Journal of System Simulation

Volume 31 | Issue 10 Article 22

12-12-2019

Dynamic Inventory Routing Optimization Based on Deep Reinforcement Learning

Jianpin Zhou
1. School of Navigation, Jimei University, Xiamen 361021, China; ;

Shuliu Zhang
2. Jilin Power Supply Company of State Grid, Jilin 132000, China;

Dynamic Inventory Routing Optimization Based on Deep Reinforcement Learning

Abstract

Abstract: Aiming at the dynamic stochastic inventory routing problem with periodic fluctuation of demand, *a novel simulation optimization approach based on deep reinforcement learning is proposed to achieving periodic steady strategy.* Firstly a dynamic combinatorial optimization model is constructed. Then, *by deep reinforcement learning and setting heuristic rules, the replenishment nodes set selection and the replenishment batch allocation weights in each period are determined.* The simulation experimental results show that the proposed method can improve the average profit of a cycle by about 2.7% and 3.9% in low fluctuating demand case and by about 8.2% and 7.1% in high fluctuating demand case compared with the two solution methods in the existing literature, and the cycle service level can be stabilized within a small fluctuation range under different demand fluctuation environments.

Keywords

inventory routing problem, heuristic rules, deep Q-learning, dynamics, periodic steady strategy

Recommended Citation

Zhou Jianpin, Zhang Shuliu. Dynamic Inventory Routing Optimization Based on Deep Reinforcement Learning[J]. Journal of System Simulation, 2019, 31(10): 2155-2163.

第 31 卷第 10 期 系统仿真学报© Vol. 31 No. 10
2019 年 10 月 Journal of System Simulation Oct., 2019

基于深度强化学习的动态库存路径优化

周建频[1]，张姝柳[2]
(1. 集美大学航海学院，厦门 361021；2. 国网吉林供电公司，吉林 132000)

摘要：针对具有周期性波动需求的动态随机库存路径问题，*提出了基于深度强化学习进行仿真优化并实现周期平稳策略的新方法。*所研究问题构建动态组合优化模型，*通过深度强化学习和设置启发规则来综合决定每个时期的补货节点集合和补货批量分配权重。*仿真实验结果表明，与现有文献中的两种方法相比，所提出的方法在较低波动需求情况下可分别提高一个周期的平均利润约 2.7%和 3.9%，在较高波动需求情况下提高约 8.2%和 7.1%，而周期服务水平在不同需求波动环境下都可以平稳地保持在一个较小的波动范围内。

关键词：库存路径问题；启发规则；深度 Q-学习；动态；周期平稳策略
中图分类号：TP391.9 文献标识码：A 文章编号：1004-731X (2019) 10-2155-09
DOI: 10.16182/j.issn1004731x.joss.18-0820

Dynamic Inventory Routing Optimization Based on Deep Reinforcement Learning

Zhou Jianpin[1], *Zhang Shuliu*[2]
(1. School of Navigation, Jimei University, Xiamen 361021, China; 2. Jilin Power Supply Company of State Grid, Jilin 132000, China)

Abstract: Aiming at the dynamic stochastic inventory routing problem with periodic fluctuation of demand, *a novel simulation optimization approach based on deep reinforcement learning is proposed to achieving periodic steady strategy.* Firstly a dynamic combinatorial optimization model is constructed. Then, *by deep reinforcement learning and setting heuristic rules, the replenishment nodes set selection and the replenishment batch allocation weights in each period are determined.* The simulation experimental results show that the proposed method can improve the average profit of a cycle by about 2.7% and 3.9% in low fluctuating demand case and by about 8.2% and 7.1% in high fluctuating demand case compared with the two solution methods in the existing literature, and the cycle service level can be stabilized within a small fluctuation range under different demand fluctuation environments.

Keywords: inventory routing problem; heuristic rules; deep Q-learning; dynamics; periodic steady strategy

引言

供应链上下游协同要求各驱动要素的组合决策优化，库存与运输是供应链系统的两个主要驱动要素，将库存补货决策与车辆路径规划问题(Vehicle Routing Problem，VRP)进行集成研究就形成了库存路径问题(Inventory Routing Problem，IRP)，随着供应链协同的实践发展，库存路径问题越来越受到重点关注。

由于库存路径问题理论和实践的背景条件不同，对该问题的学术研究包括很多不同的变种[1]。

收稿日期：2018-12-10 修回日期：2019-02-15;
基金项目：福建省自然科学基金(2017J01797，2017J01796)；
作者简介：周建频(1968-)，男，福建，博士，副教授，研究方向为人工智能与供应链系统仿真；张姝柳(1989-)，女，吉林，硕士，助理工程师，研究方向为电气工程与项目管理。

