Reinforcement Learning in Robot Path Optimization

Qian Zhang
China University of Mining & Technology Xuzhou, 221116, China
Email: zhangqian374@126.com

Ming Li
China University of Mining & Technology Xuzhou, 221116, China
Email: liming@cumt.edu.cn

Xuesong Wang
China University of Mining & Technology Xuzhou, 221116, China
Email: wangxuesong@126.com

Yong Zhang
China University of Mining & Technology Xuzhou, 221116, China
Email: zhangyong@126.com

***Abstract*—Along with the development of robot technology, a robot not only need to complete a specific task, but aslo need to do path planning in the process of performing the task. So, path planning is widly studied. This paper introduce a method of robot path planning based on reinforcement learning,which aimed at Markovian decision process. In this paper, we introduce the basic concept, principle and the method of reinforcement learning and some other algorithms.Then,we do research from single robot's path planning in the static invironment based on Q-learning, and describe the application of this algorithm on the path planning by setting off state space and action space reasonablly and designing reinforcement function.By edditting Matlab program,we do some simulation experiments,which incarnate the algorithm visually and get the optimal path.**

***Index Terms*—reinforcement learning, markov decision process, Q-learning**

I. INTRODUCTION

With the development of modern industrial technology and computer technology, robotics has also been rapid development. Due to the mobility, mobile robot replace human in space and deep-sea operation, precision operation, high temperature and other aspects of the key technology and equipment, and have important military and civil value. However, the mobile robot will encounter a variety of obstacles in explore environment. The key measure of its performance indicators is escape these obstacles flexibly and rapidly.

To adapt to unknown environments, learning ability is the key to mobile robot intelligence. Currently, in the field of machine learning, according to the different of feedback, the learning technologies can be divided into three categories: supervised learning, unsupervised learning and reinforcement learning. Since late eighties of last century, reinforcement learning has gradually become a research focus, and widely used in intelligent control, robotics, and decision analysis and other fields[1].

II. MOBILE ROBOT

A. Classification of Mobile Robot Path Planning

When path planning, mobile robot often require a certain criteria along an optimal (or suboptimal) path to walk in the work space. Robot path planning problem can be modeled as a constrained optimization problem, and complete the path planning, positioning, obstacle avoidance and other tasks. According to environmental information and different obstacles, mobile robot path planning can be divided into the following categories: the known environment of static obstacles; unknown environment of static obstacles; known environment for dynamic obstacles; unknown environment for dynamic obstacles.

B. Mobile Robot Path Planning Research

With environmental information, path planning method is the collision-free path to the robot from start

Manuscript received Mar. 20, 2011; revised Apr. 1, 2011; accepted Jan. 12, 2011.

Project number: the National Natural Science Foundation of China(60804022, 60974050)， Specialized Research Fund for the Doctoral Program of Higher Education of China (GrantNo.: 20070290537)， the 'Qing-Lan' Program for Young Excellent Teachers of Jiangsu Province and the Scientific and Technological Foundation of China University of Mining and Technology (Grant Nos.: 0C080302).

Corresponding author:zhangqian374@126.com, 13952182374

doi:10.4304/jsw.7.3.657-662

point to target point. Researchers in mobile robot path planning has been done a lot of research. Method[2] used visibility graph to view the problem of searching optimal path into the target point from the start point to a straight line through the shortest distance between these visual problems. Lee proposed an high-level robot fuzzy navigation method in unknown environment, with the ultrasonic sensor to provide environmental information, and the navigation based on fuzzy control to calculate the information[4]. Lebedev proposed high efficiency path planning method using a discrete time and dynamic environment neural network in dynamic environment, and got environmental information in learning process[5]. Davies modeling using grid method, and design a method in 3D static and dynamic environment with a genetic algorithm[6].

III. DISSIPATIVE STRUCTURE MODEL

A. The Concept and Principle of Reinforcement Learning

Reinforcement learning (RL) is a real-time, online learning methods. Interacting with the environment, robot select and perform actions and affect environment. At the same time, according to the reinforcement signal which is environment given to, robot constantly adjust their actions and find the optimal movement strategy through trial and error method, and the system behavior obtain the maximum value of the accumulated reward from the environment. Fig.1 shows the schematic of reinforcement learning, in which the environmental must be Markov.

