doi:10.11887/j.cn.202105015 http://journal.nudt.edu.cn

强化学习在多阶段装备组合规划问题中的应用*

张骁雄[1],丁 松[2],李明浩[3],丁 鲲[1],王 龙[1],义余江[4]
(1. 国防科技大学 第六十三研究所,江苏 南京 210007;2. 浙江财经大学 经济学院,浙江 杭州 310018;
3. 国防科技大学 系统工程学院,湖南 长沙 410073;4. 西南电子电信技术研究所,四川 成都 610041)

摘 要:针对多阶段武器装备组合规划中的选择难、规划难问题,提出基于多目标优化算法以及强化学习技术的混合优化方法。在各个阶段以装备组合效能最大和成本最小为准则,构建单阶段多目标优化模型,并设计基于非支配排序遗传算法的求解算法以生成各阶段的Pareto解,在此基础上建立多阶段的组合优化模型。通过强化学习的Q-Learning方法,在各阶段的Pareto解中采用探索或者利用两种模式,生成各阶段的装备组合,并指导下一阶段的装备选型,从而生成整个周期内的规划方案。通过对比实验分析,验证了所提模型和算法的有效性,能够为多阶段武器装备组合规划提供辅助决策。

关键词:武器装备;组合规划;非支配排序遗传算法;强化学习;Q-Learning
中图分类号:O22; N94 **文献标志码**:A **文章编号**:1001-2486(2021)05-127-10

Application of reinforcement learning in multi-period weapon portfolio planning problems

ZHANG Xiaoxiong[1], *DING Song*[2], *LI Minghao*[3], *DING Kun*[1], *WANG Long*[1], *YI Yujiang*[4]
(1. The Sixty-third Research Institute, National University of Defense Technology, Nanjing 210007, China;
2. School of Economics, Zhejiang University of Finance & Economics, Hangzhou 310018, China;
3. College of Systems Engineering, National University of Defense Technology, Changsha 410073, China;
4. Southwest Electronics and Telecommunication Technology Research Institute, Chengdu 610041, China)

Abstract: Aiming at the difficulties in the choosing and planning in multi-period weapon systems development problems, an optimization simulation approach combining multi-objective optimization algorithm and reinforcement learning technique was proposed. A multi-objective optimization model was built to maximize the capability and minimize the cost of weapon portfolios in each period. Moreover, a solving algorithm based on the non-dominated sorting genetic algorithm-Ⅲ was presented to obtain the Pareto set in each period, based on which an optimization model for multi-period problem was built. The Q-Learning method, one of the reinforcement learning algorithms, searches within the Pareto set using two different ways for the selection of weapon portfolios in each period, whose outcome is used for the selection in the next period and the optimization of the portfolios over the entire planning horizon. An illustrative example was studied to demonstrate the effectiveness of the proposed model and hybrid algorithm, which can support the decision making on the weapons development and planning.

Keywords: weapon; portfolio planning; non-dominated sorting genetic algorithm-Ⅲ; reinforcement learning; Q-Learning

装备组合规划选择是武器装备体系顶层发展规划中的重要问题,旨在对一定规划期内装备的具体建设发展进行总体规划安排[1]。当前战争形态的变化,要求决策者们更多关注装备组合作为一个整体发挥的效能,而不再局限于单一装备的性能。同时,在考虑涉及多个阶段的装备组合方面,任何单一阶段的最优装备组合无法保证整个规划周期内的最优性。因此,需要合理权衡规划不同阶段、不同周期的装备组合选择,从而更好地满足未来作战能力需求和完成多元化的任务。

装备组合选择源于项目组合选择问题,Markowitz[2]最早提出了组合的概念来处理投资组合问题,旨在最大化投资收益的同时降低投资的市场风险,奠定了金融领域的投资组合理论。后来该理论又逐渐被应用到项目管理中辅助组合方案的比较和选型。针对军事领域的组合选择问题,常见的研究方法有多准则决策分析、专家评审法、价值分析法、风险分析法和资源分配方法等。例如,Kangaspunta等[3]在考虑装备之间相互关联的条件下,提出了一种费用-效能分析方法,辅助

* 收稿日期:2020-01-18
基金项目:国家自然科学基金资助项目(71901215,71901191);国防科技大学校科研计划资助项目(ZK20-46)
作者简介:张骁雄(1990—),男,江苏淮安人,高级工程师,博士,硕士生导师,E-mail:zxxandxx@163.com

装备组合选型；Yang 等[4]对复杂军事大数据环境下的武器装备组合选择优化问题进行了建模，并设计了一种自适应的遗传算法对模型进行求解；Li 等[5]基于能力规划的思想，提出了一种基于异质网络模型的高端装备组合选择方法；Dou 等[6]提出了一种基于偏好基线值的方法，对装备组合中冗余装备的取舍进行了研究；王等[7]运用 epoch-era 思想，构建了区间型需求下的装备组合多阶段随机规划模型；孙等[8]提出了面向作战需求的卫星装备组合优化算法，对不同装备组合的作战效能进行了评估。

此外，还有一些比较流行的概念和方法论，被用来指导武器装备组合选择与优化，包括美国国防部提出的基于能力的规划（Capability Based Planning，CBP）[9]、麻省理工学院提出的多属性权衡空间探索（Multi-Attribute Tradespace Exploration，MATE）方法[10]、美国军方提出的将费用当作独立变量的方法[11]等。同时，装备组合选择与评估优化问题也引起了国内如军事科学院[12-13]、国防大学[14]、国防科技大学[15-16]等高校与研究机构的广泛关注，并取得了一定的研究成果。

不同学者对军事领域的组合选择进行了不同的探索和尝试，然而现实中这种建模对数据要求较高，目前仍然缺少较为定量的模型与算法，在支撑装备体系顶层规划和决策方面仍略有不足。同时，随着考虑的场景、规划的装备数目以及规划周期的增多，传统的数学方法以及多目标优化算法在求解效率上往往捉襟见肘。例如，对于一个具备 K 个场景和 T 个优化周期的规划问题来说，决策者需要至少同时考虑 $K \cdot T$ 个优化目标，大大增加了求解难度。近年来，深度学习在图片识别等任务上取得了前所未有的效果，强化学习也在 AlphaGo 方面效果显著，它通过学习和选择动作改变外界环境，并使用一个累计回报来定义任意动作序列的质量，正适用于解决多阶段下的装备组合选择问题。