对于较复杂的 IRP 模型，如具有非线性和非凸的目标函数及非线性约束条件等，采用精确算法较难以实现，因此应用启发式算法来求解 IRP 模型成为研究的主要趋势[2]，如文献[3]对于有产能约束的 VRP，提出了基于机会约束和蒙特卡罗仿真的方法，文献[4]采用了改进的 C-W 节约(Clarke & Wright Savings，CWS)算法，文献[5]将原 IRP 分割为两个子问题，然后对每个问题应用启发式方法求解。仿真启发方法(Simheuristics)在启发式方法框架中结合仿真优化方法，可用于处理不确定条件下大规模的和动态的组合优化问题[6-7]，是实现随机 IRP 策略优化的有效途径。

需求随机性在目前研究中受到重视，当前研究重点已从优化一个所有信息都是先验已知的 IRP 转移到具有随机需求的 IRP 问题[8-9]，如文献[8]提出了一种在初始解基础上进行可变邻域搜索方法求解不确定需求下的 IRP 模型，文献[9]分析了在不同约束条件下解决随机需求库存路径问题的最优平稳策略。但在实际中需求模式常随着时间的推移不断变化，即需求信息不仅存在随机的不确定性，还具有动态性和周期性变化的特点，由此产生了动态随机库存路径问题(Dynamic Stochastic Inventory Routing Problem，DSIRP)的概念[10-12]。DSIRP 模型使用动态信息输入，输出随着时间的推移应采取的行动方案。文献[10]采用了一种基于问题分解和特定规则的启发算法(Problem Decomposing and Rule-Specified Algorithm，PD-RSA)，将 DSIRP 分解为 3 个子决策问题：配送客户集合选择、每个客户补货数量决策和配送路径决策，其中前 2 个决策问题是根据设定的特定库存策略和具体规则进行的，但这种方法需加强补充库存与路径规划之间的决策关联。文献[12]针对 DSIRP 问题提出了一种动态前瞻策略，使用一种值函数逼近的动态规划方法并通过模拟未来的需求预期来自动适应需求模式的变化。

动态库存路径问题可以被建模为一个马尔可夫决策过程(Markov Decision Process, MDP)，而强化学习方法如 Q-学习(Q-Learning)是基于马尔可夫决策过程的适应与学习方法，在有限阶段的 MDP 中，Q-Learning 方法被证明具有收敛性，是最终可以找到最优策略的，因此是解决多时期动态随机 IRP 问题的一个可行途径。深度 Q-网络(Deep Q-Networks，DQNs)是 Q-学习和深度神经网络的结合，使用深度神经网络估计 Q 值的合适近似解[13-14]。在 DSIRP 问题中面临的主要挑战是应对大规模需求节点的决策变量状态空间，本论文采用深度强化学习方法并结合基于启发规则(Heuristic Rules，HRs)指导的决策变量空间局部搜索方法，通过启发排序规则对各需求节点的补货批量进行概率分配调整，这些规则根据优化实际决策行为的启发方法建立。在各阶段需求动态变化的情况下，结合 HRs 和 DQNs 进行仿真优化，找到适应环境动态性和周期波动性的周期平稳策略。

1 问题描述

1.1 问题结构与环境

本论文针对需求变化具有周期波动模式的库存路径问题进行研究，按照文献[1]中对 IRP 结构的分类，本论文研究所对应的 IRP 结构是多时期的时间跨度、一对多的供应链结构、每条路径到访多个需求节点、对每个需求节点采取柔性补货策略(补货批量取决于启发规则确定的当期概率分布权重)、允许缺货、使用相同的配送车辆并有车辆容量和数量约束，以及多车辆多回路的运输路径。

假设供应网络由 n 个节点构成，由一个供应节点供应满足 n–1 个需求节点的需求，有一个包括 m 个具有相同载重量车辆的运输车队，每个需求节点有一个最大库存容量限制。假设供应商有足够的库存来满足其客户的需求，客户需求模式随着时间推移呈现动态的周期性变化，以一周为一个循环变化周期，且以一天作为一个阶段时期。设需求节点 i 在时期 t 的需求为 d_{it}，各节点每天产生的需求构成为：需求=(水平需求项+趋势项)×周期波动系数+

随机变化项，对于趋势项，假设不同节点有不同的变化趋势，上升或下降且程度不同；对于周期波动系数项，假设所有节点具有相同的周期季节性，即在周六和周日是需求高峰期，在周二和周三是低峰期；对于随机项用方差估计，假设所有需求节点都有相同的方差系数。

