文章编号：2095-2163(2023)10-0060-06 中图分类号：TP181；F253 文献标志码：A

基于 DDQN 的生鲜农产品零售商库存成本控制模型

李姣姣，何利力，郑军红
（浙江理工大学 计算机科学与技术学院，杭州 310018）

摘 要：针对生鲜农产品零售商库存成本控制问题，将该问题转换为马尔可夫决策过程，引入三参数 Weibull 函数，描述生鲜农产品的损腐特征，并考虑过期、损腐、缺货、订货和持有等成本，从供应链视角建立生鲜农产品库存成本控制模型，使用深度强化学习中深度双 Q 网络（Double Deep Q Network，DDQN）优化订货，以控制库存总成本。实验结果表明，相比单周期随机型库存成本控制模型和固定订货量库存成本控制模型，DDQN 模型的总成本分别降低约 6% 和 11%，具有实际应用价值。

关键词：生鲜农产品；深度强化学习；深度双 Q 网络；库存成本控制；供应链；Weibull 分布

Inventory cost control model of fresh agricultural retail based on DDQN

LI Jiaojiao，HE Lili，ZHENG Junhong
（College of Computer Science and Technology，Zhejiang Sci-Tech University，Hangzhou 310018，China）

【Abstract】 In this paper，we solve the inventory cost control problem of fresh agricultural retail by transforming it into a Markovian decision process. A three-parameter Weibull function is introduced to describe the spoilage characteristics of fresh agricultural products，and the costs of expiry，rot，out-of-stock，ordering and holding are considered. We establish a fresh agricultural product inventory cost control model from the perspective of supply chain，and use the Double Deep Q Network（DDQN）in deep reinforcement learning to optimize ordering to control the total inventory cost. The experimental results show that the total cost when using DDQN inventory cost control model is reduced by about 6% and 11% respectively compared with that when using the single-cycle stochastic inventory cost control model and the fixed order quantity inventory cost control model.

【Key words】 fresh agricultural products；deep reinforcement learning；DDQN；inventory cost control；supply chain；Weibull distribution

0 引 言

生鲜农产品包括果蔬、肉类以及水产品等初级产品，在居民日常生活消费中占据重要地位[1]。然而，生鲜农产品具有保质期短、储存困难和损耗率高等特性。发达国家生鲜产品的损腐率约为 5%，而中国果蔬、肉类、水产品损腐率则分别高达 15%、8%、10%，大幅提高了生鲜农产品的成本[2]。冷链物流可以使生鲜农产品在加工、运输、储藏等过程中保持低温状态，从而保证产品质量，减少损耗。而中国果蔬、肉类、水产品冷链流通率仅为 35%、57%、69%[3]。特别的，相较批发商，零售商还存在诸如库存管理粗放，冷库设施不足等问题。因此，建立一个以生鲜农产品为核心的零售商库存成本控制模型具有现实意义。

传统的供应链库存管理模型能够降低库存成本，但在实际运用中存在较大局限性。如：供应商管理库存模型、协同式库存管理模型和联合库存管理模型[4]等管理成本高、操作难度大，ABC 库存管理法和 CVA（Critical Value Analysis）库存管理法均无法给出科学定量的库存控制方案，经济订货批量模型的前提条件较为苛刻等。

强化学习方法可用于研究序贯决策和最优控制问题，近年来有学者将其引入供应链库存控制研究中。蒋国飞等[5]提出基于计数器的直接探索策略，并将该策略和 Q 学习（Q-learning）相结合，解决具有连续状态和决策空间的库存控制问题。Yu 等[6]将多智能体强化学习方法用于解决两级备件库存控制问题，结果表明优于 (s, S) 策略的遗传算法。Bharti 等[7]使用 Q-learning 算法求解一个四阶段串

