Mobile Robot Planning Based on RDL-Q Learning Algorithm

Shengmin Wang [a, *], Wei Lin

Faculty of Computer, Guangdong University of Technology, Guangzhou, 510006 China.

[a, *] 1546003584@qq.com

Keywords: complex environment; mobile robot; path planning; Q learning algorithm; state trajectory; exploration strategy.

Abstract. Aiming at the problem of Q value update is slow for traditional Q learning algorithm in complex unknown environment, resulting in low learning efficiency and low real-time performance of mobile robot. A Reverse Double Linker Q (RDL-Q) learning algorithm is proposed. According to the state trajectory of the mobile robot, two state linkers are established to record the current sate-action pair and current state-reverse action pairs, from the value of the tail of a single chain, the current state, is traced back to the Q value at the end of a single linker head until the target is reached. Meanwhile, the Boltzmann search strategy combined with heuristic search strategy is used to guide the action selection strategy of the mobile robot learning process. The simulation results show that the algorithm can effectively speed up the convergence of learning algorithm and improve the learning efficiency in complex unknown environment and achieve the robot navigation task with the best path.

1. Introduction

With the rapid development of service, storage and logistics industry and the upgrading of related industries, autonomous mobile robots have a wide range of applications. Path planning is one of the key technologies for mobile robot to achieve autonomous navigation. Path planning means finding a collision free optimal path from initial pose to target position according to certain evaluation criteria in an obstacle environment [1].

Recent years, the improved Q learning path planning problem of mobile robot can be divided into four categories: (1) redefine the state space environment (2) random selection strategy action (3) initialization strategy for Q tables (4) initialization strategy update strategy of Q function (5) design of the penalty function. Gao Hui [2] proposed a Q learning algorithm based on approximate action space model, which effectively reduced the dimension of Q value table. However, it needed to dynamically add alternate actions on the basis of basic actions, which greatly increased the running time. Tijsma A D [3] compared four exploration strategies of Q learning, which proved the validity of Boltzmann's exploration strategy. Yong Song [4] uses the initial values of the Q table based on neural networks to improve the performance of Q learning, but the need to create known environmental information to train the network. Xu Ya [5] proposed a chain Q algorithm, effectively speed up the convergence, but in the process of generating single stranded in the removal of original state in the path of all rings, cause the algorithm to find a feasible solution, but may not be the optimal solution. Liang Quan [6] divides the robot into four states according to the distance from the mobile robot to the obstacle. It sets the reward function through the transformation between different states, which helps the robot learn faster and better.

In this paper, RDL-Q learning algorithm is proposed to improve the lag of the traditional Q learning data transmission by establishing a bidirectional state chain, so that the current state of the action decision quickly affects the previous state action pairs. The strategy with Boltzmann exploration favorable to reach the goal of knowledge to inspire and guide mobile robot balance exploration and learning process, to avoid the Q learning algorithm is easy to fall into local optimal solution, thus speeding up the running time of the algorithm and reduce the learning iterations, improve the convergence speed of the algorithm.

2. Design of RDL-Q Learning Navigation Algorithm

2.1 Heuristic Exploration Strategy for Unknown Environment

ε-greedy is a widely used exploration strategy, and ε-greedy introduces random probability values to ensure more exploration and learning of mobile robots in the learning process. ε-greedy strategy can balance the exploration and learning process to some extent, but mobile robot always explores all actions by equal probability, which increases the chance of small probability action selection.

Aimed at the above defects, a new strategy for autonomous random detection of unknown environment is proposed in this paper. This strategy combines Boltzmann and heuristic search strategy, which has the advantage of avoiding local optimal and accelerating the convergence of expected Q value. The azimuth diagram of a robot's action is shown in Figure 1.

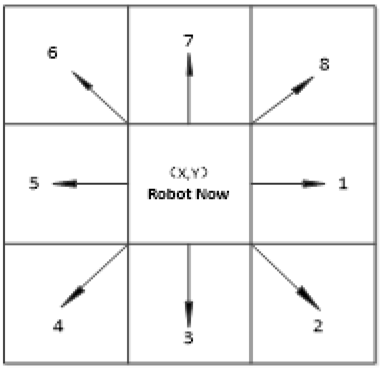


Figure 1. The Azimuth Diagram of a Robot's Action

Assuming that (x, y) and (yg, yg) represent the current location and final destination of the mobile robot, θ represents the angle between the two points and X axis, and a represents action. As shown in Figure 1, the heuristic search rules are as follows:

