文章编号:1673-3819(×)0×-0001-0

基于深度强化学习的装备组合运用方法

文东日[1,2], 陈小虎[1], 李 文[1], 杜二锋[1]
(1.国防大学, 北京 100091;2.中国人民解放军 63936 部队, 北京 102202)

摘 要:为实现装备自动匹配,尽可能高效利用装备资源,发挥整体作战效能,提出并分析了装备组合运用问题。探讨了深度强化学习的原理,建立了基于深度强化学习的装备组合运用方法的概念、模型、框架。以飞机反舰突防案例进行实验,验证了基于深度强化学习的装备组合运用方法的可行性,表明该方法能够有效解决装备运用领域的"组合爆炸"问题,可为装备指挥和决策部门制定装备运用方案提供理论和技术支持。

关键词:深度强化学习; 装备组合运用; 马尔科夫决策
中图分类号:TJ01 **文献标志码**:A **DOI**:10.3969/j.issn.1673-3819.×.×.00

Method of Equipment Combination Application Based on Deep Reinforcement Learning

WEN Dong-ri[1,2], CHEN Xiao-hu[1], LI Wen[1], DU Er-feng[1]
(1. National Defense University of PLA, Beijing 100091;2. Unit63936 of PLA , Beijing 102202, China)

Abstract: In order to realize automatic equipment matching, use equipment resources as efficiently as possible, and give full play to the overall operational effectiveness, the problem of equipment combination operation is put forward and analyzed. This paper discusses the principle of deep reinforcement learning, and establishes the concept, model and framework of equipment combination application method based on deep reinforcement learning. The feasibility of equipment combination application method based on deep reinforcement learning is verified by the experiment of aircraft anti-ship penetration, which shows that this method can effectively solve the problem of "combination explosion" in equipment application field and provide theoretical and technical support for equipment command and decision-making departments to formulate equipment application plans.

Key words: deep reinforcement learning; equipment combination application; Markov decision making

由于装备是离散存在,因此需要组合运用。组合运用是装备运用的内在要求,是装备释放体系作战效能的主要形式,是实现装备自动匹配的核心问题。但由于组合问题的复杂性,以及作战意图与作战环境的不确定,求解装备组合运用问题非常困难,各界对此进行了大量理论探索。如围绕不同领域,孙盛智等[1]研究了面向作战需求的卫星应用装备组合优化问题,李雄等[2]探索了面向目标中心战的自适应装备保障指挥方式。如基于不同的方法,宋春龙[3]提出"使命任务—能力需求—体系设计—结构优化"的装备组合问题研究框架,豆亚杰[4]提出用差分进化算法优化武器系统组合选择问题,孙建彬[5]提出基于遗传算法的武器系统组合优化方法,杜波[6]提出基于代理模型的武器装备体系优化方法,于少波[7]将多 Agent 评估方法用于电子信息装备体系作战效能评估,等等。2020 年,美国陆军在项目融合演习中试验了"烈火风暴"(FireStorm[8])人工智能软件,该软件能够根据打击目标自动匹配最优的"射手",表明在技术上实现装备自动匹配已成为可能。为此,本文着眼人工智能技术的最新发展,构建基于深度强化学习[9-10]的装备组合运用方法,探索装备自动匹配的实现途径。

文章分析了装备组合运用问题,建立了基于深度强化学习的装备组合运用方法的概念、模型、框架,进行了实验验证,在~~战机~~突破舰艇防空系统的作战想定下,智能体均能按照作战意图推荐较合理的装备运用方案,取得较理想的实验效果,表明了基于深度强化学习的装备组合运用方法的可行性。