因此，拟借鉴强化学习的思想，研究多阶段情形下的装备组合优化问题。以装备组合的效能和成本为目标，建立武器装备组合规划问题的多目标优化模型，并基于智能算法生成各阶段的最优装备组合。相比传统研究，本文采用强化学习对不同阶段的装备组合进行寻优，生成整个规划周期内的最优装备组合方案。目前，鲜有研究将强化学习应用于多阶段的装备规划研究方面，且该方法可以大大提高求解效率。

1 问题分析及建模

重点面向多个作战场景，研究多阶段情形下的装备组合选择问题。在横向上突出面向不同场景的优化，纵向上突出时间维度，并非将单阶段单场景下的装备组合方案进行简单叠加。任何针对单一场景或固定效能值的装备组合选择往往具有一定的片面性。图 1 为多阶段装备组合发展示意图，该问题研究的难点在于阶段之间相互关联，上一阶段的决定直接影响后续阶段的选择，即每个阶段装备的解空间都发生变化，且装备不能被重复选择[17]。

聚焦单一阶段的优化求解问题。假定 $x_i \in X$ 为当前可选装备集合 X 中第 i 个装备，发展该装备需要耗费成本 c_i。假设装备的组合发展需要同时考虑和应对 K 个不同的场景。因为不同的作战场景等外部因素，装备发挥的效能各不相同。因此，令 r_i^k 代表装备 x_i 在场景 k 下可发挥的效能值。

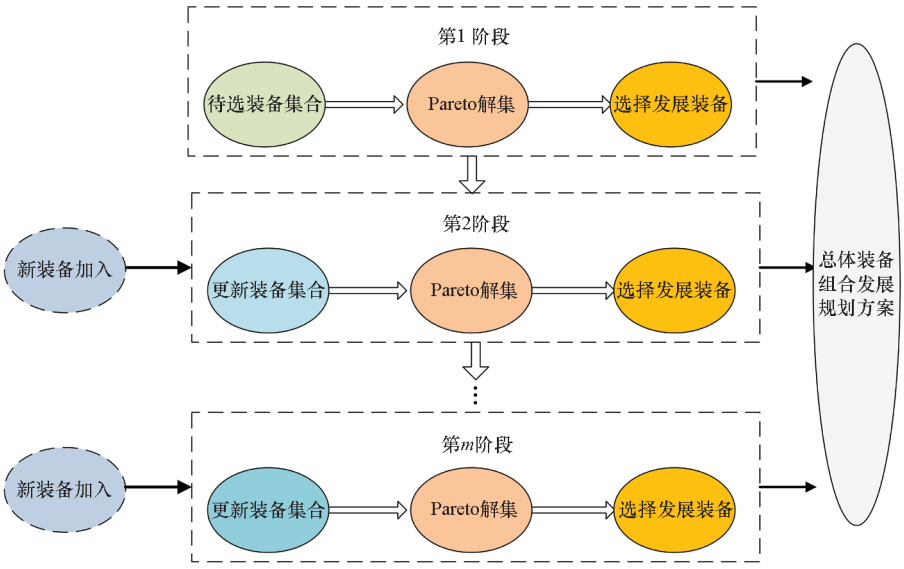


图 1 多阶段装备组合发展示意图

Fig. 1 An illustration of multi-period weapon portfolio selection

决策变量 x_i 数学形式的定义为:

$$x_i \in X = \begin{cases} 1, \text{如果 } x_i \text{ 被选中发展} \\ 0, \text{其他} \end{cases} \tag{1}$$

基于上述分析,针对单一阶段的装备组合优化问题,需要同时考虑 K 个可能场景,从当前可选装备集合中选取合适的装备组合,达到最大化装备组合效能以及最小化装备组合成本的目标。由此,可构建如式(2)所示目标函数。

$$\begin{aligned} &\max \sum_i r_i^k x_i, \ \forall k \\ &\min \sum_i c_i x_i, \\ &\text{s.t.} \ \sum_i c_i x_i \leqslant B(1+\delta) \\ &x_i \in \{0,1\} \end{aligned} \tag{2}$$

其中,$k \in [1,K]$ 表示任意场景,K 为场景总数,B 表示总预算限制,δ 为预算违背阈值。不等式约束限制了选中装备组合的总成本。显然,上述问题存在 $K+1$ 个待优化目标。

武器装备发展规划需要对一个周期(通常为 10 a)内的装备进行统筹安排,并需要考虑装备的更替。进一步,将上述优化问题扩展到多个阶段,即决策者需要选择能够在整个规划周期内最大化装备组合效能并最小化装备组合成本的方案。形式上,决策变量 x_{it} 被定义为:

$$x_{it} \in X_t = \begin{cases} 1, \text{如果 } x_i \text{ 在时刻 } t \text{ 被选中发展} \\ 0, \text{其他} \end{cases} \tag{3}$$

其中,X_t 代表 t 阶段可选装备集合。

此时,目标函数在多阶段多场景下变更为:

$$\begin{aligned} &\max \sum_i r_{it}^k x_{it}, \ \forall k,t \\ &\min \sum_i c_{it} x_{it}, \ \forall t \\ &\text{s.t.} \ \sum_i c_{it} x_{it} \leqslant B_t(1+\delta_t), \ \forall t \\ &x_{it} \in \{0,1\} \end{aligned} \tag{4}$$

其中,r_{it}^k 表示阶段 t、场景 k 下装备项目 x_i 所具备的效能,c_{it} 为装备 x_i 在阶段 t 下对应的开发成本,B_t 为阶段 t 下的经费预算,δ_t 代表阶段 t 下的预算违背阈值。

针对本节构建的多阶段不确定性模型,可通过综合使用多目标优化算法以及强化学习来处理。决策者可以有效应对未来阶段的不确定性,并在每个阶段产生的最优解中进行动态优化。

为使构造的模型更加合理,限定如下基本假设:

1)初始阶段装备项目已知,并在未来每一阶段会有新装备加入;

2)不同场景下各装备的效能服从一定的分布,假定为正态分布;

3)装备之间相互独立,可并行发展;

4)各装备发展成本已知且固定;