1) If $0 \le \theta < \frac{\pi}{2}$, $a = rand(1,7,8)$

2) If $\frac{\pi}{2} \le \theta < \pi$, $a = rand(5,6,7)$

3) If $\pi \le \theta < \frac{3}{2}\pi$, $a = rand(3,4,5)$

4) If $\frac{3}{2}\pi \le \theta \le 2\pi$, $a = rand(1,2,3)$

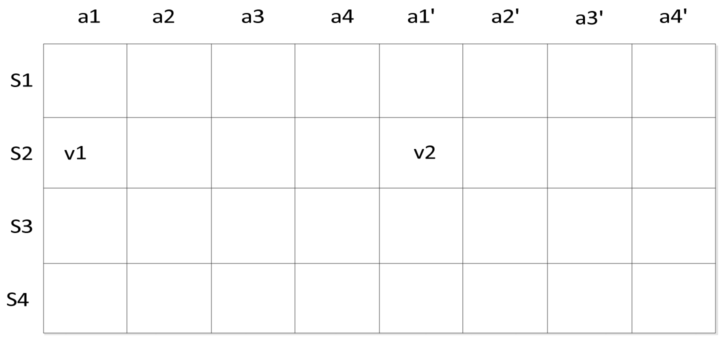
Therefore, the unknown environment exploration strategy is as follows:

Step1 randomly generates a number δ, $0 < \delta < 1$;

Step2 if $\delta < p(s,a_i)$ happens randomly, a should adopt Boltzmann strategy to accept random action a_i; otherwise, the random action a should follow the heuristic search strategy.

2.2 RDL-Q Algorithm and Step

This article uses a reverse double-chain Q value update strategy. Reversal means that at each time t, if the mobile robot accepts a reward for taking an action in a given states, then at this moment the mobile robot can also receive punishment for the opposite action in the same state without need to take the opposite action to train. It will update two Q values at the same time, as shown in Figure 2, respectively v1, v2. v2 is the opposite action value of v1 state. Therefore, the mobile robot can simultaneously explore the action and the reverse action, update the Q value corresponding to the action opposite to the selected action for each time step, and the "double update" speeds up the learning process.



	a1	a2	a3	a4	a1'	a2'	a3'	a4'
S1								
S2	v1				v2			
S3								
S4								

Figure 2. Q Matrix

Define $M(t) \leftarrow [s_t, a_t, r_t]$ to record the state-action pairs experienced by the mobile robot, where rt is the return value at time t and defines $M'(t) \leftarrow [s_t, a_t', r_t']$. To record the state-reversal pairs experienced by the mobile robot. The Q value is updated by tracing the state-action pair in the memory matrix. The update formula is as follows:

$$Q(s_k, a_k) \leftarrow (1-\lambda)Q(s_k, a_k) + \lambda[r_k + \gamma \max_{a_{k+1} \in A} Q(s_{k+1}, a_{k+1})] \tag{1}$$

$$Q(s_k, a_k') \leftarrow (1-\lambda)Q(s_k, a_k') + \lambda[r_k' + \gamma \max_{a_{k+1} \in A} Q(s_{k+1}, a_{k+1})] \tag{2}$$

Where s is the state, a is the action, and a' is the reverse actions.

Based on the above exploration strategy, improved Q-value iterative formula and reward and penalty function, the flow chart of the mobile robot navigation Q algorithm in the unknown environment proposed in this paper is shown in Figure 3:

The algorithm steps are as follows:

Step 1: Initialize the starting position, goal position, Q matrix and discount coefficient, learning rate β, and temperature initial value τ_0 of the mobile robot.

Step 2: Execute a loop for each episode and loop through each step in each episode.

Step 3 Measure the distance of the nearest obstacle of the mobile robot by Lidar.

Step 4 Determine if the robot is safe. If it is safe, proceed directly to the target and go to step 7. If it is not safe, go to step 5.

Step 5 Select the action according to the strategy given in section 1.1 and perform the Q-value update according to equation 1) and 2).

Step 6 Select the next action to hit the obstacle before reaching the target. Adjust the temperature parameter to step 2. If there is no collision, go to step 7.

Step 7 The current position of the mobile robot is the end point, go to step 8; the current position of the mobile robot is not the end position. If the max step is exceeded, the episode fails and go to step 3 to re-learn. Otherwise, step increases by 1 and goes to step 3 to learn.

Step 8 End this episode study, update the temperature control parameter T, go to step 2 until the maximum episode.

