doi:10.3772/j.issn.1002-0470.2021.04.004

基于深度强化学习的舰载机在线调度方法研究①

于彤彤② 董婷婷 肖创柏③

(北京工业大学信息学部 北京 100124)

摘 要 针对传统调度算法在高危多变环境下实现多目标在线调度所面临的困境,提出基于深度强化学习的调度优化算法,并应用于大规模舰载机出动回收多目标在线调度问题中。该方法以减小舰面位移、减少会遇次数、均衡设备利用率和稳定调度周期作为调度决策目标,依照马尔可夫决策过程(MDP),构造以舰载机和各设备状态作为输入,调度行为动作作为输出,带权特征向量作为奖赏的在线调度即时决策模型。搭建用于训练的优化深度强化学习网络,改进动作选择策略和网络结构以提升性能,从而实现在线调度决策优化。实验结果表明,利用该方法得到的决策模型能够在线解决突发状况,在静态和动态调度方面,相对于启发式算法和调度规则本文方法在安全性和高效性方面具有优势。

关键词 深度强化学习;舰载机出动回收;在线调度;多目标决策

0 引 言

航空母舰是一个国家航海作战能力的体现,而航母的作战能力则主要取决于航母上舰载机的出动能力。由于当代航母体量大、甲板环境复杂、设备操作困难和无人设备逐渐增多等现状,导致舰载机需要在具有动态不确定性的有限空间环境中出动和回收来完成作战任务[1]。在舰载机出动回收调度问题中,一个可靠的任务动作决策者需要为任务中包含的每一架舰载机规划可行的调度方案,涉及从舰载机弹射出动前的保障到着舰回收后的停机整个调度过程[2]。

目前国内外学者大多是采用传统的启发式智能算法解决舰载机调度问题[3],如遗传算法[4]、粒子群算法[5]、混合差分算法[6]等。这类算法将提前制定固定批次大小的任务放入算法模型中,通过复杂的计算得到对应该批次任务的指定调度方案。由于该类算法计算量大,时间复杂度高,大多数国内外学者都应用该类算法研究小批量舰载机群的调度问题,这在实际的作战过程中是不合理的。同时传统的智能算法在调度计算中会花费大量的时间,不具有灵活快速地解决实时状况的能力,不适合解决实时在线调度问题[7]。马尔可夫决策过程(Markov decision process, MDP)适用于构建在线调度决策模型,能够实现舰载机出动回收调度单步即时决策,利用人工智能算法中的强化学习方法训练得到决策模型。强化学习的过程是通过智能体与环境不断交互,最终得到在指定状态下采取最优动作的策略。针对传统强化学习算法无法解决大规模状态动作空间求解的问题,利用深度学习的感知能力和强化学习的决策能力的深度强化学习方法得到了广泛关注和应用[8]。

为了能够得到解决舰载机出动回收多目标在线调度问题较为安全高效的策略,本文提出将该调度问题构造为马尔科夫决策过程,利用深度强化学习算法决策。目标是实现在高峰出动情况下,缩短舰载机甲板移动距离、减少舰载机在舰面的会遇次数

① 国家自然科学基金(61701009)和教育部-中国移动科研基金(MCM20180503)资助项目。

② 女,1994 年生,硕士生;研究方向:深度强化学习;E-mail:xtina988@hotmail.com/tongtong.yu@ia.ac.cn

③ 通信作者,E-mail:cbxiao@bjut.edu.cn

(收稿日期:2020-04-10)

并稳定调度周期,同时均衡各设备的利用率。本文利用马尔可夫决策过程构造舰载机出动回收调度模型,利用优化后的深度 Q 学习(deep Q-learning,DQN)算法对模型进行深度强化学习训练。构造在线环境进行仿真实验,本文方法与启发式算法和调度规则相比,得到的结果具有更好的高效性和安全性。

1 舰载机出动回收调度模型

1.1 舰载机甲板环境建模

舰载机甲板空间狭窄,机位和固定设备分布密集,并承载着危险系数较高的移动设备和舰载机。因此,为了满足实际需求,本文首先对甲板环境空间进行合理建模。舰载机在出动前必须进行常规的保障任务,目前比较先进的是一站式保障,即每个机位都在所需的各种保障资源覆盖范围内,只需要保障组移动即可,这种保障方式大大降低了事故风险系数[9]。除了一站式保障位,甲板上还包括起飞位、着舰跑道、升降机以及移动的保障组、用于牵引舰载机在甲板上移动的牵引车等移动设备[10]。本文对某航母甲板上的固定设备为 18 个保障机位、4 个起飞位、2 个升降机和 1 个跑道进行编号,构建简化的甲板环境如图 1 所示。

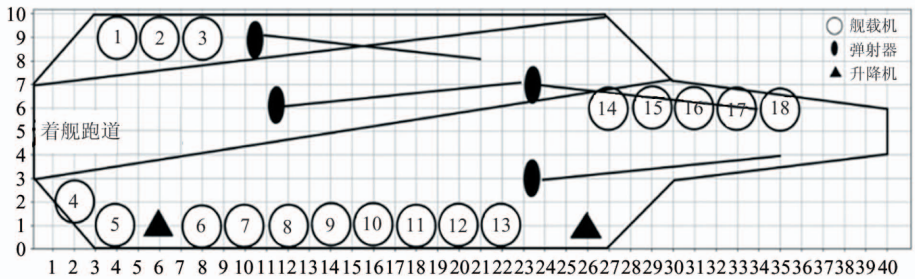


图 1 甲板环境仿真图

1.2 舰载机出动回收过程

舰载机按照批次编队出动回收,根据任务不同,每次安排同一型号的若干舰载机出动到空中完成任务[11]。某舰载机在甲板持续出动的流程如图 2 所示,其中动作执行和转移的时间遵循正态分布[12]。