5)装备一旦被选中发展则不可剔除。

2　模型构建求解

针对多阶段装备组合规划问题,本节给出基于多目标优化算法以及强化学习的求解框架,并分小节阐述。

2.1　基于 NSGA-Ⅲ的多目标优化算法

针对任一阶段的装备组合选型,需要在给定的决策空间中,最大化所选择装备组合的效能。由于考虑 K 个不同场景,且不同场景下装备组合的效能无法进行简单的叠加。故而,将其转变为 $K+1$ 个多目标优化问题,包括 K 个不同场景下装备组合的效能以及装备组合的成本。随着场景数目以及装备数目的增多,该多目标优化问题具备 NP-hard 性质。传统的搜索方法效率低下,且使用范围有限。

非支配排序遗传算法(Non-dominated Sorting Genetic Algorithm-Ⅲ,NSGA-Ⅲ)[18]是一种新型智能优化算法,算法沿用了 NSGA-Ⅱ的框架,但临界层选择方法采用参考点方法选择个体,以使种群具有良好的分布性,保证更加准确的全局搜索能力。

针对上述待优化模型,首先初始化种群 A,经过与 NSGA-Ⅱ相同的选择、交叉、变异后,选择生成非支配个体 A′。在对约束部分进行处理时,算法采用罚函数将个体违反约束的部分累加到目标函数中。之后,NSGA-Ⅲ将主要执行如下步骤。

步骤 1:归一化。假设共存在 M 个待优化函数,确定每一个目标 i 上的最小值构成集合 $z_i^{\min}$,并针对每一维度目标函数进行标量化操作。

$$f_i'(x) = f_i(x) - z_i^{\min}, \ x \in S_t \tag{5}$$

式中,S_t 为种群的个体集合。

之后寻找极值点,定义函数 ASF。

$$ASF(x,W) = \max_{i=1:M} \frac{f_i'(x)}{W_i}, \ x \in S_t \tag{6}$$

遍历每个函数,找到 ASF 数值最小的个体,即为极值点,再根据这些点计算出每个坐标点在对应坐标轴上的坐标值 α_i。之后,采用式(7)进行归一化。

$$f_i^n(x) = \frac{f_i'(x)}{\alpha_i - z_i^{\min}} = \frac{f_i'(x) - z_i^{\min}}{\alpha_i - z_i^{\min}}, \ i = 1,2,\cdots,M \tag{7}$$

步骤2:参考点确定。NSGA-Ⅲ的参考点可在归一化的超平面内进行。在目标空间中,一个维度为 $M-1$ 的标准超平面对所有目标轴都有相同的倾斜度。若考虑沿着每个目标方向进行 p 等分,则参考点 $\boldsymbol{H}$ 的总数为 $\boldsymbol{H}=\begin{pmatrix}M+p-1\\p\end{pmatrix}$。

步骤3:关键层解的选择策略。通过定义参考线的方式,计算种群每个个体到参考线的垂直距离,并将种群中的个体分别关联到相应的参考点。假设与参考点 j 关联的解的数量为 ρ_j。从关键层选取 ρ_j 最小的参考点 j 加入种群中。若 $\rho_j=0$,则从关键层里选取一个距离该参考点 j 最小的解加入种群,否则将该参考点从当前代中去除;若 $\rho_j\geqslant 1$,则从关键层里面随机挑选一个关联到该参考点的解加入种群。

2.2 Q-Learning 强化学习方法

通过对单一阶段的求解,可以获取每个阶段的 Pareto 解。然而任意单阶段的最优解未必是整个规划周期里的最优选择。同时,当前阶段的选择又直接影响着下一个阶段的决策空间和选择。

强化学习[19]是一种重要的机器学习方法之一,它明确考虑了目标导向的智能体与不确定环境交互的整个问题,旨在最大化期望积累奖励。强化学习的特点正适用于解决多阶段的装备组合选择与规划问题。图2为强化学习示意图。主要包括如下几个关键要素:环境、回报、动作和状态。

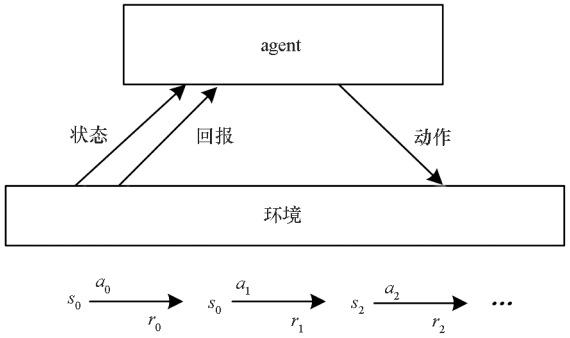


图2 强化学习示意图

Fig. 2 Illustration of reinforcement learning

1)状态(state)。描述当前 agent 所处的状态,所有可能的状态称为状态空间。状态 s 对应不同的决策阶段。不同的决策阶段对应不同的选型空间,即截至当前阶段所有未被选中发展的装备集合。

2)行动(action)。指 agent 根据每次所处的状态以及上一状态的回报确定当前要执行的动作。本研究中的动作即决策者在各个时刻选取装备组合的行为。

3)回报(reward)。强化学习是 agent 可以学习行为以实现最大化其累计奖励的方式,即在发生状态转移的同时,环境反馈给 agent 的奖赏,且回报是一个标量。

4)策略(policy)。策略用来描述 agent 在不同状态下执行的动作。常见的策略分为确定性策略以及随机性策略。确定性策略描述在状态 s 下执行确定动作 a,随机策略描述状态 s 下执行动作 a 的概率。本模型中,动作 a 代表在 t 时刻选取 $x_{it}=1$ ($x_{it}\in X_t$)的装备选择行为。

5)价值函数(value function)。强化学习是一个连续决策的过程,当下的行为是否正确需要经过一定的时间才能得知,因此需要用未来一段时间的收益来作为当下行为的评判。如果仅仅关注当前阶段收益的最大化,容易导致决策的片面性。因此建立当前状态下的价值函数:

$$V_\pi(S)=E_\pi[R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\cdots|S_t=s] \tag{8}$$

式中:γ 为奖励衰减因子,且取值区间为[0,1]。γ 越接近1,则考虑越长远;若为0,则表示只考虑一步的奖励。

6)状态转移模型。使用状态转移模型来预测接下来的动作行为,即在当前状态下执行某一动作导致的状态以及产生的回报。采用动作转移概率与动作状态回报来描述该模型。

$$\begin{cases}P^a_{SS'}=P[S_{t+1}=s'|S_t=s,A_t=a]\\R^a_S=E[R_{t+1}|S_t=s,A_t=a]\end{cases} \tag{9}$$