1.2 系统成本构成与决策目标

供应网络的收入来自于各需求节点的销售收入，产生的成本是供需不匹配成本和运输配送成本的总和，其中运输配送成本基于运输距离来计算，供需不匹配成本包括库存过剩成本和库存缺货成本。假设库存过剩成本包括由于在当期没有售出而造成的滞销风险和商品新鲜度下降等，库存过剩会产生单位过剩成本 C_o，其损失远大于单位库存持有成本，因而库存持有成本相对较小可以忽略不计。

供应网络中各需求节点满足需求的程度用周期服务水平(Cycle Service Level，CSL)来衡量，在本论文中用 L_{csl} 表示。对于需求节点，最优周期服务水平的计算公式为：$C_u/(C_u+C_o)$，其中 C_u 为单位缺货成本，C_o 为单位过剩成本，本论文假设只考虑一种产品，在各需求节点的成本价格相同，因此各需求节点具有相同的最优周期服务水平。

需要构建供应网络动态随机库存路径问题的决策模型，使循环周期的总利润最大。模型的决策变量是在一个周期内的各个时期对各需求节点的补货批量，以及在每个时期完成配送任务的车辆路径规划。

2 模型构建

2.1 基本模型

设各需求节点在每个时期的需求由水平项(基础项)、趋势项、随机项和周期波动系数构成，需求水平项、趋势项和随机项分别用 L_{it}、T_{it} 和 R_{it} 表示，波动系数用 $S_1, S_2, \cdots, S_p$ 表示，p 为一个周期内的波动时期数量。则节点 i 在第 t 期的需求由公式(1)给出：

$$d_{it}=\left(L_{it}+T_{it}\right)\times S_t+R_{it} \tag{1}$$

在给定预期的需求均值 d_{it} 和标准差 σ_{it} 条件下，由期望的周期服务水平可确定对应的期望库存水平，需求节点的期望库存水平可表示为期望周期服务水平的反函数，由公式(2)给出：

$$I_{itL}=F^{-1}\left(L_{csl},d_{it},\sigma_{it}\right) \tag{2}$$

各需求节点在第 t 期的利润为其销售收入与库存过剩成本的差额，取决于 t 期的实际库存水平变量 I_{it} 以及需求随机变量 d_{it}，在 t 期各需求节点的利润总和 P_{It} 由公式(3)给出：

$$P_{\mathrm{It}}=\sum_{i=1}^{n-1}Min\left(d_{it},I_{it}\right)C_u-\max\left(0,\left(I_{it}-d_{it}\right)\right)C_o \tag{3}$$

在 t 时期需求节点 i 的补货需求变量 q_{it} 为期望库存水平与 t-1 期末剩余库存的差额，由公式(4)给出：

$$q_{it}=\max(0,I_{itL}-I_{\mathrm{i}(t-1)}) \tag{4}$$

对于每一个补货水平的组合，需要进行一个相应的 VRP 规划，VRP 可被定义在一个完整的无向图 $\boldsymbol{G}$ 上：$\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{A})$，节点集合用 $\boldsymbol{V}=\{0,1,\ldots,n-1\}$表示，其中节点 0 表示供应节点，其余为需求节点。连接节点的线路(弧)用集合 $\boldsymbol{A}$ 表示，用 $\boldsymbol{C}=(c_{ij})$表示与 A 对应的距离成本矩阵，c_{ij} 表示单位距离成本，该矩阵是对称的，即对于任何弧有 $c_{ij}=c_{ji}$。设车辆集合用 $\boldsymbol{K}$ 表示，VRP 的决策变量 x_{ijkt} 取值由公式(5)给出：

$$x_{ijkt}\in\{0,1\},\quad \forall(i,j)\in V,\forall k\in K \tag{5}$$

表示在 t 时期当车辆 k 访问节点 i 后直接访问节点 j 时，变量 x_{ijkt} 取值为 1，否则取值为 0。则 t 时期的总路径成本由公式(6)给出：

$$C_{R\mathrm{t}}=\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}c_{ij}x_{ijkt} \tag{6}$$

总目标函数是使整个供应网络在 T 个时期利润之和最大，应等于 T 个时期的需求节点总利润(需求节点销售收入与库存过剩成本的差额)减去总路径成本，由公式(7)给出：

$$\max\quad P_T=\sum_{t=1}^{T}(P_{\mathrm{It}}-C_{Rt}) \tag{7}$$

决策变量是在各时期对各需求节点的补货批量 y_{it} 和运输路径，运输路径由 VRP 模型的决策变量 x_{ijkt} 确定，补货批量决策与补货需求量、最优周期服务水平、最大库存容量限制以及需求变化趋势等因素相关，根据启发规则设置补货批量的概率分布。根据分配概率 p_{it} 选择和分配节点补货批量由公式(8)给出：

$$y_{it} = q_{it} \times p_{it}\left(r\text{ules}: x_i \in \Omega_t \subset V\right) \tag{8}$$