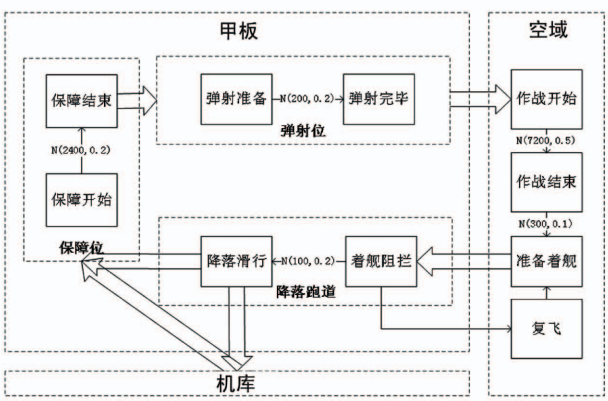


图 2 舰载机出动回收流程图

1.3 不确定性因素及调度目标构建

舰载机出动回收过程中,会产生很多不确定性因素。例如起飞时,弹射器可能由于某些机械性原因无法使用,导致舰载机无法按计划起飞;由于保障组人员车辆需要在狭窄的甲板上快速移动,这个过程中的不确定性也会导致舰载机保障时间受到影响;着舰时,阻拦装置出现状况会导致舰载机着舰失败,着舰跑道可能会被舰面移动的舰载机或保障组占用暂时无法使用;作战过程中,舰载机可能会被敌方击中导致损坏或坠亡等[13]。

因此,调度规划必须在规定的任务时间内完成并且能够在遇到突发状况时及时做出合理的调度决策。本研究通过衡量舰面位移、舰面会遇次数和各设备利用次数标准差来衡量舰载机的出动回收能力。设计的多目标函数 f 如下:

$$f = \min \sum_{k \in P} Dis_k + \min \sum_{i \in C} Con_i + \min \sum_{j \in B} Balan_j \tag{1}$$

$$Balan_j = \sigma(B\_tn)_j + \sigma(T\_tn)_j \tag{2}$$

其中,$\sum_{k \in P} Dis_k$ 为所有舰载机在甲板移动的总位移,$\sum_{i \in C} Con_i$ 为舰载机在舰面移动时产生会遇状态的次

数，$\sum_{j\in B} Balan_j$ 为各设备的利用次数的标准差，$\sigma(B\_tn)$ 为各保障机位资源利用量的标准差，$\sigma(T\_tn)$ 为各弹射器利用量的标准差。

1.4 马尔可夫决策过程模型构建

马尔可夫决策过程提供了一个用于对决策情景建模的数学框架，在利用强化学习解决舰载机出动调度的问题上，首先需要将问题转化为有限状态下的 MDP。MDP 由一个五元组（S, A, T, γ, R）表示，其中 S 代表状态集，A 代表动作集，T 代表状态转移，$\gamma \in (0,1)$ 代表折扣函数，R 代表奖赏函数，以下对 MDP 过程中的关键要素进行阐述。

1.4.1 环境和智能体

通过对本文问题背景的分析，可以确定模型的环境类型。由于舰载机出动回收调度不能只根据当前状态来确定结果，因此属于随机性环境；调度的过程是在任何时候都可以被确定的，因此属于完全可观测环境；状态转移的行为状态是有限个数的，因此归属于离散环境；由于执行的所有行为都是相关的，当前的行为会影响到今后的行为，因此属于非情景环境；本文选用单智能体环境，即环境中只有一个智能体，智能体确定为调度的决策者。

1.4.2 状态集

以调度决策者作为智能体，依照马尔可夫决策过程构建状态空间如式(3)所示。

$$s\_space = \left\{ \begin{matrix} 'E_0':op,\cdots,'E_m':op, \\ 'A_0':[op, fuel, pos], \\ \cdots,'A_n':[op, fuel, pos] \end{matrix} \right\} \tag{3}$$

设备组为 $[E_0, E_1, \cdots, E_m]$，包括保障区域保障设备以及弹射器；舰载机群表示为 $[A_0, A_1, \cdots, A_n]$；设备及舰载机可用状态用 op 表示，用布尔值表示可用或不可用；舰载机所剩油量 $fuel$ 可以离散的划分为 0 ~ 3 四个等级；舰载机所在位置集合可以表示为 $loc = [pos_0, pos_1, \cdots, pos_l]$，包括停机位、位置不同并互相影响的弹射起飞位、着舰跑道、着舰渐进航线、作战的空域和机库。

由此可以得到 n 架舰载机在承载 m 个设备的甲板上完成调度任务的解空间的大小如式(4)所示。

$$(2op)^m \times (4fuel \times 2op)^n \frac{loc!}{n!} \tag{4}$$

当 $m = 21$，$n = 30$，$loc = 25$ 时，可以算出状态解空间达到 10^{27}，因此如果使用人工方法或传统智能算法都将面临着巨大的计算量。

1.4.3 动作集

目前部分学者在应用强化学习解决调度算法的研究中，为了缩小动作空间，将执行的动作设置为启发式算法或分配规则[14]，这样的做法大大限制了强化学习探索最优策略的能力。本文中将动作定义为调度直接执行的动作，即某一舰载机移动到某一位置，动作空间表示如式(5)所示。

$$a\_space = [[A_0, P_0],[A_0, P_1],\cdots,[A_{n-1}, P_{m-1}]] \tag{5}$$

其中，$A = [0,1,2,\cdots,n-1]$ 为舰载机集合，$P = [0,1,2,\cdots,m-1]$ 为位置集合，动作空间矩阵大小为 $m \times n$。

1.4.4 状态转移

状态转移是指智能体从当前状态执行动作转移到下一状态的过程。本问题考虑如下约束[15]。

$$D(A_i) = (F_i = 0) \wedge (P_i \notin S) \tag{6}$$

$$\forall (ET_{T_j} < ST_{T_{j+1}}) \tag{7}$$

$$\forall (T_{ij} \cap T_{kj} = \varphi) \tag{8}$$

$$(T_L \cap T_{T3}) \wedge (T_L \cap T_{T4}) \wedge (T_{T3} \cap T_{T4}) = \varphi \tag{9}$$

$$\forall (ST_{P_j} < ST_{P_{j+1}}) \tag{10}$$

其中，式(6)为油量约束，A_i 表示舰载机，S 表示保障机位集合，F_i 表示油量等级，P_i 表示舰载机当前位置，舰载机油量等级为 0 时只能位于保障机位上，否则该舰载机坠亡；式(7)为技术约束，T 是按照执行顺序排列的任务集合，ST 是任务开始时间，ET 是任务结束时间，舰载机任务必须按照“保障→弹射→作战→渐进航线→着舰→保障/停机/维修”的顺序执行，且任务时间戳不能重叠；式(8)为互斥约束，T_{ij} 是舰载机 i 在停机位 j 的时间戳，T_{kj} 是舰载机 k 在停机位 j 的时间戳，对于任意的两个舰载机都不能共用停机位，即各舰载机在同一停机位的时间戳不能重合；式(9)为空间约束，$T_L/T_{T3}/T_{T4}$ 分别是着舰跑道和 3 号、4 号弹射器被使用的时间戳集合，着舰跑道和 3 号、4 号弹射器不能同时使用；式(10)为任务约束，P 是任务优先级集合，舰载机舰面会遇

