doi: 10.3969/j.issn.1007-7375.220247

基于改进 DQN 算法的无人仓多 AGV 路径规划

谢　勇[1]，郑绥君[1]，程念胜[2]，朱洪君[1]

(1. 华中科技大学 人工智能与自动化学院，湖北 武汉 430074；2. 航天信息股份有限公司，北京 100195)

摘要: 针对无人仓中多 AGV 路径规划与冲突问题，以最小化总行程时间为目标，建立多 AGV 路径规划模型，提出一种基于动态决策的改进 DQN 算法。算法设计了基于单 AGV 静态路径规划的经验知识模型，指导 AGV 的学习探索方向，提前规避冲突与障碍物，加快算法收敛。同时提出基于总行程时间最短的冲突消解策略，从根本上解决多 AGV 路径冲突与死锁问题。最后，建立无人仓栅格地图进行仿真实验。结果表明，本文提出的模型和算法较其他 DQN 算法收敛速度提升 13.3%，平均损失值降低 26.3%。这说明该模型和算法有利于规避和化解无人仓多 AGV 路径规划冲突，减少多 AGV 总行程时间，对提高无人仓作业效率具有重要指导意义。

关键词: 多 AGV；路径规划；DQN 算法；经验知识；冲突消解

中图分类号: F406.2; TP24　　**文献标志码:** A　　**文章编号:** 1007-7375(2024)01-0036-09

Multi-AGV Route Planning for Unmanned Warehouses Based on Improved DQN Algorithm

XIE Yong[1], ZHENG Suijun[1], CHENG Niansheng[2], ZHU Hongjun[1]

(1. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China;
2. Aerospace Information Co., Ltd, Beijing 100195, China)

Abstract: To solve the problem of multi-AGV route planning and conflicts in unmanned warehouses, with the objective of minimizing the total travel time, a multi-AGV route planning model is established, and an improved DQN algorithm based on dynamic decision-making is proposed. An empirical knowledge model based on static route planning of a single AGV is designed to guide the learning and exploration direction of AGVs. It avoids conflicts and obstacles for AGVs in advance, and accelerates the convergence of the proposed algorithm. Also, a conflict resolution strategy based on the shortest total travel time is proposed to fundamentally solve the problem of multi-AGV route conflicts and deadlocks. Finally, a grid map of an unmanned warehouse is established for simulation experiments. Results show that, compared with other DQN algorithms, the convergence speed of the proposed model and algorithm is increased by 13.3%, and the average loss value is reduced by 26.3%. This result indicates that the model and algorithm are conducive to avoiding and resolving the conflicts of multi-AGV route planning in unmanned warehouses, reducing the total travel time of multiple AGVs and having important guiding significance to improve the efficiency of unmanned warehouse operations.

Key words: multiple AGVs; route planning; DQN algorithm; empirical knowledge; conflict resolution

随着科技的发展和人们生活水平的提高，物流产业蓬勃发展，据统计，2021 年全国社会物流总额 335.2 万亿元，同比增长 9.2%。物流产业的飞速发展促进了物流设备智能化和生产制造柔性化的转型，各大物流业龙头纷纷在各地建立起自己的自动化分拣仓库 (无人仓) [1]。在调度系统的支持下，大量的移动机器人在无人仓中有序并高效地完成货物搬运任务。本文的研究对象是无人仓中的移动机器

收稿日期: 2022-12-08

基金项目: 国家自然科学基金资助面上项目 (71771096)；国家自然科学基金创新群体资助项目 (71821001)

作者简介: 谢勇 (1974—)，男，湖北省人，副教授，主要研究方向为智慧物流、优化调度、智能制造。

通讯作者: 郑绥君 (1998—)，女，湖南省人，硕士研究生，主要研究方向为多智能体的调度优化与路径规划。
Email：zheng893724451@163.com

人——自动导引小车 (automated guided vehicle, AGV) 实现货物的搬运。通过优化多 AGV 的搬运路径与冲突消解大大提高 AGV 的搬运效率、缩短搬运时间，这将对整个调度系统的运行效率和经济效益产生深远影响。

目前常用的 AGV 路径规划算法包括遗传算法、蚁群算法、Dijkstra、A*算法及其混合算法等。王秀红等[2] 针对仓储物流移动机器人路径规划提出改进 A*算法，但不考虑机器人碰撞情况。Yang 等[3] 结合蚁群算法和动态窗口算法，提出一种增强混合算法解决移动机器人在复杂动态环境下的路径规划问题。Zhong 等[4] 提出一种结合 A*算法和自适应窗口方法的混合路径规划方法，用于大型动态环境下移动机器人的全局路径规划、实时跟踪和避障。启发式算法在解决单 AGV 路径规划的问题上体现出很大的优势，但在解决多 AGV 的路径规划问题中，启发式算法大多需要与时间窗相结合以解决多 AGV 间的冲突，时间窗的计算大大增加了计算量，而强化学习对环境的试错式探索与反馈评价的方式，使多 AGV 系统具有环境学习和适应性，利于多 AGV 进行全局路径规划。Yang 等[5] 针对无人仓调度问题，提出 A*算法与强化学习算法相结合的多机器人路径规划算法。Guo 等[6] 将深度强化学习与人工势场相结合，为无人船舶在未知海上环境下进行智能路径规划提供解决方案。Gao 等[7] 基于随机抽样建立无冲突的路径图，并结合 Q-learning 强化学习算法进行多机器人的路径规划。但是上述文献基本都是在环境信息全部已知或特定环境下进行的，依赖地图信息准确性。而本文旨在保留强化学习算法的环境学习能力，在只提供部分环境信息的前提下，通过环境学习获取障碍物和路径信息，研究一种更具有环境适应性的多 AGV 路径规划的方法，为无人仓多 AGV 路径规划提供一种不需要实时高精度地图信息的低成本解决思路与方案。

本文针对无人仓多 AGV 路径规划与冲突问题，在只提供起点和目标点位置的前提下，提出一种基于经验知识和冲突消解策略的改进 DQN 算法进行多 AGV 路径规划，AGV 通过静态路径规划的环境学习获取路径和障碍物信息以构建经验知识模型，通过冲突消解策略解决多 AGV 冲突，并设计实验验证算法的有效性。

