【后勤保障与装备管理】　　doi：10.11809/bqzbgcxb2019.08.022

基于增强学习的航材二级库存优化配置研究

徐常凯，周家萱，杜加刚

（空军勤务学院 航材四站系，江苏 徐州　221000）

摘要：为了解决编制体制调整后，二级库存条件下航材供应的横向调配供应问题，提出了增强学习思想下使用马尔科夫决策过程对二级库存航材供应过程建模，利用增强学习中的策略评估和策略迭代方法求解，得到在任意库存状态下，对应的最优横向供应策略。通过验证，该模型可以在二级库存条件下，减小各单位库存不平衡对航材保障效率的影响，延长保障时间，增强航材横向供应的提前性和科学性。

关键词：二级库存；横向供应；增强学习；MDP

本文引用格式：徐常凯，周家萱，杜加刚. 基于增强学习的航材二级库存优化配置研究[J]. 兵器装备工程学报，2019，40(8)：106－110.

Citation format：XU Changkai，ZHOU Jiaxuan，DU Jiagang. Study on Air Material Allotment for Dual Echelon Inventory System with Reinforce Learning[J]. Journal of Ordnance Equipment Engineering，2019，40(8)：106－110.

中图分类号：TP312；E939　　**文献标识码**：A　　**文章编号**：2096－2304(2019)08－0106－05

Study on Air Material Allotment for Dual Echelon Inventory System with Reinforce Learning

XU Changkai，ZHOU Jiaxuan，DU Jiagang

（Department of Material and Four Station，Air force Logistic College，Xuzhou 221000，China）

Abstract：Aiming at the problem of dual echelon air material inventory system with horizontal supply after the adjustment of authorized strength. The model of dual echelon air material inventory system using markov decision processes were established，then solved with policy evaluation and policy iteration in reinforce learning，reaching the best horizontal supply policy for every inventory states. This model can reduce the influence of air material supply efficiency caused by inventory imbalance，strength the lead time and scientificity of air material horizontal supply.

Key words：dual echelon inventory system；horizontal supply；reinforce learning；MDP

如何对航材库存进行科学的优化配置，寻求航材保障良好率与航材保障经济性之间的最佳平衡，是航材库存优化的重要问题。空军后方仓库编制体制调整后，航材库存从空军、战区空军、航材股三级变为器材仓库和队属仓库二级，解决二级库存体制下的航材库存优化配置，成为航材库存决策人员的现实难题。陈砚桥等[1]使用蒙特卡罗方法求解了允许横向供应的库存系统器材配置问题；刘少伟等[2]利用排队论系统，建立了可修件两级库存模型；Engin Topan 等[3]通过分支与定价算法求解了多备件两级库存系统模型。本文提出一种基于马尔科夫决策过程建模，利用强化学习的策略迭代方法求解，寻求保障效率与保障经济性的平衡，最终得到在航材不断消耗过程中的库存优化配置模型。

1　MDP 与策略迭代理论

1.1　马尔科夫决策过程

马尔科夫决策过程是描述动态系统决策优化的数学模型，通常由五元组 $\{S,A,P_{sa},\gamma,R\}$ 表示，其中 S 表示状态空

收稿日期：2019－01－19；**修回日期**：2019－02－20

作者简介：徐常凯(1971—)，博士，教授，硕士生导师，主要从事控制工程与科学研究。

通讯作者：周家萱(1995—)，硕士研究生，主要从事控制工程与科学研究。

间,A 表示行动空间,P_{sa} 为在状态 s 下执行行动 a 的概率,γ 为折扣因子,R 为奖励函数[4]。一个动态马尔科夫决策过程可表述为:从状态空间 S 中的某一状态 s_0 开始,在行动空间 A 中选择行动 a_0 执行后,马尔科夫决策过程的状态随机转移到状态 s_1,记为 $s_1 \sim P_{s0a0}$,类似的,整个决策过程可表示如下[5]:

$$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \xrightarrow{a_3} \cdots$$

定义策略 π 是由状态到行动的映射,即 $\pi: S \rightarrow A$,当决策过程处于状态 s 时,执行行动 $a = \pi(s)$ 时,定义价值函数 V^{π}:

$$V^{\pi}(s) = E[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots \mid s_0 = s, \pi] \tag{1}$$

