Research on optimization method of routing buffer linkage based on Q-learning

Zhonghua Han1,2,3,4, Hongfeng Qi1*, Daliang Chang2,3,4, Wei Gong4, Shilong Dai1
1School of Information and Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China.
2Key Laboratory of Networked Control Systems, Chinese Academy of Sciences, Shenyang 110016, China.
3Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China.
4Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China.
*Corresponding author’s e-mail: xiaozh1977@sjzu.edu.cn

Abstract—According to the characteristics of the painting process of passenger car manufacturing enterprises, by formulating the routing buffer linkage rules based on the total renewal cost, the linkage process of the bus in the routing buffer is controlled, and an improved Q-learning (Q-The routing buffer of learning) algorithm quickly finds the optimal path method. According to the actual production situation, this method improves the dynamic parameters of the algorithm on the basis of the traditional Q-learning algorithm, and improves the optimization speed and accuracy of the algorithm by establishing the correlation between the work-in-process and its neighboring work-in-process in the current state. Through multiple sets of example simulation tests, the effectiveness of the Q-learning algorithm in solving the optimization problem of routing buffer linkage is verified.

1. Introduction

1.1 Background and Significance of Research
In passenger car manufacturing enterprises, according to the customized production orders provided by customers, different types of passenger cars correspond to different processing procedures. At the same time, the processing man-hours in the same processing process are different, and in the coating process such as automobile welding and spraying, there are multiple processes in parallel, so the production workshop with routing buffer in automobile manufacturing enterprises has the characteristics of a typical flexible flow shop with multiple processes and multiple parallel stations. The bus volume is usually large, and the physical space of the buffer zone of the manufacturer cannot be expanded indefinitely. It can only be temporarily placed in the limited parking spaces of the buffer zone. Therefore, the buffer movement problem in the bus manufacturer belongs to the optimization problem of the production scheduling of the flexible flow shop. At the same time, it is also a question of path selection.
1.2 Research status
According to the limitations of some factors in the actual situation and the actual needs of the manufacturers in the production scheduling process, the research on the limited buffer zone has attracted more and more attentions. In 2015, Aldowaisan T, Allahverdi A [1] used genetic algorithm combined with simulated annealing algorithm to formulate scheduling rules to solve the optimization problem of flow shop scheduling with modified operations. Czeslaw [2] uses a branch and bound method to solve a two-stage replacement flow shop scheduling problem with buffer constraints. In recent years, a large number of scholars have conducted research on related scheduling optimization problems. For the limited buffer pipeline scheduling problem, Wardono [3] and others used heuristic algorithms to solve the buffer-constrained multi-process parallel machine flow shop scheduling problem. In 2016, Guanlong Deng [4] and others used artificial bee colony algorithm and limited buffer to study the flexible flow shop with limited buffer. In 2017, Cheng Zhang [5] and others used a batch processor to study the limited buffer. QING etc [6] proposed a two-stage method based on embedded differential evolution algorithm. Rohaninejad M, Kheirkhah AS [7], etc. used the firefly algorithm combined with the domain search method to solve the scheduling optimization problem of the scheduling workshop with modified operation. In 2017, Sioud A, Gagné C [8] used the migratory bird algorithm to solve the problem of time optimization for aircraft modification. Zhang et al. [9] used the batch processor to study the limited buffer.

1.3 Problem description
The model of the routing buffer based on Q-learning studied in this paper is described as follows in the workshop model of a passenger car manufacturer.

This problem description is shown in Figure 1. It can be described as a passenger car manufacturer with M workpieces produced in the order of daily production orders, and there are N processing procedures. One of the procedures has multiple parallel stations, and in one of the two procedures. There is a routing buffer in the buffer zone. There are multiple parallel lanes in this buffer zone. The passenger car can move forward along the lane, and the skid can move in parallel between the lane and the lane to change the route. When the space in the routing buffer is released, the work piece of the previous process enters the routing buffer, and this station can work again. When the workpiece in the buffer enters the next process, when the information of the workpiece that is about to be online in the second station is inconsistent with the workpiece just processed, the machine needs to be changed. According to the processing time of the known process in the production process, by optimizing the movement route of the workpiece in the routing buffer, better scheduling results can be obtained.