式中：Ω_t 表示在 t 时期进行补货的节点集合。

模型构建需考虑两个主要方面的协同关系，一个方面是总库存不匹配成本和总配送成本之间的协调平衡关系，即在每个时期内使各需求节点的补货量尽量满足最优服务水平对应的批量要求，同时考虑使补货批量方案能够以较低的运输路径成本完成；另一个方面是在一个决策周期中各个时期之间的协调互动关系，主要是需求淡季和需求旺季之间补货量水平的互动影响，如在需求高峰期到来之前增加库存，而在需求低峰期到来之前减少库存，以使整个周期的供需动态平衡和总利润最大，实现循环周期最优的周期平稳策略。

2.2 应用DQNs方法

由于多时期的动态库存路径问题是多个阶段的组合决策优化，可以被建模为一个马尔可夫决策过程，因此可采用 Q-Learning 方法寻找最优组合决策策略，通过对 Q 值函数的学习和训练来逼近最优策略。Q-Learning 以一个元组(s_t, a_t, r_t, s_{t+1})为训练样本进行循环递归学习方法，其中 s_t 为当前时期的状态，a_t 为当前状态下选择的决策活动(action)，r_t 为执行 a_t 后获得的回报，s_{t+1} 为下一个时期的状态。Q-Learning 训练过程如公式(9)所示：

$$Q_{new} \leftarrow (1-\partial)\bullet Q(s_t, a_t) + \partial\bullet\left(r_t + \gamma\bullet\max Q(s_{t+1}, a_t)\right) \tag{9}$$

式中：γ 为对未来最大 Q 值的折扣系数；a 为学习速率，其中每个时期的回报值 r_t 采用公式(7)计算。

DQNs 将深度神经网络和 Q 强化学习相结合，用深度神经网络代替 Q 值表(Q-table)进行 Q 值估计，可极大地增强 Q-Learning 中对状态-活动(s_t,a_t)空间的处理能力。本论文采用 DQNs 对多时期动态库存路径问题进行决策优化。系统状态(state)由所有节点当期的最优库存水平、初始库存、需求均值、需求方差、需求变化追踪信号以及当期的序列号构成。决策活动(action)包括两部分内容：选择各启发规则的概率权重分配；对各需求节点补货批量分配的概率权重进行调整。

载有 DQNs 系统的智能体(Agent)可从供应网络获得决策活动的状态反馈和回报，进行 DQNs 模型的循环训练和参数更新。在 DQNs 的结构设计中，本研究采用了经验重演(Experience Replay)、双 DQN(Double DQN)和分流 DQN(Dueling DQN)技术进行优化设计，以提高样本利用率、减少输出波动使训练过程稳定，以及将环境对 Q 值影响与决策活动对 Q 值影响区分开来，使学习目标更明确。

3 求解算法

3.1 启发规则描述

根据供应网络情况设置可选的 HRs 如下：

HR1：按各节点最大库存水平与剩余库存量之差的相对水平排序；该水平反映了补货需求的最大程度，对于对应节点的销售收入、库存相关成本和周期服务水平有较大影响。

HR2：按各节点最优库存水平与剩余库存量之差的相对水平排序；该水平反映了补货需求的相对程度，对于对应节点的销售收入、库存相关成本和周期服务水平有较大影响。

HR3：按各节点需求的周期波动程度排序；在不同时期需求分布均衡程度不同，需求分布均衡程度对于利润有较大影响，在需求高峰期大量需求发生的确定性较大。因此，对于周期波动确定性较大的节点应在高峰期增加补货批量的概率权重，在低峰期减少补货批量的概率权重。

HR4：按各节点需求均值变化趋势的追踪信号排序；需求追踪信号反映了节点需求的非周期性变化，即趋势性变化的预期，对于显现上升趋势的节点应分配更大的概率权重。

模型输出的决策 actions 给每个启发规则赋予一个概率选择权重，各规则权重之和等于 1。设有 m 个启发规则，第 j 个规则的概率权重为 ω_j，则选择该规则的概率为：