时按照当前舰载机任务优先级决定通过顺序。

1.4.5　奖赏函数

由于本文解决的是一个多目标问题，目标函数受多个因素影响，无法用强化学习常见的单一值形式来表示奖赏函数，因此本文设计线性化奖赏函数 $R(s)$，定义如下：

$$R(s) = \boldsymbol{\varpi}^{\mathrm{T}}\boldsymbol{\phi}(s) \tag{11}$$

其中，$\boldsymbol{\varpi}^{\mathrm{T}}$ 为权重向量 $\boldsymbol{\varpi}$ 的转置，$\boldsymbol{\phi}(s)$ 为状态 S 下的特征向量。特征向量中的每一个特征值都影响调度算法获得最优的策略。其中权重向量的值由用户定义，状态特征向量定义为以下 7 个特征值：1 号特征值为濒临坠亡的舰载机数量，2 号特征值为坠毁的舰载机数量，3 号特征值为完成目标任务的舰载机数量，4 号特征值为舰面冲突次数，5 号特征值为成功弹射起飞的舰载机数量，6 号特征值为舰载机舰面位移值，7 号特征值为各保障组工作频次的标准差与弹射器工作频次的标准差之和。

1.5　舰载机出动回收持续在线调度

为了实现持续出动回收状态下的在线调度，本研究将在线调度模型设计为随时间变化的出动回收模型。在态势更新方面，除利用马尔科夫模型持续选择动作改变智能体状态外，还增加了一个控制任务进程的计时器参与状态更新。计时器通过推进当前任务序列中各任务完成的进度更新状态，计时器在指定时间间隔后触发更新指令，对任务序列中执行任务的时间进行更新。当任务完成时对当前状态进行更新并将当前任务从任务序列中删除，当执行新的有效动作时将形成的新任务加入任务序列中。这样的操作能够实现舰载机持续出动回收，同时连续进行动作选择将缩短调度时间并提升调度效率。持续出动回收的在线调度机制如图 3 所示。

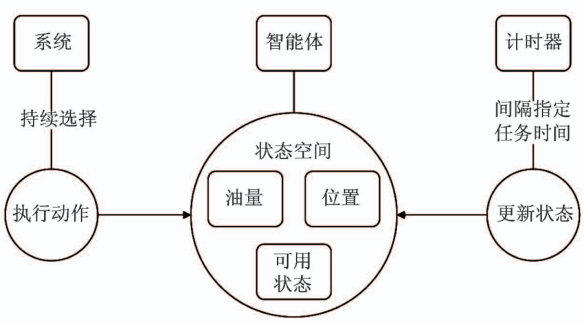


图 3　持续出动回收的在线调度机制

由于深度强化学习神经网络计算的时间相对较长，会影响任务执行时间，因此在推进任务进程计算时间时需要去掉系统运行时间。各个任务的执行时间以实际时间为基础等比缩小，并服从正态分布，各任务执行时间如表 1 所示。

表 1　任务执行所用时间表

任务	时间/s
甲板移动	N(6, 0.05) × 移动位移
保障任务	N(2400, 0.2)
弹射任务	N(200, 0.2)
作战任务	N(7200, 0.4)
回收任务Ⅰ(渐近线)	N(300, 0.1)
回收任务Ⅱ(降落跑道)	N(100, 0.1)
回收任务Ⅲ(返回空域)	N(300, 0.1)

在线调度过程中，舰载机在甲板移动过程的会遇情况会严重影响调度决策，因此需要持续对每一个舰载机的位置追踪，并监测舰载机间的位移是否小于安全位移。因此在计时器每次触发更新指令时，除了要做任务进度的更新，还要对舰载机的位置进行更新。如果某架舰载机与其他的一架或几架舰载机的距离小于安全距离，就要根据舰载机任务的优先级选择通过的舰载机，另外的舰载机原地等待，进度不更新。任务优先级如表 2 所示。

表 2　甲板任务优先级

甲板任务	位置变化	优先级
特殊任务		Ⅰ
着舰任务	渐近线→着舰跑道	Ⅱ
起飞任务	保障机位→起飞位	Ⅲ
保障任务	任意位置→保障机位	Ⅳ
维修任务	任意位置→升降梯	Ⅴ

2　基于深度强化学习的舰载机在线调度决策

2.1　算法原理

本文讨论的舰载机出动回收在线调度问题的状态解空间过大，因此应用传统的强化学习算法将会花费大量的时间。深度神经网络是有效处理大量数

据的工具,因此本文研究以深度强化学习中的 DQN 算法为基础,通过优化的 DQN 算法来解决大规模舰载机出动回收多目标在线调度问题。DQN 算法流程如图 4 所示。