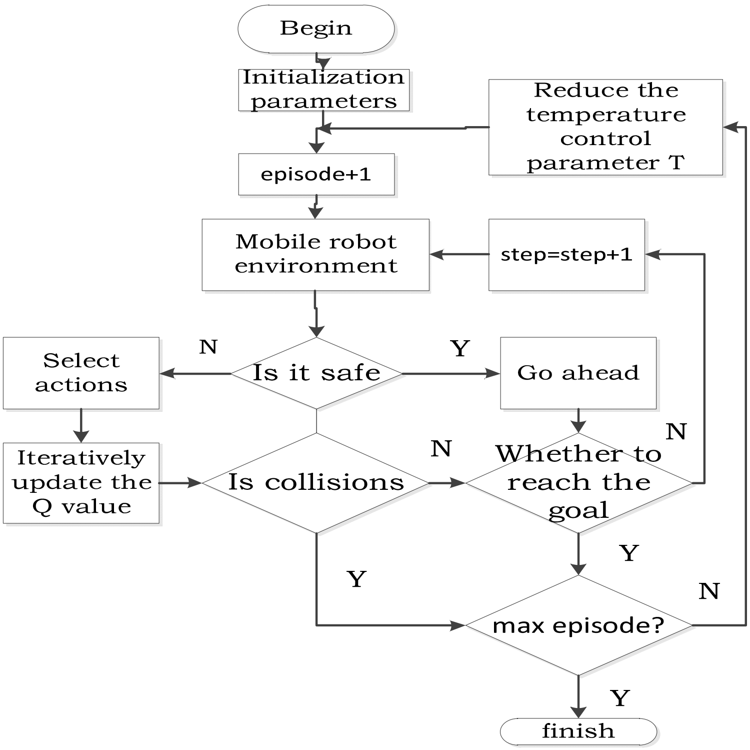
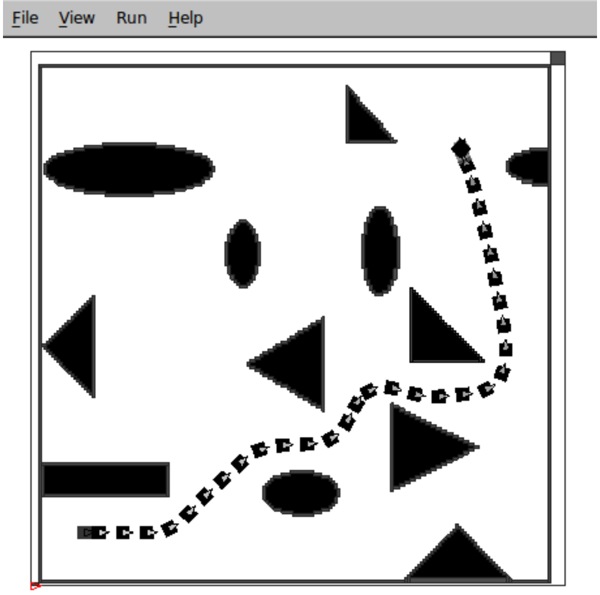

Begin
Initialization parameters
Reduce the temperature control parameter T
episode+1
Mobile robot environment
step=step+1
Select actions
N
Is it safe
Y
Go ahead
Iteratively update the Q value
Is collisions
N
Whether to reach the goal
N
Y
Y
N
max episode?
Y
finish