$$p_{rj} = \omega_j \Big/ \sum_{i=1}^{m} \omega_i \tag{10}$$

在 t 时期对所有需求节点的补货优先度分别按不同规则进行排序，对每个规则分配一个概率权重 p_{rj}，则每个需求节点的补货优先度总权重由公式(11)给出：

$$S_t = p_{r1}S_{r1t} + p_{r2}S_{r2t} + p_{r3}S_{r3t} + p_{r4}S_{r4t} \tag{11}$$

式中：S_{rit} 为对应需求节点排序数的倒数；p_{ri} 为对应规则的权重分配；S_t 为该节点在补货优先度的总权重。以优先度权重作为选择概率，从所有需求节点中选择一个在 t 期进行补货的节点子集合，子集合的总补货批量应满足不超过总车辆运载能力的约束条件。按照补货优先度总权重确定各节点配送需求的优先顺序，将各节点的优先度总权重转换为满足其补送需求量的概率分布权重，并设置一个最小补货批量阈值，补货需求量在阈值之上的节点按概率权重分配其补货批量。最小补货批量阈值需对配送路径成本和需求节点安全库存成本进行权衡设定。确定了各需求节点的补货批量之后，IRP 问题就转化为一个 VRP 问题。

3.2 算法主要步骤

结合启发规则的深度强化学习模型(Heuristic Rule-based Deep Q-Networks，HR-DQNs)进行循环仿真训练的算法步骤如下：

Step 1：初始化供应网络各节点的最大库存水平、最优库存水平、初始库存、需求均值、需求方差、需求变化追踪信号，由各初始值构成节点集合的初始状态(state)。

Step 2：对各节点配送需求优先度进行排序。

Step 2.1：分别按 4 个规则分别对 t 时期的各节点配送需求优先度进行排序。

Step 2.2：为每个规则赋予决策对应的权重，利用公式(10)~(11)计算各节点的总排序值。

Step 3：选择需配送节点和调整配送量。

Step 3.1：按总排序值的优先次序选择配送节点集合并设置各节点的初始配送量，需满足集合中补货量大于最小补货阈值的要求。

Step 3.2：按概率分布调整各节点配送量，根据 actions 对应的调整方式对初始配送量按概率分布进行调整，当累积计划配送量小于总配送能力时，按分配概率权重产生下一节点配送量。

Step 4：根据配送量计划应用启发算法确定 VRP 方案，并按公式(6)计算总配送成本。

Step 5：更新网络状态信息并计算系统总回报。

Step 5.1：根据配送量更新各节点库存水平，根据需求均值和方差产生当期的各节点需求量。

Step 5.2：计算库存过剩或缺货数量，按公式(3)计算需求节点的总赢利，并利用公式(7)更新总回报值(reward)。

Step 5.3：根据市场环境变化更新需求变化差异度，更新各节点当期期末的库存水平，根据需求分布更新下一期的最优库存水平，更新需求均值变化追踪信号，这些更新内容构成更新的 state，返回更新的 reward 和 state 到 DQNs。DQNs 输出仿真决策 action 到下一步的补货批量决策。

Step 6：如果 $t<T$，继续进行下一步仿真循环；否则，计算一个周期的总回报，即供应网络在 T 个时期的总利润和平均 CSL，如果仿真循环数小于总循环数，开始下一个循环周期，否则，仿真结束。

3.3 算法的时间复杂度分析

动态库存路径问题对于时效性要求较高，因此需要考虑算法的时间复杂度问题。在每一步训练或测试中，DQNs 更新状态-活动的 Q 值，输出决策活动选择，其中卷积神经网络层是深度强化学习时间复杂度最大的部分，每一卷积层的时间复杂度是 $O(M^2K^2C_{in}C_{out})$，其中 M 为每个卷积核(kernel)输出特征矩阵的边长，K 为每个卷积核的边长，这里假设输入矩阵和卷积核的形状都是正方形。C_{in} 为输入通道数，也是上一层的输出通道数。C_{out} 为本卷积层具有的卷积核个数，也是输出通道数。回报

(reward)由各需求节点利润和总配送路径成本构成，其复杂度主要体现在求解 VRP 方案，本论文采用 CWS 节约算法，其中确定可节约列表的函数模块为复杂度最大部分，时间复杂度是 $O(N^2)$，N 为网络需求节点数量。由于模型是库存与路径优化的集成，所以求解 HR-DQNs 的时间复杂度为：

$$Time \sim O\left(N^2 N_a \sum_{l=1}^{D} M_l^2 K_l^2 C_{l-1} C_l \right) \tag{12}$$

式中：D 为 DQN 的深度，在本模型 DQN 中使用了 4 个卷积神经网络层，因此 D=4，N_a 为可选决策活动的数量。可见在 DQNs 确定情况下，求解 VRP 的算法对整体的时间复杂度有较大影响。对于 HR-DQNs 方法，可以先将模型进行离线训练，然后使用训练好的模型进行实时应用，这是本方法在应用时效性方面的一个显著优势。