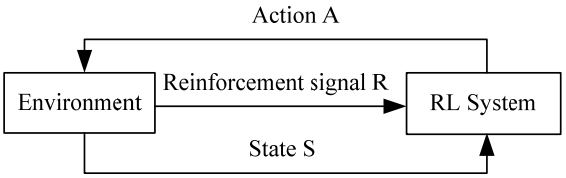


Fig.1 Schematic of Reinforcement Learning

The purpose of robot learning is to maximize the sum of reinforcement function, therefore rational allocation helps to improve learning efficiency. Reinforcement signal is usually a scalar, often expressed positive encouragment, negative punishment. Living environment for robot is described as a set of possible states S, which may perform probable action set A, each in a state of the implementation of an action a_t, the robot receives a real return r_t. Robot's task is to learn a control strategy $\pi : S \to A$ to maximize the expected value of these returns, and the followed return value reduced as delay of index.

B. Q Learning

To learn Q, the key is to find a reliable way to estimated training value on the basis of immediate return sequencesed expand in the timeline only.

$$Q(s,a) = r(s,a) + \gamma \max_{a'} Q(\delta(s,a), a') \qquad (1)$$

In this algorithm, learning machine represent assumptions Q with a table, with each state-action pair a entry, which stored the current assumptions of $Q(s,a)$. Agent repeatedly observe the current state s, select and execute a certain action a, observe the result return $r = r(s,a)$ and the new state $s' = \delta(s,a)$. Then agent updated each such entry with the conversion. Rules are as follows:

$$\hat{Q}(s,a) \leftarrow r + \gamma \max_{a'} \hat{Q}(s', a') \qquad (2)$$

This training algorithm used the agent to refined its previous state estimate with current Q value of the new state. Although (1) described with function $\delta(s,a)$ and $r(s,a)$, but the agent training rules with (2) need not know these general functions.

The Q learning algorithm of deterministic MDP is accurate described in Table 1. Using this method, agent estimated Q value converging to the actual Q function in limit.

TABLE I. Q-learning Algorithm Under the Certainty Return and Action

Q learning algorithm
initialized entries for each s,a are 0
observe the current state
has been repeated:
- Select an action a and execute it
- receive immediate reward r
- observe new state s'
- update the entry accordance to the following formula:

$$\hat{Q}(s,a) \leftarrow r + \gamma \max_{a'} \hat{Q}(s', a')$$
$$s \leftarrow s'$$

C. Non-deterministic Reward and Action

We considered the Q-learning in deterministic environment above. Here we consider non-deterministic case, in which return functions $r(s,a)$ and actions conversion functions $\delta(s,a)$ may have probability outputs. In non-deterministic case, we must restate the target learning machine to consider the output is no longer a certainty situation. Obviously, a general approach is to redefine the value V^{π} of a strategy for expected of discounted cumulative return with this strategy before. Return series r_{t+i} generated in the strategy π by starting with the state s. Define the optimal strategy π^* as strategy π that to maximize $V^{\pi}(s)$ in all state s.

$$Q(s,a) = E[r(s,a)] + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s', a') \qquad (3)$$

The training rule in deterministic case above can not convergence in non-deterministic conditions. By

modifing training rule, using $\hat{Q}_n$ in place of agent's estimation in n-times cycle, the following revised rules ensure $\hat{Q}$ converge to Q.

$$\hat{Q}_n(s,a) \leftarrow (1-a_n)\hat{Q}_{n-1}(s,a) + a_n\left[r + \gamma \max_{a'} \hat{Q}_{n-1}(s',a')\right] \quad (4)$$

$$which \quad a_n = \frac{1}{1+visits_n(s,a)}$$

In this revised rule, the key idea is updating $\hat{Q}$ more smoothly than deterministic case. Value a_n decreases with the increase of n, and update frequency become smaller gradually during training, this can achieve the purpose of convergence to right Q function.