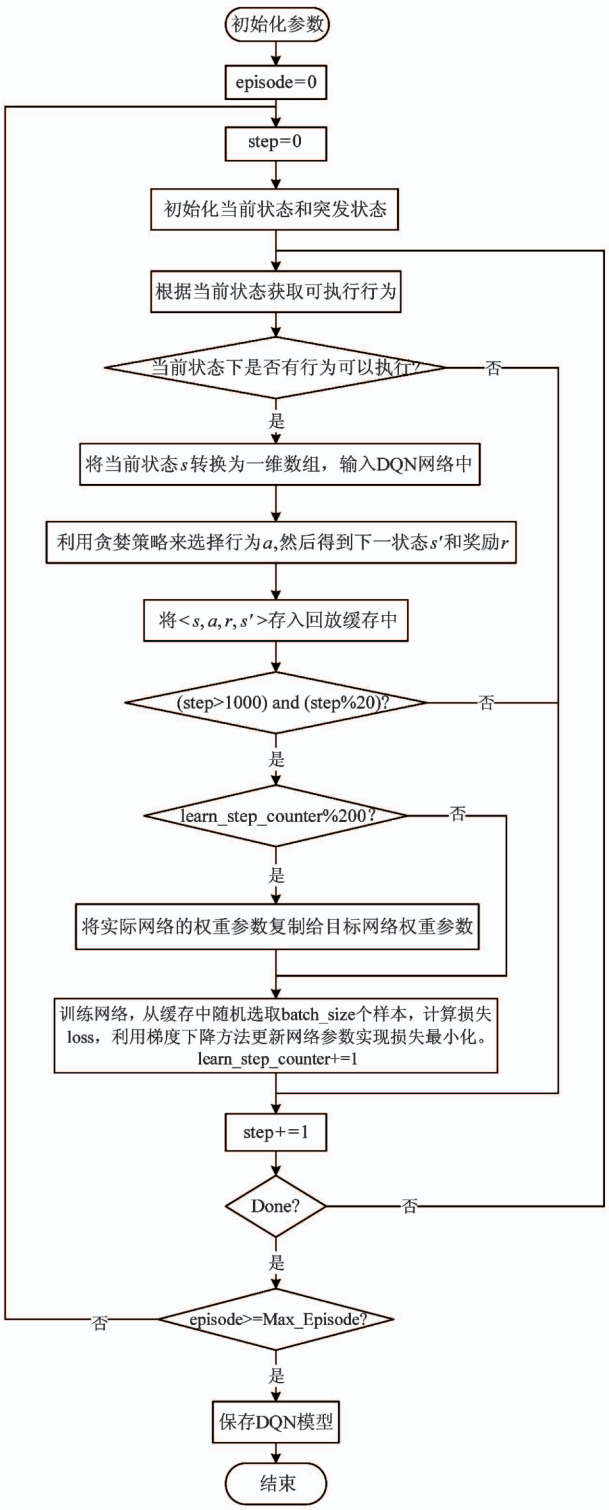


图 4　DQN 算法流程图

强化学习通过不断选择动作根据反馈更新策略得到最优决策模型。其中 Q 值就是评估模型在某一状态下某行为的最优值。Q-learning 算法通过建立一个 Q 表来存储所有可能的“状态-行为”对的 Q 值,通过不断选择动作更新 Q 值最终使算法结果收敛,得到最优策略。在 Q-learning 中 Q 值的更新公式如式(12)所示。

$$Q(s,a) = Q(s,a) + \alpha(r + \max Q(s') - Q(s,a)) \tag{12}$$

式中,α 为学习率,$r + \max Q(s')$ 为目标值,$Q(s,a)$ 为预测值。在深度强化学习中,利用一个权重为 θ 的神经网络近似每个状态下发生动作的 Q 值,即有 $Q(s,a;\theta) \approx Q^*(s,a)$。通过梯度下降的方法更新神经网络参数权重 θ 最小化损失,使神经网络预测的 Q 值逐渐逼近目标值。由于直接利用深度神经网络逼近 Q 值的算法常常会不稳定且难以收敛,因此 DQN 算法针对这一问题进行了两方面的优化。首先,建立两个结构相同的神经网络,用于预测 Q 值的实际网络 θ 使通过梯度下降来学习正确的权重,而用于计算目标 Q 值的目标网络 θ' 在滞后若干时间后再通过复制实际 Q 网络中的参数进行更新。定义目标值和预测值的均方差作为神经网络的损失函数如式(13)所示。

$$loss = ((r + \gamma \max Q(s';\theta')) - Q(s,a;\theta))^2 \tag{13}$$

除此之外,DQN 算法建立了用于存储之前经验 $< s, a, r, s' >$ 的经验回放记忆库,由于 Q-learning 是 off-policy 学习的方法,既可以学习当前的经历也可以学习过去的经验,因此在 DQN 神经网络更新的时候可以从记忆库中随机抽取指定大小批次的经验进行学习,这样的做法打乱了经验之间的相关性,同时避免神经网络过拟合问题。DQN 神经网络更新流程如图 5 所示。

2.2　算法优化

2.2.1　探索与利用

在强化学习的过程中,每次会选择利用 Q 值最大的动作,这就是利用贪婪策略执行动作选择。但在强化学习的最初阶段智能体并不能掌握 Q 值,因此需要通过随机的方式来探索选择未知的动作,在

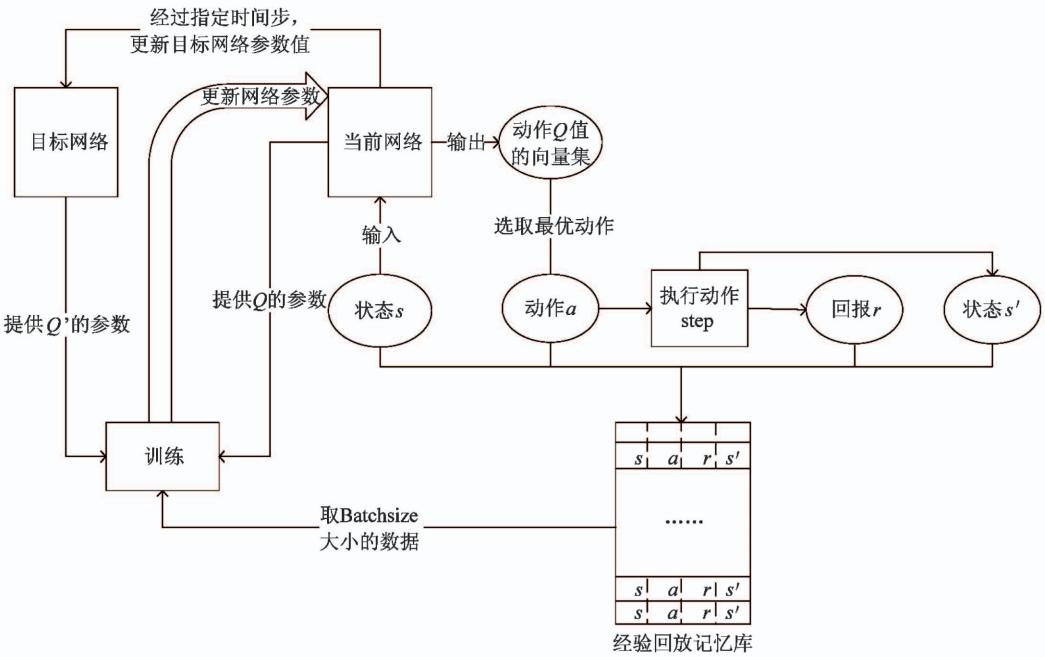


图 5　DQN 网络更新流程图

经历一段时间的学习后，就能够获取一定量的 Q 值。但此时是继续探索未知动作还是利用当前 Q 值最大的动作，这就是强化学习面临的探索利用平衡问题。为了解决这个问题，本文在 ε-贪婪策略的基础上，设计了可变 ε 贪婪策略方法来选择动作，如图 6 所示。初始设置一个非 0 概率 ε，在概率 ε 下，随机探索不同动作，并以概率 $1-\varepsilon$ 选择具有最大 Q 值的行为。随着时间推移，智能体的学习能力越来越强，更新得到的状态动作值越来越好，此时逐渐降低 ε 的值，更多地利用学到的 Q 值选择最优行为。