基金项目：浙江省重点研发计划（2022C01238）。

作者简介：李姣姣（1998-），女，硕士研究生，主要研究方向：供应链库存管理、强化学习；何利力（1966-），男，博士，教授，博士生导师，主要研究方向：数据分析、企业智能；郑军红（1978-），男，博士，讲师，主要研究方向：商务智能、人工智能。

通讯作者：郑军红 Email：zdzhengjh@sohu.com

收稿日期：2022-10-19

行供应链模型，解决订单管理问题。考虑到易腐品不易储存、易腐烂、储藏时间短等特性，Kara 等[8]将强化学习用于易腐烂产品的库存订货策略，结果证明 Q-learning 和 SARSA 算法性能都优于遗传算法。

在前人工作的基础上，本文针对生鲜农产品零售商库存成本问题，将其转换为马尔可夫决策过程，更加全面地考虑费用项。其中包括过期费、缺货费、订货费、持有费和损腐费。另外考虑到在现实生活中生鲜农产品损腐率并非一成不变，使用三参数 Weibull 分布描述损腐率。从供应链的视角，对由一个批发商和一个零售商构成的单级供应链进行分析，运用深度强化学习领域中的 DDQN 方法，制定订货策略以控制库存总成本。

1　算法理论与方法

1.1　强化学习

马尔可夫性质是指将来的状态仅取决于当前状态，而与过去状态无关。马尔可夫决策过程（Markov Decision Process，MDP）满足马尔可夫性质。MDP 状态转移函数为

$$p(s' \mid s,a) = P(S_{t+1} = s' \mid S_t = s, A_t = a) \quad (1)$$

MDP 是强化学习的数学基础，强化学习是基于智能体和环境的交互式学习方法。智能体进行试错学习，通过与环境交互获得的奖励指导动作，找到最优策略，以最大化累计奖励[9]。智能体与环境交互过程如图 1 所示。

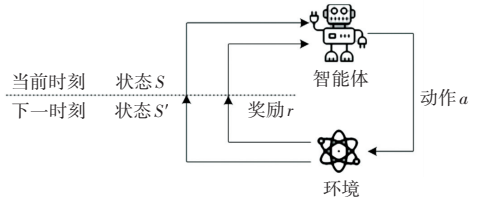

图 1　智能体与环境交互图

Fig. 1　Diagram of interaction between agent and environment

1.2　深度双 Q 网络

Watkins 等[10]提出的 Q-learning 方法，是强化学习中经典的价值迭代算法。Q-learning 通过观测、动作、奖励的历史序列，使智能体能够在马尔可夫域中学习，选择最优行动。Q-learning 中用贝尔曼最优方程进行估值更新。贝尔曼最优方程为

$$Q^*(s,a) = E_{s' \sim p(\cdot \mid s,a)}\left[R(s,a) + \gamma \max_{a'} Q^*(s',a') \mid s,a\right] \quad (2)$$

通过求解该方程，寻找最优价值函数和最优策略。Q-learning 算法的动作价值函数更新迭代式为

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a)) \quad (3)$$

Q-learning 用于复杂的现实问题不仅存在维度灾难问题，还存在自举和最大化导致的非均匀高估问题。因此，Q-learning 在现实中表现并不理想。

Mnih 等[11]提出的深度 Q 网络（Deep Q Networks，DQN）将神经网络和 Q-learning 相结合，其中目标网络和经验回放的设计可以缓解 Q-learning 自举导致的非均匀高估。经验回放降低了样本间的相关性，目标网络则可以减弱预测 Q 值和目标 Q 值间的相关性。DQN 预测网络的优化目标为

$$\tilde{y} = r + \gamma Q(s', \arg\max_{a'} Q(s',a';\theta');\theta') \quad (4)$$

Van Hasselt 等[12]在 DQN 的基础上，将动作选择和动作 Q 值估计分离，提出了 DDQN，进一步缓解了 Q-learning 最大化造成的高估。DDQN 中使用目标网络获取最优动作，再通过预测网络估计该动作的 Q 值。DDQN 预测网络优化目标：

$$\tilde{y} = r + \gamma Q(s', \arg\max_{a'} Q(s',a';\theta);\theta') \quad (5)$$