VI. Simulation Results

To validate the effectiveness of the proposed algorithm, we use two different types of grid world, in which agent exploration walk based on algorithm. Two cases for experiment, deterministic MDP and non-deterministic MDP. In deterministic MDP, agent walking under the algorithm to seek a path with the least number of steps. In another complex environment, after implementing action, the position is impacted by environment, and we can only try to find a suboptimal path. There are two environment state in simulation, one is simple aggregation obstacles problem, another is the scattered obstacles problem.

Q learning algorithm parameters are set as follows: in deterministic MDP, the discount factor γ = 0.9, with greedy exploration strategy algorithm. In non-deterministic MDP experiment, the discount factor γ = 0.9, with ε-greedy strategy algorithm, where ε = 0.5, and random probability reach to next state.

A. Deterministic MDP

1) Simple Aggregation Obstacles Environment

Figure 2, in 12*12 two-dimensional grid environment, " * " indicates environmental boundaries and barriers, "○" indicates the agent's start point, "☆" indicates the target point, agent can walk along the grid lines with each walking step to reach a grid line intersection.

Before learning, due to the completely unknown environment information, how to arrive from one state to another is decided by the action. Agent can choose a straight line walk from four actions,that up, down, left and right. Thus, each state corresponding to the four actions, there will be a four state-action pairs. Agent performs an action will get a return value, it will return 100 if get to the end, and reset agent to start point; and other actions are executed a 0 return value. When the agent access to the boundaries or barriers, discount accumulated return is 0. After several visits, each discount accumulated return of state-action pairs are converge to a determined value. Under greedy policy, agent get an optimal path by selecting the action that the maximum value of discount accumulated return.

Circles "○" marked in Fig.3 indicate the path agent passed. In this environment, agent have multiple optimal paths. It can be seen, agent selected a walking path along the border which far away from obstacles, and had not back phenomenon. This a very good proof of the effectiveness of this algorithm.

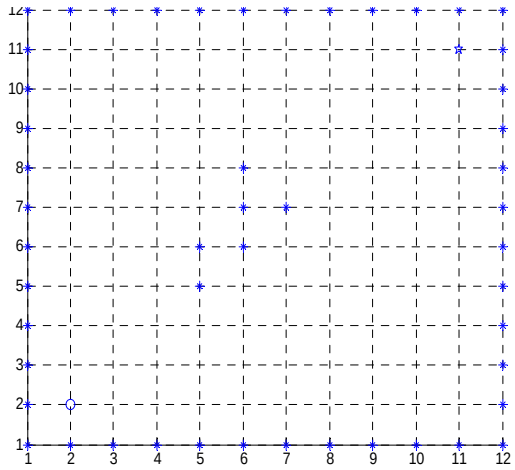


Fig. 2 Simple aggregation obstacles environment

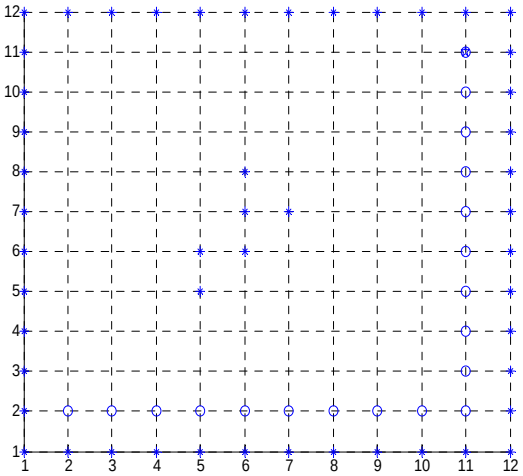


Fig.3 Path planning in above environment

2) *Scattered Obstacles Environment*

The environment as Fig.4 represents are same with the above experiment, except the number and distribution of obstacles. In this environment, the obstacles are not just focused on the central of environment, but rather to disperse, and this test barrier capacity more explicit.

