文章编号:1002-0640(2021)04-0072-06

基于深度强化学习的无人战车自主行为决策*

张 耀,武富春,王 明*,段 宏,张 昭,王海龙
(北方自动控制技术研究所,太原 030006)

摘 要:针对高动态强对抗战场环境下,无人战车面临的自主行为决策问题,分析了未来陆战场无人战车实际作战需求,构建了基于马尔可夫决策过程的自主行为决策模型,提出了一种深度强化学习结合行为树的方法,利用行为树的逻辑规则与先验知识降低强化学习问题的难度,保证收敛性和鲁棒性,同时使行为决策模型具有学习能力。构建典型作战场景,验证深度强化学习结合行为树的无人战车自主行为决策方法的有效性。

关键词:无人战车,火力打击决策,强化学习,行为树
中图分类号:TJ811 **文献标识码**:A DOI:10.3969/j.issn.1002-0640.2021.04.013
引用格式:张耀,武富春,王明,等.基于深度强化学习的无人战车自主行为决策[J].火力与指挥控制,2021,46(4):72-77.

Autonomous Behavior Decision of Unmanned Combat Vehicle Based on Deep Reinforcement Learning

ZHANG Yao, WU Fu-chun, WANG Ming*, DUAN Hong, ZHANG Zhao, WANG Hai-long
(*North Automatic Control Technology Institute*, *Taiyuan* 030006, *China*)

Abstract: Aiming at the problems of autonomous behavior decision-making faced by unmanned combat vehicles in a highly dynamic and strong confrontation battlefield environment, the actual combat requirements of future unmanned combat vehicles on land battlefields are analyzed, and an autonomous behavior decision-making model based on the Markov decision-making process is constructed. The method of deep reinforcement learning combined with behavior trees is proposed, the logic rules and prior knowledge of the behavior tree are used to reduce the difficulty of reinforcement learning problems and to ensure convergence and robustness, and to make behavior decision models have learning capabilities to deal with emergency situations. A typical combat scenario is constructed to verify the validity of the autonomous behavior decision- making method of unmanned combat vehicles combined with deep reinforcement learning and behavior trees.

Key words: unmanned combat vehicles, fire strike decision, reinforcement learning, behavior tree
Citation format: ZHANG Y, WU F C, WANG M, et al. Autonomous behavior decision of unmanned combat vehicle based on deep reinforcement learning [J]. Fire Control & Command Control, 2021, 46(4): 72-77.

0 引言

随着科技的不断进步,战争形态将发生深刻的变革。在军事智能化的趋势之下,无人化作战将成为基本形态[1]。无人战车是未来陆军实施地面突击作战的主要装备,其自主系统主要由感知、决策和

收稿日期:2020-01-15 **修回日期**:2020-03-24
* **基金项目**:兵器工业联合基金资助项目(6141B011504)
作者简介:张 耀(1996-),男,山西汾阳人,硕士研究生。研究方向:人工智能。
*通信作者:王 明(1984-),男,山西汾阳人,博士,硕士生导师,副研究员。研究方向:火力与指挥控制,无人系统,人工智能等。

控制等子系统组成。其中,自主行为决策模块,是无人战车在复杂陆战场环境中对敌目标实施快速、精确、有效打击的核心[2]。

在传统装甲装备作战过程中,完成作战任务主要靠驾驶员、车长、炮长分工协作,存在目标搜索速度慢,操作随机性大,决策时间长的问题,同时作战效能受乘员心理素质、生理状况以及战场环境的影响较大[3]。随着人工智能的不断发展,由于机器对海量信息处理能力强,反应速度快,面对动态战场环境具有独特的优势,逐渐替代人类乘员成为可能,推动无人战车的出现。无人战车按照系统控制方式可以分为遥操作控制、半自主控制和全自主控制[4]。随着无人战车智能化水平的提高,全自主控制成为其未来发展的必然方向,自主行为决策技术必将成为主要的技术推动力。