1　无人仓多 AGV 路径规划模型

1.1　问题描述与模型假设

本文研究无人仓多 AGV 全局路径规划，主要解决 3 个问题：1) 每台 AGV 在路径规划过程中避免与货架 (可视为静态障碍物) 碰撞；2) 在路径规划过程中，考虑 AGV 之间的路径冲突，避免 AGV 与其他 AGV (可视为动态障碍物) 碰撞；3) 多 AGV 总路径行程时间最短。根据实际无人仓环境，借鉴文献 [1]，抽象出一类典型的无人仓环境模型，并用栅格地图形式表示，如图 1 所示。无人仓环境模型主要包括：1) 包裹到达区，传输并放置待搬运的包裹；2) 停靠区，所有的 AGV 从停靠区装载包裹后出发，空闲 AGV 停在停靠区；3) AGV，可在无人仓的道路上通行，并完成包裹搬运任务；4) 道路，白色栅格，供 AGV 通行；5) 货架，根据包裹的分类规则，在无人仓中有规律地设立多个货架。

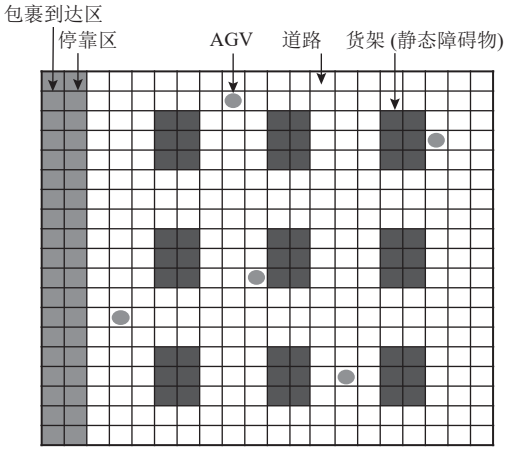


图 1　无人仓环境模型

Figure 1　An unmanned warehouse environment model

结合上述环境模型与实际问题背景，提出如下假设。

1) 所有 AGV 在搬运过程中速度一致，匀速通行。

2) 所有 AGV 只能横向或纵向通行，不能斜向通行。

3) 不考虑 AGV 转弯的时间和半径，即转弯和直行的时间与路径长度一致。

4) 每一台 AGV 只能搬运一个包裹，只搬运到一个目标点。

5) 不考虑 AGV 返程，先到达目标点的 AGV 将原地卸载并等待其他 AGV 完成搬运任务。

6) 不考虑包裹装载时间，所有 AGV 在零时刻

出发。

7) AGV 搬运任务事先已分配好，即 AGV 起点和目标点位置已给定。

1.2 模型构建

基于上述假设，本模型定义以下符号与集合。

s_t^i：第i台 AGV 在t时刻状态，$s_t^i=(x_t^i,y_t^i)$，x,y为位置坐标，$i\in I=\{1,2,\cdots,n\}$，n为 AGV 总数；

k：时间索引号，$k\in K=\{1,2,\cdots\}$，t_1表示初始时刻，t_{k_i}表示第i台 AGV 到达目标点的时刻；

η_i：第 i 台 AGV 搬运包裹过程中静止等待的次数；

S_t：t时刻的 AGV 联合状态集，$S_t=\{s_t^1,s_t^2,\cdots,s_t^n\}$；

$\hat{A}$：AGV 可选动作集，$\hat{A}=\{a_1,a_2,\cdots,a_m\}$，$m$为 AGV 可选动作总数，本文设置 AGV 可选动作集为$\hat{A}=\{a_1,a_2,a_3,a_4,a_5\}$，其中，$a_1$为静止等待，$a_2$为向上，$a_3$为向下，$a_4$为向左，$a_5$为向右；

A：AGV 联合动作集合，$A=\{(a_1^1,a_1^2,\cdots,a_1^n),(a_2^1,a_1^2,\cdots,a_1^n),\cdots\}$，$t$时刻 AGV 的联合动作$A_t\in A$；

R_t：t时刻所有 AGV 获得奖励的集合，$R_t=\{r_t^1,r_t^2,\cdots,r_t^n\}$，$r_t^i$为第$i$台 AGV$t$时刻与环境交互后获得的奖励值，具体见式 (10)；

G：静态障碍物集$G=\{g_1,g_2,\cdots,g_w,\cdots\}$，$g_w=(x_{g_w},y_{g_w})$为障碍物位置坐标；

t_c：AGV 路径规划中状态转移的单位时间步长，即单步搬运或静止等待的时间，为一常数；

$\bar{p}^i$：第i台 AGV 静态最优路径，$\bar{p}^i=(\bar{s}_1^i,\bar{s}_2^i,\cdots,\bar{s}_t^i,\cdots,\bar{s}_k^i)$，其中，$k$为第$i$台 AGV 的最优路径长度；

p^i：第i台 AGV 的最优路径，$p^i=(s_1^i,s_2^i,\cdots,s_t^i,\cdots,s_e^i)$，其中，$e$为第 i 台 AGV 的最优路径长度。

基于以上符号，设路径规划总耗时为T，最小化多 AGV 的总行程时间可表示为

$$\min T=\min\sum_{i=1}^{n}\left(\sum_{k=1}^{k_i}h_{t_{ik}}t_c+\eta_i t_c\right)。\tag{1}$$

s.t.

$$(x_{\text{end}}^i,y_{\text{end}}^i)\neq(x_{\text{end}}^j,y_{\text{end}}^j),i,\ j\in I,\ i\neq j;\tag{2}$$

$$h_{t_{ik}}=\begin{cases}1, & t_k\text{时刻,第}i\text{台AGV进行单步搬运;}\\ 0, & \text{其他;}\end{cases}\tag{3}$$

$$s_t^i\notin G,\ i\in I。\tag{4}$$

式 (1) 为目标函数，表示最小化所有 AGV 的路径规划完成时间，其中，括号中的表达式为每台 AGV 的搬运时间与等待时间之和。式 (2) ~ 式 (4) 为约束条件，式 (2) 表示每台 AGV 的目标点互不相同；式 (3) 用于计算 AGV 的搬运时间；式 (4) 表示任何 AGV 在任何时刻都不能与货架 (静态障碍物) 相撞。

