind基于监督学习、模仿学习、强化学习推出了AlphaStar[2]，在“星际争霸II”中达到了人类大师级的水平，并且在官方排名中超越了99.8%的人类玩家。智能体“OpenAI Five”也用类似方法在“刀塔II”游戏中击败世界冠军[3]。最近大火的ChatGPT和GPT-4，也经历了语言模型预训练和强化学习等多个阶段。

兵棋是复杂的智能体决策环境，需要控制多个同类型或者不同类型的作战实体(棋子)，使之协同配合，达到整体的作战目的。根据文献[4]报道，美国的JSAF/ONESAF兵棋智能体可以指挥实体自动化行动，但需要任务计划来指定每个实体或单位的任务和目的地以及对应的路径，在运行中模拟实体基于有限数量的行为来感知和反映其周围环境，无法处理超出其预定义行为的新情况。国内学者探索将深度强化学习等技术应用到兵棋环境。文献[5]利用深度强化学习在“墨子·未来指挥官系统”中进行了一系列智能博弈的研究并取得了突出成果，但这种学习型智能体能应对的作战场景还相对简单。文献[6]则从局部机动角度，提出基于深层神经网络从复盘数据中学习战术机动策略模型的方法，对于态势认知研究具有重要参考价值。文献[7]分析了游戏博弈到作战指挥的决策差异，提出态势能否理解、知识如何运用等设计智能体面临的问题。

针对复杂作战场景和作战目的，本文探讨了兵棋推演智能体的概念，提出以人类拟定的作战方案为智能体设计的基本遵循：从作战方案框架中确定智能体决策空间，通过深层神经网络实施关键决策空间探索。主要工作：①通过有限状态机来体现作战方案中前后依赖的作战阶段；②通过分析兵棋设计的要点抓住最需要智能体迭代决策的变量，确定决策空间，设计可学习的神经网络，迭代探索人类也难以充分梳理清楚的多棋子最优部署和协作问题；③给出先验知识模仿学习和深度强化学习相结合的智能体策略迭代机制。

1 兵棋推演智能体的基本概念

兵棋，是使用形象化的棋子，用经验方式提炼的规则以随机概率的方式进行行动裁决，在各种模拟地形上通过决策对抗来进行人员训练或作战论证的一种作战模拟工具[8]。根据棋子模拟的兵力单位分辨率和战争级别不同，现代计算机兵棋系统分为了战略兵棋、战役兵棋、战术兵棋、分队兵棋等。面向集团作业模式和指挥所编组作业模式，兵棋推演的指令形式也各有不同。兵棋的用途主要有指挥训练、作战方案评估论证和作战创新[9]。无论哪种用途、什么指令形式、何种分辨率，传统兵棋推演都需要人在回路，扮演红蓝两方进行操作。首先，军事人员对本方作战方案充分理解，制定本方兵力单位的行动指令集；技术保障人员在军事人员的基础上，形成计算机兵棋系统的棋子行动指令，将指令不断输入；兵棋系统依据规则对红蓝双方棋子行动进行裁决，形成对抗过程中的态势，并进行可视化展示。

兵棋推演需要同时组织红方和蓝方开展上述工作，花费大量人力物力。若智能体能指挥单方的全部兵力，将显著降低工作量和工作难度；若通过复盘能不断修正计算机决策策略及其知识，那么利用计算机算力，可以不断迭代进化。本文将这种针对推演过程的智能体称之为兵棋推演智能体，其基本定义为：一种计算机程序，能接收兵棋系统的当前推演态势(即环境状态)，根据知识进行推理运算，做出决策形成棋子行动格式化指令，交给兵棋系统执行，改变下一步推演态势，往复循环；推演结束后，能根据对抗结果数据“复盘”学习改进原有知识。

(1) 兵棋推演智能体“依附于”兵棋系统

文献[8]指出“不要寄希望研制万能兵棋”，兵棋推演智能体也一定是在某种兵棋系统之上的，它与该兵棋“营造的训练战场环境”相对应。智能体实施推演，离不开兵棋系统的程序接口支持。智能体需要兵棋系统提供与人类操作类似的态势

DOI: 10.16182/j.issn1004731x.joss.23-0477

数据，智能体为了驱动棋子，还必须编码生成兵棋系统所能接收的棋子行动指令，并调用兵棋系统接口函数，将指令作为参数传递给兵棋系统实施。

兵棋推演智能体迭代学习的对抗策略也与依附的兵棋特点相对应，不大可能训练出直接对各种兵棋系统都有效的智能体。针对同一兵棋，其训练的目标是统一的，兵棋推演智能体的形式化知识也具有较强相似性，在不同想定下，智能体通过较小规模的迁移训练应能得到较优的对抗策略知识。若通过兵棋推演智能体迭代演进进行作战创新，它必定是针对兵棋所模拟的特定作战场景的。

(2) 兵棋推演智能体“松耦合于”兵棋系统

智能体可以与人类进行对抗推演，也可以与其他智能体进行推演，这些智能体可存在于不同的计算机上，只需智能体能远程获取所依附的兵棋态势、棋子命令、裁决等信息。从该意义上，兵棋推演智能体从物理空间上可以“松耦合于”其所依附的兵棋系统。正是这种松耦合特点的存在，为智能体并发、分布式学习，优化己方策略奠定了基础。例如，基于开发接口 PySC2[10]，DeepMind 公司构建了数千个星际争霸智能体，每个智能体都使用 32 个 TPU 进行复盘学习，经过 44 天的训练最终得到了超强智能体 AlphaStar。