2.3 基于 Q-Learning 的多阶段组合优化模型求解

多阶段装备组合规划选型旨在从每一阶段的非支配解中选取合适的方案构成整个规划周期内的装备组合,并使装备组合效能和成本总体达到最优。任何单一阶段的最优解的集合未必在多个阶段仍然最优,同时需要综合考虑每一阶段,决策对未来的影响。结合 Q-Learning 的算法,基于强化学习的多阶段装备组合规划问题的求解步骤如下。

步骤1:在各阶段,删除之前阶段已被选中发展的装备组合,同时增加新型待发展的装备集合(代指可供选择发展的新增装备),更新并生成当前可供选择发展的装备集合,即当前阶段的解空间。

步骤2:针对 K 个场景的选择规划问题,每个阶段存在 $K+1$ 个目标待优化,采用 NSGA-Ⅲ算法对当前阶段的目标进行求解,生成当前阶段可供选择的非支配 Pareto 解。

步骤 3:采用随机探索或者利用最优 Q 值的方式,从上阶段的 Pareto 解中选取一个装备组合,并采用式(10)中的 Q-Learning 公式更新当前阶段下选择该装备组合的 Q 值。

$$Q(S_t,a_t)\leftarrow(1-\alpha)\cdot Q(S_t,a_t)+\cdots$$
$$\alpha[R_t+\gamma\max_{PS}Q(S_{t+1},a_{t+1})] \tag{10}$$

式中,$Q(S_t,a_t)$ 表示在状态 S_t 下采取动作 a_t 产生的 Q 值,$\alpha\in[0,1]$ 表示学习率,描述控制新信息被采用的程度。该公式评估了在某个特定状态采取某个特定行动的价值。

步骤 4:重复迭代,直至达到停止标准。

如步骤 2 所述,需要针对每一阶段求解生成该阶段的非劣解,并从中选取一个装备组合作为该阶段的动作行为。步骤 3 中基于探索或利用的策略,从当前阶段的 Pareto 解中随机选择或者选择 Q 值最高的装备组合。常见的 Q-Learning 引入一个参数 τ 来控制在两种选择策略之间的权衡关系。一般来说,将 τ 设置为 0.5,即允许算法在两种策略之间随机选择。

步骤 5:回报函数的构建是衡量和计算非劣解中方案 Q 值的重要依据。采用式(11)来衡量当前阶段 S_t 选择方案 a_t 的回报值。

$$R_t=w_1R_E+w_2R_C \tag{11}$$

式中:R_E 和 R_C 分别代表装备组合在效能以及成本方面的回报,默认为二者都已经过归一化处理;w_1 和 w_2 是针对效能和成本的权重,且满足两者之和为 1,此处将二者都设置为 0.5。

具体来说,R_E 与当前所选装备组合以及下一阶段可能选择的装备组合的效能息息相关,采用式(12)进行度量。

$$R_E=\frac{1}{2}\left[\frac{\sum_K E_{a_k}}{K}+\frac{\sum_{P_{t+1}}\sum_K E_{a'_k}}{N_{P_{t+1}}K}\right],$$
$$\forall k\in K,\ a'\in P_{t+1} \tag{12}$$

其中,K 代表场景的个数,等式右边括号中前半部分代表当前所选择装备组合 a 在 K 个场景中效能的算术平均,后半部分代表下一阶段所有可能装备组合 a' 在 K 个场景中效能的算术平均,P_{t+1} 为下一阶段的最优 Pareto 解,$N_{P_{t+1}}$ 代表该 Pareto 解的个数。

对于 R_C,决策者希望在每个阶段 t 所选择的装备组合 a_t 的成本能尽可能贴近当前阶段给定的总成本约束 B_t,同时下一阶段的装备组合非劣解中每个方案的成本也尽可能与下阶段的成本约束相近,由此,采用式(13)来衡量与成本相关的回报。

$$R_C=-\frac{1}{2}\left[\frac{|C_a-B_t|}{B_t}+\frac{\sum_{P_{t+1}}|C_{a'}-B_{t+1}|}{N_{P_{t+1}}B_{t+1}}\right],$$
$$\forall a'\in P_{t+1} \tag{13}$$

其中,C_a 表示当前阶段装备组合的成本。等式右边括号中前半部分对当前阶段的选择进行了衡量,后半部分则对未来阶段的可能性进行了衡量,以此来凸显当前选择可能对未来的影响。由于决策者希望任一阶段的装备组合成本更加贴近给定的预算,即与给定预算之间的差值越小越好,因此对两边的加和进行取反操作,以保证 R_C 越大越好。获得方案的当前回报值 R_t 后,采用式(10)中的 Q-Learning 公式对 Q 值进行更新。

3　实验与结果分析

3.1　数据说明

本节开展示例研究,将 Q-Learning 思想应用于多阶段装备组合规划中。采用随机生成的方法产生装备的效能以及成本数据。具体参数设置如下:

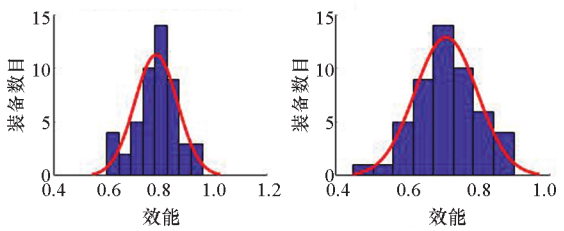
1)装备项目:假定存在 50 个初始装备,之后每年增加 5 个。

2)场景(K):假定存在 3 个不同场景。

3)规划阶段(T):假定整个规划周期为 10 a,该数值可根据需要进行调整。

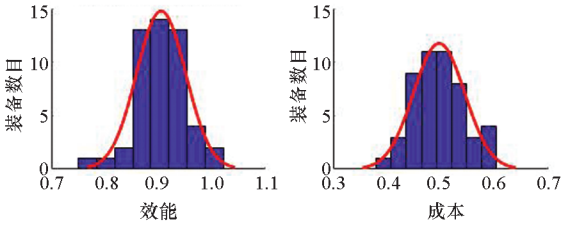
4)效能与成本:通过正态分布模拟装备在不同场景下发挥的效能以及发展成本,如图 3 所示。表 1 给出了初始阶段的装备效能以及成本数值,且假设装备效能和成本取值均已经过归一化处理。