2 改进 DQN 算法的多 AGV 路径规划

2.1 多 AGV 路径规划算法框架

本文将针对上述模型基于深度强化学习算法研究多 AGV 的路径规划问题。深度强化学习算法 DQN (deep Q-learning network, DQN) 综合了神经网络的强感知能力和 Q-learning 算法[8]的强决策能力，能从高维、大量数据中提取特征，指导决策的制定和执行。DQN 算法使用神经网络映射 Q 值表，无论状态和动作空间有多大，都可以用神经网络的输出得到 Q 值，解决多 AGV 的 Q 值表维数爆炸问题。为了减少 AGV 通信的计算成本，同时考虑 AGV 间的协调工作，考虑使用联合动作和联合状态进行环境学习与路径规划，在此过程中多 AGV 间没有进行通信，所有信息都通过集中控制系统进行共享。设第i台 AGV 在t时刻的状态为s_t^i，它执行动作a_t^i后达到另一状态s_{t+1}^i，同时获得环境反馈的奖励值r_t^i。t时刻，每台 AGV 采取的动作组成联合动作$A_t=\{a_t^1,a_t^2,a_t^3,\cdots,a_t^n\}$，AGV 的联合状态为$S_t=\{s_t^1,s_t^2,\cdots,s_t^n\}$，获得的奖励集合为$R_t=\{r_t^1,r_t^2,\cdots,r_t^n\}$，$Q$ 值的迭代公式为

$$Q^{(K+1)}(S_t,A_t)\leftarrow Q^{(K)}(S_t,A_t)+\alpha\left[R_t+\gamma\max Q^{(K)}(S_{t+1},A_{t+1})-Q^{(K)}(S_t,A_t)\right]。\tag{5}$$

其中，α为学习率；γ为折扣因子；K为迭代次数。以联合状态S_t为输入，输出每个动作的 Q 值，即输出一个包含所有动作的 Q 值向量，如式 (6) 所示。

$$Q(S_t,A_t)=\{Q\{s_1,a_1\},Q\{s_1,a_2\},\cdots,Q\{s_1,a_m\},Q\{s_2,a_1\},\cdots,Q\{s_n,a_m\}\}。\tag{6}$$

DQN 算法核心是优化一个损失函数[9]，即最小化目标值与神经网络输出值的偏差，损失函数定义为

$$L(\omega)=E[(R_t+\gamma\max Q(S_{t+1},A;\omega^-)-Q(S_t,A_t;\omega))^2]。\tag{7}$$

令$y=R_t+\gamma\max Q(S_{t+1},A;\omega^-)$为目标 Q 值，$Q(S_t,A_t;\omega)$为当前神经网络输出的 Q 值，使用梯度下降法训练神经网络权重参数ω。DQN 算法为了满足神经网络训练数据独立同分布的要求，引入经验回放

和双层网络结构[10]的方法将强化学习和深度学习算法更好地结合起来。但直接运用传统 DQN 算法并不能解决神经网络训练数据庞大、环境探索难度大、算法收敛速度慢、无法从根本上解决多 AGV 冲突与死锁等问题。对此，本文在 DQN 算法中引入动态决策来解决这些问题：1) 增加经验知识，进行预先路径规划，避免静态障碍物碰撞，同时降低动作选择随机性和环境探索的难度，加快算法收敛；2) 改进奖惩函数、设计 AGV 冲突消解策略，以解决多 AGV 冲突和死锁问题，并使多 AGV 的路径规划完成时间最短。基于 DQN 算法的多 AGV 路径规划算法架构如图 2 所示。

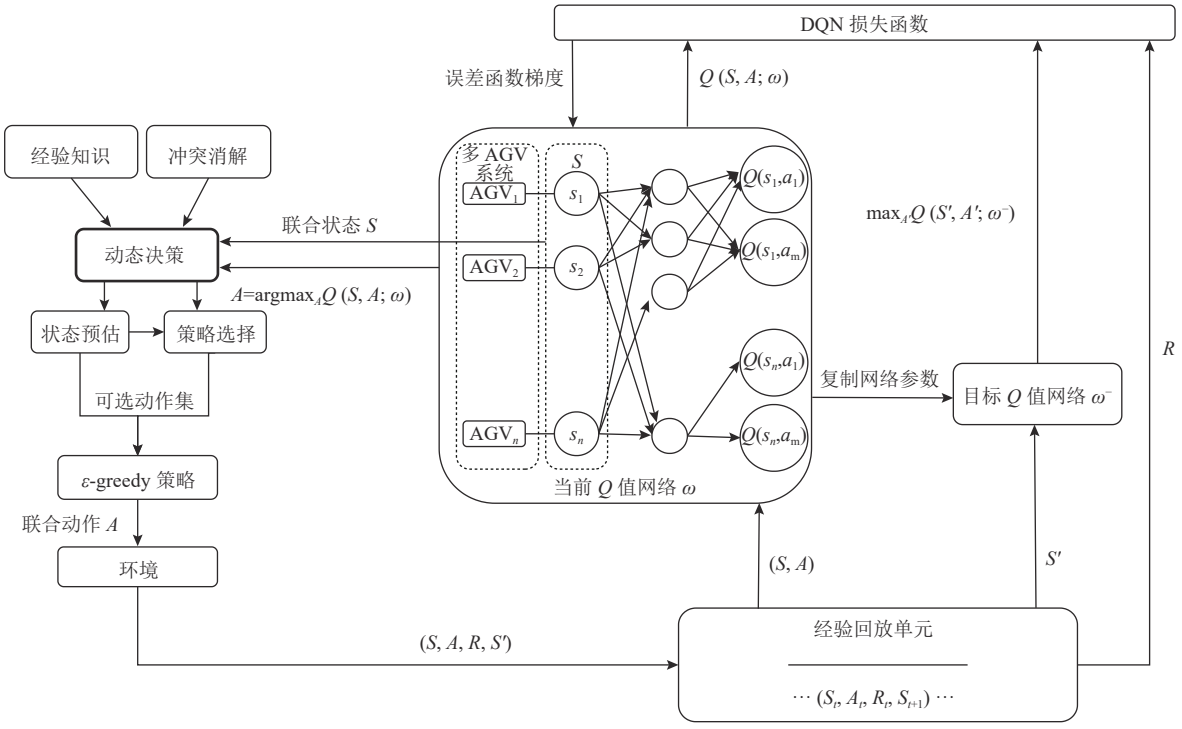


图 2　基于 DQN 算法的多 AGV 路径规划算法架构

Figure 2　The framework of the multi-AGV route planning algorithm architecture based on DQN algorithm