目前,对于无人战车自主行为决策的研究还处于起步阶段。相比之下,无人战斗机自主空战决策[5]学术研究成果丰硕,虽然应用场景不同,但是其行为决策方法有很大参考价值,主要包括基于规则的方法以及基于学习的方法两大类。基于规则的行为决策方法根据大量数据,以及专家知识构建动作规则库,针对不同态势制定对应的行为决策;基于学习的方法则以强化学习方法为代表,文献[6]采用强化学习中的 Actor-Critic 构架,通过神经网络学习,解决连续状态空间上的空战决策问题。但是该方法存在收敛速度慢、算力要求高的缺点。

因此,本文提出一种深度强化学习结合行为树的方法解决无人战车自主行为决策问题,利用行为树的逻辑规则与先验知识降低强化学习的问题复杂度,保证算法稳定收敛,同时使行为决策模块具有学习能力。本文从未来陆战场单车的实际作战需求出发,研究无人战车自主行为决策技术。分析了无人战车自主行为决策问题;建立了自主行为决策模型;提出一种深度强化学习结合行为树的自主行为决策方法;最后,针对典型作战场景,利用无人战车对战模拟仿真环境,验证所提出的自主行为决策方法的有效性。

1　无人战车自主行为决策问题分析

无人战车是一种能感知环境并与环境交互,具有自主地面机动能力、自主精确火力打击能力的智能化装甲装备。作为无人战车“大脑”的行为决策模块,直接体现了无人战车的自主水平,对于战车快速、精确、有效地打击敌方起着决定性作用。根据无人战车执行作战的任务流程分析,无人战车采用类人自主行为决策结构,如图 1 所示。无人战车自主行为决策包括自主机动决策与自主火力决策。

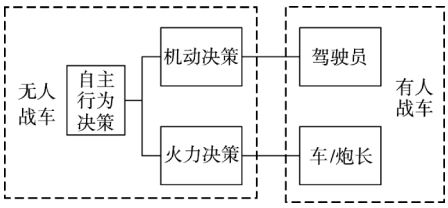

图 1　无人战车类人自主行为决策

机动决策是自主行为决策的基础。无人战车的机动决策是指根据实时感知的环境信息、自身行驶状态,对随时可能出现的敌方行为作出避险或迎敌反应,无人战车自主产生合理驾驶策略的过程。自主机动决策模块对应于传统驾驶员的决策行为,连接环境感知模块和车辆底盘运动控制模块,共同实现自主行驶。主要包括自适应巡航,避障避险,通过特定区域等内容,需要不断地适应战场环境变化,调整自己的机动速度和方向,从而能够快速安全驶向目标区域。从战术层面考虑,机动能够使我方占据有利地形,无论攻守都处于有利态势。

火力决策是自主行为决策的核心。自主火力决策模块对应于传统车长和炮长的决策行为,完成目标搜索、目标瞄准、火炮控制和目标打击等使命,直接决定了无人战车整体的战斗力。现有的武器火控系统在人的操作下能够实现目标探测、火控解算和控制武器射击等使命。无人战车则需要自主火力决策与武器火控系统相配合,利用快速处理信息的能力以及学习能力,实现自主目标瞄准和自主目标打击决策,以提高首发命中率和射击反应时间。

由于战场环境复杂,动态时变,特别是对抗场景下,敌方行为具有不确定性,无人战车自主行为决策模块需要根据环境态势选择动作,通过作出多步决策,实现作战任务。其所面临的问题称为序贯决策优化问题(Sequential Decision Problem)。基于规则构建决策模型,难以适应这种动态不确定性;强化学习方法可以通过对策略的迭代,找到最优策略,对于突发事件也有较好的响应,故采用强化学习方法解决自主行为决策问题。

2　基于马尔可夫决策过程的自主行为决策模型

2.1　马尔可夫决策过程

强化学习方法是一种基于马尔可夫决策过程(Markov Decision Process,MDP)的序贯决策方法,其核心思想是交互试错[7]。马尔可夫决策过程可由元组 $<S,A,P,R,\gamma>$ 表示,其中:S 是一个有限状态集,