对于某一确定的策略 π,其价值函数满足 Bellman 方程为[6]:

$$V^{\pi}(s) = R(s) + \gamma \sum_{s' \in S} P_{s\pi(s)}(s') V^{\pi}(s') \tag{2}$$

式(2)中,$P_{s\pi(s)}(s')$ 表示在状态 s 时,执行策略 π,转移至状态 s' 的概率。Bellman 方程用于快速求解所有 $V^{\pi}(s)$ 中的最优值,定义最优价值函数 $V^{*}(s) = \max_{\pi} V^{\pi}(s)$,即为奖励函数总和的最佳期望,则对应有 V^{*} Bellman 方程为[7]:

$$V^{*}(s) = R(s) + \max_{a \in A} \gamma \sum_{s' \in S} P_{sa}(s') V^{*}(s') \tag{3}$$

以及最佳策略 $\pi^{*}: S \rightarrow A$:

$$\pi^{*}(s) = \arg\max_{a \in A} \sum_{s' \in S} P_{sa}(s') V^{*}(s') \tag{4}$$

即在达到式(3)最大值的策略。

1.2 策略迭代

从 MDP 的定义中可以看到,求解最优策略的目的是在状态空间 S 和行动空间 A 中,选择恰当的状态和行动序列,达到最优化价值函数的目的,即使 V 收敛至 V^{*}。显然,状态空间和行动空间的大小决定了 MDP 问题的求解难度,策略迭代方法用于解决有限状态的 MDP 问题,即 $|S| < \infty$,$|A| < \infty$。策略迭代算法的描述如下[8]:

1) 随机初始化策略 π

2) 重复至收敛{

a. 令 $V := V^{*}$;

b. 对每一状态 s,令

$$\pi(s) = \arg\max_{a \in A} \sum_{s' \in S} P_{sa}(s') V^{*}(s') \tag{5}$$

其中 a 过程可以通过 Bellman 方程求解,b 过程通常称为策略迭代中的贪婪算法,经过有限次迭代后,最终 V 和 π 会收敛至 V^{*} 和 π^{*}。

2 库存优化配置 MDP 模型构建

库存模型由基地级和基层级构成。一个基地级仓库可同时供应多个基层级仓库,基层级仓库之间具有横向供应能力,当基层级出现缺件时,可以根据实际需求,选择由基地级仓库供应,或是由其他基层级仓库横向供应,具有横向供应的航材二级库存模型结构如图 1 所示[9]。

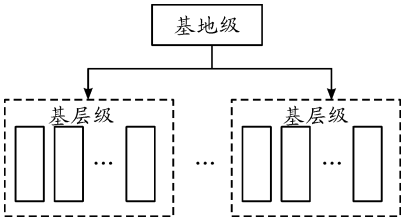

图 1 航材二级库存结构框图

2.1 模型假设

1) 为减少 MDP 求解的时间复杂度,模型简化为 1 个基地级仓库(D_1)供应 2 个基层级仓库(B_1,B_2),基层级仓库间可以横向运输;

2) MDP 按时间离散,以 1 天为离散单位时间,即各级仓库的消耗和供应按天计数;

3) 基层级仓库正常消耗获得收益(R_1,R_2),供应损失(L_1,L_2);

4) 各基层级对航材的消耗需求独立,服从泊松分布,参数分别为 λ_{11},λ_{21};

5) 各基层级接受供应的航材数量独立,服从泊松分布,参数分别为 λ_{12},λ_{22};

6) 各基层级仓库的最大存储数量(S_{MAX_1},S_{MAX_2}),最大供应数量($MOVE_{\mathrm{MAX}}$),整个系统内某器材的总数量(S_{MAX})有上限[10]。

2.2 MDP 参数设定

状态空间 S:2 个基层级仓库各自的器材数量;

行动空间 A:由基地级仓库的直接供应数量和基层级仓库的横向供应数量;

奖励函数 R:正常消耗获得的收益与供应造成的损耗之差;

折扣因子 γ:目前策略对后续策略的影响程度,根据具体需要设定;