2.2　经验知识

DQN 算法是无模型的学习方法，AGV 需要对无人仓环境进行增量式学习，而刚开始学习时，多 AGV 系统对环境的了解很少，通过反复试错的学习方式效率很低。因此，考虑在开始学习之前，给予多 AGV 系统一些经验知识，令多 AGV 系统提前对环境有一定的了解，减少无用的探索。常用的获取经验知识的方法有启发式规则法、启发式算法搜索法和改造奖励函数法等[11-12]。为了保持算法对环境的学习和适应能力，本文在进行多 AGV 路径规划前，先让单 AGV 基于 Q-learning 强化学习算法进行静态路径规划，得到每台 AGV 在静态环境下的最优路径，同时综合每台 AGV 与障碍物相撞时的状态，组成障碍物位置信息集，共同指导多 AGV 学习探索过程中对动作的选择。在多 AGV 路径规划时，每台 AGV 都倾向于选择利用经验知识避免与静态障碍物碰撞和更快到达目标点的最优策略，当 AGV 之间发生冲突时才会根据其他策略选择动作，这使多 AGV 提前对环境有所了解，缩短学习探索时间，提高算法收敛速度。

设第 i 台 AGV 的静态最优路径为 $\bar{p}^i=(\bar{s}_1^i,\bar{s}_2^i,\cdots,\bar{s}_t^i,\cdots,\bar{s}_k^i)$，综合 n 台 AGV 的探索结果构成静态障碍物信息集 $G=(g_1,g_2,\cdots,g_k,\cdots)$。设第 i 台 AGV 的马尔可夫四元组为 $\{s_t^i,a_t^i,s_{t+1}^i,r_t^i\}$，定义经验知识函数 E 为

$$E(s_t^i,a_t^i)=\begin{cases}10\varepsilon, & s_t^i=\bar{s}_h^i, s_{t+1}^i=\bar{s}_{h+1}^i\neq s_{t+1}^j (i\neq j);\\ -10, & s_{t+1}^i=g_k\in G;\\ 0, & \text{其他。}\end{cases} \tag{8}$$

其中，AGV 在状态 s_t^i 时执行动作 a_t^i 后状态转移为 s_{t+1}^i。$\varepsilon(0<\varepsilon<1)$ 为 $\varepsilon-$greedy 策略的探索因子。当 ε 很大时，主要根据经验知识选择动作以防盲目探索；当 ε 减小时，主要根据历史学习的 Q 值进行路径规划，Q 值的大小取决于 AGV 获得奖励的大

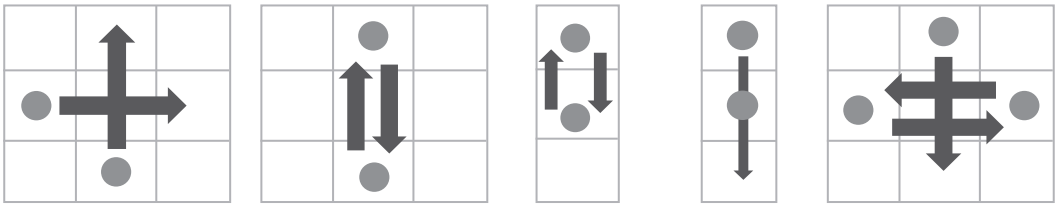
小。本文设置奖励值大小为±10，为使经验值与Q值大小可比，保证其数量级一致，将经验知识函数E的系数设置为 10。使用经验知识时 AGV 动作选择规则策略为

$$\pi(s_t^i)=\arg\max_{a_t}[(1-\delta)Q(s_t^i,a_t^i)+\delta E(s_t^i,a_t^i)]。\tag{9}$$

其中，$\delta(0\leqslant\delta<1)$为一常数，代表经验知识的权重。由于本文在有经验知识时，动作选择更偏向于经验知识，设置$\delta>0.5$。AGV 的状态很大程度上决定是否使用经验知识选择动作，如果第i台 AGV 出现$s_t^i\in\overline{p}$或$s_{t+1}^i\in G$时，则使用经验知识，否则使用$\varepsilon-$greedy策略进行动作选择。

2.3 多 AGV 路径规划的冲突消解

多 AGV 路径规划冲突主要有路口冲突、相向冲突、追击冲突、节点冲突和多 AGV 死锁[13]。由于本文假设所有 AGV 速度一致，可以避免 AGV 追击冲突，本文重点讨论 AGV 间的冲突类型如图 3 所示。



(a) 路口冲突 (b) 相向冲突 (c) 节点经过冲突 (d) 节点占用冲突 (e) 多 AGV 死锁

图 3 AGV 间的冲突类型

Figure 3 Types of conflicts among AGVs

对于任意时刻t，AGV_i与$\mathrm{AGV}_j(i\neq j)$发生冲突，定义 5 种类型的冲突集合$C=\{C_1,C_2,C_3,C_4,C_5\}$。

C_1：路口冲突，$s_{t+1}^i=s_{t+1}^j$，且$(|x_{t+1}^i-x_t^i|,|y_{t+1}^i-y_t^i|)\neq(|x_{t+1}^j-x_t^j|,|y_{t+1}^j-y_t^j|)\neq(0,0)$。

C_2：相向冲突，$s_{t+1}^i=s_{t+1}^j$，且$(|x_{t+1}^i-x_t^i|,|y_{t+1}^i-y_t^i|)=(|x_{t+1}^j-x_t^j|,|y_{t+1}^j-y_t^j|)\neq(0,0)$。

C_3：节点经过冲突，$s_{t+1}^i=s_t^j$且$s_{t+1}^j=s_t^i$。

C_4：节点占用冲突，AGV_i与$\mathrm{AGV}_j(i\neq j)$中有一台 AGV 保持静止，即$s_{t+1}^i=s_{t+1}^j$，且$(|x_{t+1}^i-x_t^i|,|y_{t+1}^i-y_t^i|)=(0,0)$或$(|x_{t+1}^j-x_t^j|,|y_{t+1}^j-y_t^j|)=(0,0)$。