兵棋推演智能体既依附又松耦合的特性，使其理论上一定能比人类在创新优化上更高效。下面论述智能体实例时，以“庙算 • 陆战指挥官”[11](下面简称庙算平台)为对象，其兵力单元分辨率为连/排，可归类为战术兵棋，以作战方案创新优化为智能体研究的主要目的。

2 兵棋推演智能体的设计框架

2.1 兵棋推演智能体的设计要点

作战有明确但又复杂的意图，比如明确要先打击哪些敌方目标，达成什么样的效果，摧毁还是致瘫等。最终达到哪些既定的目的，比如夺控某要点等。兵棋推演智能体也应指挥棋子实现类似的过程。根据 AlphaStar、ChatGPT 等智能体，深层神经网络将是不可或缺的学习功能承载体。但神经网络进行哪些决策，如何学习得到对应最优策略的神经网络参数是设计智能体必须要考虑的，即动作空间设计[12]。最终，希望智能体通过兵棋系统给出的初始态势，学习得到到达最终目的最优行动序列。面临强大的人类对手和巨大的状态动作组合，让神经网络决策所有兵棋系统设定的棋子动作，并采用试错强化方式学习，仍有很大挑战。为此，提出针对想定、以人类设计的作战方案为智能体基本遵循的方法。

(1) 依据作战方案框架，确定智能体决策空间

AlphaStar 智能体有 Who、When、What、Where 四类决策项[2]，构成其神经网络的决策空间。在兵棋中，应根据作战方案，选择人类不易决策的且能够迭代试验的决策项，作为智能体深层神经网络的决策空间。

作战方案是“根据首长决心拟制的对作战进程和战法的设想”，通常包括情况判断结论、上级企图和本部队任务、友邻任务及作战分界线，各部队的编成、配置和任务，作战阶段划分，各阶段情况预想及处置方案，保障措施、指挥的组织等[13]。这些条框描述了作战方案的框架。兵棋推演就是要论证评估特定想定下作战方案框架的具体内容以及训练参训人员确定作战方案框架内容的能力。

目的是作战要达到的预期结果，任务是对作战目的的具体化，目的的实现是要通过一系列任务的完成来达成的[14]。作战阶段决定了任务完成的前后条件即时间关系，与 AlphaStar 的 When 决策项类似。作战阶段的设计对作战任务的完成至关重要，且前后逻辑关系十分复杂，智能体学习这种复杂逻辑仍比较困难。人类设计完成复杂作战目的的各作战阶段更为合适。

在兵棋推演中，想定给定作战编成，但编成内的兵力编组及其在各作战阶段的任务分配是作

战方案的重要内容。任务相当于AlphaStar的What决策项，如分配作战目标；兵力编组相当于AlphaStar中的Who决策项，即分配兵力棋子组合成编组。可以采用给各个兵力棋子分配作战目标，也可固定作战目标分配实现目标的棋子组合。这里采用后者，即固定What，决策兵力编组Who。某个阶段某编组的棋子完成打击出现在某点位对手的任务，关键是该编组的棋子要能机动到能打击该点位的有利位置，并且机动过程尽可能不暴露我方意图。决策从起点到终点的整条机动路径，相当于AlphaStar中的Where决策项。由智能体神经网络决策Who和Where项，便于发挥机器快速迭代学习的优势。

由上，完成各个作战目标的棋子编组、棋子的机动路径等构成神经网络的待决策变量，它们的取值范围就是智能体的决策空间。这种从作战方案框架中获取智能体决策空间的做法，也是提高兵棋推演智能体可解释性的重要抓手。具体设计中应进一步估算当今计算机算力，决定决策空间大小，避免决策空间过于海量、计算机无法在有效时间内迭代出好策略的情形。

(2) 抓住兵棋系统自身设计特点，确定智能体的学习型决策突破点

根据作战方案得到的智能体决策空间，还可能过大，这种情况下可进一步聚焦兵棋系统。每一款兵棋在设计时都有自己的训练目标或关注的作战研究(论证或创新)点[8]，反映了特定场景的训练任务和研究课题。理解把握兵棋自身对受训者的关注点，是兵棋推演智能体的率先突破点。

兵棋规则是兵棋设计的重要内容，而规则设计中的关键是行动分析和影响因素分析，往往在进行轻重取舍后，确定兵棋重点关注的行动与裁决规则。兵棋行动设计中体现出的取舍考虑，也决定兵棋推演智能体应率先突破的关键决策环节。由此，兵棋训练设计的要点动作对应决策变量和决策空间，就是兵棋推演智能体当前需要迭代学习的；其他的决策变量，在有限的计算算力下，可暂不考虑其迭代学习。

例如，在某个阶段棋子部署的路径涉及多个点位，决策空间指数级增长。为了减少计算规模，若起点已定，可仅考虑终点，这个终点位置下面称为部署点位，部署点位就是当前算力下需要通过深层神经网络迭代学习的。