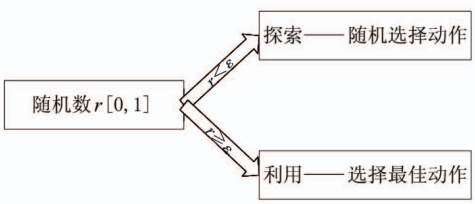


图 6　可变 ε-贪婪策略

2.2.2　动作集约束

由于互斥约束的限制（式（8）），各舰载机在同一停机位的时间戳不能重合，同一舰载机任务的时间戳也不能重合。因此在指定状态下智能体只能执行动作集中的部分动作。神经网络输出动作 Q 值集合时没有考虑当前动作是否能够执行，当利用学习到的策略进行动作选择时，很可能选择到一个根本无法执行的动作。这样会降低强化学习的学习效率，导致神经网络难以收敛。本文设计了 Action-Mask 机制对动作空间进行处理，在每次执行动作选择前，先判断动作空间中的每一个动作涉及的舰载机或机位是否可用，若有一方不可用那么这个动作就无法执行，无法执行的动作需要进行处理，将无法选择的动作的值加上负无穷大，将其值降为最小，相当于给该动作加上一个面罩，避免该动作被选择。

2.2.3　神经网络收敛速度优化

强化学习是一个不断试错的过程，解决本文问题需要巨大的数据量，因此在学习过程中用来训练的同一批数据之间差距有时会较大，导致在训练过程中容易出现梯度爆炸的情况。解决梯度爆炸的方法有很多，最简单的方式是降低学习率。但是强化学习训练的不确定性和随机性导致学习率只有降低到 1×10^{-7} 甚至更低的时候才能避免梯度爆炸，这种级别的学习率后会导致神经网络需要很长的训练时间才能收敛。本文用一种叫做批归一化（batch-normalization，BN）的方法对数据进行归一化，通过对每一批输入数据归一化，使数据间差距减小。神经网络结构如图 7 所示。

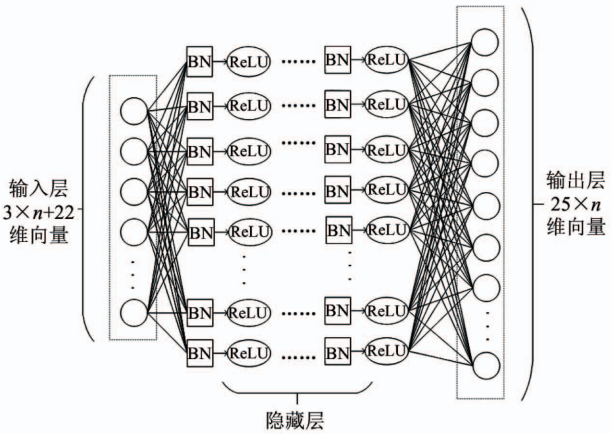


图 7　神经网络结构

3　仿真实验

3.1　实验环境及参数设置

本文实验提出将深度强化学习模型应用在不同规模舰载机调度问题上,并验证在线调度方法效果。实验将舰载机规模分为 3 类,舰载机数量为 n = [10, 20, 30],保障机位设置为 18 个,标号为[0, 1, 2,…,17]。本研究中采用双周期空中作战,其中一个周期的时间设置为 3600 s,在模型训练阶段将一次任务设计为 4 个周期。为了能够在短时间内出动更多架次的舰载机执行任务,本文中实验环境考虑为 3 个编队的舰载机同时进行任务调度,并分批次出动回收。初始状态下,各组舰载机分别位于保障位、渐近线和空中,并根据表 3 设置对应油量等级。一轮任务完成的标志为所有舰载机都经历至少一次“出动-回收”过程,同时为了满足在线调度的突发性,设置随机时间区间内会产生一种随机的突发情况。突发情况分为以下 4 种,即固定设备损坏、移动保障组和资源延迟、舰载机损坏和阻拦着舰失败。

表 3　舰载机初始位置对应油量

舰载机所处位置	油量等级
保障机位/甲板	3
空域	2
渐近线	1

应用 Python 3.7 在 JetBrains Pycharm 平台上调用 TensorFlow 库编码实现,运行于 3.60 GHz Intel i7 处理器 Windows x64 系统。编写舰载机甲板出动回收环境类,包括构造、重置、执行等方法仿真调度过程,利用 TensorFlow 构造深度神经网络进行学习训练。

深度强化学习神经网络部分超参数的设置是根据一般性原则,经过多次实验调整选取最优结果。由于目标是使最终总奖赏最大,因此折扣因子 γ 越大越好,本实验中设置 γ = 0.97;用可变的 ε-贪婪策略实现探索利用平衡,在训练的初始阶段充分探索策略,最终阶段选取学习到的最优策略,因此初始设置 ε = 1,并以 0.9996 的折扣率衰减至 ε = 0.1;将学习率设置为 α = 0.001,记忆库最大容量为 N = 20 000,采样批次大小为 batch _ size = 32。神经网络隐藏层神经元个数设为 400,参数初始化采用随机策略。总训练次数为 2500 次。

3.2　结果分析与讨论

下面将从静态离线调度和动态在线调度两个方面对算法效果进行分析评估。

3.2.1　静态离线调度分析

(1) 算法收敛性分析。设计 3 种规模的舰载机群分别利用深度强化学习模型训练,得到决策模型的误差曲线如图 8 所示。3 种规模均能通过深度强化学习模型近似收敛,认为学到了好的策略。