状态转移概率 P_{sa}:状态转移行动共包括 4 种,分布是 2 个基层级仓库的消耗(C_1,C_2)和接受供应(S_1,S_2),其概率服从泊松分布,参数 λ 为消耗和接受供应数量均值的倒数,可由航材业务数据统计后得出。

步长:MDP 系统的步长为 1 天,即所有数据按天计算[11]。

3 实例仿真分析

3.1 参数设定

系统内有 1 个基地级仓库,2 个基层级仓库,基层级仓库间满足横向供应条件。基本参数设置如表 1 所示。

表 1 算例参数

参数	参数值	参数	参数值	参数	参数值
R_1	10	λ_{11}	3	S_{MAX_1}	10
R_2	10	λ_{21}	4	S_{MAX_2}	10
L_1	2	λ_{12}	3	$MOVE_{MAX}$	5
L_2	2	λ_{22}	2	S_{MAX}	10
γ	0.9				

3.2 算法流程

算法分为策略评估和策略迭代两部分。首先,生成包含 2 个基层级航材股仓库所有器材配属情况的状态矩阵 $\boldsymbol{S}$,在状态矩阵 $\boldsymbol{S}$ 中的每一个状态 s 上,对每一个可能的行动进行策略评估,计算执行行动产生的最终奖励值,选择每一状态的所有横向供应行动中,结果最优的横向供应行动集合,作为下一次迭代的初始策略。算法退出的条件为两轮策略迭代的奖励值变化小于设定的阈值。算法流程见图 2。

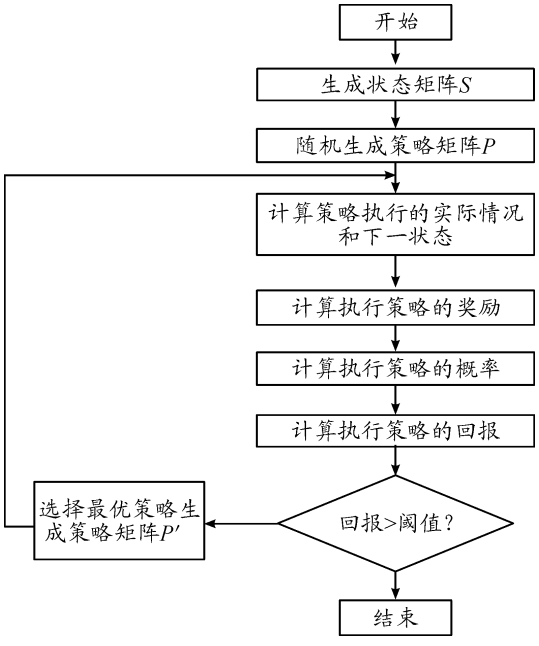

图 2 算法流程框图

3.3 实例计算

3.3.1 计算细节说明

本例需要迭代的次数过多,文章中无法将所有过程完整表示,因此本节选取第一次迭代过程进行计算细节说明。

1) 生成状态矩阵 $\boldsymbol{S}$ 和初始策略矩阵 $\boldsymbol{P}$

在本例的 2 个航材股仓库中,存放该器材的最大数量为 10,即每个航材股的器材存储状态各有 0 – 10 共 11 种,则状态矩阵 $\boldsymbol{S}$ 为 11 × 11 维矩阵。

$$S=\begin{bmatrix}
(0,0) & (0,1) & (0,2) & (0,3) & (0,4) & (0,5) & (0,6) & (0,7) & (0,8) & (0,9) & (0,10)\\
(1,0) & (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) & (1,7) & (1,8) & (1,9) & (1,10)\\
(2,0) & (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) & (2,7) & (2,8) & (2,9) & (2,10)\\
(3,0) & (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) & (3,7) & (3,8) & (3,9) & (3,10)\\
(4,0) & (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) & (4,7) & (4,8) & (4,9) & (4,10)\\
(5,0) & (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) & (5,7) & (5,8) & (5,9) & (5,10)\\
(6,0) & (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) & (6,7) & (6,8) & (6,9) & (6,10)\\
(7,0) & (7,1) & (7,2) & (7,3) & (7,4) & (7,5) & (7,6) & (7,7) & (7,8) & (7,9) & (7,10)\\
(8,0) & (8,1) & (8,2) & (8,3) & (8,4) & (8,5) & (8,6) & (8,7) & (8,8) & (8,9) & (8,10)\\
(9,0) & (9,1) & (9,2) & (9,3) & (9,4) & (9,5) & (9,6) & (9,7) & (9,8) & (9,9) & (9,10)\\
(10,0) & (10,1) & (10,2) & (10,3) & (10,4) & (10,5) & (10,6) & (10,7) & (10,8) & (10,9) & (10,10)
\end{bmatrix}$$