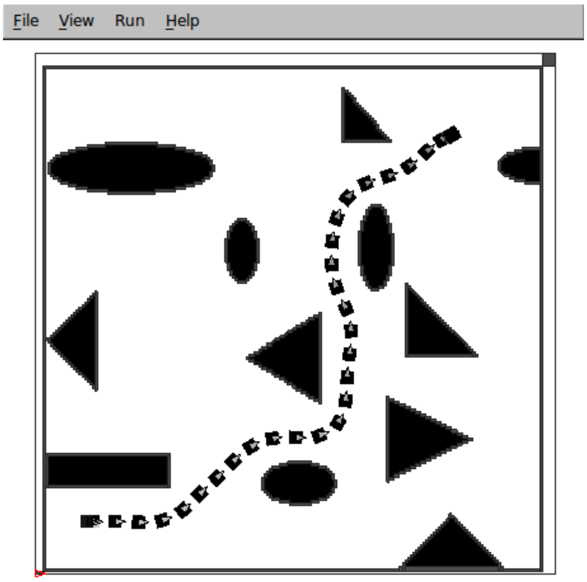
Figure 3. Algorithm Design Flow Chart

3. Experiment and Analysis

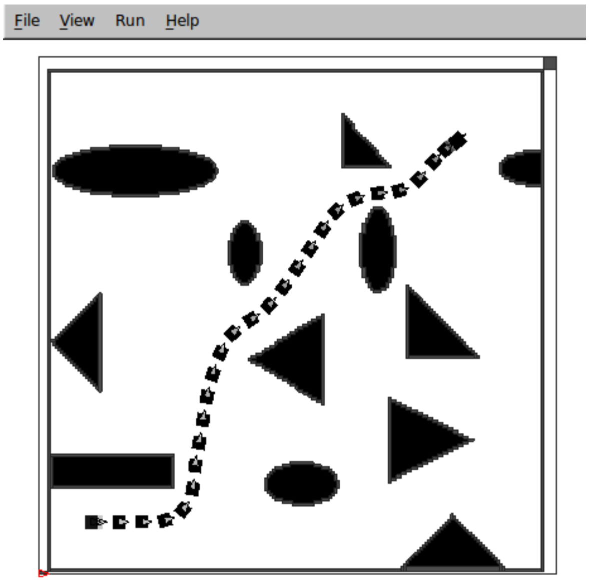
In this paper, an unknown environment for robot navigation is built on the ROS robot simulation platform. The environment includes obstacles such as walls, boxes, and baffles. The shape and position of the obstacle are relatively unknown to the mobile robot. The relevant parameters in the experiment were set as follows: learning factor λ=0.4, discount factor γ=0.95, τ_0=200, maximum number of iterations maxepisode=150, maximum number of steps maxstep=800. Each episode of the mobile robot starts training from the specified starting position and ends the entire training process when the number of iterations reaches the maximum. The following three algorithms are all performed under the same conditions. According to the converged Q, the mobile robot completes the final navigation, and each algorithm obtains an optimal collision-free path, as shown in Figure 4.



(a)Traditional Q Learning Algorithm



(b) Q(λ) Learning Algorithm



(c) The algorithm of this paper

Figure 4. Robot Navigation Path in a Complex Environment

As can be seen from Figure 4, the navigation paths of each algorithm in the figure are different. The traditional Q-learning algorithm takes 32 steps. The Q (λ) learning algorithm takes 29 steps and the algorithm of this paper takes 27 steps.

Table 1. Convergence Performance Comparison

	Simple environment		Complex environment	
	N1	N2	N1	N2
Traditional Q learning algorithm(TQ)	60	9006	98	16897
Q(λ) learning algorithm	39	4681	53	9698
The algorithm of this paper	20	1947	25	4628

N1 represents the number of convergence steps, and N2 represents the total number of steps. It can be seen from Table 1 that the convergence speed of our algorithm is 1.95 times faster than Q (λ) and 3 times faster than TQ in a simple environment. the convergence speed of this algorithm is 2.12 times faster than Q(λ) and 3.92 times faster than TQ in a complex environment.

4. Conclusion

Although the traditional Q learning algorithm can find a collision-free path to the target point in an unknown environment, the convergence speed is slower. This article aims at the slow convergence speed of traditional algorithms in complex and unknown environments and cannot meet the real-time requirements. Based on the traditional Q learning, according to the actual situation, it is proposed that the RDL-Q learning algorithm also improves the action selection strategy in learning and the simulation environment. It has been verified that the experimental results show that the improved method has superiority and can learn an optimal path faster in a complex unknown environment. When the mobile robot is in a dynamic and large-scale environment, how to reduce the dimension and dynamic path planning of the Q-table is the next step to be studied.

Acknowledgments

At the time of this thesis, i would like to thank Associate Professor Lin. Can be successfully completed the writing of the paper, all embodies the teacher's effort and sweat. The teacher in the paper topics, research programs to determine and the specific implementation process have given careful guidance, their rigorous attitude and systematic research ideas I benefit for life.

References

[1]. Li S, Xu X, Zuo L. Dynamic path planning of a mobile robot with improved Q-learning algorithm[C]// IEEE International Conference on Information and Automation. IEEE, 2015:409-414.

[2]. Gao Hui. Research on Path Planning of Mobile Robot Based on Reinforcement Learning [D]. Southwest Jiaotong University, 2016.

[3]. Tijsma A D, Drugan M M, Wiering M A. Comparing Exploration Strategies for Q-learning in Random Stochastic Mazes[C]// IEEE SSCI. IEEE, 2016.

[4]. Song Y, Li Y B, Li C H, et al. An efficient initialization approach of Q-learning for mobile robots [J]. International Journal of Control, Automation and Systems, 2012, 10(1):166-172.

[5]. Xu Ya. Research on Path Planning of Mobile Robot Based on Reinforcement Learning [D]. Shandong University, 2013.

[6]. Liang Quan. Path Planning of Mobile Robot Based on Reinforcement Learning in Unknown Environment [J]. Mechatronic Engineering, 2012, 29(4):477-481.