2.2 基于作战方案的兵棋推演智能体架构

兵棋推演智能体从功能上讲，是要根据每个时间步或回合兵棋系统给予的态势数据，决策产生行动指令，不断驱动己方的兵力单位棋子，达成作战目的。

智能体采用游戏人工智能中的有限状态机(finite state machine, FSM)[15]技术实现作战方案中的阶段及其转换逻辑。FSM是具有内部记忆功能的抽象机器，表示有限离散状态以及这些状态之间转移的数学模型[16]。它总是处于有限状态集合中的某一个状态，但满足转移条件时，会从当前状态转移到另一个状态。FSM的状态与作战方案中的作战阶段相对应，一个作战阶段对应一个状态。这种状态区别于每个时间步的环境状态，称为高层状态。FSM的状态转移与作战方案中的阶段转换类似，使用阶段转换中的可量化条件作为转移条件。

在高层状态存续期间，智能体将驱动棋子执行与棋子类型对应的特色任务。这由产生式规则知识推理机在博弈计算模型“指导”下实现。由此，得到由FSM、博弈计算环节、产生式规则推理模块构成的智能体架构，如图1所示。图1中假定只有3个作战阶段，即3个高层状态，实际中依情况确定。

博弈计算环节仅在高层状态发生改变的第一个时间步(step)前或此时间步时使用，如图2的右半部分所示。该环节的核心是博弈计算模型，它作为智能体的灵魂用于决策针对各个作战目标的棋子编组和本阶段棋子的部署点位等决策变量取