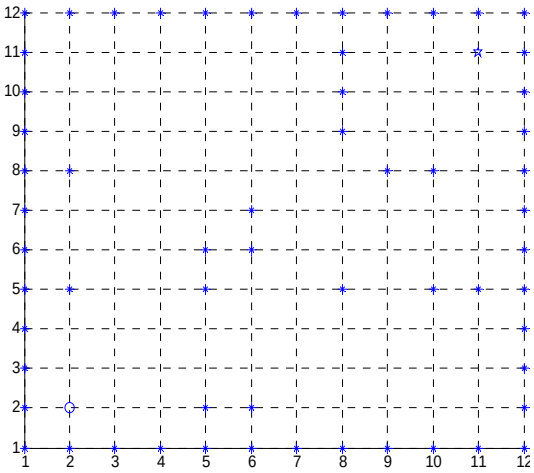


Fig.4 Scattered obstacles environment

As shown in Fig.5, agent still be able to avoid obstacles and find an optimal path to reach the target point, this indicating that the algorithm can effectively achieve barrier and exploration in ideal environment.

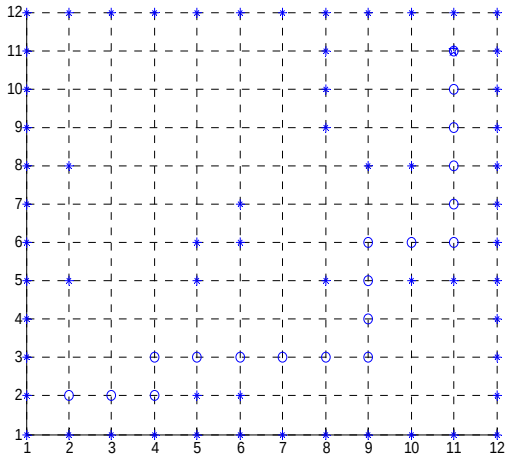
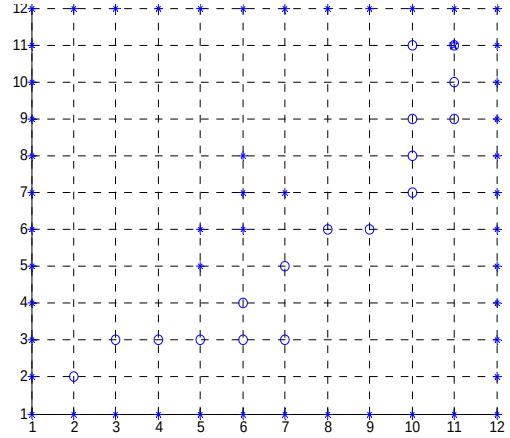


Fig.5 Path planning in above environment

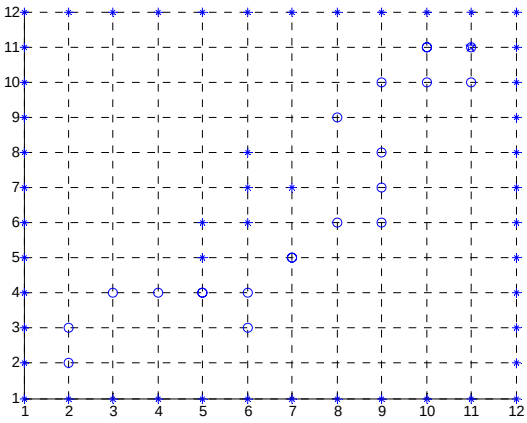
B. Non-deterministic MDP

Environment settings are same with deterministic MDP. After implement the ation, agent are not be able to reach the desired next state, might deviate from this state to the other state. By several exploration, each discount accumulated return of non-deterministic state-action pairs are converge to approximate value. The results are as follows.