本例中每天横向供应的最大件数 $MOVE_{MAX}$ 为 5,对于航材股 1 而言,有[– 5,5]共 11 种情况,其中正数表示航材股 1 向航材股 2 横向供应,负数反之,0 为该天未发生横向供应,航材股 2 亦是如此,为优化计算,只使用航材股 1 作为策略执行的主体,则策略矩阵同样为 11 × 11 维矩阵,但矩阵的每一行都相同。

$$P[,1]=[-5\quad -4\quad -3\quad -2\quad -1\quad 0\quad 1\quad 2\quad 3\quad 4\quad 5]$$

2) 计算策略实际执行情况和下一状态

以状态矩阵 $\boldsymbol{S}$ 中的状态 $s=(8,6)$ 为例,执行策略矩阵 $\boldsymbol{P}$ 的实际情况和执行后状态如表 2 所示。

3) 计算执行策略产生的回报

在本例中,执行策略的回报包含 2 个部分:横向供应造成的损耗和正常器材供应产生的收益。其中,横向供应的损耗可由表 2 中的实际执行数量与供应损失(L_1,L_2)求出;正常供应产生的收益也可由实际供应数量与供应收益(R_1,R_2)求出。在本例中,仓库正常供应和横向供应的数量是一个泊松过程,实际消耗值(C_1,C_2)分别服从参数为 λ_{11},λ_{21} 的泊松分布,实际横向供应值(S_1,S_2)分别服从参数为 λ_{12},λ_{22} 的泊松分布,即在状态 s 下执行行动 a 的状态转移概率 P_{sa} 为:

$$P_{sa}=\frac{\lambda_{11}^{C_1}}{C_1!}e^{-\lambda_{11}}\times\frac{\lambda_{12}^{C_2}}{C_2!}e^{-\lambda_{12}}\times\frac{\lambda_{21}^{S_1}}{S_1!}e^{-\lambda_{21}}\times\frac{\lambda_{22}^{S_2}}{S_2!}e^{-\lambda_{22}}\qquad(6)$$

表 2 状态 $s=(8,6)$ 执行后状态

行动	– 5	– 4	– 3	– 2	– 1	0	1	2	3	4	5
实际执行	– 2	– 2	– 2	– 2	– 1	0	1	2	3	4	4
下一状态	(10,4)	(10,4)	(10,4)	(10,4)	(9,5)	(8,6)	(7,7)	(6,8)	(5,9)	(4,10)	(4,10)

根据 Bellman 方程(式(2)),执行行动后的奖励函数 R 可表示为(第 1 次迭代):

$$R = \sum P_{sa} \times (C_1 \times R_1 + C_2 \times R_2 - L_1 \times S_1 - L_2 \times S_2) \tag{7}$$

根据 V^* Bellman 方程(式(4)),第 2 次迭代至收敛的奖励函数 R 可表示为:

$$R = R^{-1} + \sum P_{sa} \times (C_1 \times R_1 + C_2 \times R_2 - L_1 \times S_1 - L_2 \times S_2 + \gamma \times R^{-1}(s_{B_1}, s_{B_2})) \tag{8}$$

式(8)中,R^{-1} 表示上一轮迭代得到的最优奖励函数;(s_{B_1}, s_{B_2}) 表示在本轮迭代执行行动后,2 个仓库的实际库存状态;$R^{-1}(s_{B_1}, s_{B_2})$ 表示在上一轮迭代后,执行最优策略得到奖励函数中,状态为 (s_{B_1}, s_{B_2}) 的函数值。将 C_1,C_2 组合,求得在状态 $s=(8,6)$ 中执行行动 $a=2$ 的奖励函数值为 8.825。类似的,在状态 $s=(8,6)$ 下,执行策略矩阵 $\boldsymbol{P}$ 的最终奖励函数如表 3 和图 3。