A 是一个有限行为集，$\boldsymbol{P}$ 是集合中基于行为的状态转移概率矩阵，R 是基于状态动作对(s,a)的奖励函数，γ 是奖励折扣因子，取值范围为 $\gamma\in[0,1]$。

无人战车的自主行为决策过程可建模为一个基于 MDP 的序贯决策问题。在作战过程中，无人战车需要不断观察环境态势，获取状态 s，并依据策略 $\mu(a|s)$选择动作 a，通过多步决策，最终完成火力打击任务。在与战场环境的交互中，无人战车通过交互试错最大化累积奖励 r，使其自主行为策略收敛至最优，如图 2 所示。

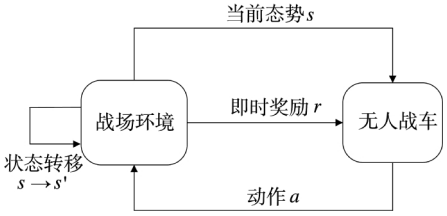

图 2　强化学习原理图

2.2　状态空间与动作空间

针对无人战车自主机动决策问题，本文选取我方战车位置、战车速度、机动目标位置、碰撞检测等参数描述状态空间 S_m，如表 1 所示。动作空间 A_m 由方向、油门、刹车的连续控制量组成，如表 2 所示。

表 1　自主机动决策状态输入

状态量	数量	取值范围
战车位置	2	[-2 000,2 000](m)
战车速度	2	[0,60](km/h)
机动目标位置	2	[-2 000,2 000](m)
碰撞检测	1	{0,1}

表 2　自主机动决策动作输出

控制量	数量	取值范围
方向	1	[-1,1]
油门	1	[0,1]
刹车	1	[0,1]

针对无人战车自主火力问题，本文选取目标距离、战车方位角、目标毁伤程度、我方战车毁伤程度等参数描述状态空间 S_f，如表 3 所示。动作空间 A_f 由火炮高低角、方位角和是否开火描述，如表 4 所示。

2.3　奖励函数与约束条件

奖励函数是强化学习算法评估当前动作好坏的直接指标，也是指导策略迭代优化方向的关键因素。奖励函数的设计，直接影响到算法的学习效率和最终策略。

表 3　自主火力决策状态输入

状态量	数量	取值范围
战车方位角	1	[-180°,180°]
战车速度	2	[0,60](km/h)
目标距离	1	[0,3000](m)
目标方位角	1	[-180°,180°]
目标速度	2	[0,60](km/h)
目标毁伤程度	1	[0,1]
战车毁伤程度	1	[0,1]

表 4　自主火力决策动作输出

状态量	数量	取值范围
火炮高低角	1	[-10°,30°]
火炮方位角	1	[-180°,180°]
开火	1	{0,1}

针对无人战车自主机动决策问题，本文选取奖励函数为 $r_m=I_{\text{collision}}+\frac{1}{t_1}+\frac{1}{D}$。其中，$I_{\text{collision}}$ 表示是否发生碰撞，若未发生碰撞取 0，若发生碰撞，取 -50，t_1 为自主机动总时长，D 为无人战车与机动目标之间的距离。

针对无人战车自主火力决策问题，考虑首发命中率与反应时间，本文选取奖励函数为：$r_f=r_{\text{hit}}+r_{\text{damage}}+\frac{1}{t_2}$。其中，$r_{\text{hit}}$ 为命中奖励，若命中敌方取 400，默认将敌方击毁，若未命中取 -50，r_{damage} 为被毁奖励，若被敌方击毁取 -100，t_1 为火力决策反应时间，从发现敌方目标开始计算。