1.3　三参数 Weibull 函数描述易损腐物品

三参数 Weibull 分布是概率论中的一种连续型分布，被广泛应用于电子元器件的失效情况、物品的变质和拟合度模拟等诸多方面[13]。本文采用三参数 Weibull 函数描述生鲜农产品的损腐特征。

三参数 Weibull 累积分布函数和概率密度函数的公式分别为：

$$F(t) = 1 - e^{-\alpha(t-\gamma)^{\beta}} \quad (6)$$

$$f(t) = \alpha\beta(t-\gamma)^{\beta-1} e^{-\alpha(t-\gamma)^{\beta}} \quad (7)$$

式中：α、β、γ 分别是三参数 Weibull 函数的尺度因子、形状因子和位置因子，t 为时间。

1.4　单周期随机型库存成本控制模型

单周期随机型库存成本控制模型是运筹学存储论中的一种库存模型[14]。单周期是指上一期剩余的库存不会转结到下一期，而多周期则与之相反。

设：货物需求 r 是连续随机变量，密度函数为 $\Phi(r)$，k 为单位货物进价，p 为售价，C_1 为存储费。分布函数为

$$F(a) = \int_0^a \Phi(r)\,dr (a > 0) \quad (8)$$

货物存储费为

$$C_1(q) = \begin{cases} C_1(q-r) & r \leqslant q \\ 0 & r > q \end{cases} \quad (9)$$

最佳订货量 q 满足

$$F(q) = \int_0^q \Phi(r)\,dr = \frac{p-k}{p+C_1} \quad (10)$$

此时,库存总成本最低。

2 研究内容

2.1 业务模型

供应链管理协调成员企业间合作关系,控制物流、资金流和信息流 3 个关键流,涉及需求、生产运作、物流及供应 4 个领域,具有交叉性、需求导向性、动态性等特征[15]。

如图 2 所示,整条供应链由供应商、制造商、批发商、零售商和顾客组成,本文主要研究批发商、零售商和顾客这 3 个实体。

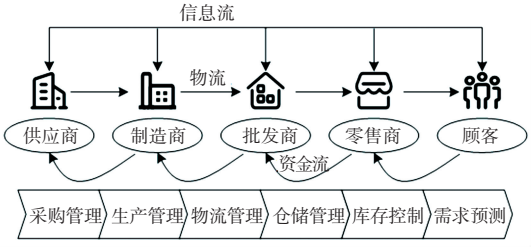

图 2 供应链模型

Fig. 2 Supply chain model diagram

假设:顾客需求 $D \sim N(\mu, \sigma^2)$,商品售价为 p,则期望总收入为 $p\mu$,是一个与库存数量无关的常数。因缺货导致失去销售机会而未实现的收入是潜在损失,是一种机会成本,定义为缺货成本。定义库存总成本为缺货成本加实际成本,则利润等于期望总收入减去库存总成本。因期望总收入为常数,则库存总成本越低,利润越高。

为满足顾客需求,零售商每天向批发商提交订货订单,每天都更新一次库存。批发商每天向零售商提供货物,批发商的商品数量无限。商品订货提前期用 m 表示,订货提前期表示零售商发出订单到收到货物的时间。商品生命周期用 l 表示,商品被零售商接收后,就进入生命周期,生存期也开始增加。

销售产品使用先进先出策略,即先卖生存期大的产品以满足客户需求。若商品生存期大于生命周期 l, 就产生过期成本;若商品生存期在损腐时期内就产生损腐成本;若商品无法满足顾客需求,就产生缺货成本。DDQN 库存成本控制模型如图 3 所示。