DOI: 10.16182/j.issn1004731x.joss.23-0477

值。这要根据态势数据以及对手可能的选择来计　　算，因此将其称为博弈计算模型。

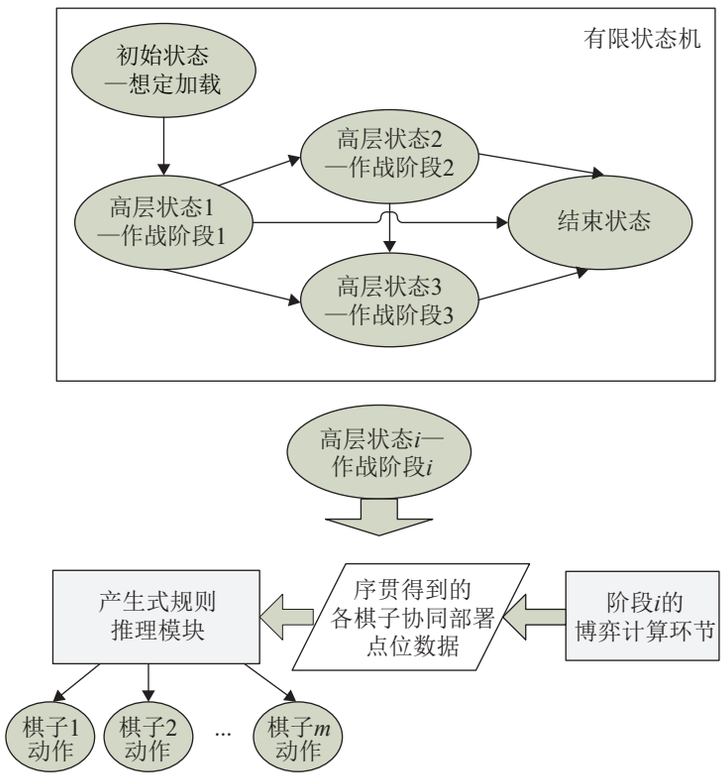


图1　智能体的总体框架

Fig. 1　Overall framework of agent

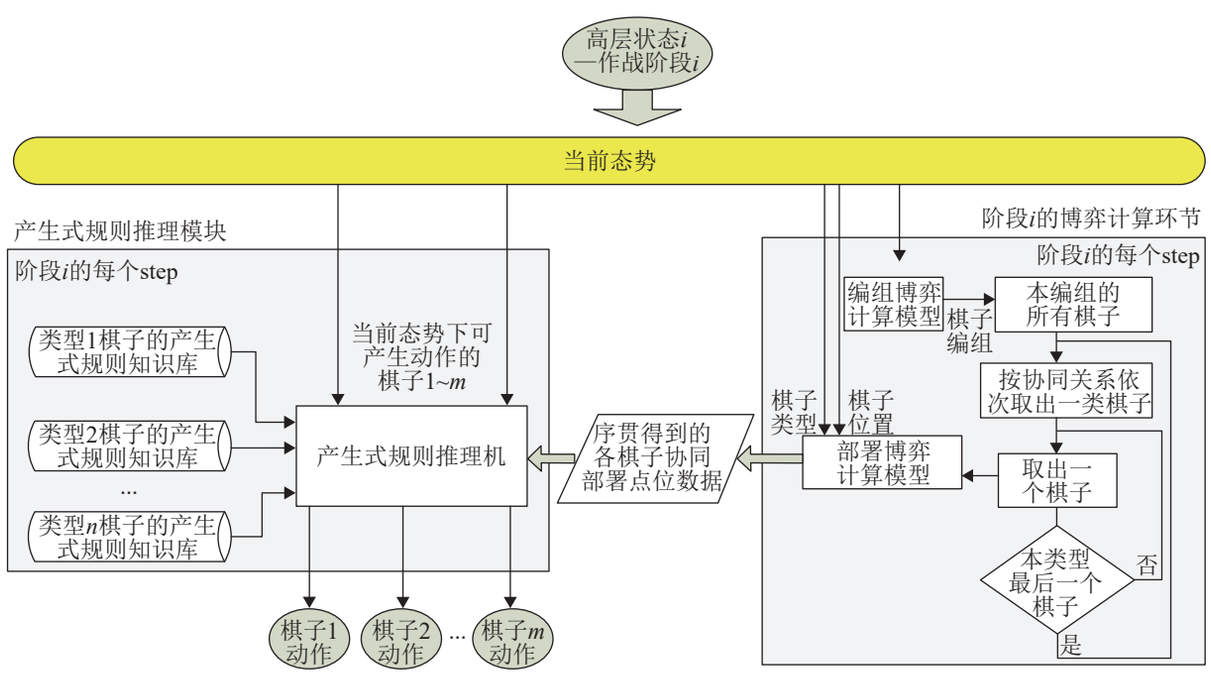


图2　阶段 i 的产生式规则推理模块和博弈计算环节

Fig. 2　Generative rule inference module and game calculation link for phase i

博弈计算环节每调用一次博弈计算模型，仅输出一个棋子的决策结果，该结果存储在协同部署点位数据变量中，在整个高层状态持续期间沿用，这样保证一个棋子在一个阶段对应一个任务。通过对多类多个棋子序贯调用博弈计算模型，推理输出当前棋子的部署点，同步考虑了其他已经决策的棋子，体现出阶段内多棋子的部署协同。

产生式规则推理模块用于单个高层状态持续期间生成各个棋子的具体行动指令。它包括了规则知识库和推理机两部分，如图2的左半部分所示。不同类型棋子的知识库使用不同知识规则排列。推理机根据规则知识库确定产生式规则的后件行动类型，而行动参数根据棋子编组和部署点位等信息计算生成。例如，产生式规则知识确定当前要执行机动动作，但具体向哪个方向机动的参数，则由能到达部署点的路径决定，避免每步机动方向也需要学习。在每个时间步，产生式规则推理模块将按编组、按棋子类型、按存活棋子数量实施推理机调用，每调用一次输出一个棋子的行动指令(可能是空指令)。仍以机动为例，贯序输出多个棋子的机动指令时，可通过产生式规则知识库中协同前件条件，实现阶段内多棋子的行动相互协同。

需要强调的是，深层神经网络是博弈计算模型的核心，是前述智能体迭代学习的依托，可用于建模各种难以精确总结的决策问题，如棋子在阶段任务中的最佳部署点。但存在的问题是，深层神经网络在随机参数冷启动下，决策效果较差，也难以迭代学习优化。在AlphaStar中采用人类数据监督学习和模仿学习来解决冷启动问题，在ChatGPT中使用了超大规模语料的预训练模型。而在兵棋推演智能体中，构建大量人类数据的成本很大。

为此，本文提出构建基于先验数学模型的神经网络冷启动方法。先通过策略博弈[17]模型构建编组及部署点位等阶段决策变量值的求解算法，得到博弈对抗场景下各个变量取值的概率，作为局部博弈均衡解下的混合策略。然后，将混合策略作为深层神经网络输出值的模仿对象，通过模仿学习得到红方深层神经网络参数初值，改善冷启动问题。同时，混合策略可直接视为较高水平的蓝方，当作红方的初始对手，红方通过强化学习与上述蓝方对抗，求解最优响应解[18]；得到最佳响应解的红方再作为蓝方的对手，使用强化学习求解蓝方的最佳响应，如此往复，博弈学习不断迭代，升级两方智能体的决策水平。

针对复杂作战场景，可构造作战方案中的每个阶段独立的博弈计算模型。不同阶段使用不同的神经网络，将使得神经网络训练难度降低、收敛变快。

3 兵棋推演智能体的设计实例

3.1 有限状态机

在庙算平台下，以分队级想定设计实现兵棋推演智能体。针对该想定，作战方案可划分为机动渗透阶段、中远攻击阶段、夺控战斗阶段。这样智能体对应的有限状态机有“机动渗透”状态、“中远攻击”状态、“夺控战斗”状态等3个高级状态。庙算平台在机动渗透阶段前，专门提供了初始的筹划部署阶段，此阶段推演不开始计时、不驱动棋子产生实际动作，可将其用于比较耗时的机动渗透博弈计算环节，因此，初始状态与“机动渗透”高级状态互相交叠。此外，庙算平台有固定长度的推演时长，到点后无论双方处于什么阶段都会退出，因此在有限状态机中，还设立了结束状态，该状态下也不需要驱动棋子产生动作。各状态间切换条件如图3所示。

在“机动渗透”“夺控战斗”两个高级状态下，智能体的决策空间由棋子编组变量和棋子的部署点位变量取值范围决定。在中远攻击状态下，认为各个棋子一般应坚守上一状态下各自的点位，对于躲避等动作可以使用产生式规则知识实现，而无需改变上一状态下的决策结果。

DOI: 10.16182/j.issn1004731x.joss.23-0477

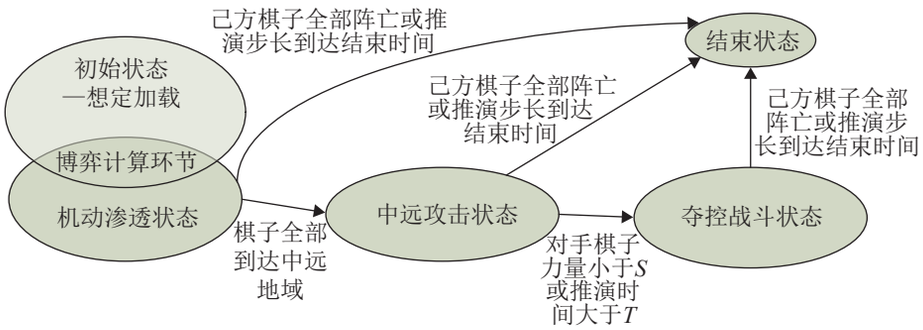


图3　庙算平台分队级下有限状态机

Fig. 3　Finite state machine of element level under miaosuan platform

3.2 博弈计算模型与反馈学习

“机动渗透”状态和“夺控战斗”状态博弈计算模型采用两套相互独立的深层神经网络，这是因为两个阶段任务有显著区别，独立的模型更容易训练，相互间的关联可以通过设计态势表征张量来传递。