3　深度强化学习的自主行为决策方法

3.1　深度强化学习

深度强化学习算法有众多分支，本文采用 DDPG(Deep Deterministic Policy Gradient)算法，针对连续的状态空间和动作空间，适用于无人战车的自主机动控制、火力打击等连续变量控制问题。DDPG 算法采用了双神经网络的 Actor-Critic 框架[8]，其算法框架如下页图 3 所示。

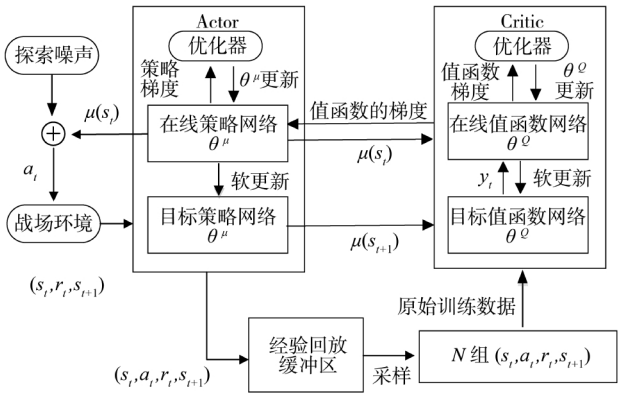

图 3　DDPG 算法框架图

首先,Q 函数(也称为动作值函数)定义为针对特定动作 a_t 的累积奖励期望:

$$Q^{\pi}(s_t,a_t)=E[\sum_{i=0}^{\infty}\gamma^i r_i \mid s_0=s_t,a_0=a_t]$$

对于 Critic 网络部分,使用一个卷积网络来近似值函数,故 Critic 网络又称值函数网络,其输入为动作 a 和当前状态 s,输出为 $Q(s,a)$,而值函数网络的参数更新向着估计的 Q 值和实际的 Q 值的平方损失最小的方向进行,其中估计的 Q 值可以通过状态估计网络(输入为当前状态 s_t 和动作估计网络输出 a_t)得到,实际的 Q 值为状态现实网络(输入为下一状态 s_{t+1} 和动作估计网络的动作输出 a_{t+1})输出的 Q 值和即时收益 R 的加和得到。

$$y_t=r(s_t,a_t)+\gamma Q(S_{t+1},\mu(S_{t+1}\mid\theta^{Q}))$$

$$\min_{\theta^{Q}}\mathrm{Loss}=\min_{\theta^{Q}}\frac{1}{N}\sum_{i=1}^{N}(y_i-Q(S_i,a_i\mid\theta^{Q}))^2$$

其中,θ^{Ω} 为 Critic 网络的权重参数,θ^{μ} 为 Actor 网络的权重参数,s_t 为 t 时刻的输入状态,a_t 为针对 t 时刻的输入状态输出的动作,$r(s_t,a_t)$为在状态 s_t 下执行动作 a_t 的即时奖励,$Q=Q(s,a\mid\theta^{Q})$表示 Critic 网络的输出,$\mu=\mu(s\mid\theta^{Q})$表示 Actor 网络的输出。

对于 Actor 网络,算法同样利用卷积网络来近似策略函数,故 Actor 网络又称为策略网络,其输入为当前状态 s,输出为动作 a。而策略网络的参数更新主要是朝着值函数网络输出增大的方向进行,梯度可以近似转化计算。

为了加速 DDPG 算法稳定收敛,本文采用经验回放机制[9],先将交互数据存入缓存区中,再均匀随机采样,打破数据之间的关联,使策略更快地稳定收敛。通过引入软更新技巧,对 Actor 和 Critic 分别设置在线网络和目标网络,实现目标网络参数的平稳更新。公式如下:

$$w'_{t+1}=\tau w_t+(1-\tau)w'_t$$

$$\theta^{\mu'}_{t+1}=\tau\theta^{\mu}_{t}+(1-\tau)\theta^{\mu'}_{t}$$

其中,w_t 表示 t 时刻 Critic 在线网络的参数,w'_t和w'_{t+1}分别表示 Critic 目标网络在 t 和 t+1 时刻的神经网络参数。θ^{μ}_{t+1}表示 t 时刻 Actor 在线网络的参数,$\theta^{\mu'}_{t}$和θ^{μ}_{t+1}分别表示 Actor 目标网络在 t 和 t+1 时刻的神经网络参数。