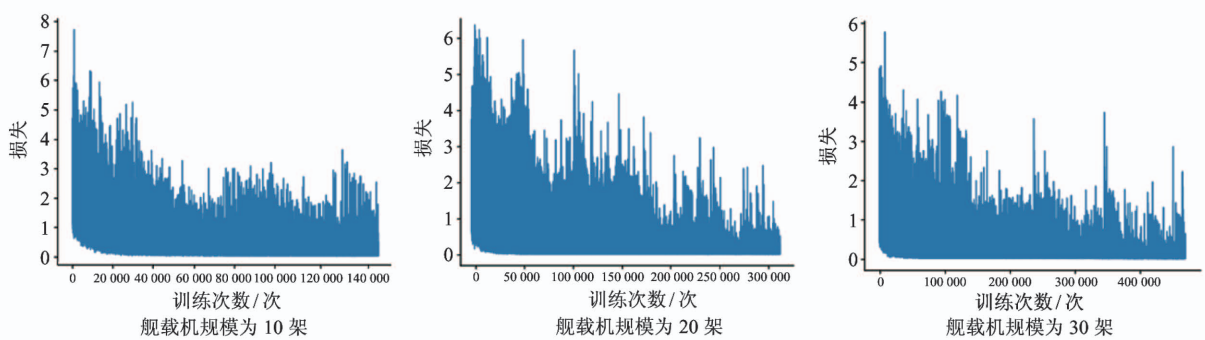


图 8　误差 loss 曲线图

对比总奖赏函数的变化曲线如图 9 所示。可以看出规模越小的舰载机规模学习奖赏函数增长的趋势越大，说明该深度强化学习算法对于小规模舰载机调度问题处理的效果比大规模要更好。

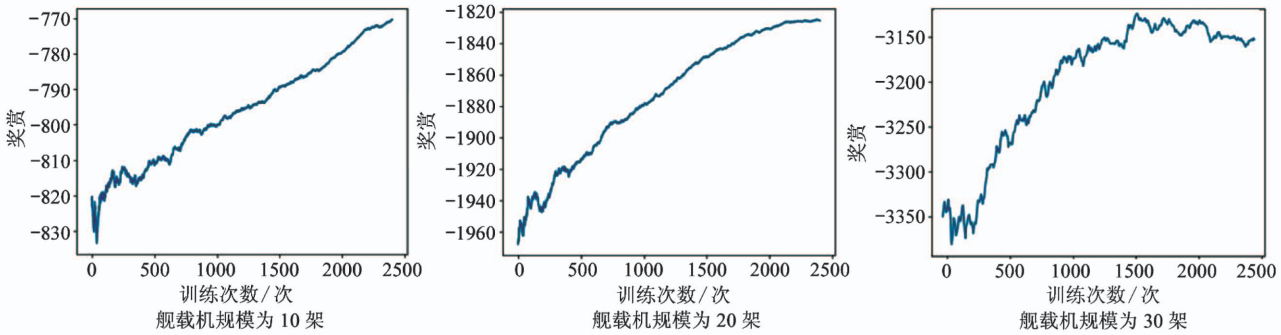


图 9　奖赏 reward 曲线图

（2）执行静态离线调度并与遗传算法对比分析。

以遗传算法（genetic algorithm，GA）为代表的传统启发式算法在解决调度问题时，是将提前设置好的任务固定输入模型，经过不断迭代计算，得到最优策略结果。该类算法在动态在线调度问题中无法实现满足舰载机调度的即时性和高效性，但作为传统普适的算法，该类算法在静态离线调度问题中能够求得较优解，因此将深度强化学习算法在静态离线调度的结果与遗传算法进行对比。

$$A_{mn}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{14}$$

遗传算法模型的染色体编码 A_{mn} 如式（14）所示，m 为舰载机数量，n 为保障机位及弹射起飞位数量，a_{ij} 为 0 或 1 的值，表示舰载机在是否在该位置上完成任务。由于舰载机是编队执行任务，一批舰载机在某一时间段会集中完成某种类型的任务，因此舰载机执行任务的顺序按照先入先出（first in first out，FIFO）规则执行。遗传算法的适应度函数是最小化舰面移动位移，交叉概率设置为 $P_C=0.6$，变异概率设置为 $P_m=0.06$，初始种群规模设置为 $S_Z=60$，经过多次实验证明最大迭代次数 $Gen\_max=1000$ 就能够找到最优策略。本实验使用规模为 30 架的舰载机进行实验。随机初始化各机群舰载机数量，舰载机所在位置和油量，生成 20 个初始样本，分别应用遗传算法 GA 计算 20 次，并应用深度强化学习 DQN 算法训练模型 2500 次后对 20 个初始样本执行调度决策，应用先入先出 FIFO 调度规则执行 20 个样本的调度。调度任务执行平均总时长结果对比如图 10 所示。

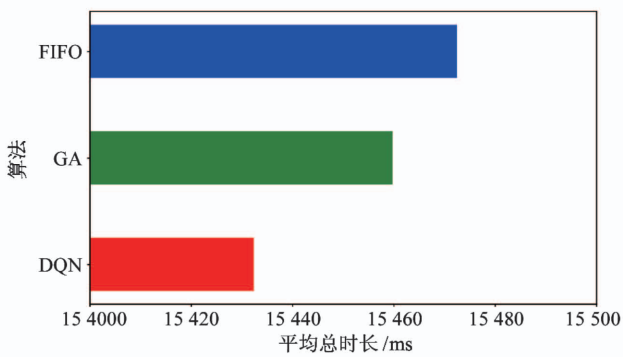


图 10　任务执行平均总时长结果对比

由于遗传算法只能进行离线调度决策，因此只能考虑舰面移动位移。若考虑更多约束，如动态情况下舰面会遇影响调度时间的问题，则会使计算时间过长。由上图能够看出，深度强化学习算法在任务完成时间优化上有突出显著的优势。同时，遗传算法在完成决策任务时，需要输入任务统一计算得到完整的调度决策结果后才能执行，而深度强化学习算法载入训练好的调度模型后，只需要输入当前状态就可以得到当前所需要执行的动作，在动作执行的过程中再计算下一动作，节省了大量的时间。计算 20 个任务的平均决策净消耗时间对比如表 4 所示，明显看出遗传算法决策任务所消耗的时间是深度强化学习算法近 100 倍。因此遗传算法解决的