1 装备组合运用问题描述

冷兵器时代,装备的组合运用主要是简单的数量与功能组合,体现在"阵"的形式与变换之中。商周时期,车阵编成中士兵以 5 人为一伍,分别执戈、戟、殳、矛、弓,形成 5×5 的步卒阵,是谓"两"(如图 1 所示[11])。在与商周同时代的西亚,出现亚述军事帝国及亚述阵(如图 2 所示[12])。此外,还有雁行阵、锥形阵、钩形阵、玄襄阵、鸳鸯阵等。通过"阵"的形式,发挥各个兵器的用途,形成整体的战斗能力。热兵器时代,

收稿日期: 2021-06-02 修回日期: 2021-06-17
作者简介: 文东日(1986—),男,湖南攸县人,硕士研究生,研究员实习员,研究方向为军事装备和智能化。
陈小虎( )

“阵”逐步退出历史舞台,散兵战术、机动作战成为主要特征,装备的组合在形式上没有冷兵器时代那样规整,但仍然是装备运用的精要。信息化时代,基于网络信息体系,各种装备能够在更广阔的时空范围内进行组合,表现出体系对抗的特点。1999 年科索沃战争,南联盟由于没有信息系统的支撑,其米格-29 战机在对阵 F-16 战机(性能与米格-29 相当)时,既搜索不到敌机目标,又不知面临的威胁,以致其空军司令员亲自驾机升空也只有殒命蓝天的悲壮。可见,装备组合运用是一个既古老又崭新的课题,随着时代发展,其组合的方法、形式、特点、规律又各不相同。

发现装备最优的组合形式,实现装备体系运用的效果,关键在于科学的装备组合运用方法。在智能化时代,创新装备组合运用方法,实现装备根据任务自动匹配,是增强决策优势,重塑作战流程、作战组织和作战理念的基础,甚至是推动智能化军事革命的关键所在。美军开发的所谓的马赛克战、决策中心战、分布式海上作战、联合全域指挥控制等作战概念,都是首先基于装备自动匹配问题的解决。为此,进行智能化作战,需加强装备组合运用问题研究,探索智能化的装备组合运用方法。

为定量研究装备组合问题,首先对问题进行形式化的描述。一般认为:装备组合运用问题,是在一定的作战条件和作战目标下,军事指挥人员为发挥最大的作战效能,研究如何组合运用多个装备的问题。用数学表述为:在一定的作战条件和作战目标下,设有 m 型装备,每型装备的数量依次为 $x_1,x_2,x_3,\cdots x_m$,用向量 $\boldsymbol{x}$ 表示,运用到的每型装备的数量依次为 $a_1,a_2,a_3\cdots a_m$ $(0\leqslant a_j\leqslant x_j,1\leqslant j\leqslant m)$,用向量 $\boldsymbol{a}$ 表示,$J(\boldsymbol{a})$ 表示装备组合运用的作战效能,求解使得 $J(\boldsymbol{a})$ 最大的 $\boldsymbol{a}$。

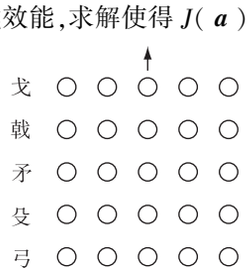

图 1 “两”的示意图

由于装备在运用中是“活”的,具有各种属性,比如时间先后顺序、空间方位路径、武器挂载方案、电磁管控措施等。考虑装备的运用属性,进一步把装备组合运用问题表述为:在一定的作战条件和作战目标下,设有 m 型装备,每型装备的数量依次为 $x_1,x_2,x_3,\cdots x_m$,用向量 $\boldsymbol{x}$ 表示,每型装备都有 n 个属性,每个属性可选的值的个数依次为 $s_1,s_2,s_3,\cdots s_n$,用向量 $\boldsymbol{s}$ 表示,运用到