3.2　深度强化学习结合行为树自主行为决策方法

虽然深度强化学习方法具备上述优点,但在解决无人战车自主行为决策问题时面临着收敛速度慢、奖励函数设计难度高、对训练数据和计算能力要求高等缺点,故考虑将基于规则的方法行为树与深度强化学习方法相结合,划分任务空间,提高鲁棒性。

行为树是一种遵循一定遍历顺序的决策模型,具有逻辑清晰、模块化、可扩展性好等特点[10]。如图 4 所示即为行为树的遍历方式。每个控制周期开始时,都将行为树的根节点作为起始点,自顶向下、从左到右遍历行为树的各节点。行为树节点可分为:行为节点、选择节点、顺序节点、并行节点。其中,行为节点是行为树中的叶节点,负责输出最终的行为;选择节点会在其子节点中选择一个遍历;串行节点按照从左到右的顺序遍历其全部子节点;并行节点每次同时遍历所有子节点。

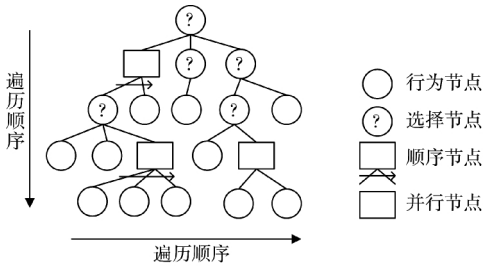

图 4　行为树遍历方法

本文提出了深度强化学习结合行为树的无人战车行为决策方法,实现基于规则方法和基于学习方法的优势互补。该决策方法将 DDPG 决策模块作为行为树的行为节点,利用行为树模型的逻辑规则和先验知识将任务划分为多个子任务,模块化地使用强化学习方法,降低状态空间和动作空间的复杂程度,加速策略收敛,降低算力要求;同时,采用 DDPG 决策模块,具备一定的泛化能力和学习能力[11]。

下页图 5 所示即为所提出的无人战车行为决策方法原理图。在每一个行为决策周期中,无人战车首先会更新自身状态,并按照自顶向下、从左到右的顺序遍历整个行为树。1)当运行到机动顺序节点时,无人战车首先从环境中获取态势信息,判断是否到达目标位置。若到达,则停止机动,若未到