调度问题大多是不紧急并任务较为固定的问题,而并不适用于舰载机出动回收调度这类紧急多变的问题,而深度强化学习在该类调度问题中应用的优势十分显著。

表 4　决策净消耗时间对比

决策算法	时间/s
GA	15.636
DQN	0.175
FIFO	0.028

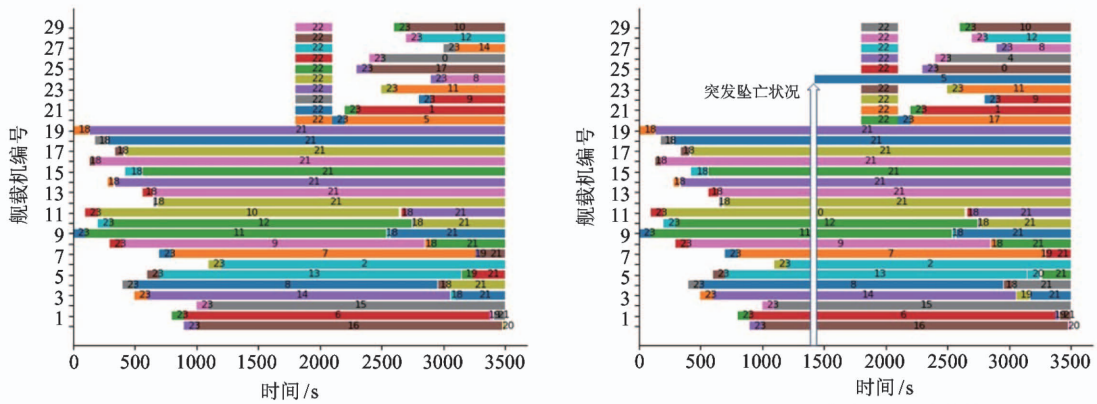
遗传算法模型只能根据指定任务求得对应决策结果,每次任务更新都需要重新计算决策,而深度强化学习算法只需要离线训练一次,模型就可以针对不同任务进行快速决策。在舰载机出动回收调度问题中,任务类型由多种因素影响,若任务中关键变量改变,遗传算法和深度强化学习算法得到的模型是否需要重新计算模型的对比如表 5 所示。由于每次更换任务,遗传算法都要经过长时间的计算,说明该种算法灵活性较差,难以迁移并广泛应用。

表 5　舰载机出动回收任务的变量影响

变量	GA 模型重新计算	DQN 模型重新计算
甲板类型	是	是
舰载机总数	是	是
舰载机各编队数量	是	否
各舰载机油量	是	否
各舰载机初始位置	是	否
任务执行周期	是	否
任务总时长	是	否

3.2.2　动态在线调度

将学到的规模为 30 架舰载机的调度模型载入,测试在面临突发状况时模型的处理能力。无突发状况和遇到突发状况的调度甘特图如图 11 所示。



(a) 无突发状况调度甘特图　　(b) 有突发状况调度甘特图

图 11　调度甘特图

图中所示的实验突发状况设计为在任务进行到第 1420 s 时 24 号舰载机坠亡,此时调度决策执行从机库调度一架舰载机并重新编号为 24 号的动作,此舰载机移动至保障位开始保障,等待出动。由调度甘特图可以看出,后续调度并没有受到突发状况的影响,可以按时完成任务,因此该算法模型能够快速高效地解决在线调度问题。

本文方法与调度规则的对比分析如下。在调度规模为 30 架舰载机的情况下,分别应用深度强化学习 DQN 算法和调度规则对 20 个不同初始环境进行测试,采用的调度规则见表 6。

表 6　调度规则

算法	描述
SJF	短时优先算法,选择任务所用时长最短的行为。当多个最短时长任务时,利用均匀随机方法选择行为
FIFO	先入先出算法,按照进入等待队列的顺序依次执行行为
MOPNR	剩余最多优先算法,选择剩余任务最多的舰载机执行任务。当多个舰载机剩余任务最多时,利用均匀随机方法选择行为

各方法得到的舰面移动位移、舰面会遇次数、设备利用率标准差和调度任务完成总时间的箱线图如图 12 所示。

可以看出，DQN 算法在实验中得到的舰面冲突次数结果与 SJF 方法接近，都优于其他两种方法，但是 DQN 算法在某些初始状态下能够得到更少的会遇次数；在舰面总位移的对比中，能够看到 SJF 算法的结果最好且最稳定，由于该方法就是要选择最短时间也就是移动位移最短的动作，而 DQN 得到的结果也比较好，并能够稳定在位移总值较低的区域；DQN 方法在设备利用次数衡量上优于 SJF 和 FIFO，说明各个设备的使用较为均衡，MOPNR 方法由于采取了随机均匀分布，所以各设备使用率的标准差较低；任务完成总时间上 4 种方法的平均值接近，但是 SJF、FIIFO 和 MOPNR 都出现了某些实验完成时间过长，是由于这 3 种算法对于突发状况没有较快的反应和处理能力，影响了后续任务的执行，这里再次体现了 DQN 算法在解决在线调度上的优势。