C_5：多 AGV 死锁，为以上 4 种冲突的组合。

强化学习让 AGV 通过与环境交互获取奖励值作为反馈信号，使 AGV 一直向累计奖励最大的方向学习和探索。如果 AGV 当前选择的动作是有利的，应给予正反馈，奖励值为正；如果是不利的，应给予负反馈，奖励值为负。鉴于奖励值r的大小不影响算法的收敛性[14]，为简化计算和理解，本文设奖励值r为

$$r=\begin{cases}10,\ s_t\text{为目标状态；}\\-10,\ s_t\text{为AGV与静态或动态障碍物相撞；}\\-1,\ s_t\text{为其他状态(等待或通行)。}\end{cases}\tag{10}$$

其中，AGV 到达目标点时，获得奖励值$r=10$可以使 AGV 向累计奖励最大的方向学习和探索。AGV 与静态 (货架) 或动态 (其他 AGV) 障碍物发生冲突时，设置$r=-10$可以大程度降低冲突发生。多 AGV 死锁是由多 AGV 相互等待而造成多 AGV 静止的死循环，一旦有 AGV 跳出循环，就能解决死锁问题，因此当 AGV 进入静止等待时设置单步惩罚$r=-1$。但 AGV 也可能因不等待而选择绕路，造成更大的损耗 (如电量、搬运时间等)。对此，设置单步惩罚$r=-1$，一方面可以促使 AGV 尽快到达目标点，最小化总行程时间；另一方面如果当 AGV 到达目标点时才给予正奖励，奖励矩阵将十分稀疏，不利于 AGV 向累计奖励最大的方向学习探索，增加单步奖励可以加快算法收敛速度。

但奖励函数是对 AGV 发生冲突后的反馈，以降低 AGV 再次来到冲突地段的概率，起规避冲突的作用，并没有提供 AGV 发生冲突时的消解方案，所以还需设计冲突消解策略来处理多 AGV 冲突问题。设 AGV 可选动作集$\hat{A}=\{a_1,a_2,a_3,a_4,a_5\}$，其中，$a_1$为静止等待；$a_2$为向上；$a_3$为向下；$a_4$为向左；$a_5$为向右。设任意时刻$t$，$\mathrm{AGV}_i$与$\mathrm{AGV}_j$发生冲突的消解策略为$U=\{u_1,u_2,u_3\}$，冲突消解策略为 AGV 重新确定可选动作集$\hat{A}$，最后具体选择执行哪个动作取决于$\varepsilon-$greedy策略。冲突消解时将以概率$\rho$选择某一 AGV 静止等待或让行，概率$\rho$的大小受 AGV 优先级的影响。经过冲突消解策略确定可选动作集后，AGV 选择某一动作的概率发生变化，

记可选动作为单个的概率为$\bar{P}_u$，AGV 以概率$\bar{P}_u$选择该唯一动作，记可选动作集中有$o(o>1)$个动作的概率为P_u，根据$\varepsilon-$greedy策略，AGV 随机选择可选动作集中每一个动作的概率为$P=\frac{1}{o}P_u\varepsilon$，选择 Q 值最大的动作的概率为$P=P_u(1-\varepsilon)$。

1) 当AGV_i与AGV_j发生路口冲突C_1或相向冲突C_2时，对应的冲突消解策略为u_1，u_1表示以概率ρ选择AGV_i保持静止等待，AGV_j正常通行，即$P(a_t^i=a_1|C_1\text{或}C_2)=\rho$，$a_t^j\in\{a_2,a_3,a_4,a_5\}$。

2) 当AGV_i与AGV_j发生节点经过冲突C_3时，对应的冲突消解策略为u_2，u_2表示以概率ρ选择AGV_i让行，AGV_j正常通行。若冲突发生在x方向，$P(a_t^i\in\{a_2,a_3\},a_t^j\in\{a_4,a_5\}|C_3)=\rho$，表示$\mathrm{AGV}_i$选择$y$方向的动作让行，$\mathrm{AGV}_j$在$x$方向上继续通行，$y$方向上的节点经过冲突同理。

3) 当AGV_i与AGV_j发生节点占用冲突C_4时，对应的冲突消解策略为u_3，u_3分以下两种情况进行讨论。(1) 若其中有已到达目标点的 AGV，u_3表示不论优先级高低，到达目标的 AGV 保持静止，另外的 AGV 选择其余非静止的可行动作。若AGV_i到达目标点，则$P(a_t^i=a_1|C_4,s_t^i=s_{\text{end}}^i)=1,a_t^j\in\{a_2,a_3,a_4,a_5\}$，反之同理。(2) 若没有已到达目标点的 AGV，则若t时刻AGV_i静止，u_3表示以概率ρ选择AGV_j保持静止，否则AGV_j选择其余绕行动作，则$P(a_t^i=a_1|C_4,a_t^i=a_1)=\rho$，反之同理。

4) 当多 AGV 发生死锁C_5时，先消解高优先级 AGV 间的冲突，低优先级保持静止，依次消解。若优先级相等，由于不知目标点的距离信息，根据 AGV 当前位置与起点的曼哈顿距离$d=|x_t-x_0|+|y_t-y_0|$，选择距离起点最远的两 AGV 先进行冲突消解；若与起点的距离均相等，则随机选择两 AGV 先消解。

综上所述，本文针对传统 DQN 算法收敛速度慢、初始探索效率低等问题，增加经验知识，减少无用探索，提前规避静态障碍物碰撞；针对多 AGV 路径规划过程中可能出现的冲突问题进行探讨，改进奖惩函数，给出冲突消解方案，前者规避冲突，后者从根本上解决冲突，并使总路径行程时间最短。本文根据经验知识模型和冲突消解策略指导多 AGV 系统进行动态决策，动态决策过程如图 4 所示。