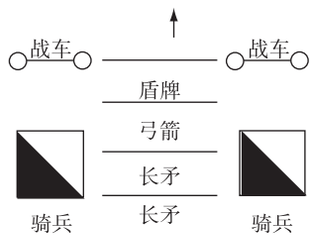

图 2 亚述阵示意图

的每型装备的数量依次为 $a_1,a_2,a_3\cdots a_m$ $(0\leqslant a_j\leqslant x_j,1\leqslant j\leqslant m)$,用向量 $\boldsymbol{a}$ 表示,其中,某个装备的运用属性依次为 $b_{11},b_{12},b_{13},\cdots b_{1n}$,用向量 $\boldsymbol{b}_i$ 表示,所有装备的运用属性依次为 $\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3\cdots$,用矩阵 B 表示,$J(\boldsymbol{a},B)$ 表示装备组合运用的作战效能,求解使得 $J(\boldsymbol{a},B)$ 最大的 $\boldsymbol{a}$ 及 B。

2 深度强化学习原理

深度强化学习也称深度增强学习,是融合深度学习与强化学习的一类人工智能算法。由于综合利用了深度学习的感知表示能力和强化学习的决策规划能力,深度强化学习更接近人类的思维方式,具有处理各种复杂问题的能力。

深度强化学习主要遵循强化学习的框架,采用马尔科夫决策过程,形式化地描述智能体与环境的交互过程,如图 3 所示[13]。

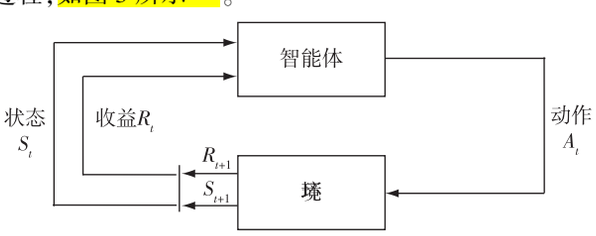

图 3 马尔科夫决策过程中的“智能体-环境”交互

在强化学习中,智能体的目标是最大限度地获取长期收益。假设在时刻 t 采取动作 A_t 后接受的收益序列为 $R_{t+1},R_{t+2},R_{t+3},\cdots$,引入折扣因子为 γ 表示 A_t 对后续收益的贡献衰减程度,G_t 表示期望回报。

$$G_t=R_{t+1}+\gamma^1R_{t+2}+\gamma^2R_{t+3}+\cdots\sum_{k=0}^{\infty}\gamma^kR_{t+k+1} \tag{1}$$

智能体期望回报的大小取决于智能体选择的动作,选择动作的根据称之为策略,用 π 表示。v_π 称为策略 π 的状态价值函数,$v_\pi(s)$ 表示在策略 π 下状态 s 的价值。

$$v_\pi(s)=\mathbb{E}\mid G_t\mid S_t=s=\mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^kR_{t+k+1}\mid S_t=s\right] \tag{2}$$

q_π 称为策略 π 的动作价值函数,$q_\pi(s,a)$ 表示在状态 s 时根据策略 π 采取动作 a 的价值。

$$q_\pi(s,a)=\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\mid S_t=s,A_t=a\right] \quad (3)$$

为解决一般的强化学习算法对高维状态空间和动作空间的难题,运用深度神经网络近似表示值函数或策略函数,即为深度强化学习。最早将深度学习和强化学习结合,是郎齐(Lange)等人[14]将深度自编码网络应用到强化学习中,解决路径规划寻优问题。而深度强化学习的真正开端是尼曲(Mnih)等人[15]在2013年提出深度Q学习算法(DQN),直接从视频图像中学习玩Atari游戏。当前,深度强化学习的算法主要有深度确定性策略梯度算法(DDPG)、异步的基于优势函数的"行动器—评判器"算法(A3C)、信赖域策略优化算法(TRPO)等、近端策略优化算法(PPO),以及分层深度强化学习、多智能体深度强化学习、多任务迁移深度强化学习等前沿研究方向。

深度强化学习解决复杂问题的能力在围棋、星际争霸、刀塔(Dota)等游戏中得到充分体现,启发广大研究人员利用其解决军事问题。而且应用深度强化学习解决现实问题,具有无需数据样本从而摆脱数据依赖,无需环境模型从而超越经验知识,无需提取特征从而绕过特征工程等优势。