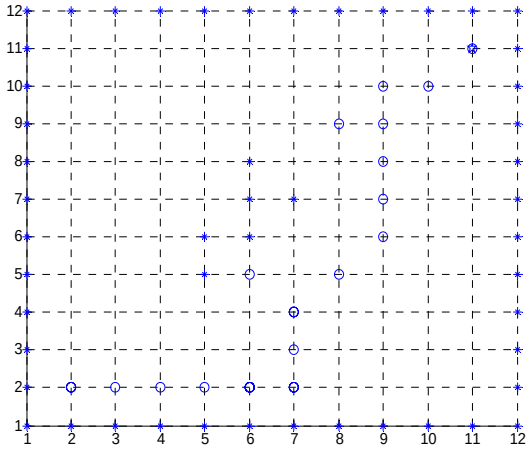
In Fig.6, agent used 15, 19 and 24 steps from start point to target point. Although the environment is uncertain, but according to discount accumulated return updated from the state-action pairs, agent can find the optimal path also in non-deterministic situation.



(a) 15 Steps discover



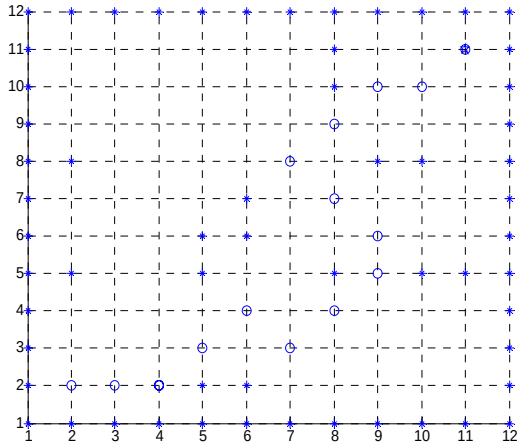
(b) 19 Steps discover



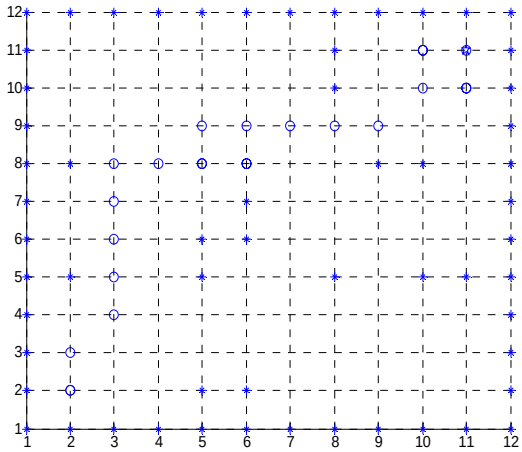
(c) 24 Steps discover

Fig.6 Simple aggregation obstacles environment

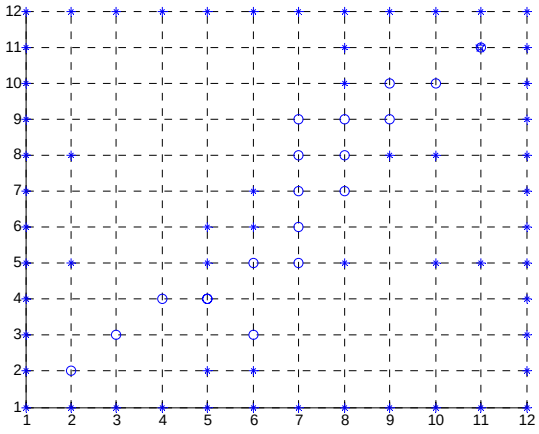
In Fig.7, like aggregation obstacles situation, there simulation results are also different. Despite the different distribution of obstacles, agent still explore the optimal path in non-deterministic MDP.



(a) 15 Steps discover



(b) 22 Steps discover



(c) 17 Steps discover

Fig.7 Scattered obstacles environment

V. CONCLUSION

Mobile robot path planning problem is to find an optimal path from initial state to target state avoiding obstacles in work space, based on some optimization criterion. In order to solve the path planning problems, this paper introduces proposed a Q learning algorithm, and carried out experiments under Matlab. The simulation results show that this method can achieve good effects.

