Improved reinforcement learning algorithm for mobile robot path planning

Teng Luo*

School of Information and Control Engineering, Liaoning Petrochemical University, Fushun, China

Abstract. In order to solve the problem that traditional Q-learning algorithm has a large number of invalid iterations in the early convergence stage of robot path planning, an improved reinforcement learning algorithm is proposed. Firstly, the gravitational potential field in the improved artificial potential field algorithm is introduced when the Q table is initialized to accelerate the convergence. Secondly, the Tent Chaotic Mapping algorithm is added to the initial state determination process of the algorithm, which allows the algorithm to explore the environment more fully. In addition, an ε-greed strategy with the number of iterations changing the ε value becomes the action selection strategy of the algorithm, which improves the performance of the algorithm. Finally, the grid map simulation results based on MATLAB show that the improved Q-learning algorithm has greatly reduced the path planning time and the number of non-convergence iterations compared with the traditional algorithm.

Keywords: Reinforcement learning, Greedy strategy, Mobile robot, Path planning.

1 Introduction

Today, with a variety of robots in life, in order to make robots play a role in more fields, the problem of robot mobility has gradually attracted more attention. The primary problem in the whole process of mobility is the path planning problem. For mobile robots, the amount of environmental information can directly determine the method for path planning. If all obstacles in the path including the starting point and the end point can be understood and modeled, then the global path planning algorithm such as A* algorithm[1], dijkstra algorithm based on viewable[2], swarm intelligence optimization algorithm[3-4] can be used and have obvious advantages compared with other algorithms. If the understanding of the overall environment is limited, local path planning algorithms such as dynamic window method[5] vector histogram algorithm[6], artificial potential field algorithm[7] and reinforcement learning algorithm[8-11] are more advantageous. Among them, the reinforcement learning algorithm widely used in robot path planning is Q-learning algorithm. However, the traditional Q-learning algorithm has a lot of invalid exploration in the early stage and slow convergence in the later stage.

* Corresponding author: 1830173869@qq.com

Aiming at the problem of poor early search ability and slow convergence speed of Q-learning algorithm, this paper proposes an improved artificial potential field function as the initialization function of Q value, so that the algorithm can effectively explore the environment in the initial stage. At the same time, the ε-greed strategy is improved, the greedy factor is dynamically adjusted according to the number of iterations of the algorithm, and the convergence speed of the algorithm is accelerated on the basis of fully increasing of the environmental exploration. The Chaos mapping algorithm is introduced in the training process, so that the algorithm has better randomness and more uniform ability to explore the environment in the training process. Simulation results show that the improved algorithm can effectively find the optimal path and indeed accelerate the convergence rate.

2 Improved Q-learning algorithm

2.1 Q-learning algorithm

Q-learning algorithm is a model-independent off-line strategy sequential detection algorithm. The algorithm first initializes the Q value, and then lets the robot randomly select a starting state s . According to the ε-greed strategy, the action a is selected. After the action is selected, the reward value r and the next state s' can be obtained. According to the action estimation value Q of the next state s', the maximum value Q is used as the estimated value of the current state-action pair (s,a) . In this way, it is continuously iterative until the state is updated to the target state. The update formula is as follows:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max Q(s',a') - Q(s,a)] \tag{1}$$

where $Q(s,a)$ denotes the estimated value of the current state-action pair ; $\max Q(s,a)$ denotes the maximum estimated value of the state-action pair of the next state ; r denotes the reward for selecting execution under state s ; α denotes the learning rate and γ denotes the discount factor. Generally, the ranges of α and γ are $(0,1]$.

2.2 Q value initialization

The artificial potential field algorithm is a common robot path planning algorithm, which includes the gravitational field used to guide the robot to move to the target and the repulsion field generated by obstacles. In this paper, in order to improve the generality of the algorithm, the influence of the repulsion field of obstacles is not considered. The gravitational potential field function is:

$$U_{\text{att}} = \frac{1}{2}\eta\rho^2 \tag{2}$$

where η is the gravitational constant, and ρ is the distance between the robot and the target. However, this function produces less gravity when the robot is close to the target point, resulting in the difficulty of reaching the target point. Therefore, we propose the following improved function:

$$V(s') = \eta e^{-\frac{1}{2}[(x-\mu_1)^2+(y-\mu_2)^2]} \tag{3}$$

where η is the gravitational constant, and μ_1 , μ_2 is the transverse and longitudinal coordinates of the target state, and x , y are the transverse and longitudinal coordinates of the current stateand the Q value is initialized by formula:

$$Q(s,a) = r + \gamma \sum_{s' \in S} P(s' \mid s,a) V(s') \tag{4}$$

where r denotes the reward for selecting execution under state s and $V(s')$ is the value function where the algorithm is located, $P(s' \mid s,a)$ represents the probability of transition from s to s' in the state.