3　装备组合运用模型及仿真框架

3.1　装备组合运用模型

本文采用马尔科夫决策过程的框架,构建装备组合运用的基本模型:指挥人员从作战环境中获取战场态势,作出装备运用决策,获得战绩得分,持续以上循环过程直至任务结束,如图4所示。与强化学习对应的核心概念如下:

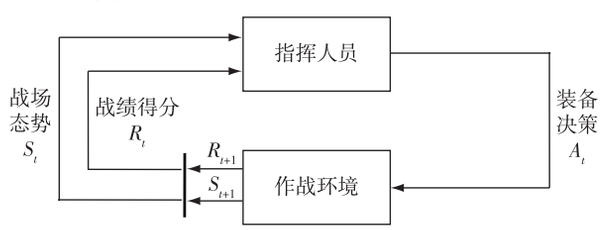

图4　马尔科夫决策过程中的装备组合运用的基本模型

指挥人员即为智能体,是作出装备组合运用决策的主体。军事问题有战略、战役、战术等不同的层次,不是所有层次的指挥员都被统一视为智能体,如在考虑战术行动的装备组合运用问题时,战术层次的指挥员可被视为智能体,而战役层次的指挥员则视为环境的一部分;考虑战役行动的装备组合运用问题时,战役层次的指挥员可被视为智能体,而战略层次的指挥员则视为环境的一部分。

在强化学习中,智能体之外所有与其相互作用的事物都被称为环境。对于装备组合运用问题,对应所指的主要是作战环境。如上文把有的层次的指挥人员视为环境的一部分,基于深度强化学习的装备组合运用方法所指的环境,不仅包括部队所处的自然环境、社会环境等,还包括本层次指挥员所不能控制的其他所有的部分,如作战对手、友方部队、上级单位等。

状态是任何对决策有帮助的信息,可以把战场态势作为状态。指挥人员根据战场态势作出装备决策。状态信息主要来源于作战对手(知彼)、己方部队(知己)、客观环境(知天知地)等三个方面。由于存在"战争迷雾",装备组合运用问题中的状态不是完全可观测的,属于不完全信息决策问题。

装备决策是指挥员的动作,包括决策动用装备的型号、数量以及各种属性等。决策动用一件装备可以看作是做出一个动作,多个动作决策形成装备组合运用方案,从而把装备组合问题转变为序贯决策问题。由于装备是离散的,因而对于其组合运用问题,可认为是在离散动作空间的强化学习问题。与收益存在延迟现象类似,指挥员作出装备动用的决策可能是计划的,因而其实际效果也可能存在延迟。

战绩得分是指挥员的目标,可被视为智能体的收益,衡量装备运用的整体效能。战绩得分根据仿真推演评分标准(评分标准由作战想定事先确定)计算得到,体现的是想定作业的任务要求。例如,设摧毁敌方指挥所为胜,得1分,否则得-1分。评分标准也可以是多种指标的综合。以战绩得分作为智能体的奖惩函数,计算智能体的状态或动作的价值函数,引导指挥员学习最佳装备运用策略,评估装备整体作战效能。为避免奖励稀疏的问题,可根据具体情况,适当调整智能体的奖惩函数,以引导智能体更快收敛。

基于马尔科夫决策过程的框架,利用一个主体—智能体、一个客体—环境、三个要素—状态、动作、收益,抽象地描述了装备运用的决策过程。智能体在仿真环境的交互中不断收集状态、动作、下一个状态、收益以及是否结束的经验数据,从而利用强化学习算法探索实现战绩得分最大的装备运用策略。基于马尔科夫决策过程的装备组合运用模型体现了博伊德的"OODA环"模型(如图5所示)的内在逻辑。"OODA环"模型强调了指挥人员在"智能体—环境"交互中的主体地位,无论是装备决策的输出还是战场态势的输入,指挥人员都可以能动地施加影响。这说明,基于马尔科夫决策过程描述装备组合运用现象,符合一般军事规律,甚至可以扩展至普遍的军事决策问题。