其他方面,设置总经费 $S=25$ 亿元,一般情形下保证年度经费分配相对平均,并允许在一定范围 $\delta=0.1$(10%)内波动,即每年的年度经费波动范围为 $[(1-\delta)S/T,(1+\delta)S/T]$[20];回报函数中,学习率 $\alpha=0.1$,折算率 $\gamma=0.9$。



(a) 场景 1 下装备效能分布　(b) 场景 2 下装备效能分布

(a) Weapon effectiveness distribution of scenario 1　(b) Weapon effectiveness distribution of scenario 2



(c) 场景3下装备效能分布
(c) Weapon effectiveness distribution of scenario 3

(d) 装备成本分布
(d) Weapon cost distribution

图3 效能成本分布

Fig. 3 Distribution for weapon effectiveness and cost

表1 不同场景下装备效能与成本

Tab. 1 Weapons effectiveness and costs under various scenarios

装备项目	场景1：效能	场景2：效能	场景3：效能	成本
1	0.802 5	0.669 2	0.890 9	0.538 3
2	0.814 8	0.829 6	0.854 6	0.487 2
3	0.827 9	0.726 1	0.961 3	0.466 9
4	0.749 7	0.659 8	0.874 5	0.492 3
5	0.887 1	0.617 3	0.859 1	0.519 7
6	0.988 3	0.648 3	0.836 9	0.555 2
7	0.601 0	0.740 0	0.826 6	0.505 1
8	0.815 8	0.784 7	0.913 8	0.413 8
9	0.875 6	0.704 8	0.883 4	0.541 7
10	0.736 3	0.650 7	0.939 4	0.472 7
11	0.807 2	0.733 5	0.897 2	0.592 3
12	0.813 6	0.767 1	0.889 5	0.530 5
13	0.752 0	0.657 8	0.813 9	0.532 7
14	0.856 5	0.844 3	0.861 2	0.482 9
15	0.897 2	0.663 6	0.905 2	0.457 8
16	0.817 9	0.789 8	0.878 6	0.463 5
17	0.829 6	0.796 8	0.963 8	0.435 2
18	0.756 3	0.705 7	0.808 4	0.458 8
19	0.889 5	0.691 5	0.890 9	0.559 1
20	0.792 8	0.787 2	0.905 7	0.470 1
21	0.835 3	0.519 6	0.863 9	0.497 0
22	0.894 0	0.703 4	0.905 6	0.577 9
23	0.776 2	0.706 9	0.983 1	0.475 1
24	0.784 5	0.736 6	0.847 5	0.489 9
25	0.807 2	0.611 5	0.882 5	0.534 6
26	0.782 1	0.559 0	0.829 0	0.510 0
27	0.698 5	0.436 2	0.936 4	0.523 1
28	0.838 6	0.754 5	0.839 7	0.481 1
29	0.684 7	0.807 8	0.857 1	0.491 2
30	0.775 6	0.655 6	0.971 1	0.581 0

续表

装备项目	场景1：效能	场景2：效能	场景3：效能	成本
31	0.687 3	0.937 9	0.941 2	0.601 1
32	0.904 1	0.635 9	0.862 2	0.498 3
33	0.806 8	0.568 9	0.950 8	0.461 3
34	0.838 3	0.685 4	0.867 0	0.495 5
35	0.779 0	0.710 6	0.905 3	0.471 1
36	0.926 9	0.768 2	0.991 7	0.464 1
37	0.767 7	0.712 1	1.012 9	0.401 3
38	0.892 1	0.623 6	0.877 8	0.422 5
39	0.892 3	0.547 7	0.898 5	0.498 5
40	0.758 0	0.641 3	0.952 0	0.467 6
41	0.834 3	0.563 6	0.930 4	0.566 1
42	0.810 1	0.772 4	0.888 3	0.407 4
43	0.969 2	0.689 6	0.889 5	0.453 6
44	0.720 5	0.514 7	0.860 8	0.525 1
45	0.673 2	0.736 6	0.786 4	0.523 5
46	0.824 9	0.607 8	0.892 2	0.478 0
47	0.722 4	0.835 0	0.840 1	0.449 8
48	0.752 9	0.705 2	0.888 6	0.381 4
49	0.826 5	0.722 9	1.023 4	0.476 2
50	0.696 2	0.621 3	0.901 7	0.523 5

本次实验仿真采用MATLAB 2017软件，运行于Windows 7 64位系统中，软件环境见表2。

表2 实验硬件环境

Tab. 2 Experimental hardware environment

序号	参数	值
1	CPU核数	4
2	CPU主频	3.6 GHz
3	内存容量	32 GB
4	显卡容量	2 GB
5	硬盘容量	1TB×2

另外，由于每年会增加一些新的待选装备，而之前已经被选中发展的装备在未来规划阶段内不能作为待选装备出现，因此需要对每个阶段的可选装备组合进行更新。具体装备信息生成、更新方法如图4所示。

3.2 结果分析

基于所述算法，采用探索和利用相结合的方式，设置$\tau=0.5$，对示例进行20次运行。每次运行需要考虑整个规划周期内每个年度的优化目标。将所采用的多目标优化算法(NSGA-Ⅲ)的种