ACKNOWLEDGMENT

We gratefully acknowledge the financial support from the National Natural Science Foundation of China(60804022, 60974050)， Specialized Research Fund for the Doctoral Program of Higher Education of China (GrantNo.: 20070290537)， the ‘Qing-Lan’ Program for Young Excellent Teachers of Jiangsu Province and the Scientific and Technological Foundation of China University of Mining and Technology (Grant Nos.: 0C080302)

REFERENCES

[1] C.L. Chen, Autonomous Learning and Navigation Control for Mobile Robots Based on Reinforcement Learning.Dissertation for Ph.D.2006.

[2] X.D. Zhuang, Q.C. Meng, B. Yin, A Method of Robot Path Searching in Dynamic Enviroment Based on Fuzzy Control. Robot, 2001,05.

[3] T.H. Lee, H.K. Lam, F.H.F.Leung, P.K.S.Tam, Apraetieal Fuzzy Logie Controller for the Path Tracking of Wheeled Mobile Robots[C].IEEEControl System Magazine,2003.

[4] C.Davies, P.Lingras.Grentie Algorithms for Rerouting in Dynamic and Stoehastie Networks[J]. European Journal of Operational Research,2003.

[5] W.G. Liu, Z.P. Chen, Y. Zhang. MATLAB Program Design and Application. Higher Education Press.2003.

[6] J.H. Chu, improvement of Q-learning Reinforcement learning algorithm and its application. Master thesis.2009.

[7] E. Yang, D.B. Gu. A Multiagent Fuzzy Policy Reinforcement Learning Algorithm with Application to Leader-Follower Robotic Systems. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems.2006.

[8] R.Y. Sun, G. Zhao, An Efficient Multi-Agent Q-learning Method Based on Observing the Adversary Agent State Change. IEEE International Conference on Systems, Man, and Cybernetics.2006.

[9] Y. Zhang; J.P. Liu, Research on Improvement of Q-Learning and Its Simulation Experiments. Computer Simulation.2007,10.

Qian Zhang AnHui Province, China. Birthdate: December, 1981. is Control theory and control engineering doctorates, graduated from School of Information and Electrical Engineering China University of Mining & Technology. And research interests on reinforcement learning and robot path optimization.

Ming Li AnHui Province, China. Birthdate: Mar, 1962. is Ph.D., Professor, Associate Dean of China University of Mining Electrical letter. Colonel young academic leaders, Jiangsu Automation Society, member of China Coal Society of General Automation.And research interests on data mining, complex networks, business intelligence.

His research interests include pattern

recognition and intelligent systems, power electronics and power transmission. Involved in 863 projects a bear more than a dozen companies commissioned by the project. Research by the China Coal Industry Science and Technology Progress Award 1, Jiangsu Provincial Science and Technology Progress Award 1, the State Coal Industry Bureau of Science and Technology Progress Award for a second, third prize of scientific and technological progress of China Coal Industry 1. Published a monograph, published more than 30 academic papers, which were SCI, EI retrieve more than 10 articles.

Xuesong Wang AnHui Province, China. Birthdate: Dec, 1974. is Ph.D., China University of Mining and Technology Professor. Party branch secretary of Automation Institute , was selected by "New Century Excellent Talents"of Ministry of Education, outstanding young teachers in Jiangsu Province "project blue". University outstanding young academic leaders. Research interests include machine learning, intelligent robotics, bioinformatics and so on.

Prof. Wang is the National Natural Science Foundation and the Doctoral Fund of Ministry of Education letter of evaluation experts, Jiangsu Provincial Department of Education dissertation quality assessment experts; as IEEE Journal of Automation, Computers, Electronics of more than 20 international and domestic well-known journals Contributing reviewer.

Yong Zhang ShanDong Provice, Chian. Birthdate: Sep, 1979. is Ph.D., Master Instructor. Lecturer for School of Information and Electrical Engineering China University of Mining & Technology. Research interests include swarm intelligence and multi-robot coordination control.

Dr. Zhang is contributing reviewer for domestic and foreign journals like "International Journal of Electrical Power and Energy Systems" and "Automation Technology". Recently published (employment) of more than 20 academic articles, which were retrieved 14 SCI and Ei.