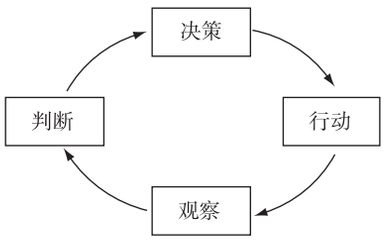

图 5 “OODA”模型

3.2 仿真框架

为实现基于深度强化学习的装备组合运用方法,仿真框架主要包括开发平台、仿真平台、计算平台等三个部分。其中,开发平台主要实现智能体的功能,搭建神经网络,实现算法模型,输入仿真数据,输出决策指令;仿真平台主要实现环境的功能,提供支撑推演所需的作战想定、装备模型、作战规则等,输入决策指令,输出仿真数据;计算平台提供基础运行环境,并为算法模型优化及仿真引擎运行提供支持。为实现开发平台与仿真平台的交互,需仿真平台 AI 开发包提供各类函数接口以及通信接口。总体框架如图 6 所示。

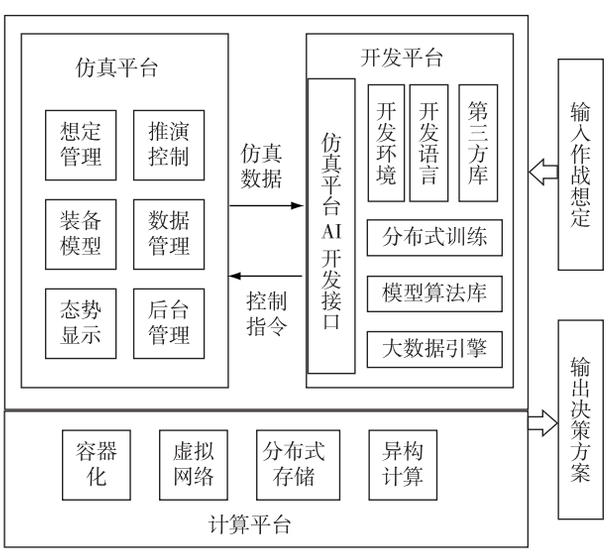

图 6 基于深度强化学习的装备组合运用方法的总体框架

4 实例验证

4.1 作战想定

想定红方航空母舰在太平洋某海域战斗归航途中,发现蓝方导弹驱逐舰。蓝方导弹驱逐舰以 37 km/h 的速度向西北方向逃窜。红方有三种作战意图:1)以最少的兵力摧毁蓝方导弹驱逐舰;2)不惜一切代价尽快摧毁蓝方导弹驱逐舰;3)最少的兵力、尽快摧毁蓝方导弹驱逐舰。红蓝双方兵力编成如表 1、表 2 所示。要求红方根据评分标准(表 3、表 4、表 5),派出合理的飞机编队,摧毁蓝方导弹驱逐舰,达成作战意图。想定时间 100 min。

表 1 蓝方兵力编成

序号	单元名称	数量	单元主要武器	数量
1	阿里伯克级导弹驱逐舰	1	RIM-162A 型“海麻雀”舰对空导弹	30

表 2 红方兵力编成

序号	单元名称	数量	单元主要武器	数量
1	F-35C 型“闪电Ⅱ”战斗机	8	AGM-154C 联合防区外武器	6
2	F/A-18A 型“大黄蜂”战斗机	12	GBU-54(V)2/B 激光联合直接攻击弹药	4
3	EA-18G“咆哮者”电子战飞机	4	AN/ALQ-99F-V 型主动干扰电子对抗吊舱	4
			AN/ALQ-126B 型防御电子对抗吊舱	1

表 3 评分标准(一)