Figure 1 Mathematical model of production
2. Method research

2.1. Formatting author affiliations
The Q-learning algorithm is a typical model-independent reinforcement learning algorithm. It uses the corresponding state-action as an estimation function. It continuously updates the Q value and modifies the Q table, so that the probability of choosing positive reward is increasing, with the exchange of information with the actual situation, changing the action strategy set, so that the difference of the Q value between adjacent states reaches a certain convergence condition, and finally tends to the optimal action set.

The algorithm formula of Q-learning is:

\[
Q(s, a) = Q(s_t, a_t) + \alpha \left[ r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]
\]

In formula (1), \(s_t\) is the state of the agent at time \(t\), when the action \(a_t\) is executed in the state \(s_t\), the state of the agent becomes the estimate of \(s_{t+1}\), represents the reward value obtained by the agent from performing an action from the current state to the next state; action \(a \in A\), \(A\) is the direction of movement, state \(s_t, s_{t+1} \in S\), \(S\) is the state space; \(\alpha\) is the learning rate, the larger \(\alpha\) is, the faster the Q value will converge, but the result is prone to shocks; \(\max_a Q(s_{t+1}, a)\) indicates that an action \(A\) is selected from the action set \(A\) to maximize the value of \(Q(s_{t+1}, a)\); \(\gamma\) is the attenuation factor, which indicates the degree of influence of future rewards on the current action.

2.2 Algorithm improvement
In general, reinforcement learning algorithm is a random selection strategy, through the greedy strategy to clearly explore the purpose, but the greedy strategy is easy to fall into the local extremum, especially in a more complex environment, it is difficult to complete, and even often jump into "traps" and terminated. In this paper, by changing the exploration factor \(\epsilon\) in Q-learning, according to the interaction between the agent and the environment, and according to the policy constraint Demo(), judge whether the exploration is completed, and by continuously adjusting the exploration factor, decide to randomly select or choose the behavior corresponding to the smallest Q value, when the \(\epsilon\) is larger, the exploration range is larger, and vice versa, the method is to avoid falling into the local extreme value to a large extent.

2.3 Movement rules of routing buffer
In the routing buffer scheduling optimization problem studied in this paper, a 4*4 routing buffer mathematical model is constructed, which is composed of multiple parallel moving skid lanes with limited space. Each skid can carry one vehicle to complete the previous process, each skid between two adjacent parallel lanes is equipped with a translation car, the workpiece to be processed can be driven by the translation car to move the skid to the adjacent lane, so as to realize the buffering of the passenger car parallel movement in the area, but the passenger car has the characteristic of moving forward as a whole. This way of path selection has the characteristics of routing.

2.4 Algorithm process
Specific steps are as follows:
Figure 2 Algorithm flow chart

3. Simulation test and comparative analysis

3.1 Construct simulation data

In order to test the effectiveness of the improved Q-learning algorithm to solve the production scheduling problem in the routing buffer in compliance with the local scheduling rules, and to further illustrate the effectiveness of the algorithm for solving the routing buffer based on the simulation results, four solutions were design.

The Dijkstra algorithm, SPF algorithm, Q-learning algorithm and improved $\epsilon$-Q-learning algorithm are used as local optimization algorithms, and the combination of local assignment rules of multi-sequence buffers is used to solve the scheduling optimization problem of a passenger car manufacturer mentioned above, and the effect of Q-learning algorithm in solving the scheduling problem of flexible flow workshop with routing buffers is further analyzed. The information of the 4 groups of simulation schemes is shown in Table 3.

| Simulation solution | Local optimization algorithm | Local assignment rule |
|---------------------|-----------------------------|-----------------------|
| plan 1              | Dijkstra algorithm          | MQBCR, SST, FAM, FCFS|
| plan 2              | SPF algorithm               |                       |
| plan 3              | Q-learning algorithm        |                       |
| plan 4              | $\epsilon$-Q-learning algorithm |                   |
3.2 Simulation results and evolution curve analysis