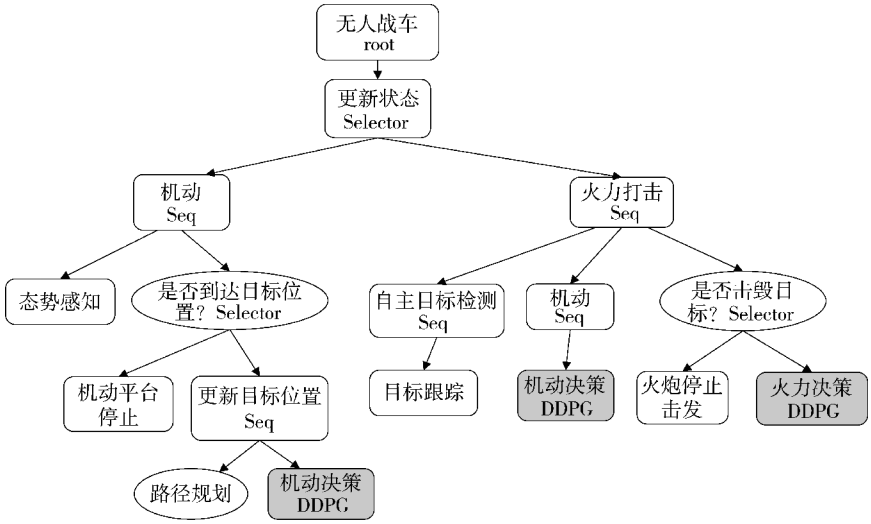

图 5　深度强化学习结合行为树的决策方法

达，则更新目标位置，顺序执行路径规划和机动决策 DDPG 节点，实现全局规划和局部避障功能。2）当运行到火力打击顺序节点时，无人战车首先从环境中获取目标信息并跟踪目标，之后进入机动决策 DDPG 节点，对敌方行为作出避险或迎敌动作。最后，判断目标是否被击毁。若尚未被击毁，则进入火力决策 DDPG 节点，自主瞄准目标，并控制火炮的高低、方位角以及开火动作。与单独使用强化学习方法完成作战过程中的自主机动和火力决策相比，该方法利用行为树机制将机动任务和火力打击任务解耦，简化状态空间与动作空间的元素组成，降低复杂程度，从而加速算法收敛，提高鲁棒性。

所提出的深度强化学习结合行为树的决策方法算法流程如下：

算法 1　深度强化学习结合行为树的决策方法

获得完整行为树，初始化深度强化学习神经网络各参数
While 任务未完成：
　　更新战车状态，遍历行为树子节点
　　if 机动节点：
　　　　顺序执行子节点
　　　　if 未到达目标位置：
　　　　　　顺序执行路径规划、DDPG 机动决策
　　if 火力打击节点：
　　　　顺序执行目标检测、机动决策自动避险
　　　　if 未摧毁目标：
　　　　　　执行 DDPG 火力决策
　　　　if 已摧毁目标：
　　　　　　break
任务完成

DDPG 机动决策节点与火力决策节点算法流程如下：

算法 2　DDPG 机动决策、火力决策算法

随机初始化 Actor 和 Critic 的在线网络参数
将在线网络参数赋值给 Actor 和 Critic 的目标网络
初始化经验回放缓存区
for episode = 1……M：
　　初始化随机探索噪声 N_t
　　接收初始战场态势 s_1
　　for t = 1……T：
　　　　根据当前策略和探索噪声，选择动作

$$a_t = \mu(s_t \mid \theta) + N_t$$

　　　　执行动作 a_t，得到奖励 r_t，更新状态 s_{t+1}
　　　　将状态转移序列(s_t, a_t, r_t, s_{t+1})存入经验回放缓存区中
　　　　从经验回放缓存区中随机采样 N 个状态转移序列

$$(s_i, a_i, r_i, s_{i+1})$$

　　　　设置 $y_t = r_t + \gamma Q'(s_{t+1}, a_{t+1})$
　　　　更新 Critic 网络，最小化 Loss 函数：

$$L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i))^2$$

　　　　更新 Actor 策略网络：

$$\nabla_\theta \mu \mid s_i = \frac{1}{N}\sum_i \nabla_a Q(s_i, \mu(s_i) \mid \theta^Q) \nabla_{\theta^\mu} \mu(s_i \mid \theta^\mu)$$

　　　　软更新目标网络参数
　　End for
End for

4　仿真验证

4.1　仿真场景搭建

本文基于 UE4 仿真环境构建了无人战车作战场景，以单对单无人战车作战想定为例，验证本文提出深度强化学习结合行为树的无人战车自主行为决策方法的有效性。敌方目标战车采用基于规则的行为决策方法，我方无人战车采用本文所提出的

深度强化学习结合行为树自主行为决策方法,完成机动和火力打击任务。

首先,以城市作战为背景,搭建模拟仿真环境。设置敌方无人战车围绕基地自主巡逻,在侦察到我方无人战车后能够及时反应,给予火力打击。我方无人战车在基地附近,与敌方无人战车开展对抗,多轮次运行来训练无人战车自主行为决策模块。然后,加快敌方无人战车反应速度,统计 10 次敌我对抗测试结果。