4 仿真实验

设供应网络共有 51 个节点，包括一个供应节点和 50 个需求节点。按照需求的周期性波动程度从低波动到高波动，将仿真环境设置为场景 1、场景 2、场景 3 和场景 4 四个实验场景。为便于比较所采用方法的实验效果，仿真实验应采用已有研究文献中的典型案例数据作为参照标准，但本论文所研究的具有周期波动需求的 DSIRP 难以在已有研究文献中找到适合的案例数据，因此本论文以一个典型 VRP 案例的节点位置和需求数据的一组数据集作为基础样本数据，通过增加需求变化的周期性、趋势性和随机性生成适合本论文研究情况的训练数据集，再以这个案例的另一组数据集作为基础样本数据，以相同的方法生成研究所需的测试数据集。

以一周(7 天)为一个周期，以各节点一天的样本需求数据作为基期需求数据。为产生每天的需求均值，首先对各节点的样本需求加上变化趋势项，以反映不同的趋势性，然后再乘以各天对应的周期波动系数，以反映需求变化的周期波动性。对于随机性用方差反映，为重点突出需求的动态周期性特点，将需求方差设置为较低值，设为需求均值的 0.25 倍。以各时期各节点的需求均值和方差为参数，以截断(需求大于 0)正态分布产生对应的需求数据。各需求节点在 7 个时期(天)的周期波动系数如表 1 所示。采用一个周期内各天需求波动系数的方差系数作为判断波动程度的标准，4 个场景对应的方差系数分别为 0.12、0.24、0.32 和 0.40。

表1 各期需求波动系数

Tab. 1 Demand fluctuation coefficients in periods

时期	1	2	3	4	5	6	7
场景 1	1	0.9	0.9	1	1.1	1.2	1.2
场景 2	1	0.7	0.8	1	1.1	1.4	1.3
场景 3	0.9	0.7	0.8	1	1.3	1.6	1.5
场景 4	0.9	0.6	0.7	1	1.2	1.8	1.6

各需求节点的最大库存容量设为对应基期需求的 1.5 倍，各需求节点的初始库存统一设为 10 个单元。由于供应网络总收益由需求节点总盈利和运输路径成本两部分决定，而路径成本由运输距离乘以一个成本系数确定，为使总库存不匹配成本和 VRP 成本处于相近的数量级以具可比性，经测试将运输成本系数设为 5。根据对配送成本和需求节点安全库存成本影响的权衡，将最小补货批量阈值设为 4 个产品单元。在各需求节点产品销售价为 30 元/单元，进货成本为 13 元/单元，剩余产品残值为 10 元/单元。根据这些成本设置以及考虑 VRP 成本可知，需求节点的最优 *CSL* 应在 0.85 稍偏上的水平。

为对比实验效果，分别采用文献[10]提出的基于问题分解和特定规则算法(PD-RSA)、由文献[8]可变邻域搜索方法和文献[12]动态前瞻规划方法结合而成的基于前瞻的可变邻域搜索(Look Ahead based Variable Neighborhood Search，LA-VNS)方法，以及本论文所提出的 HR-DQNs 方法进行仿真实验。LA-VNS 方法是在每个时期进行下一时期的需求预测，然后以各节点补货需求相对水平的排序结果为概率权重，以此确定各节点补货批量的初始解，在初始解基础上对每个时期进行可变邻域搜索寻找优化方案，搜索迭代循环次数设置为 3 000。为便于比较，3 种方法的 VRP 求解部

分都采用较快速的 CWS 启发算法。对于提出的 HR-DQNs 方法使用 Python 编程语言和 TensorFlow 软件平台建立和训练模型，对 DQNs 模型的训练过程设置为 5 000 个循环周期，每个周期包括 7 个时期，对应一周中的 7 天。其它具体仿真参数设置如表 2 所示。

表2 DQNs训练的参数设置

Tab. 2 Set-up of training parameters for DQNs

学习速率	处理批量/unit	更新频率/cycle	折扣系数	退火步数/step
0.001	32	4	0.99	10 000

系统仿真评估的绩效指标为平均周期总利润(mean-profit)和平均周期服务水平(mean-CSL)。应用训练数据集，HR-DQNs 模型对应于 4 个场景的仿真训练结果分别如图 1~4 所示。

在 4 个仿真场景中分别应用 PD-RSA 算法、LA-VNS 算法和训练后的 HR-DQNs 模型，采用独立的测试数据集进行 50 个周期的测试运行，测试输出的平均绩效结果由表 3 中给出。