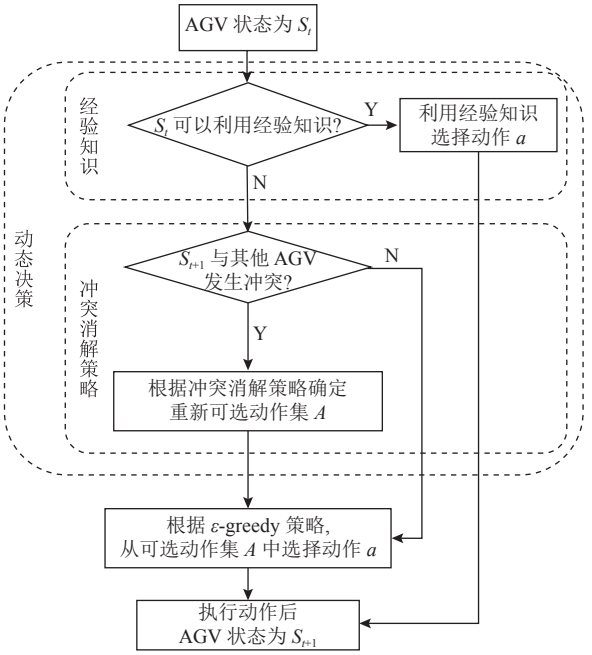


图 4　AGV 动态决策过程

Figure 4　Dynamic decision procees of AGVs

2.4　算法流程

基于经验知识和冲突消解策略的改进 DQN 算法的多 AGV 路径规划算法步骤如下。

Step1　初始化。建立栅格地图，设置 AGV 初始状态S、目标点、探索因子ε、Q 值和参数α、γ。初始化经验回放池D，初始化当前 Q 值网络和目标 Q 值网络，并随机生成权重参数ω，令$\omega^-=\omega$。

Step2　每台 AGV 通过 Q-learning 算法进行静态路径规划，集合所有 AGV 的最优路径生成经验知识$\bar{P}=(\bar{p}^1,\bar{p}^2,\cdots,\bar{p}^i,\cdots,\bar{p}^n)$，并建立静态障碍物信息集$G=(g_1,g_2,\cdots,g_k,\cdots)$。

Step3　AGV 在当前状态s_t根据动态决策和$\varepsilon-$greedy策略执行动作a，获得奖励值r，状态转移至S_{t+1}，将所有 AGV 以上信息组成向量(S_t,A_t,R_t,S_{t+1})，作为样本存入经验回放池D。每探索C步，从D中随机抽取少量样本，用当前 Q 值和目标 Q 值网络分别计算$Q(S_t,A_t;\omega)$和$y=R_t+\gamma\max Q(S_{t+1},A;\omega^-)$，其中若$S_{t+1}$为目标状态，则$y=R_t$，对损失函数$(y-Q(S_t,A_t;\omega))^2$用梯度下降法更新权重$\omega$。

Step4　更新目标 Q 值网络权重参数$\omega^-=\omega$，更新探索因子ε。

Step5　若超过最大寻路步数N_t或 AGV 均到达目标点，跳转 Step6；否则，跳转 Step3。

Step6　若达到最大迭代次数N，算法终止；否

则，AGV 回到初始位置，跳转 Step3。

3 实验验证

由于传统 DQN 算法缺乏经验知识的指导和冲突消解策略，直接用于多 AGV 路径规划最终算法很难收敛，无法得到路径规划的最优结果，因此，为验证本文提出的改进 DQN 算法 (后文简称 IDQN) 的有效性，将其与结合 *A**算法的 DQN 算法[5] (后文简称 *A**-DQN) 进行实验对比分析。

3.1 实验参数设置

本文使用 Tensorflow 深度学习框架在 PyCharm 平台进行实验，使用处理器 AMD Ryzen 75800H，内存 16GB，Windows10 操作系统。以图 1 所示的无人仓环境模型为仿真环境，构建20×20栅格地图，设置 3 台 AGV，AGV 动作集为$A=\{0,1,2,3,4\}$，分别表示保持静止、向上、向下、向左和向右。3 台 AGV 的优先级依次减小，高优先级 AGV 在发生冲突时选择静止或让行动作的概率$\rho=0.25$。神经网络结构为三层全连接神经网络，网络模型训练参数参考文献 [11]，如表 1 所示。所有结果均在尽量保证系统运行环境一致的情况下取 10 次实验的平均值。

表 1 参数设置

Table 1 Parameter setting

参数	值
学习率 α	0.01
衰减系数 γ	0.9
单次训练样本/个	32
经验回放池样本/个	2000
最大寻路步数 N_t /(步·台$^{-1}$)	300
目标 Q 更新频率/步	200
总迭代次数 N /次	3 000
ε更新频率 K / 成功寻路次数	5

3.2 实验结果与分析

3.2.1 算法有效性验证

基于本文提出的 IDQN 算法的路径规划结果如图 5 所示。下面从收敛效果和收敛速度两个方面对 IDQN 和 A*-DQN 算法的性能进行比较，表 2 给出 IDQN 和 A*-DQN 的性能数据。结果表明 IDQN 在损失值、奖励值、探索步数上均优于 A*-DQN，说明 IDQN 算法的收敛效果优于 A*-DQN。这主要是因为 IDQN 算法采用了动态决策，经验知识加速算法收敛，冲突消解策略能有效解决 AGV 冲突。

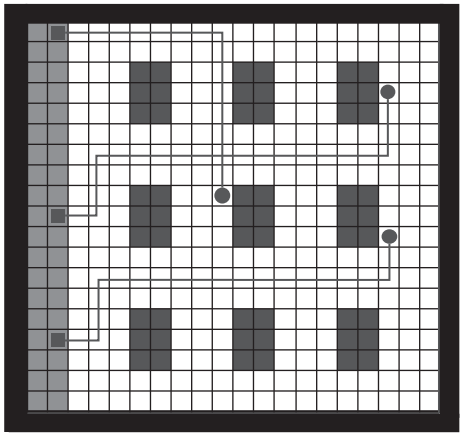


图 5 基于 IDQN 算法的路径规划结果

Figure 5 Route planning results based on IDQN algorithm

表 2 算法性能对比

Table 2 Comparison of algorithm performance

性能评估指标	IDQN	A*-DQN	算法改进率/%
平均损失值	**0.244**	0.331	26.3
平均奖励值	**−29.22**	−36.76	20.5
探索步数/步	**69 874**	73 233	4.6
总路径行程时间	**60**	**60**	0
算法耗时/s	1 256	**1 197**	−4.93
获取经验知识耗时/s	196	**105**	−86.7