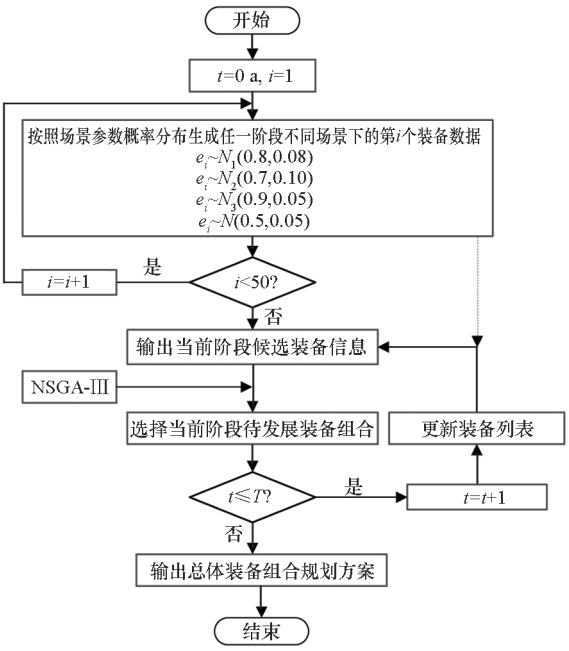


图 4　装备更新策略

Fig. 4　Weapon update strategy

群规模设为 100,迭代次数设为 50,交叉概率设为 0.8,变异概率设为 0.02。

经过 100 次学习,可以获得 100 组 Q 值矩阵,对应不同的装备组合方案。选取总体效能最大的方案,各个年度对应装备组合的 Q 值如图 5 所示。

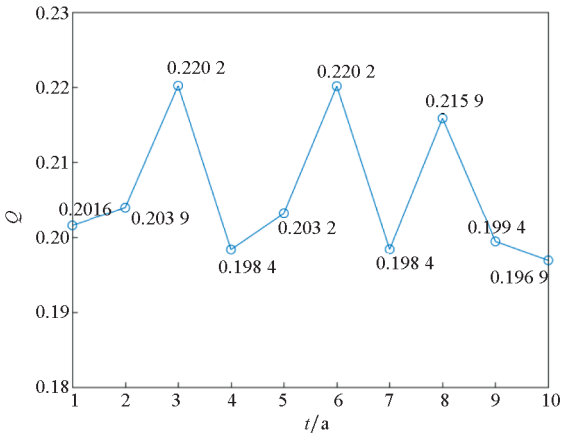


图 5　最优规划方案对应的 Q 值

Fig. 5　Corresponding Q value of the best portfolio solution

$\boldsymbol{Q}$ 矩阵中每一行代表一种装备组合规划方案,而每一元素代表该方案在当前阶段下装备组合产生的 Q 值。图 5 中,第 1、4、7 以及第 10 阶段,采用随机探索的方式选取装备组合方案,其他年度按 Q 值最大值选取装备组合方案。

总的规划周期内,各个规划阶段的装备组合选择方案如图 6 所示。图 6 中,黄色部分代表整个规划周期内被选中发展的装备。由图 6 可知,

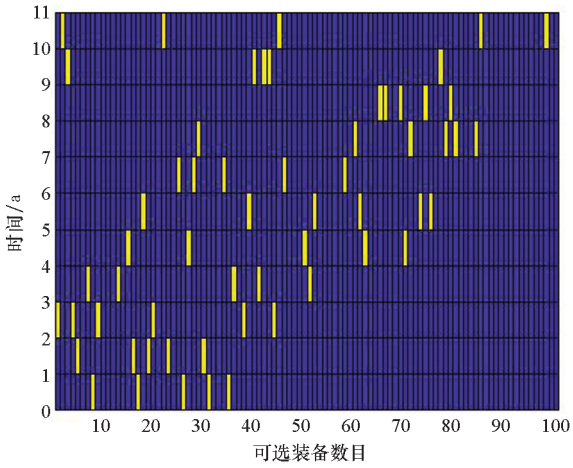


图 6　最优装备发展方案

Fig. 6　Optimal weapon development solution

得益于每年新装备的加入,此次规划方案一共选择发展 58 个装备,每一年被选中发展的装备数目相对平均,除了第 3 年、第 6 年和第 8 年,选择发展了 6 个装备,其他阶段都选择发展了 5 个装备,且每年装备投入的成本也相对均衡,满足现实约束的需要。

为突出 NSGA-Ⅲ参数对算法结果的影响,对算法中主要参数进行敏感性分析。分别独立运行各种情形 20 次,并对各情形下的装备组合方案效能值以及成本取平均值进行分析,结果见表 3。

表 3　NSGA-Ⅲ参数敏感性分析

Tab. 3　Parameter sensitivity analysis on NSGA-Ⅲ

参数		效能 1	效能 2	效能 3	成本
种群规模	50	44.56	39.57	49.74	27.16
	100	45.41	40.15	50.59	27.55
	150	45.54	39.91	50.60	27.33
	200	45.73	40.88	50.58	27.66
迭代次数	50	45.41	40.15	50.59	27.55
	100	44.40	39.41	49.60	26.75
	150	45.14	39.28	49.78	27.26
	200	45.37	39.91	50.78	27.64
交叉概率	0.20	44.65	39.43	49.99	26.84
	0.40	45.07	39.48	49.73	26.91
	0.60	44.59	40.28	49.85	27.28
	0.80	45.41	40.15	50.59	27.55
变异概率	0.02	45.41	40.15	50.59	27.55
	0.04	45.33	40.21	50.61	27.30
	0.06	45.53	39.88	50.72	27.54
	0.08	45.71	40.27	50.83	27.66

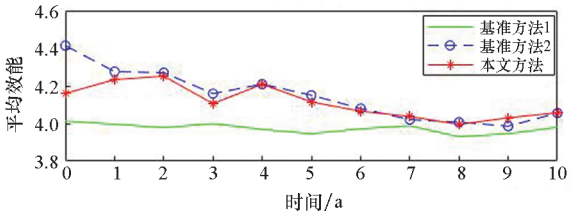
由表 3 可知,不同情形下最终方案效能以及

成本各异。总体来说,随着种群规模和迭代次数的增大,最后生成的方案在效能上更优,但方案成本以及算法运行时间也随之增大。随着交叉概率的增大,各最终生成方案总体更优,主要表现为方案的效能总和不断增加,因为较大的交叉概率可以较好地保证进化时种群的丰富性。随着变异概率的增大,各情形下生成的方案结果差异性不大,主要因为总体变异幅度相对较小。

3.3 对比分析

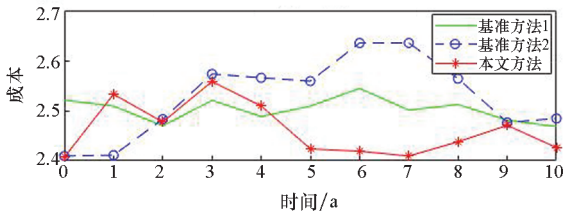
为验证本文方法的有效性,分别设置两个传统解决多阶段问题的基准方法进行对比分析。其中,基准方法 1 在满足经费约束条件下随机生成各阶段的装备组合方案,并实时更新下一阶段的可选装备集合空间。基准法 2 与基准方法 1 相似,但在各阶段选取装备时按照当前可选装备在各场景中发挥效能均值的大小从高到低依次进行选择,生成当前选择装备组合,并更新下一阶段的可选装备组合空间。重复上述步骤直至生成整个规划周期内的装备组合方案。两种方法都更加注重短期内各阶段的选择,没有考虑多个阶段之间的权衡选择问题,且这两种方法都没有选择智能优化算法对多目标优化问题进行求解[19]。分别运行上述算法以及本文方法 20 次,图 7 给出了不同方法策略下的最优装备组合方案在各阶段的效能均值以及成本均值。