2.3 Chaotic mapping training

In the numerical training process of Q-learning, it requires a large number of random assignments to the initial position, so that the robot can obtain optimal strategies of multiple initial positions to find the global optimal strategy. Therefore, it is very important to make each initial state randomly. In this paper, chaotic mapping algorithm is used to select the initial state, and the formula is as follows:

$$s' = \begin{cases} s/\beta & s \in (0,\beta] \\ (1-s)/(1-\beta) & s \in (\beta,1] \end{cases} \tag{5}$$

where β is a random factor greater than 0, which is the occurrence rate of state s , and the occurrence probability of the next state s' at the same time.

2.4 Dynamic adjustment of ε value

Q-learning algorithm is an algorithm using ε-greed strategy. If a larger ε value is selected, the later convergence will slow down, and a smaller ε value will increase the difficulty of early exploration. In order to solve this contradiction, an improved ε-greed strategy is proposed in this paper. By dynamically adjusting the ε value, the algorithm can explore the environment with a small value in the early stage, and make full use of the known strategy in the later stage. The calculation formula of the ε value is as follows:

$$\varepsilon = \varepsilon_{\max} - (\varepsilon_{\max} - \varepsilon_{\min}) * (\varepsilon_{\min} * \mathrm{e} - \mathrm{e}^{-50(\frac{t}{T})^3}) \tag{6}$$

Where t is the current iteration number, T is the maximum iteration number, $\varepsilon_{\max}$ and $\varepsilon_{\min}$ are the maximum and minimum ε values set by the operator respectively.

3 Simulation and analysis

3.1 Simulation environment

In this paper, the 40 × 40 grid map builted by MATLAB shown in Figure 1 is used as the

simulation environment. The starting point at (1, 1) is the lower left point in Figure 1, and the end point at (40, 40) is the upper right point in Figure 1. The obstacles in the map are black grids, and other grids are accessible spaces. All grids correspond to a state in the algorithm. Each state has two executable action spaces, up and down.

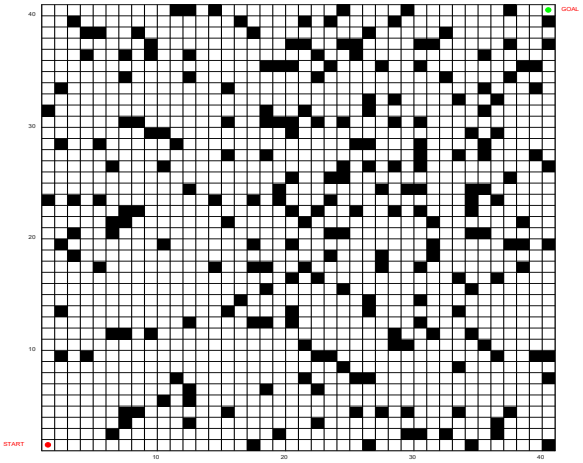

Fig. 1. Path planning simulation environment.

3.2 Simulation parameter

In the grid map above, the following two algorithms are simulated and compared. Trad_QL represents the traditional Q-learning algorithm, tepg_QL represents the improved algorithm proposed in this paper. The parameter β in the training of chaotic mapping algorithm is set to 0.6. The dynamic greedy factor of ε-greed strategy are as follows : $\varepsilon_{\max} = 0.9$, $\varepsilon_{\min} = 0.6$, $T = 10000$. For the gravitational function of the artificial potential field, the distance d is calculated by the actual coordinate and $\eta = 4$. Other parameters are all the same in all algorithms, where $\alpha = 1$ and $\gamma = 0.9$. Maximum number T of iterations is 10000. When the standard deviation of 10 consecutive iterations is less than 100, the algorithm is convergent. The reward function is set as:

$$r = \begin{cases} -100 \text{ if } s' \text{is an obstacle} \\ 100 \text{ if } s' \text{ is the target} \\ 0 \quad \text{if } s' \text{ is other position} \end{cases} \tag{7}$$

3.3 Results and analysis

By comparing the results of simulation experiments, we can make the following conclusions. After enough iterations, both two algorithms can present good results. The result of final path is represented in Figure 2. It shows the shortest path that the algorithm can obtained.

The picture in Figure 3 shows the convergence process of the above algorithms. Table 1 compares the performance of the two algorithms in detail. The data in Table 1 takes the average value of each algorithm after running for 10 times.