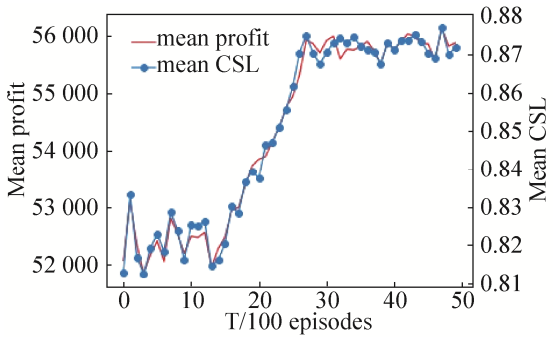

图 1 场景 1 中 DQNs 的训练过程

Fig. 1 Training process of DQNs in scenario 1

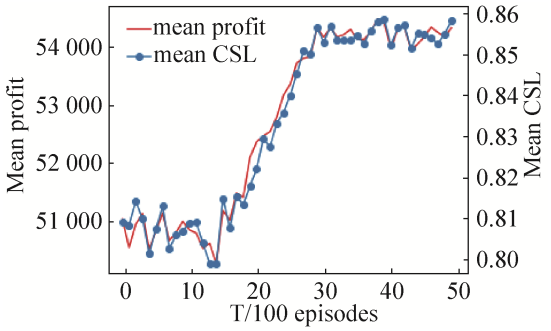

图 2 场景 2 中 DQNs 的训练过程

Fig. 2 Training process of DQNs in scenario 2

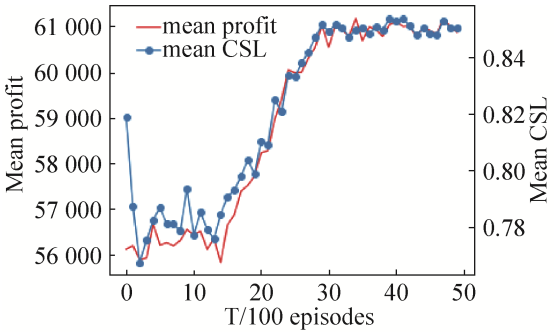

图 3 场景 3 中 DQNs 的训练过程

Fig. 3 Training process of DQNs in scenario 3

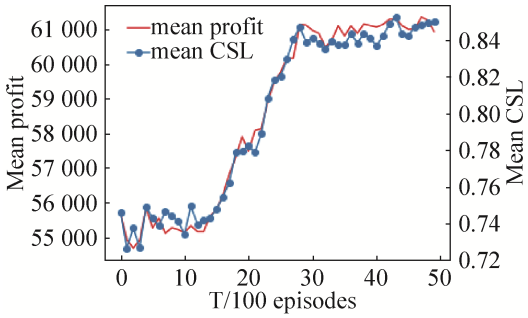

图 4 场景 4 中 DQNs 的训练过程

Fig. 4 Training process of DQNs in scenario 4

表3 仿真测试绩效比较

Tab. 3 Performance comparison of simulation tests

采用方法	场景 1		场景 2		场景 3		场景 4	
	平均利润/元	周期服务水平	平均利润/元	周期服务水平	平均利润/元	周期服务水平	平均利润/元	周期服务水平
PD-RSA	54 932	0.903	58 056	0.834	57 053	0.797	54 047	0.852
LA-VNS	54 287	0.908	58 791	0.831	57 659	0.806	53 989	0.850
HR-DQNs	56 458	0.870	61 384	0.861	61 756	0.862	55 694	0.864

由表 3 得出，在 4 个场景中本文提出的 HR-DQNs 方法获得的平均总利润均高于 PD-RSA 方法和 LA-VNS 方法。由表 1 可以看出，场景 3 和场景 4 的总需求分别比场景 1 和场景 2 的总需求是有所增加的，由表 3 看出，在场景 3 和场景 4 中 PD-RSA 方法和 LA-VNS 方法的利润水平增加较少，而 HR-DQNs 方法的利润增幅较大。在需求波动较低的场景 1 中采用 HR-DQNs 模型的周期平均利润比 PD-RSA 方法和 LA-VNS 方法分别提高了 2.7%和 3.9%，而在需求波动较高的场景 4 中 HR-DQNs 模型的周期平均利润比 PD-RSA 和 LA-VNS 方法分别提高了 8.2%和 7.1%。说明当需求周期波动较大时，本文提出的周期平稳策略效果更显著。另一方面，在 4 个场景中，PD-RSA 方法和 LA-VNS 方法的平均周期服务水平变化较大，随着需求波动程度的增大，服务水平下降较明显，这是因为模型没有考虑到不同时期库存补货决策的关联影响，不能适应需求的波动变化模式。采用 HR-DQNs 模型的平均周期服务水平的波动程度较小，在需求低波动和高波动的不同场景中都保持在 0.86～0.87 之间，是因为模型经过训练后建立了不同时期之间的决策关联关系，能够在需求由淡季到旺季的转换过程中动态调整决策模式，基本上实现了周期平稳策略的效果。