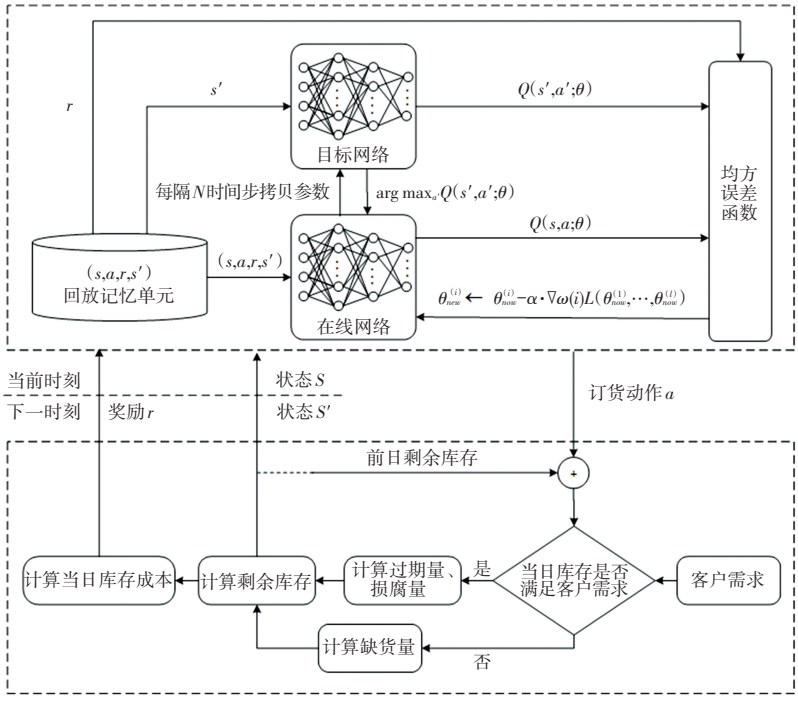

图 3 DDQN 库存成本控制模型图

Fig. 3 DDQN inventory cost control model diagram

具体业务流程如下:

(1)零售商将上一日订购的商品入库,并更新库存。

(2)零售商接收客户需求,如果能满足需求则计算是否产生过期量和损腐量;否则产生缺货量。

(3)零售商计算当日剩余库存量和库存成本,并更新库存。

(4)零售商根据 DDQN 库存成本控制模型制定

的订货策略，向批发商发送次日订货量

生鲜农产品属于易损腐类商品，损腐率使用非线性函数 $\mu(t)$，其计算公式为

$$\mu(t)=\frac{f(t)}{1-F(t)}=\alpha\beta\,(t-\gamma)^{\beta-1} \tag{11}$$

其中，$1<\beta<2$ 且 $\gamma>0$，损腐率变化情况如图 4。物品进入库存系统不会立即损腐，而是经过一段时间才会损腐，该参数设置适合时滞或者易损腐物品库存模型。

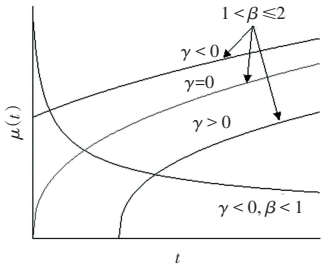

图 4　三参数 Weibull 函数损腐率随时间变化情况

Fig. 4　Change of decay rate of three-parameter Weibull function with time

2.2　DDQN 算法模型

本文使用 DDQN 方法解决生鲜农产品零售商库存成本控制问题，下面分别对状态空间、动作空间、回报函数进行设计。其中数学符号定义见表 1。

表 1　数学符号表示

Tab. 1　Mathematical symbol

数学符号	含义
s_i^t	第 t 天生存期为 i 的产品数量
p	单位售价
k	单位进价
c_e	单位过期费
c_p	单位缺货费
c_o	一次订货固定订货费
c_h	单位持有费
c_s	单位损腐费
n_e^t	第 t 天的过期数量
n_p^t	第 t 天缺货数量
d^t	第 t 天顾客需求数量
x_i^t	第 t 天生存期为 i 的损腐数量