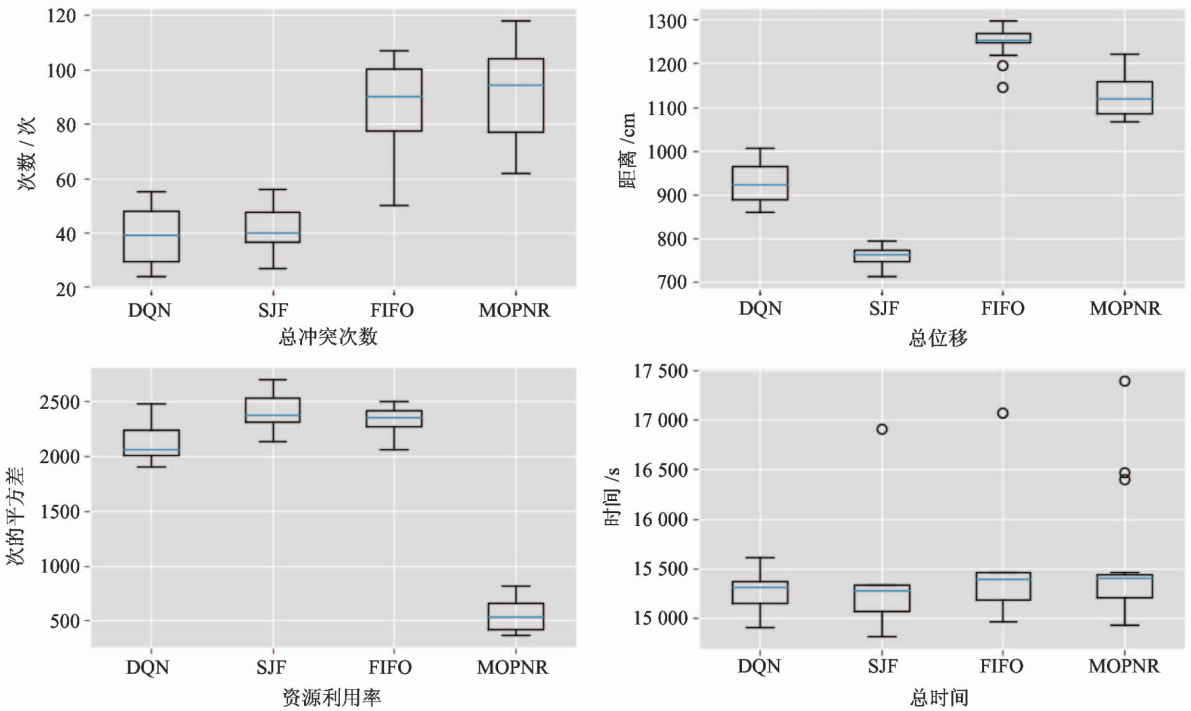


图 12　算法目标结果对比箱线图

4　结　论

本文将舰载机出动回收过程抽象为马尔可夫决策过程，模型的状态设置为全部舰载机和设备的当前状态，动作设置为某舰载机移动至某位置，建立带权重的特征向量作为奖赏，构造模型与实际调度过程更贴近，更利于投入实际应用。强化学习模型和深度学习网络的参数能够根据需求直接进行修改，可以广泛应用于各种规模和约束下的舰载机出动回收调度问题，体现了方法的灵活性。利用该模型进行在线调度能够极大缩短舰载机在线调度应急处理时间，做到及时高效处理突发状况，体现了方法的实时性。通过优化 DQN 算法中神经网络的结构和强化学习的动作选择方法，使算法在舰载机出动回收调度问题的学习上能够近似收敛，并且在优化多目标函数时能得到较好的效果，获得安全高效的调度策略。

本文研究是深度强化学习在舰载机出动回收调度领域的大胆探索，能够为航母作战指挥人员规划调度方案提供快速高效的决策支持，在人工智能理论不断完善和应用的背景下，该算法具有很好的发展前景。未来在马尔可夫决策过程模型构建时，可以考虑将状态空间进一步扩大，使智能体能够对当前状态有更全面的了解，有助于学习效果的提升。

参考文献

[1] Qi C, Wang D. Dynamic aircraft carrier flight deck task planning based on HTN[J]. *IFAC-Papers OnLine*, 2016, 49(12):1608-1613

[2] 刘翱,刘克. 舰载机保障作业调度问题研究进展[J]. 系统工程理论与实践, 2017, 37(1): 49-60

[3] Ryan J, Cummings M, Roy N, et al. Designing an interactive local and global decision support system for aircraft carrier deck scheduling[C]// AIAA Infotechat Aerospace, Conference and Exhibit, Louis, USA, 2011: 1337-1348

[4] 杨炳恒, 毕玉泉, 张彪, 等. 航母多机出动甲板作业流程研究[J]. 舰船电子工程, 2016,36(8): 150-152

[5] 司维超, 韩维, 宋岩,等. 基于多种群协作混沌智能算法的舰载机出动调度[J]. 计算机应用研究, 2013,30(2): 454-457

[6] 苏析超, 李聪颖, 陈志刚. 混合差分进化算法在舰载机出动调度中的应用[J]. 计算机仿真, 2015,32(4): 74-78

[7] Xanthopoulos A S, Koulouriotis D E, Tourassis V D, et al. Intelligent controllers for bi-objective dynamic scheduling on a single machine with sequence-dependent setups [J]. *Applied Soft Computing*, 2013,13(12): 4704-4717

[8] Mnih V, Kavukcuo K, Silver D, et al. Human-level control through deep reinforcement learning[J]. *Nature*, 2015, 518(7540): 529-533

[9] 刘相春. 美国"福特"级航母"一站式保障"技术特征和关键技术分析[J]. 中国舰船研究, 2013, 8(6): 1-5

[10] Michini B, How J. A human-interactive course of action planner for aircraft carrier deck operations[C]// AIAA Infotechat Aerospace, Conference and Exhibit, Louis, USA, 2011: 1326-1336

[11] 刘相春, 卢晶, 黄祥钊. 国外航母舰载机出动回收能力指标体系分析[J]. 中国舰船研究, 2011, 6(4): 1-7

[12] Ryan J C, Cummings M L. A systems analysis of the introduction of unmanned aircraft into aircraft carrier operations[J]. *IEEE Transactions on Human-Machine Systems*, 2016, 46(2): 209-220

[13] Ryan J C, Banerjee A G. Comparing the performance of expert user heuristics and an integer linear program in aircraft carrier deck operations[J]. *IEEE Transactions on Cybernetics*, 2014, 44(6): 761-773

[14] 肖鹏飞, 张超勇, 孟磊磊, 等. 基于深度强化学习的非置换流水车间调度问题[J]. 计算机集成制造系统, 2021, 27(1): 192-205

[15] 冯强, 曾声奎, 康锐. 不确定条件下舰载机动态调度仿真与优化方法[J]. 系统仿真学报, 2011, 23(7): 1497-1501, 1506

Research on carrier aircraft online scheduling method based on deep reinforcement learning

Yu Tongtong, Dong Tingting, Xiao Chuangbai

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124)

Abstract

Aiming at the predicament faced by traditional scheduling algorithms to achieve multi-objective online scheduling in high-risk and variable environments, a scheduling algorithm based on deep reinforcement learning is proposed. This method applies to the multi-objective online scheduling problem of large-scale carrier aircraft dispatch and recovery. This approach takes the reduction of deck displacement and the number of meets, the stability of the scheduling cycle, the balance of equipment utilization as the scheduling multi-objective. According to the Markov decision process (MDP), the model is built, which uses the aircraft and equipment's states as inputs, the behavior value function as the outputs and weighted feature vectors as the reward. Then, an optimized deep reinforcement learning network is built for training, improving action selection strategies and neural network's training performance. Experiment results show that the decision model obtained can be processed immediately with the emergency, while compared with the scheduling rules and heuristic algorithm for scheduling multi-objective decision-making problems, the method has significant advantages on security and flexibility.

Key words: deep reinforcement learning, aircraft dispatch and recovery, online scheduling, multi-objective decision-making