表 3 执行 $a=2$ 奖励函数值

行动	奖励函数值	行动	奖励函数值
-5	-5.774	1	7.486
-4	-4.609	2	8.825
-3	-2.892	3	8.385
-2	0.062	4	7.254
-1	3.129	5	6.774
0	6.893		

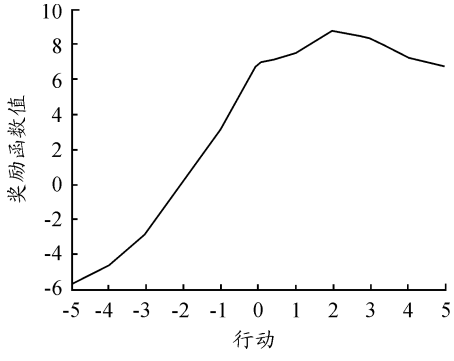

图 3 执行 $a=2$ 奖励函数值

得到在该次迭代下状态 $s=(8,6)$ 的最优策略为 $a=2$。

3.3.2 计算结果

算例经过 3 轮迭代求得最优解,分别得到 1 个策略矩阵和对应的奖励函数值。第一轮迭代中经历 10 次迭代,奖励函数值从 74.910 02 至 111.987 2,误差值从 314.609 7 至 0。第二轮迭代中经历 16 次迭代,奖励函数值从 314.273 132 至 475.865 417,误差值从 959.625 916 至 0。第三轮迭代中经历 14 次迭代,奖励函数值从 483.712 769 至 490.000 885,误差值从 3.641 499 至 0,具体变化如图 4 所示,本轮迭代的奖励函数值差为 0,得到最优策略矩阵 $\boldsymbol{P}^{(1)}$ 为:

$$\boldsymbol{P}^{(1)} = \begin{bmatrix} 0 & 0 & 0 & 0 & -1 & -1 & -1 & -2 & -2 & -3 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -2 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 2 & 2 & 2 & 1 & 1 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 1 & 1 & 1 & 0 & 0 & 0 \\ 3 & 3 & 3 & 2 & 2 & 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 3 & 2 & 2 & 2 & 1 & 1 & 0 \end{bmatrix}$$

即矩阵 $\boldsymbol{P}^{(1)}$ 为最优策略 $\pi^*(s)$,其数值分别对应初始状态矩阵 S 中各状态的最优行动,例如在状态 $s=(8,2)$ 时,行动为 $a=2$,即从仓库 B_1 向 B_2 供应 2 个器材。从策略矩阵的分布来看,在状态矩阵 $\boldsymbol{S}$ 主对角线两侧附近的最优策略是 0,即在两仓库的器材量接近时,不需要进行横向供应,在副对角线上附近的状态值基本对称,且左下角比右上角的横向供应量大,与假设的参数相符。

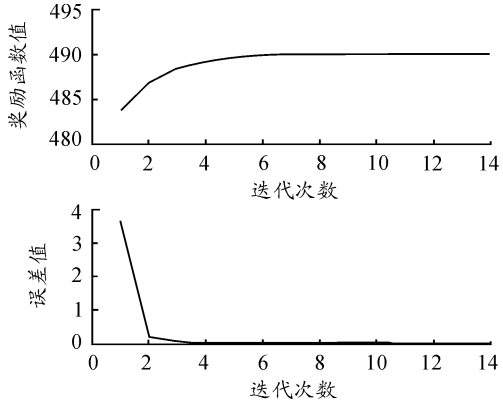

图 4 第三轮迭代中的奖励函数值与误差值

3.4 模型对比

在上节算例的基础上,将器材在仓库 B_1、B_2 的最大库存值设为 100,即初始状态矩阵和最优决策矩阵为 100×100 维。设定在平时保障状态下出现缺货时,该天无法完成正常保障,第二天从另一仓库横向供应后正常保障。两仓库的器材需求到达时间服从参数为 λ_{11} 和 λ_{12} 的泊松分布。设定库存系统的初始状态为 $s=(40,20)$,使用模型前后,各仓库库存情况随保障天数的变化分别如图 5、图 6 所示。