2.2.1　状态空间设计

在 MDP 问题中，状态信息代表智能体感知到的环境信息及其动态变化。如果产品当前处于生命周期中，但产品数量不足，则认为是缺货；如果产品有库存但不在生命周期内，则视为过期。产品生存期在损腐时期内就以一定比例进行损腐。满足库存充足和生命周期要求的产品视为可供销售。l 为产品生命周期，t 天的状态变量为 $(l+3)$ 维向量。状态空间可表示为

$$s^t=[s_0^t,\cdots,s_i^t,\cdots,s_l^t,n_e^{t-1},n_p^{t-1}] \tag{12}$$

2.2.2　动作空间设计

动作是指由智能体发出的行为和动作，以及智能体与环境之间发生的动作交互。对于特定任务而言，动作空间在事实上决定任何算法所能达到的性能上限。顾客需求 $D\sim N(\mu,\sigma^2)$，需求数据分布在 $(\mu-3\sigma,\mu+3\sigma)$ 的概率是 99.73%，因此设 q^t 为订货数量，$q^t\in[0,\mu+3\sigma]$ 取整数。动作空间可表示为

$$a^t=q^t \tag{13}$$

2.2.3　回报函数设计

在强化学习任务中，智能体根据探索过程中来自环境的反馈信号持续改进策略，这些返回信息被称为回报。零售商满足完需求后，剩余库存量为

$$n_r^t=\Big(\sum_{i=0}^{l-1}s_i^t-n_e^t-d^t-\sum_{i=\gamma}^{l-1}x_i^t\Big)^+ \tag{14}$$

回报函数可表示为

$$r^t=c_en_e^t+c_pn_p^t+c_o+kq^t+c_hn_r^t+c_s\sum_{i=\gamma}^{l-1}x_i^t \tag{15}$$

3　实验与评测

3.1　实验设计

根据上述模型与算法分析，首先对算法的神经网络进行设置。设置经验池容量大小 N 为300 000，每回合将随机从中采样；折扣率设为 0.95；更新目标网络的间隔设为 1 周期；使用 ε - greedy 探索策略，在训练开始时以概率 $\varepsilon=0.9$ 随机选择动作，此时探索力度最大；随着训练进行，ε 逐渐线性下降直至最终的 $\varepsilon=0$。

在这个过程中，DDQN 库存成本控制模型训练逐渐从“强探索弱利用”过渡到“弱探索强利用”。结合单周期随机型库存成本控制模型和固定订货量库存成本控制模型，对比 DDQN 库存成本控制模型能否有效降低生鲜农产品库存总成本。

实验以白菜为例，跟据 2022 年国家统计局数据得知白菜各种参数值见表 2。以 1 000 天为一个周期，每天仅进行一次发送订单和入库操作，库存总成本为 1 000 r^t。取 $\alpha=0.2$、$\beta=1.5$、$\gamma=1$，损腐率 $\mu(t)=0.3(t-1)^{0.5}$。

表 2 实验参数

Tab. 2 Experimental parameters

参数	数值
进价	0.65 元/500 g
售价	1.4 元/500 g
过期费	0.65 元/500 g
缺货费	1.4 元/500 g
持有费	0.2 元/500 g
损腐费	0.65 元/500 g
固定订货费	1 元/次

为了验证模型的有效性及实用价值,选择固定订货量库存成本控制模型和单周期随机型库存成本控制模型采用定期定量订货法,深度强化学习模型采用 DDQN 方法进行对比实验,3 种模型参数(如安全库存、订货提前期、产品生命周期、损腐率等)均一致。