Figure 4 shows the relationship between the fitness value of the schemes 1 to 4 and the number of iterations, it can be seen from the figure that the SPF algorithm converges very quickly in the initial stage, but as the number of evolutionary iterations increases, the search ability of SPF algorithm and Dijkstra algorithm gradually deteriorated, and fell into the local extremum in the 206th and 127th generations, respectively. The standard Q-learning algorithm has fast search characteristics, but as the probability value distribution of the probability model explored in the iterative process declines rapidly, the search performance of the solution space also declines, and it is easy to fall into the extreme value until it falls into the local extreme value. $\epsilon$-Q-learning algorithm maintains fast search performance in the initial stage. With continuous exploration, $\epsilon$-Q-learning algorithm enables the algorithm to go beyond the local extreme value and continue to find a better path by adjusting the exploration parameter $\epsilon$, and search in the 281st generation to the optimal solution.

![Figure 3 Diagram of the relationship between the number of training and the number of steps in Schemes 1 to 4](image)

3.3 Analysis of production scheduling Gantt chart

Figure 6 shows the Gantt chart of plan 8 of production scheduling. The abscissa is the time axis, and the ordinate is the station of each process. The green in the figure indicates the time that the bus stays in the buffer zone, the red indicates the preparation time when the attributes of the bus are processed before and after the station are different, and the blue indicates the blocking time at the station after the bus is processed.
4. Conclusion
In order to solve the local optimization problem in the routing buffer of the passenger car manufacturer, by establishing the RBBMS mathematical model, on this basis, according to the production specification and scheduling rules of the passenger car manufacturer, a routing buffer scheduling based on the total renewal cost is formulated rule. This article uses an improved Q-learning algorithm to locally optimize the routing buffer of the bus manufacturer. Through iterative training, the algorithm can quickly search for the optimal path from the input buffer to the output buffer at each moment in the routing buffer. By optimizing the movement path, the goal of reducing the maximum completion time is achieved. Through the actual test of the simulation model, this paper obtains the feasibility and effectiveness of the improved Q-learning algorithm for the optimization method of the routing buffer in the automobile manufacturing enterprise.

Acknowledgement
This work was supported by the Liaoning provincial key R & D plan project (No.2020jh2/10100039), Project of Liaoning Province Education Department of China (No. LJKZ0583) and Shenyang Municipal Science and Technology Project of China (No. Z18-5-015).

References
[1] Aldowaisan T, Allahverdi A. No-Wait Flowshops to Minimize Total Tardiness with Setup Times[J]. Intelligent Control & Automation, 2015, 06(1):38-44.
[2] CZESIAW S. A two-machine permutation flow shop scheduling problem with buffers [J]. OR Spektrum, 1998, 20(4): 229-235.
[3] WARDONO B, FATHI Y. A tabu search algorithm for the multi-stage parallel machine problem with limited buffer capacities[J]. European Journal of Operational Research, 2004, 155(1): 380-401.
[4] Guanlong Deng, Hongyong Yang, Shuning Zhang.,”An Enhanced Discrete Artificial Bee Colony Algorithm to Minimize the Total Flow Time in Permutation Flow Shop Scheduling with Limited Buffers,” presented at the Mathematical Problems in Engineering, 1-11, 2016.
[5] Zhang C, Shi Z, Huang Z, et al. Flow shop scheduling with a batch processor and limited buffer[J]. International Journal of Production Research, 2016, 55(11):3217-3233.

[6] QING F, APPALYER S, KUNPENG L. Optimisation of flowshop scheduling with batch processor and limited buffer [J]. International Journal of Production Research, 2012, 50(8):2267-2285.

[7] Rohaninejad M, Kheirkhah A S, Nouri B V, et al. Two hybrid tabu search–firefly algorithms for the capacitated job shop scheduling problem with sequence-dependent setup cost[J]. International Journal of Computer Integrated Manufacturing, 2015, 28(5):470-487.

[8] Sioud A, Gagné C. Enhanced Migrating Birds Optimization Algorithm for the Permutation Flow Shop Problem with Sequence Dependent Setup Times[J]. European Journal of Operational Research, 2017, 264(1).

[9] ZHANG C, SHI Z, HUANG Z, et al. Flow shop scheduling with a batch processor and limited buffer [J]. International Journal of Production Research, 2016, 55(11):3217-3233