实验场景参数设置如表 5 所示。

表 5　实验参数设置

参数	取值
场景最大运行时间	60 s
敌方战车反应时间	5 s~10 s
敌方战车巡逻速度	10 km/h 以下
我方战车最大速度	60 km/h
训练轮次	5 000
学习率	0.01
折扣因子	0.9
经验回放缓冲区容量	10 000

根据所设置的实验场景,进行仿真验证。如图 6 所示,我方无人战车训练过程中击中敌方无人战车。

图 6　无人战车仿真场景

4.2　仿真结果分析

经过 5 000 轮训练,其累计奖励如图 7 所示。

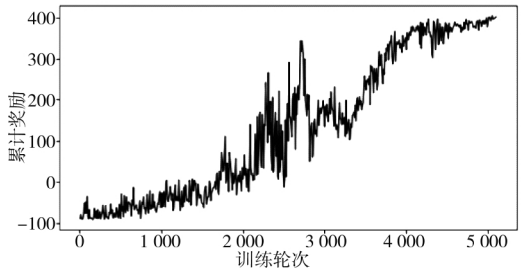

图 7　累计奖励关于训练轮次变化图

由于探索行为仍在发生,累计奖励出现波动,训练后期,无人战车自主行为决策通过策略优化迭代实现收敛。我方无人战车与加快反应速度后的敌方无人战车进行 10 次对抗测试,结果显示,我方无人战车有 9 次能够击毁敌方无人战车,由于敌方无人战车采用基于规则的方法,有一定自主能力,说明深度强化学习结合行为树的方法在本场景中优于基于规则的方法,能够有效解决无人战车自主行为决策问题。

5　结论

本文提出一种基于深度强化学习结合行为树的无人战车自主行为决策方法,解决无人战车自主行为决策问题。针对高动态强对抗环境下,无人战车完成作战任务时的序贯决策优化问题,状态空间大,策略难以稳定收敛,应用深度强化学习结合行为树方法,突破无人战车自主行为决策技术,提高无人战车的学习能力和智能化水平。通过仿真实验验证深度强化学习结合行为树的无人战车自主行为决策方法的有效性。下一步研究工作将围绕仿真环境展开,丰富作战动态场景,建立更加完善的测试机制,便于进行算法验证。

参考文献:

[1] 牛轶峰,沈林成,戴斌,等.无人作战系统发展[J].国防科技,2009,30(5):1-11.

[2] 毛宁,刘艳华,马丽媛.陆军武器火控系统的发展趋势[J].火力与指挥控制,2016,41(8):6-9.

[3] 王明,武富春,范文超,等.人工智能在装甲火力与指挥控制领域的应用[J].火力与指挥控制,2020,45(9):1-5.

[4] 李元超,毛保全,杨雨迎,等.智能辅助决策系统在武器站中的研究及应用 [J]. 兵器装备工程学报,2020,41(3):97-101.

[5] 周思羽,吴文海,张楠,等.自主空战机动决策方法综述[J].航空计算技术,2012,42(1):27-31.

[6] 孙楚,赵辉,王渊,等.基于强化学习的无人机自主机动决策方法[J].火力与指挥控制,2019,44(4):144-151.

[7] SUTTON R,BARTO A.Reinforcement learning:an introduction [M].Massachusetts:MIT Press,2018.

[8] SILVER D,LEVER G,HEESS N,et al.Deterministic policy gradient algorithms [C]//International Conference on Machine Learning,PMLR,2014:387-395.

[9] SCHAUL T,QUAN J,ANTONOGLOU I,et al.Prioritized experience replay[C]//International Conference on Machine Learning,PMLR,2014:387-395.

[10] SCHRODER J,HOFFMANN M,ZOLLNER M,et al.Behavior decision and path planning for cognitive vehicles using behavior networks [C]//2007 IEEE Intelligent Vehicles Symposium.IEEE,2007:710-715.

[11] PEREIRA,RENATODEPONTES,PAULOMARTINS ENGEL. A framework for constrained and adaptive behavior-based agents[J].Computer Science,2015,6(1):77-107.