针对分队级想定，一般只有占控主要夺控点和次要夺控点两个作战目标，棋子编组模式比较固定，因此，智能体对棋子编组不再采用神经网络进行迭代学习，而采用人类经验事前设定针对主夺控点和次要夺控点的棋子编组；智能体仅对棋子的阶段部署点位决策知识采用深层神经网络建模。

人工确定后的棋子编组和任务信息以参数的形式传给智能体，智能体据此将兵棋系统的态势数据转换成本编组的态势表征张量，如图4所示。态势表征张量体现了地图地形信息、作战目标信息、己方编组协同信息、对手信息、本次决策的棋子信息。每组信息用若干个矩阵描述，矩阵尺寸均为地图六角格的纵横编号数。首先，地图地形信息分为了8个矩阵。将地图高程归一化得到第1个矩阵；将居民地、丛林地等具有遮蔽效果的地点编码为1、其他编码为0，得到第2个矩阵；道路、河流、路障等地形因素影响机动速度，将它们统一编码为6个矩阵，分别对应从某个矩阵元素点位出发，沿着6个方向机动，速度是平地标准机动速度的倍数。在庙算平台中，作战目标指夺控点，将夺控点标识为1、周边6个控守点标识为0.5、其他非夺控点标识为0，得到本编组的作战目标信息矩阵，为1个矩阵。己方编组协同信息按棋子类型进行编码，设己方棋子类型数为num_1，则将有num_1×2个矩阵。每个类型有2个矩阵，第1个矩阵用于描述本编组中需要协同考虑的棋子，则其所在位置用兵力值(血量)来标注，由此得到第1个矩阵；第2个矩阵用来记录上述棋子已经走过的路径以及到达本阶段部署点位将要走的路径，即经过的位置矩阵值都为1，没有经过的为0。这样，在编码己方编组信息时就同步体现了协同信息，也就是说需要协同考虑的本编组的棋子才进行非0编码，否则均为0。

对手信息同样需要对手棋子类型数num_2×2个矩阵来表达，其中num_2为对手棋子类型数。每类棋子需要2个矩阵，第1个矩阵用于描述对手最近观察到的位置和兵力值，第2个矩阵用于描述可能的位置。通过最后一次观察到的位置、距离当前的时间间隔和机动速度计算得到。也就是说，第2个矩阵采用对手位置推理。

另外，对于己方棋子而言，巡飞弹、无人机等类型棋子比较特殊，不需要用神经网络决策。其余需要神经网络决策的棋子类型数记为num_3，决策对象信息也需要num_3个矩阵来表达。不同矩阵用于区别要决策的棋子类型，其所在位置用兵力值来标识，其他位置均赋值为0。

神经网络是博弈计算模型的核心。将态势表征张量输入深层神经网络，深层神经网络输出一

个棋子在全地图各个点位部署的概率值以及对这种部署的价值预测，即将策略网络和值函数网络合二为一。深层神经网络主体如图5所示，使用卷积神经网络[19]、U型网络[20]、全连接神经网络、归一化指数层等相结合方式构建。卷积层1和卷积层2主要用于对多通道特征进行合并。归一化指数层用于对各个点位的选择概率进行归一化处理。