由于基准方法 2 是在各阶段选择效能最大的装备构成当前装备组合,因此总体效能略优于其他两种方法。但由图 7(a)可知,本文方法在后续各阶段的生成装备组合效能与基准方法 2 基本持平,并在第 7 年后略优于基准方法 2。基准方法 1 生成的装备组合方案效能在各阶段均相对较低。由图 7(b)可知,基准方法 2 的成本总体较高。而本文方法除了在初始阶段成本略高于另两种方法,在后续各阶段的成本均明显低于两种基准方法,且成本总和最低。从占优的角度,本文方法优于另两种方法对应的装备组合方案。换而言之,本文方法可以在更低成本下生成总体效能更优的装备组合方案。



(a) 效能分析

(a) Effectiveness analysis



(b) 成本分析

(b) Cost analysis

图 7 三种不同方法下生成方案结果对比

Fig. 7 Average effectiveness and cost under three different methods

3.4 参数敏感性

为突出选取策略参数对模型结果的影响,在同样的参数设置下,改变每个阶段选取装备的策略:将探索和利用两种策略的控制参数 τ 从 0.1 增加到 0.9。其中,$\tau=0.5$ 对应 3.1 节中的基本设置。由于效能与成本均是归一化后的值,因此可对不同方案的结果在同一维度下进行加和比较。通过计算,五种策略对应的组合方案的三种效能值以及成本如图 8 所示。

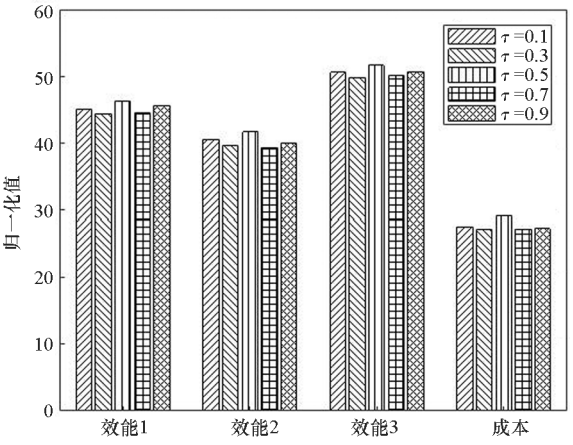


图 8 不同策略对应方案结果对比

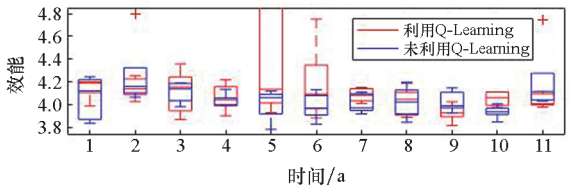
Fig. 8 Comparison of different solutions under different strategies

图 8 中的效能 1、效能 2 和效能 3 分别指装备组合方案在三种不同场景下的效能之和。对比发现,不同策略下方案的效能值以及成本各异。从占优的角度,四种方案都是非劣解,即不存在一个方案在每一项指标上都优于其他方案。但从总体效能的角度来看,方案 3($\tau=0.5$)混合策略下产生方案的效能在三种场景下皆优于其他几种方案。在效能 1 方面,方案 5($\tau=0.9$)优于方案 1($\tau=0.1$)、方案 4($\tau=0.7$)和方案 2($\tau=0.3$)。在效能 2 方面,方案 1 次优,后面依次为方案 5、方案 2 和方案 4。在效能 3 方面,方案 1 次优,后

面依次为方案 5、方案 4 和方案 2。从成本角度来看，方案 3 所产生装备组合成本相对较高，方案 4 对应装备组合方案成本最低。对比实验表明，在进行算法设计时，采取探索与利用相结合的方式选取装备，可以生成更加鲁棒的总体装备组合方案。

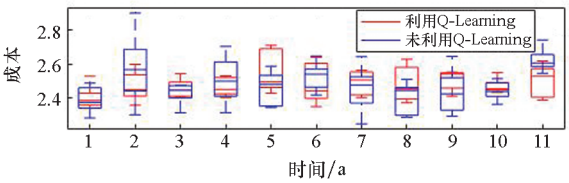
3.5　Q-Learning 效果分析

为验证模型中是否采用 Q-Learning 策略对最终选型方案的影响，继续开展对比实验。采用强化学习的策略依据 Q 函数选取各阶段的装备，而不采取强化学习的策略则在各阶段从候选 Pareto 中随机选取装备，类似于传统的优化算法。分别运行算法各 20 次，图 9 给出了是否采取强化学习策略下的最优装备组合方案运算结果分布情况。



（a）效能分析

（a）Effectiveness analysis



（b）成本分析

（b）Cost analysis

图 9　Q-Learning 策略对方案结果影响对比

Fig. 9　Comparison of different solutions with and without Q-Learning

由图 9 的盒须图可知，利用 Q-Learning 策略下生成的装备组合在大多数阶段的性能表现在最优值、均值等方面均优于未利用 Q-Learning 策略生成的方案。尤其是在第 3 至 6 阶段，利用 Q-Learning 的方案最优值显著优于未利用 Q-Learning 的方案，且方案在整个规划周期的总效能更优。在成本方面，未利用 Q-Learning 策略生成的方案在最优值方面略优于采用 Q-Learning 策略生成的方案。但在成本均值方面，两者基本相当，且在第 2、第 3、第 4、第 6 和第 10 阶段以及各阶段的总和，利用 Q-Learning 策略生成的方案在各阶段的成本之和更优。换而言之，利用 Q-Learning 策略方案可以在相对更低成本下生成总体效能更高的装备组合方案，这验证了本文模型采取 Q-Learning 策略的优势。