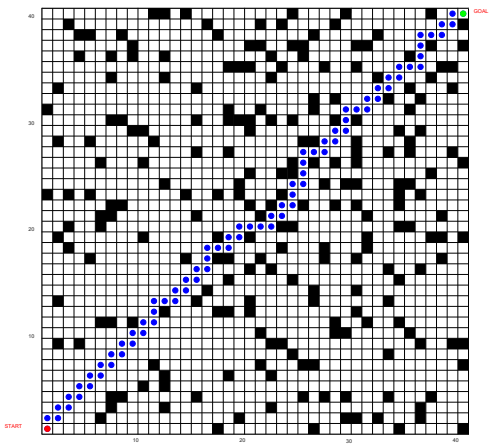

Fig. 2. Experimental results of improved Q-learning.

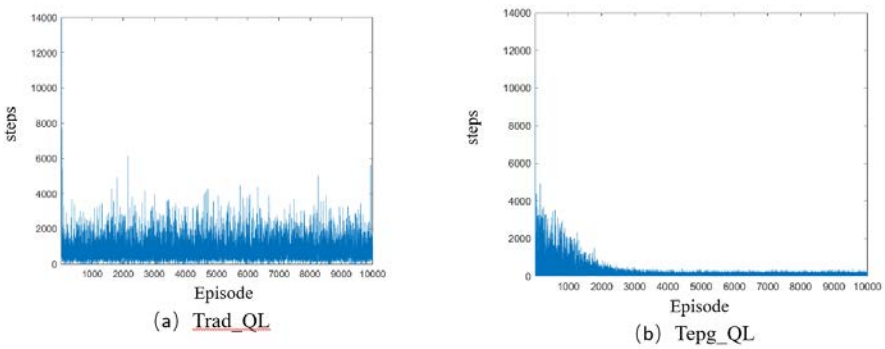

Fig. 3. Convergence Iterative Curves of Two Algorithms.

By analyzing the data in Table 1, it can be seen that although both two algorithms can find their optimal paths, the convergence of them are different. By comparing the algorithm Tepg_QL and the algorithm Trad_QL, it is clear that the convergence time of Tepg_QL is 82.5 % shorter than Trad_QL, and the average number of iteration until convergence is 82.6%

Table 1. Performance comparison of two algorithms.

Algorithm	Convergence time (s)	Number of convergences	Optimal path length
Trad_QL	576.58	644.4	79
Tepg_QL	15.84	67.3	79

4 Conclusion

In order to solve the problem of slow convergence rate of the traditional reinforcement learning algorithm in the path planning of mobile robots, this paper introduces an improved initialization of Q value in the Q-learning algorithm, and uses the Chaos mapping algorithm to select the initial position. The ε value is changing according to the number of iteration, and the greedy algorithm is used to select the action. The comparison of all the mentioned algorithms shows that the convergence efficiency of the improved algorithm is effectively improved.

References

1. Xiang Rong T, Yu kun Z, Xin Xin J. Improved A-star algorithm for robot path planning in static environment[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1792(1): 012067
2. Zhu D D, Sun J Q. A new algorithm based on dijkstra for vehicle path planning considering intersection attribute [J]. IEEE Access, 2021, 9: 19761-19775.
3. BLASI L, D'AMATO E, MATTEI M, et al. Path planning and real-time collision avoidance based on the essential visibility graph[J]. Applied Sciences, 2020, 10 (16): 5613.
4. Zhang Xu, Zeng Xiangxin, Lang Bo. Path planning of free-floating space robot based on control variable parameterization method [J]. Optical precision engineering, 2019, 27 (02): 372-378.
5. LEE H Y, SHIN H, CHAE J. Path Planning for mobile agents using a genetic algorithm with a direction guided factor [J]. Electronics, 2018, 7 (10): 212-232.
6. Fox D, Burgard W, Thrun S. Controlling synchro-drive robots with the dynamic window approach to collision avoidance[C]//Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS'96. IEEE, 1996, 3: 1280-1287.
7. Qu Yuanrang, Xue Jianru, Zhu Yaoguo, Xiao Peng. VFH algorithm for unmanned vehicle motion planning problem [J]. Computer simulation, 2018, 35 (12): 245-251.
8. Huang Bingqiang, Cao Guangyi. Research on mobile robot path planning based on artificial potential field method [J]. Computer engineering and application, 2006, 42 (27): 26-28.
9. LIU Q, SHI L, SUN L, et al. Path planning for UAV-mounted mobile edge computing with deep reinforcement learning [J]. IEEE Transactions on Vehicular Tech- nology, 2020, 69 (5): 5723-5728.
10. Zheng Bolong, Ming Lingfeng, Hu Qi, Fang Yixiang, Zheng Kai, and Li Guohui. Dynamic path planning of online Car-hailing based on deep reinforcement learning [J]. Computer research and development, 2022, 59 (02): 329-341.
11. Hu Xiaohui. A reinforcement learning action selection mechanism based on dynamic parameter adjustment [J]. Computer engineering and application, 2008, 44 (28): 29-31.