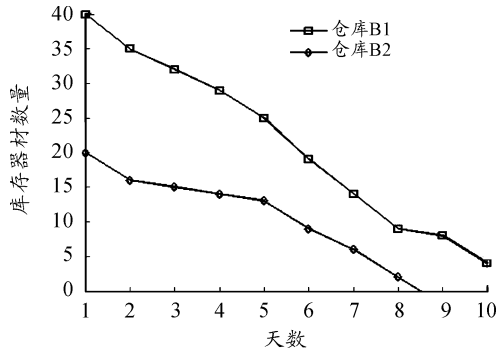

图 5 使用模型前的库存变化

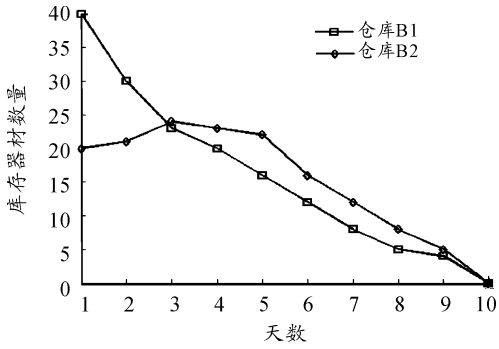

图6 使用模型后的库存变化

从图5和图6的变化情况可知,使用模型前,两仓库没有器材横向供应,库存量随时间逐渐下降,B_1 仓库库存始终大于 B_2 仓库库存,库存系统在第8天后无法进行保障;使用模型后,两仓库在第二天时就发生了横向供应,从第三天开始 B_2 仓库库存大于 B_1 仓库库存,最终在第10天时耗尽所有库存器材。使用模型后,同样的库存系统延长了25%的保障时间,有效提高了航材保障效率。

4 结论

本文利用马尔科夫决策过程建立了航材股仓库间允许横向供应的器材供应离散模型,使用增强学习的思想,对模型的进行策略评估和策略迭代,求解对应整个状态空间下所有状态的最优策略,即在不同器材配置方案下的最优横向供应方案,最终策略与算例假设情况相符,在对使用模型前后的保障情况进行仿真后,相同库存系统的保障时间提升了25%。该模型能够有效解决航材二级库存系统的器材配置问题,此外,还可以针对不同器材调整算例参数,扩展模型的适应范围,减小库存不平衡对航材保障能力的影响,提高部队整体的航材保障效率。

参考文献:

[1] 陈砚桥,魏曙寰,金家善. 允许横向供应单级多点库存系统器材配置研究[J]. 系统工程与电子技术,2013,35(7):1451-1454.

[2] 刘少伟,关娇,王洁. 具有横向供应策略的可维修备件两级库存模型[J]. 兵工学报,2015,36(7):1334-1339.

[3] TOPAN E,BAYINDIR Z P,TAN T. An exact solution procedure for multi-item two-echelon spare parts inventory control problem with batch ordering in the central warehouse [J]. Operations Research Letters,2010(38):454-461.

[4] 肖枝洪,于浩,王一超. 基于动态离差平方和准则的无监督机器学习[J]. 重庆理工大学学报(自然科学),2018(11):134-139,186.

[5] 徐昕,沈栋,高岩青,等. 基于马氏决策过程模型的动态系统学习控制:研究前沿与展望[J]. 自动化学报,2012,38(5):673-687.

[6] 王学宁. 策略梯度增强学习的理论、算法及应用研究[D]. 长沙:国防科技大学,2006:25-30.

[7] RICHARD S S,ANDREW G B. Reinforce Learning:An Introduction[M]. Cambridge:MIT Press. 1998:26-28.

[8] 黄炳强. 强化学习方法及其应用研究[D]. 上海:上海交通大学,2007:19-20.

[9] 郭日红,闫鹏程,孙江生. 可修备件横向供应多级库存模型研究[J]. 物流技术,2011,30(3):139-142.

[10] 林翔. 两主两从博弈下的装备维修器材供应链协调研究[J]. 兵器装备工程学报,2017,38(5):140-143.

[11] 张光宇,李庆民,郭璇. 基于横向转运策略的可修备件多点库存建模方法[J]. 系统工程与电子技术,2012,34(7):1424-1429.

(责任编辑 唐定国)