序号	评判事件	单位	价格(百万美元)	定基得分
1	摧毁阿里伯克级导弹驱逐舰	艘	1843	+1000
2	损失 1 架 F-35C 舰载机	架	139	-75
3	损失 1 架大黄蜂战斗机	架	50	-27
4	损失 1 架电子战飞机	架	125	-68
5	出动一个飞机架次	次	-	-10

表 4 评分标准(二)

序号	评判事件	单位	价格(百万美元)	定基得分
1	摧毁阿里伯克级导弹驱逐舰	艘	1843	+1000
2	每消耗一分钟	分	-	-10

表 5 评分标准(三)

序号	评判事件	单位	价格(百万美元)	定基得分
1	摧毁阿里伯克级导弹驱逐舰	艘	1843	+1000
2	损失 1 架 F-35C 舰载机	架	139	-75
3	损失 1 架大黄蜂战斗机	架	50	-27
4	损失 1 架电子战飞机	架	125	-68
5	出动一个飞机架次	次	-	-10
6	每消耗一分钟	分	-	-10

4.2 问题分析

红方拥有 3 种机型共 24 架飞机,在不考虑时间先后顺序的情况下,存在 351 种组合方案。若考虑飞机出击的先后顺序(假设时间离散为 30 个时间段),则大约存在 2^{116}种组合方案。由于巨大的解空间,难以通过暴力搜索最优组合方案。且不同的作战意图之下,战

机组合运用的作战效能又不相同,使得问题更加的复杂。

运用深度强化学习方法求解该问题,红方指挥员可视为智能体,目的是在各种作战意图之下获取最大战绩得分。由于红方共有 24 架飞机,设每个回合有 24 个决策步。在每个决策步,智能体决定动用某架飞机及其起飞时间。通过每个决策步的累积,最终形成红方战机组合运用方案。由于红方指挥员在制定战机运用方案中,对于蓝方的情况是未知的,其每个决策步的状态输入是己方已~~计划~~动用的战机及其出动时间。通过在仿真环境中获得战绩得分的反馈,红方指挥员经过多个回合的训练,逐渐学习获得最优的战机组合运用方案。

4.3 实验设计

实验采用深度强化学习 PPO 算法,设计 F-35C 战斗机对舰打击、F/A-18A 战斗机对舰打击、EA-18G 电子战飞机电子干扰以及空动作等 4 种动作类型,每个动作包含单元、目标、时间(时间离散化为 30 个时间段)、任务、条令 5 个要素,共 91 个动作。根据动作空间设计相应的状态空间,共设计 91 个状态变量。仿真平台根据智能体的动作进行推演,在推演结束时按照评分标准计算智能体的战绩得分。在不同的作战意图下,采用相同的网络结构及算法分别进行训练。学习率设为 0.0001,衰减因子设为 0.99998.

4.4 实验结果

实验在 windows7 操作系统进行,采用墨子联合作战推演仿真软件(包括墨子军事 AI 平台),以及 Python、PyTorch 等开发工具。实验场景如图 7 所示。

在以最少的兵力摧毁蓝方导弹驱逐舰的作战意图下,智能体推荐的战机组合运用方案为:4 架 EA-18G“咆哮者”电子战飞机、2 架 F-35 战斗机,具体策略是先运用电子战飞机干扰蓝方、引诱蓝方发射防空导弹,再运用 F-35 战斗机发射反舰导弹、摧毁蓝方驱逐舰。(实验数据如图 8 所示)。运行保存的网络模型,测试 10 个回合,智能体胜率 90%、平均战绩得分 725.5 分。可以看出,由于用兵过于谨慎,智能体推荐的装备方案有时不能达成作战意图。