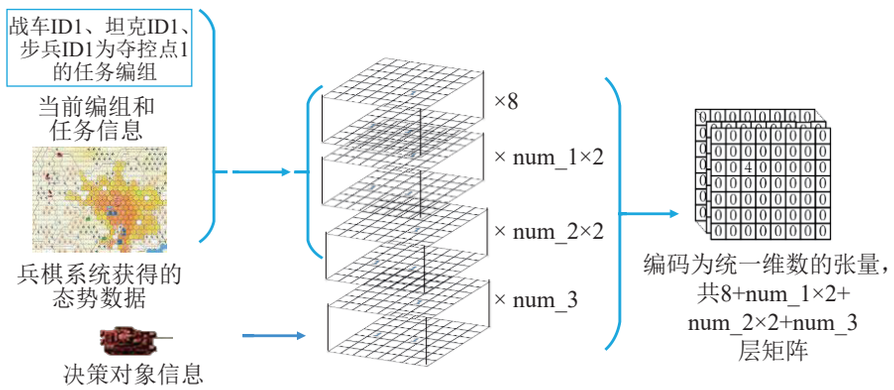


图4　态势表征张量的设计

Fig. 4　Design of situational representation tensors

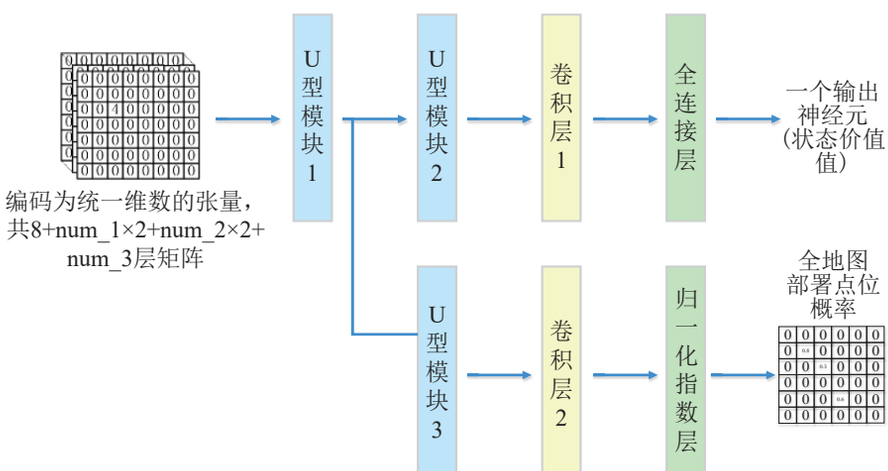


图5　卷积神经网络的设计

Fig. 5　Design of convolutional neural networks

U型模块用于对态势表征张量进行特征提取。它结合了卷积神经网络、反卷积神经网络[21]、残差网络[22]的特点，其中，反卷积层用于将特征还原至地图大小，输出部署点位的选择概率；残差模块用于降低梯度消失和爆炸问题。U型模块中卷积残差模块的个数与张量矩阵大小、卷积核大小、步长和问题求解的复杂度等有关。当卷积残差模块中卷积层步长为2时将图像大小缩减一半，对应图中梯形卷积残差模块，具体可见文献[23]。

上述的神经网络按协同关系序贯调用，这样依次将需要协同考虑的己方棋子部署信息编码到态势表征张量中，神经网络可实现阶段内多棋子的部署协同决策，并同步考虑对手博弈信息。

上述神经网络需要先模仿学习，再强化学习，总体采用了如图6所示的过程。

首先，先采用文献[24]的方法构造先验收益矩阵，求解矩阵博弈的部署点位混合策略，又称策略式博弈策略，将其作为神经网络输出的标签，

https://dc-china-simulation.researchcommons.org/journal/vol36/iss7/2
DOI: 10.16182/j.issn1004731x.joss.23-0477

形成模仿数据进行模仿学习，得到神经网络的冷启动参数值，将此时神经网络参数下对应的策略称为模仿学习策略。

在获得冷启动参数值后，再使用文献[25]中的博弈强化学习方法。其中，蓝方先采用混合策略作为冷启动红方的对手，红方使用 DeepNash 强化学习框架求解针对当前对手的最优响应策略，然后将红方最优响应策略固定，使用 DeepNash 训练此红方策略下的冷启动蓝方最优响应策略，如此交替循环，进行红蓝双方智能体的博弈学习。为降低样本数据获取时间，可以采用分布式采样，即在多台终端上使用智能体的当前策略进行对抗得到样本，回传给学习服务器，在学习服务器上进行策略更新，并更新分发到分布式采样终端上。最终学习到的策略称为博弈学习策略。