假设客户需求数据服从正态分布,每个实验周期为1 000 天,每天仅进行一次发送订单和入库操作。将成本汇总得出结论。

3.2 结果分析

图 5 和图 6 分别为在不同需求函数下,3 种库存成本控制模型在相同条件下的奖励值变化曲线。从图中可以看出,在训练初始阶段,由于动作网络均处于动作探索阶段,因此 DDQN 库存成本控制模型奖励值较低,且存在较大波动。随着智能体开始从经验池中提取历史数据进行学习,奖励值逐渐呈现明显上升趋势。图 5 中,在 50 周期左右时,DDQN 库存成本控制模型逐渐收敛于-6.94 万元;图 6 中,在 350 周期左右时,DDQN 库存成本控制模型逐渐收敛于-68.39 万元,优于固定订货量库存成本控制模型和单周期随机型库存成本控制模型。

见表 3,当需求服从正态分布 $N(100,10^2)$ 时,DDQN 库存成本控制模型的总成本相对于单周期随机型库存成本控制模型和固定订货量 100 库存成本控制模型的总成本降低 5.89%和 10.04%;当需求服从正态分布 $N(1\,000,100^2)$ 时,DDQN 库存成本控制模型的总成本相对于单周期随机型库存成本控制模型和固定订货量1 000库存成本控制模型的总成本降低 6.50%和 11.57%。固定订货量库存成本控制模型损腐量最多,DDQN 库存成本控制模型没有损腐量。可以看出,DDQN 库存成本控制模型不仅优于单周期随机型库存成本控制模型和固定订货量库存成本控制模型,且能够解决维度灾难问题。

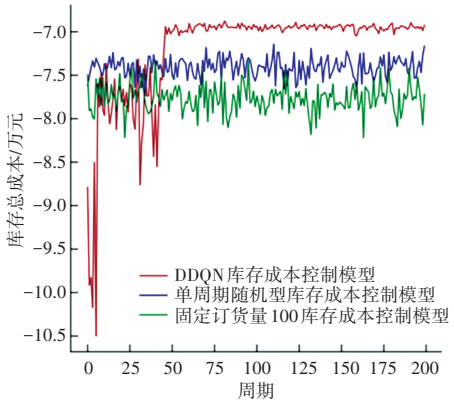

图 5 需求 $D \sim N(100,10^2)$ 时库存成本控制模型实验结果

Fig. 5 Experimental results of inventory cost control model for demand $D \sim N(100,10^2)$

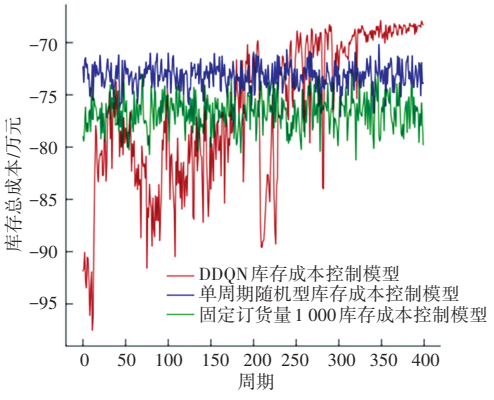

图 6 需求 $D \sim N(1000, 100^2)$ 时库存成本控制模型实验结果

Fig. 6 Experimental results of inventory cost control model for demand $D \sim N(1\,000,100^2)$

表 3 库存成本控制模型实验对比结果

Tab. 3 Comparison results of different inventory cost control models

需求函数	库存成本控制模型	库存总成本/万元	损腐量/500g
$D \sim N(100, 10^2)$	DDQN 库存成本控制模型	6.94	0.00
	单周期随机型库存成本控制模型	7.38	119.95
	固定订货量 100 库存成本控制模型	7.72	350.23
$D \sim N(1\,000, 100^2)$	DDQN 库存成本控制模型	68.39	0.00
	单周期随机型库存成本控制模型	73.15	1 283.65
	固定订货量 1 000 库存成本控制模型	77.33	4 700.01

注:表中数据为最后 10 周期数据的平均值。

(下转第 72 页)