在收敛速度方面，以损失函数值达到基本稳定时算法训练次数为评价标准进行对比，如图 6 所示。IDQN 在约 650 次训练时收敛，A*-DQN 在约 750 次训练时收敛，IDQN 收敛后波形平稳性优于 A*-DQN。IDQN 相比于 A*-DQN 收敛速度提升 13.3%，平均损失值降低 26.3%。可见，IDQN 算法在收敛效果和收敛速度上均优于 A*-DQN 算法。

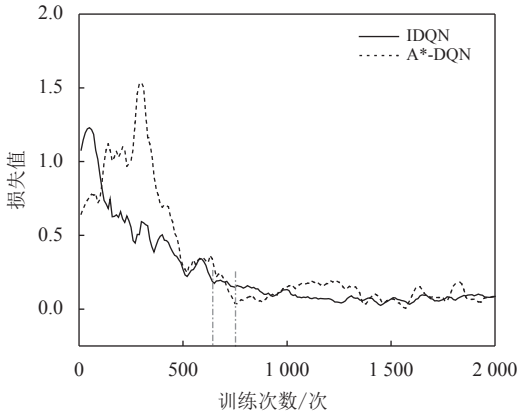


图 6 不同算法的损失函数值

Figure 6 Loss function values of different algorithms

为分析算法的稳定性，图 7 和图 8 给出 IDQN 和 A*-DQN 算法的损失值、平均奖励值箱线图。无

论是损失值还是平均奖励值，IDQN 的方差都更小，这说明 IDQN 算法具有更好的稳定性。箱线图中的异常值出现在 AGV 初始的环境学习中，表示 AGV 对环境的盲目探索产生的探索值与算法收敛时的偏差；IDQN 异常值偏离更低，这说明 IDQN 算法对环境的探索更有效，算法更具有稳定性和环境适应性。

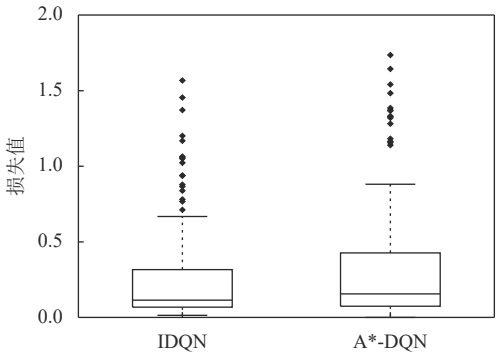


图 7　不同算法的损失函数值箱线图

Figure 7　The boxplot of loss function values using different algorithms

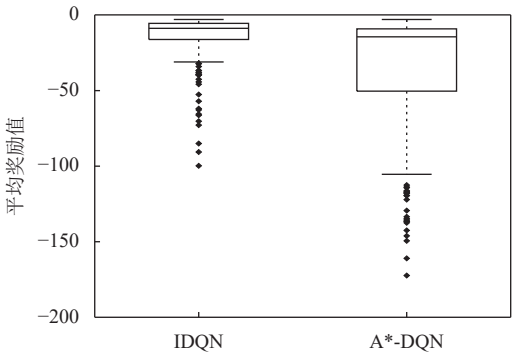


图 8　不同算法的平均奖励值箱线图

Figure 8　The boxplot of average reward values using different algorithms

3.2.2　经验知识和冲突消解策略对算法的影响分析

为了验证经验知识和冲突消解策略的有效性，将 IDQN 与带经验知识的 DQN (DQN-K) 、带冲突消解策略的 DQN (DQN-C) 进行对比实验。图 9 给出 3 种算法下 AGV 的平均奖励值。IDQN 算法的奖励值经历了由大变小最后收敛的过程，表明最开始经验知识发挥了作用奖励值大，指导 AGV 的环境探索，随着环境探索的深入 AGV 可能会发生冲突导致奖励值减小。DQN-K 算法的奖励值变化趋势与 IDQN 基本一致，在经验知识的指导下算法仍然能收敛，说明经验知识起到一定作用，其收敛速度比 IDQN 慢是因为没有加入冲突消解策略。DQN-C 算法奖励值总体偏小，且波动较大，这是因为没有经验知识的指导，AGV 只能进行盲目地探索和路径规划，算法无法收敛。

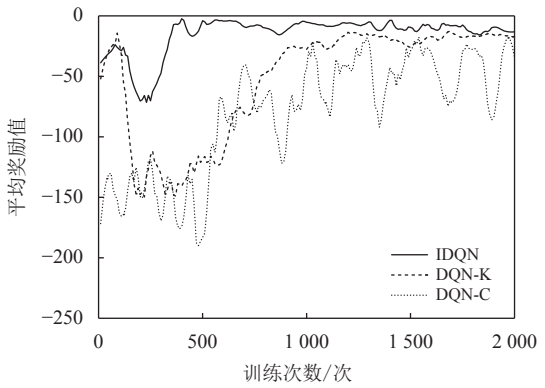


图 9　不同算法的平均奖励值

Figure 9　Average reward values of different algorithms

图 10 给出 3 种算法在不同数量 AGV 下发生冲突的情况，IDQN算法的冲突次数均远少于其他两种算法。DQN-K 算法在 AGV 数量少时，冲突次数少于 DQN-C，因为 DQN-K 算法可以通过经验知识规避冲突，但在 AGV 数量较多时，冲突难以规避，冲突次数随着 AGV 数量的增加直线上升。DQN-C 算法在 AGV数量少时缺少经验知识的指导，发生冲突的次数较 DQN-K 多，但在 AGV 数量多时，冲突消解策略发挥主要作用，发生冲突的次数远小于 DQN-K 算法。综上所述，在 AGV 数量少时，经验知识发挥主要作用，指导 AGV 路径规划和避障；随着 AGV数量增加，冲突消解策略发挥的作用越来越明显。因此，经验知识和冲突消解策略对 IDQN 算法的性能提升均具有重要的作用。