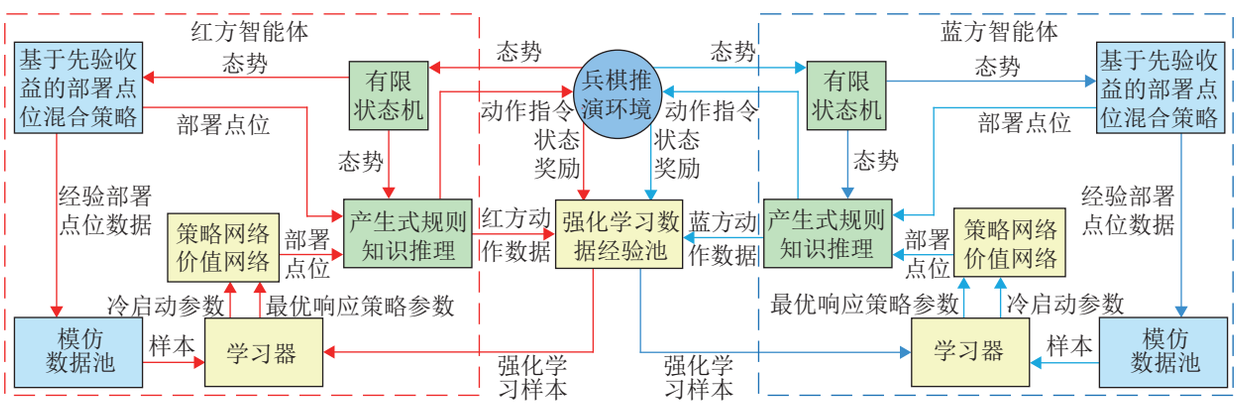


图 6　博弈计算模型中的反馈学习

Fig. 6　Feedback learning in game computational model

3.3 棋子的产生式规则知识库

产生式规则用于生成每个存活棋子的当前动作(可以为空)。应按棋子类别，分别构造其特色产生式规则知识库，配合推理机产生具体动作。

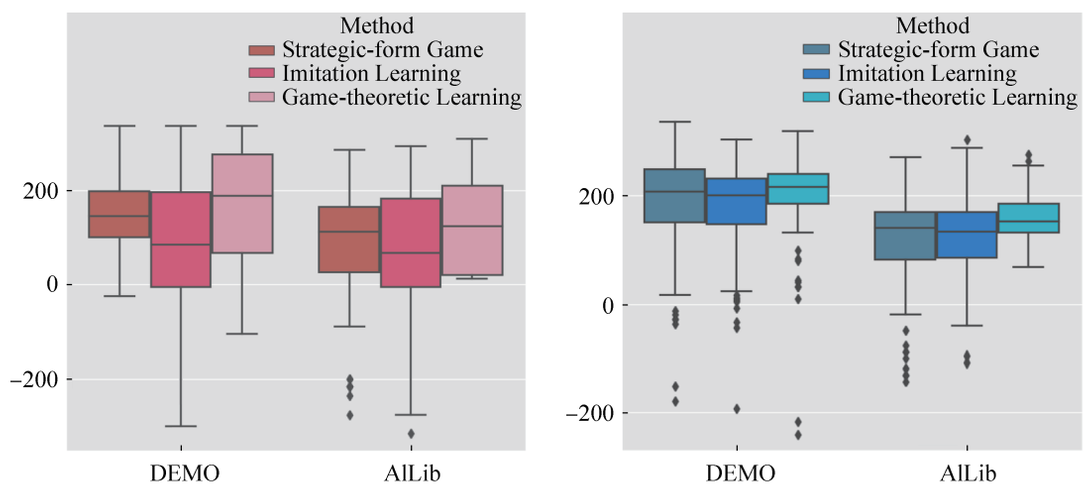
在庙算平台下，棋子的典型具体动作有射击、机动、夺控等。知识库的核心作用是排列这些动作前件条件的先后次序，推理机根据先后次序决策当前步实施何种动作。针对步兵棋子，其知识库可按射击、夺控、按阶段点位机动、隐蔽的前件条件是否满足依次排列；针对战车棋子，其知识库按照射击、夺控、躲避机动、按阶段部署点位机动、步兵下车、隐蔽的前件条件是否满足依次排列。这里增加了躲避机动和步兵下车两个特有动作，躲避机动是因为战车在遇见对手坦克或无人机后，若不具备射击条件，则应搜索附近有利地形进行躲避机动；步兵下车是战车作为兵力输送单元所特有的功能。

3.4 实验结果

庙算平台下以分队级对抗高原通道想定为场景，进行实验。在“机动渗透”高级状态下，智能体采用 3.2 节中的神经网络依次输出各个棋子机动到达点位概率，依此概率选择棋子的机动目的地，其中，战车机动目的同时是步兵的下车点位。在“夺控战斗”高级状态下，步兵棋子生命力强，由其实施夺控动作，其他棋子保持点位不变，继续射击可见的敌人。因此，在“夺控战斗”高级状态步兵采用直奔最近夺控点、交战、夺控。这样简化为不需要神经网络决策，且便于观察分析用于“机动渗透”的神经网络学习到的策略性能。

使用终局对抗的得分来度量神经网络学习到的策略性能，得分越高，代表学习到的策略越好。图 7 所示实验结果为前述策略式博弈策略、模仿学习策略和博弈学习策略下多轮对抗分数的分布箱线图，对手为庙算平台开源的 DEMO 和 AILib。

从图7中可见，在与DEMO对抗时，模仿学习策略相对于策略式博弈策略整体有所下降。这是由于神经网络是在所有可达点位范围内预测目标点位的，具有较多小概率事件，这些事件在策略式博弈策略中使用先验知识直接过滤掉了。在与AILib对抗时，模仿学习策略较策略式博弈策略有所提高，原因可在于候选点位扩充正好导致了针对AILib的优势点位增加，且测试结果受混合策略随机性和火力裁决随机性影响。相比于策略式博弈策略和模仿学习策略，博弈学习策略性能更高，并且发现该策略在与DEMO对抗时性能要略低于与AILib对抗，原因在于DEMO使用的是完全随机的策略，而AILib使用的是纯策略。



(a) 己方阵营为红方　　(b) 己方阵营为蓝方

图7　对抗得分分布箱线图

Fig. 7　Box-plots of battle score distribution

4　结论

兵棋推演智能体是实施作战方案创新、降低人员推演训练成本的有力工具，越来越受到关注。兵棋推演智能体需要具有决策策略的学习功能，但由于其模拟作战的极端复杂性，无法使用当前单一的人工智能方法端到端直接学习得到。本文在分析兵棋特点基础上，提出以作战方案为牵引，通过作战阶段构建以有限状态机为基础的智能体框架；根据兵棋训练的着重点和当前算力水平，形成以部署点位等为代表的智能体决策变量；根据决策变量的取值范围设计深层神经网络，构建博弈计算模块，确定决策变量取值，指导产生式规则知识输出各个棋子的具体动作。在博弈计算模块的神经网络学习部分，引入了模仿学习和强化学习机制，进行神经网络参数的迭代优化。实验分析了“机动渗透”高级状态下对应神经网络策略的学习训练性能。多高级状态对应多个神经网络需要分别训练，最终通过多轮次训练来探索全局作战方案的最优解。

本文给出了兵棋推演智能体的总框架。下一步应进一步细化作战样式，在对应作战样式的兵棋推演平台下，进一步找准作战结果的影响因素，将其细化为各作战阶段的决策变量。根据细分的作战样式，构建更高效的先验数学模型，实施多高级状态下多轮次模仿学习和博弈强化学习，迭代求解符合真实场景的决策变量取值，使智能体真正与作战方案相配合。

参考文献：

[1]　Schrittwieser J, Antonoglou I, Hubert T, et al. Mastering

DOI: 10.16182/j.issn1004731x.joss.23-0477

Atari, Go, Chess and Shogi by Planning with a Learned Model[J]. Nature, 2020, 588(7839): 604-609.