5 结论

本文提出了解决动态随机库存路径问题的深度强化学习启发算法，是首次将深度强化学习方法应用到动态库存路径问题研究中，通过仿真实验测试了 HR-DQNs 方法的有效性，得出主要结论如下：

(1) 针对具有循环波动需求模式的随机库存路径问题，采用 HR-DQNs 方法能够进行多时期权衡决策，比传统方法更能适应需求条件变化的随机性和波动性，并保持服务水平的基本稳定性，实现了周期平稳策略。

(2) 提出的启发规则方法能够约束决策空间的选择，减少决策活动的不确定性，实现有指导的活动空间局部搜索。建立更加系统化的启发规则体系将能够更有针对性地实现策略适应。

(3) 最小补货批量阈值的设定与供应网络的安全库存成本分布有较大关系，选择较小的补货批量阈值会使安全库存更多的集中分布在上游的供应节点，从而降低各需求节点的安全库存水平。

参考文献：

[1] Andersson H, Hoff A, Christiansen M, et al. Industrial aspects and literature survey: Combined inventory management and routing[J]. Computers & Operations Research (S0305-0548), 2010, 37(9): 1515-1536.

[2] Abdelmaguid T F, Dessouky M M, Ordonez F. Heuristic approaches for the inventory-routing problem with backlogging[J]. Computers & Industrial Engineering (S0360-8352), 2009, 56(4): 1519-1534.

[3] Chen L J, Chiang WC, Russell R, et al. The probabilistic vehicle routing problem with service guarantees[J]. Transportation Research Part E (S1366-5545), 2018, 111(3): 149-164.

[4] 林峰, 贾涛, 李然. 基于改进 C-W 算法的易腐品一体化库存路径问题研究[J]. 系统工程, 2016, 34(8): 100-107.
Lin Feng, Jia Tao, Li Ran. Integrated inventory routing problem with vehicle multi-tours for deteriorating item based on modified C-W saving algorithm[J]. Systems Engineering, 2016, 34(8): 100-107.

[5] Chitsaz M, Divsalar A, Vansteenwegen P. A two-phase algorithm for the cyclic inventory routing problem[J]. European Journal of Operational Research (S0377-2217), 2016, 254(2): 410-426.

[6] Juan A A, Grasman S E, Caceres-Cruz J, et al. A simheuristic algorithm for the Single-Period Stochastic Inventory-Routing Problem with stock-outs[J]. Simulation Modelling Practice and Theory (S1569-190X), 2014, 46(8): 40-52.

[7] Rabbani M, Heidari R, Yazdanparast R. A stochastic multi-period industrial hazardous waste location-routing problem: Integrating NSGA-II and Monte Carlo simulation[J]. European Journal of Operational Research (S0377-2217), 2019, 272(3): 945-961.

[8] Gruler A, Panadero J, De Armas J, et al. Combining variable neighborhood search with simulation for the inventory routing problem with stochastic demands and stock-outs[J]. Computers&Industrial Engineering (S0360-8352), 2018, 123(9): 278-288.

[9] 赵达, 李军, 马丹祥, 等. 直接配送下随机需求库存一路径问题最优平稳策略及其算法[J]. 中国管理科学, 2014, 22(6): 61-68.
Zhao Da, Li Jun, Ma Danxiang, et al. Computing the optimal stationary strategy of stochastic demand inventory routing problem with direct deliveries[J]. Chinese Journal of Management Science, 2014, 22(6): 61-68.

[10] Roldan R F, Basagoiti R, Coelho L C. Robustness of inventory replenishment and customer selection policies for the dynamic and stochastic inventory-routing problem[J]. Computers &Operations Research (S0305-0548), 2016, 74(10): 14-20.

[11] Sayarshad H R, Gao H O. A non-myopic dynamic inventory routing and pricing problem[J]. Transportation Research Part E (S1366-5545), 2018, 109(1): 83-98.

[12] Brinkmann J, Ulmer M W, Mattfeld D C. Dynamic Lookahead Policies for Stochastic-Dynamic Inventory Routing in Bike Sharing Systems[J]. Computers and Operations Research (S0305-0548), 2019, 106(6): 260-279.

[13] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature (S1476-4687), 2015, 518(2): 529-533.

[14] Hessel M, Modayil J, Van Hasselt H, et al. Rainbow: combining improvements in deep reinforcement learning[C]. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California USA: the AAAI Press, 2018.