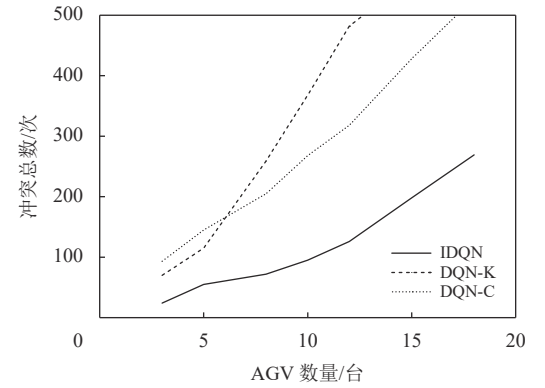


图 10　不同算法下多 AGV 冲突次数

Figure 10　Conflict times of multiple AGVs using different algorithms

3.2.3　地图环境和 AGV 数量对算法的影响

地图环境大小、复杂度和 AGV 数量是本文问题的两个重要参数，为了验证算法的适应性，对这

两个参数的影响进行实验与分析。为了分析地图规模的影响，设置地图大小分别为10×10、20×20和30×30，障碍物占比分别为 4%、13% 和 21.3% 的 3 种地图进行实验对比。图 11 显示了 3 台 AGV 在不同的地图中进行路径规划的 IDQN 算法收敛结果，算法在不同规模地图环境下均能收敛，这说明算法有良好的环境适应性。图 12 为 IDQN 算法在20×20的地图中在不同数量 AGV 下的收敛速度。可见，在不同数量 AGV 情形下，IDQN 算法均能在一定次数的训练下收敛，这说明算法能适应 AGV 数量的扩展。

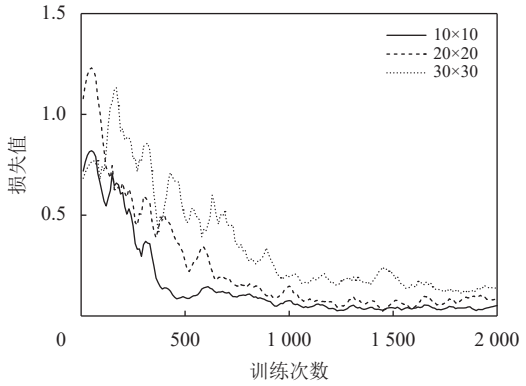


图 11 不同地图环境下算法收敛情况

Figure 11 Convergences of algorithms in different map conditions

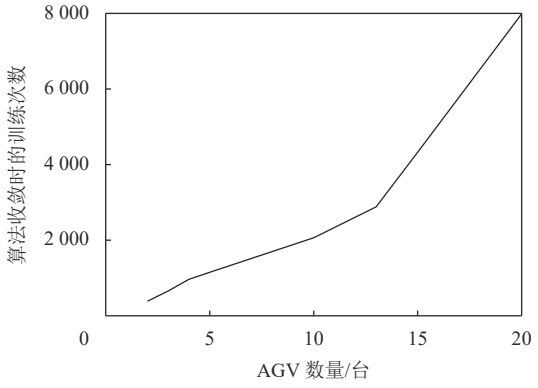


图 12 不同 AGV 数量下算法收敛速度

Figure 12 Convergence speed of algorithm with different number of AGVs

4 结论

本文针对无人仓多 AGV 路径规划问题，以总行程时间最短为目标，提出一种基于经验知识和冲突消解策略的改进 DQN 算法。针对传统的 DQN 算法用于解决多 AGV 路径规划问题存在环境探索难度大、算法收敛速度慢和无法从根本上解决多 AGV 冲突等问题，从两个方面进行改进。1) 增加经验知识：基于 Q-learning 强化学习算法进行单 AGV 路径规划，构建经验知识模型，指导多 AGV 动作的选择；2) 多 AGV 冲突消解：改进奖惩函数和设计冲突消解策略，前者有助于 AGV 预先规避冲突，后者从根本上解决多 AGV 冲突。通过实验进行验证，结果表明，本文提出的模型和算法较其他 DQN 算法收敛速度提升 13.3%，平均损失值降低 26.3%，有利于规避和化解无人仓多 AGV 路径规划冲突，对提高无人仓作业效率和经济效益具有重要指导意义。本文提出的 IDQN 算法性能优于其他 DQN 算法，且本文提出的路径规划方法是一种无模型的学习方法，既保留了强化学习对环境的学习特性，又改善了强化学习存在的学习效率低、收敛速度慢等问题，对不同环境的无人仓更具适应性，能有效提高无人仓运作效率，降低维护成本。本文仅考虑 AGV 单次搬运任务的路径规划，即多 AGV 单任务的路径规划。未来可结合多订单任务的实际情况，研究多 AGV 多任务的路径规划与调度优化。

参考文献：

[1] 余娜娜, 李铁克, 王柏琳, 等. 自动化分拣仓库中多 AGV 调度与路径规划算法 [J]. 计算机集成制造系统, 2020, 26(1): 171-180.
YU Nana, LI Tieke, WANG Bailin, et al. Multi-AGVs scheduling and path planning algorithm in automated sorting warehouse[J]. Computer Integrated Manufacturing Systems, 2020, 26(1): 171-180.

[2] 王秀红, 刘雪豪, 王永成. 基于改进 A*算法的仓储物流移动机器人任务调度和路径优化研究 [J]. 工业工程, 2019, 22(6): 34-39.
WANG Xiuhong, LIU Xuehao, WANG Yongcheng. A research on task scheduling and path planning of mobile robot in warehouse logistics based on improved A* algorithm[J]. Industrial Engineering Journal, 2019, 22(6): 34-39.

[3] YANG L, FU L, LI P, et al. An effective dynamic path planning approach for mobile robots based on ant colony fusion dynamic windows[J]. Machines, 2022, 10(1): 50.

[4] ZHONG X, TIAN J, HU H, et al. Hybrid path planning based on safe A* algorithm and adaptive window approach for mobile robot in large-scale dynamic environment[J]. Journal of Intelligent & Robotic Systems, 2020, 99(1): 65-77.

[5] YANG Y, LI Juntao, PENG Lingling. Multi-robot path planning based on a deep reinforcement learning DQN algorithm[J]. CAAI Transactions on Intelligence Technology, 2020, 5(3): 177-183.

(下转第 53 页)