[2] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster Level in StarCraft II Using Multi-agent Reinforcement Learning[J]. Nature, 2019, 575(7782): 350-354.

[3] Berner C, Brockman G, Chan B, et al. Dota 2 with Large Scale Deep Reinforcement Learning[EB/OL]. (2019-12-13) [2023-01-30]. https://arxiv.org/abs/1912.06680.

[4] Starken A, Mondesire S, Wu A. Trends in Machine Learning for Adaptive Automated Forces[C]//Iterservice/Industry Training, Simulation, & Education Conference(I/ITSEC), 2022, 22243: 1-13.

[5] 施伟, 冯旸赫, 程光权, 等. 基于深度强化学习的多机协同空战方法研究[J]. 自动化学报, 2021, 47(7): 1610-1623.
Shi Wei, Feng Yanghe, Cheng Guangquan, et al. Research on Multi-aircraft Cooperative Air Combat Method Based on Deep Reinforcement Learning[J]. Acta Automatica Sinica, 2021, 47(7): 1610-1623.

[6] 徐佳乐, 张海东, 赵东海, 等. 基于卷积神经网络的陆战兵棋战术机动策略学习[J]. 系统仿真学报, 2022, 34(10): 2181-2193.
Xu Jiale, Zhang Haidong, Zhao Donghai, et al. Tactical Maneuver Strategy Learning from Land Wargame Replay Based on Convolutional Neural Network[J]. Journal of System Simulation, 2022, 34(10): 2181-2193.

[7] 胡晓峰, 齐大伟. 智能决策问题探讨——从游戏博弈到作战指挥,距离还有多远[J]. 指挥与控制学报, 2020, 6(4): 356-363.
Hu Xiaofeng, Qi Dawei. On Problems of Intelligent Decision-making-how Far is It from Game-playing to Operational Command[J]. Journal of Command and Control, 2020, 6(4): 356-363.

[8] 俞康伦. 兵棋设计[M]. 北京: 国防工业出版社, 2018.

[9] 阳曙光. 兵棋总体设计[M]. 北京: 机械工业出版社, 2018.

[10] DeepMind. Pysc2[EB/OL]. (2017-08-10) [2023-03-20]. https://github.com/google-deepmind/pysc2.

[11] 中国科学院. 庙算•陆战指挥官[EB/OL]. (2020-09-01) [2023-03-20]. http://wargame.ia.ac.cn/main.

[12] 孙宇祥, 彭益辉, 李斌, 等. 智能博弈综述:游戏AI对作战推演的启示[J]. 智能科学与技术学报, 2022, 4(2): 157-173.
Sun Yuxiang, Peng Yihui, Li Bin, et al. Overview of Intelligent Game: Enlightenment of Game AI to Combat Deduction[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(2): 157-173.

[13] 秦晓周. 联合作战辅助决策方法研究[M]. 北京: 国防大学出版社, 2019.

[14] 马平, 杨功坤. 联合作战研究[M]. 北京: 国防大学出版社, 2013.

[15] Millington L. AI for Games[M]. 3rd ed. 北京: 清华大学出版社, 2021.

[16] 石俊杰. 基于有限状态机的游戏角色控制系统设计与实现[D]. 武汉: 华中科技大学, 2016.
Shi Junjie. Design and Implementation of a FSM-based Role Control System[D]. Wuhan: Huazhong University of Science and Technology, 2016.

[17] Michael M, Eilon S, Shmuel Z. Game Theory[M]. Cambridge: Cambridge University Press, 2013: 155-166.

[18] 周雷, 尹奇跃, 黄凯奇. 人机对抗中的博弈学习方法[J]. 计算机学报, 2022, 45(9): 1859-1876.
Zhou Lei, Yin Qiyue, Huang Kaiqi. Game-theoretic Learning in Human-computer Gaming[J]. Chinese Journal of Computers, 2022, 45(9): 1859-1876.

[19] Li Zewen, Liu Fan, Yang Wenjie, et al. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 6999-7019.

[20] Zunair H, Ben Hamza A. Sharp U-net: Depthwise Convolutional Network for Biomedical Image Segmentation[J]. Computers in Biology and Medicine, 2021, 136: 104699.

[21] Zeiler M D, Taylor G W, Fergus R. Adaptive Deconvolutional Networks for Mid and High Level Feature Learning[C]//2011 International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2011: 2018-2025.

[22] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 770-778.

[23] 王玉宾. 面向兵棋推演的智能对抗策略生成技术研究[D]. 郑州: 战略支援部队信息工程大学, 2022.
Wang Yubin. Research on Intelligent Strategy Generation Technology for Wargame[D]. Zhengzhou: PLA Strategic Support Force Information Engineering University, 2022.

[24] 王玉宾, 孙怡峰, 吴疆, 等. 陆战对抗中的智能体博弈策略生成方法[J]. 指挥与控制学报, 2022, 8(4): 441-450.
Wang Yubin, Sun Yifeng, Wu Jiang, et al. An Agent Game Strategy Generation Method for Land Warfare[J]. Journal of Command and Control, 2022, 8(4): 441-450.

[25] Perolat J, Bart De Vylder, Hennes D, et al. Mastering the Game of Stratego with Model-free Multiagent Reinforcement Learning[J]. Science, 2022, 378(6623): 990-996.