在不惜一切代价尽快摧毁蓝方导弹驱逐舰的作战意图下,智能体推荐的战机运用方案为:派出全部战机攻击蓝方驱逐舰,具体策略是先派出飞行速度快、攻击力强的 F-35 战斗机,再派出其他机型。(实验数据如图 9 所示)。运行保存的网络模型,测试 10 个回合,智能体平均战绩得分 650 分,平均耗时 35 min。显然,由于作战意图过于粗放,战机运用方案冗余过大。

在以最少的兵力、尽快摧毁蓝方导弹驱逐舰的作战意图下,智能体推荐的战机运用方案为:派出 6 架 F-35 战斗机。(实验数据如图 10 所示)。运行保存的网络模型,测试 10 个回合,智能体胜率 100%、平均战绩得分 334 分、耗时 35 min。此时,虽有 4 架战机消耗,但能保证任务完成,既不致于兵力过于冗余,又不致于作战时间过于拖沓。

通过实验,智能体在不同的作战意图之下,不利用人类经验,通过自我学习探索,均能推荐合理的装备运用方案,显示该方法具有一定的可行性和优势。但也发现该方法存在训练过程不稳定、训练时耗过长等问题。

5 结束语

装备组合运用方法研究是探索实现装备自动匹配、发挥体系效能的关键。本文分析了装备组合运用问题,构建了基于深度强化学习的装备组合运用方法。实例表明,该方法能够有效解决装备“组合爆炸”问题,可为装备指挥和决策部门制定装备运用方案提供理论和技术支持。本文主要针对给定想定下的装备运用问题进行研究,而战场环境和作战目标是持续变动的,如何在动态条件下实时地进行装备运用的组合优化是今后的研究重点。

参考文献:

[1] 孙盛智, 侯妍, 裴春宝. 面向作战需求的卫星应用装备组合优化研究[J]. 电光与控制,2018,25(5):7-11,16.

[2] 李雄, 韩战宁, 樊延平, 等. 面向目标中心战的自适应装备保障指挥方式研究[J]. 指挥与控制学报,2018, 4(2):130 813635D 3 3 1

[3] 宋春龙. 面向使命任务的装备组合方案生成方法及应用研究[D]. 长沙: 国防科技大学研究生院,2018.

[4] 豆亚杰. 武器系统组合选择问题与决策方法研究[D]. 长沙: 国防科技大学研究生院,2016.

[5] 孙建彬, 邢立宇. 基于遗传算法的武器系统组合优化方法[J]. 价值工程,2011,(29):9-10.

[6] 杜波. 基于代理模型的武器装备体系优化方法研究[D]. 长沙: 国防科学技术大学研究生院,2010.

[7] 于少波, 李新明, 刘东, 等. 基于 Multi-Agent 的电子信息装备体系作战效能评估方法[J]. 四川兵工学报, 2015,36(7):79-82.

[8] Sydney J, Freedberg Jr. A slew to a kill: Project Convergence [N/OL]. breakingdefense, 2020-09-16 [2021-06-01]. https: // breakingdefense. com/2020/09/a-slew-to-a-kill-project-convergence/

[9] 马骋乾, 谢伟, 孙伟杰. 强化学习研究综述[J]. 指挥

控制与仿真,2018,40(6):68-72.
[10] 罗荣, 王亮, 肖玉杰, 等. 深度学习技术在军事领域应用[J]. 指挥控制与仿真,2020,42(1):1-5.
[11] 江林. 战术简史[M]. 北京: 解放军出版社, 2016:12.
[12] 江林. 战术简史[M]. 北京: 解放军出版社, 2016:14.
[13] SUTTON R, BARTO A. Reinforcement learning: an introduc-tion [M]. Cambridge, Massachusetts: MIT Press, 1998.
[14] LANGE S, RIEDMILLER M. Deep auto-encode neural net-works in reinforcement learning[C]. // Proc. of the 2010 International Joint Conference on Neural Networks, 2010:1-8.
[15] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with reinforcement learning[EB/OL]. [2017-2-2]. http://arxtv.org/pdf/1312.5602.pdf.

(责任编辑:)