4　结论

武器装备组合规划是我军武器装备体系建设发展中亟须解决的现实问题，具有十分重要的战略意义。结合多目标优化算法与强化学习技术，重点回答了多个阶段多个场景下的武器装备组合选择问题，而目前仍鲜有研究将强化学习应用于多阶段的装备规划研究方面，其中，多目标优化算法用来在每个规划时间决策点内，以最大化多个场景装备组合的总效能与最小化总成本为目标，搜索非支配的装备组合方案；强化学习算法可以有效对多阶段问题进行水平搜索，形成任意阶段的策略规则，从而有效保证决策结果在整个阶段的最优性。通过具体示例验证了本文模型的可行性与求解的高效性。对比实验表明，本文方法生成的装备组合方案优于其他传统多目标决策方法，探索和利用策略的控制参数对模型结果具有一定影响，且采取强化学习生成的方案优于不采取强化学习方法生成的方案。提出的模型与算法可以支撑武器装备中长期规划决策和论证。

参考文献（References）

[1]　张骁雄，葛冰峰，姜江，等. 面向能力需求的武器装备组合规划模型与算法[J]. 国防科技大学学报，2017，39(1)：102－108.
ZHANG Xiaoxiong，GE Bingfeng，JIANG Jiang，et al. Capability requirements oriented weapons portfolio planning model and algorithm[J]. Journal of National University of Defense Technology，2017，39(1)：102－108.（in Chinese）

[2]　MARKOWITZ H. Portfolio selection[J]. The Journal of Finance，1952，7(1)：77－91.

[3]　KANGASPUNTA J，LIESIÖ J，SALO A. Cost-efficiency analysis of weapon system portfolios[J]. European Journal of Operational Research，2012，223(1)：264－275.

[4]　YANG S L，YANG M，WANG S，et al. Adaptive immune genetic algorithm for weapon system portfolio optimization in military big data environment[J]. Cluster Computing，2016，19(3)：1359－1372.

[5]　LI J C，GE B F，JIANG J，et al. High-end weapon equipment portfolio selection based on a heterogeneous network model[J]. Journal of Global Optimization，2020，78(4)：743－761.

[6]　DOU Y J，ZHOU Z X，XU X Q，et al. System portfolio selection with decision-making preference baseline value for system of systems construction[J]. Expert Systems With Applications，2019，123：345－356.

[7]　王孟，张怀强，蒋铁军. 区间型需求下基于 epoch-era 思想的武器装备组合规划模型[J]. 海军工程大学学报，2018，30(6)：36－41.
WANG Meng，ZHANG Huaiqiang，JIANG Tiejun. Model of weaponry combination planning relevant to interval demand based on epoch-era analysis[J]. Journal of Naval University of Engineering，2018，30(6)：36－41.（in Chinese）

[8]　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.

SUN Shengzhi, HOU Yan, PEI Chunbao. Optimized combination of satellite application equipment addressing operational requirements[J]. Electronics Optics & Control, 2018, 25(5): 7 – 11, 16. (in Chinese)

[9] DAVIS P K. Analytic architecture for capabilities-based planning, mission-system analysis, and transformation[M]. Santa Monica: Rand National Defense Research Institute, 2002.

[10] QIAO L, EFATMANESHNIK M, RYAN M. A combinatorial approach to tradespace exploration of complex systems: a CubeSat case study[J]. INCOSE International Symposium, 2017, 27(1): 763 – 779.

[11] SHEN Y, LI A H. Research on application of CAIV in armament demonstration[J]. Procedia Computer Science, 2015, 55: 870 – 875.

[12] 卜广志. 武器装备建设方案的组合分析方法[J]. 火力与指挥控制, 2011, 36(3): 154 – 158, 162.
BU Guangzhi. A portfolio-analysis method for selecting armament development candidates[J]. Fire Control and Command Control, 2011, 36(3): 154 – 158, 162. (in Chinese)

[13] 胡晓峰, 张昱, 李仁见, 等. 网络化体系能力评估问题[J]. 系统工程理论与实践, 2015, 35(5): 1317 – 1323.
HU Xiaofeng, ZHANG Yu, LI Renjian, et al. Capability evaluating problem of networking SoS[J]. Systems Engineering-Theory & Practice, 2015, 35(5): 1317 – 1323. (in Chinese)

[14] 王飞, 司光亚. 武器装备体系能力贡献度的解析与度量方法[J]. 军事运筹与系统工程, 2016, 30(3): 10 – 15.
WANG Fei, SI Guangya. Analysis and measurement method of contribution of weapon systems-of-systems capability[J]. Military Operations Research and Systems Engineering, 2016, 30(3): 10 – 15. (in Chinese)

[15] 李明浩. 面向能力生成的装备网络组合优化研究[D]. 长沙: 国防科技大学, 2018.
LI Minghao. Research on capability generation oriented weapon network portfolio optimization[D]. Changsha: National University of Defense Technology, 2018. (in Chinese)

[16] ZHAO Q S, LI S F, DOU Y J, et al. An approach for weapon system-of-systems scheme generation based on a supernetwork granular analysis[J]. IEEE Systems Journal, 2017, 11(4): 1971 – 1982.

[17] CHEN H K, ZHU X M, LIU G P, et al. Uncertainty-aware online scheduling for real-time workflows in cloud service environment[J]. IEEE Transactions on Services Computing, 2018: 1 – 12.

[18] 吴伟丽. 基于NSGA-Ⅲ的复杂成因变压器直流偏磁控制优化算法[J]. 电测与仪表, 2018, 55(11): 89 – 93.
WU Weili. Optimization algorithm for transformer DC bias control due to multi factors based on NSGA-Ⅲ[J]. Electrical Measurement & Instrumentation, 2018, 55(11): 89 – 93. (in Chinese)

[19] SHAFI K, ELSAYED S, SARKER R, et al. Scenario-based multi-period program optimization for capability-based planning using evolutionary algorithms[J]. Applied Soft Computing, 2017, 56: 717 – 729.

[20] 张骁雄, 姜江, 葛冰峰. 武器装备科研经费分配的规划模型与算法[J]. 系统工程与电子技术, 2015, 37(9): 2061 – 2066.
ZHANG Xiaoxiong, JIANG Jiang, GE Bingfeng. Scheduling model and algorithm for weapon equipment scientific research budgets allocation[J]. Systems Engineering and Electronics, 2015, 37(9): 2061 – 2066. (in Chinese)