International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.7 (2015), pp.285-294
http://dx.doi.org/10.14257/ijmue.2015.10.7.30

Mobile Robot Path Planning Based on Improved Q Learning Algorithm

Jiansheng Peng

Guangxi Colleges and Universities Key Laboratory Breeding Base of System Control and Information Processing Hechi University, Yizhou 546300, China
shen120410@163.com

Abstract

For path planning of mobile robot, the traditional Q learning algorithm easy to fall into local optimum, slow convergence etc. issues, this paper proposes a new greedy strategy, multi-target searching of Q learning algorithm. Don't need to create the environment model, the mobile robot from a single-target searching transform into multi-target searching an unknown environment, firstly, by the dynamic greedy strategy exploring interim to use unknown environment, improve learning ability that mobile robot learn the environment, improve the convergence of the mobile robot speed. And a large number of improved Q-learning algorithms are applied to mobile robot optimization simulation in unknown environment, by comparing with traditional Q algorithm, theory and experiment proved that improved Q-learning algorithm speed up the convergence rate of the robot, improve collision avoidance capability and learning efficiency.

Keywords: *Improved Q learning algorithm; Dynamic greedy strategy; Multi-objective; Path planning*

1. Introduction

With the vigorous development of mobile robot in applications, the demanding for mobile robot path planning have become increasingly, in a different and complex environment, mobile robots must be able to autonomous complete various intelligent tasks. Person is able to adapt to the environment, people continue to learn from the environment, explore a variety of environmental models, thereby gradually improving decision-making and behavior to get inspired.

Therefore, reinforcement learning algorithms apply human mind in an unknown environment to complete the smart task. So far, with the development of reinforcement learning algorithm that unsupervised learning has been able to study the complex physical environment, has the capable of interaction with the environment, using the basic idea of dynamic programming, gradually improve the skills, achieve the intelligent algorithms for the purpose of optimization strategy sequence. Common reinforcement learning algorithm has y algorithm, Q learning algorithm, SARSA algorithm, R learning algorithm, Q learning algorithm is one of an important enhancement learning algorithms, The structure of algorithm is simple, without the supervision of learning, be able to work in an unknown environment, and applied to various applications, such as Braga *etc*. research mobile robot in an unknown environment, based on reinforcement learning algorithm for navigation methods; literature [3] use improved Q-learning algorithm in the application of about shop scheduling. in the development process of the Q-learning algorithm; literature [4] use the Q-learning algorithm only can traffic control research for urban area. Many scholars have achieved good results in improved reinforcement learning algorithm is applied to mobile robots, such as Jing and Peng proposed multi-step Q-learning algorithm [5], have verified improved Q learning algorithm better than the traditional Q-learning

ISSN: 1975-0080 IJMUE

algorithm in certain environments, but its slow convergence, a large scale of action, occupying a large content, increasing the amount of computation; literature [6] by the greedy strategy of improved Q-learning algorithm applied to the robot arm trajectory planning. Therefore, the greedy strategy has been designed become the main study object of Q-learning algorithm, this paper based on dynamic programming Q-learning algorithm, in the situation of without needing to establish an environmental model, mobile robot from explore the environment turn to use the environment, from single target turn to multi target, and through the application of improved Q-learning algorithm for mobile robot working simulation in hazard, the simulation experiments verify the effectiveness of the improved Q-learning algorithm.

2. Q Reinforcement Learning

Reinforcement learning known as continue to learn, is an learning algorithm of unsupervised learning and autonomous learning, the algorithm by the way of trial and error to find the optimal behavioral strategies, intelligent agent after select an action in set actions, conducting evaluation the action, also give out state transition and reward r. Q learning algorithm that use dynamic programming method calculate the optimal value under numerical iterative environment. Q learning accumulated reward function $Q(s_t, a_t)$, select the appropriate action a_t by decision, when the state of s_t executed the action a_t , and get the accumulation reward r, the state transition to the next state $Q(s_{t+1}, a_{t+1})$. So $Q(s_t, a_t)$ is a two-dimensional Q table, the formula is:

$$Q(s_t, a_t)=Q(s_t,a_t)+\alpha\left[r+\gamma \max Q(s_{t+1}, a_{t+1})-Q(s_t,a_t)\right] \quad (1)$$

Where, $\alpha \in (0,1)$ is the learning rate, $\gamma \in (0,1)$ is the discount factor, s_{t+1} and a_{t+1} respectively is the next state and movement. R is the enhanced signal. After numerous learning, each appropriate state receives the corresponding reward, namely the value of each state converges to a fixed value, the formula is as follows:

$$\max_{a_t \in A} Q(s_t, a_t) \to E\{\sum_{n=1}^{\infty} \gamma^{n-1} r_{t+n}\} \quad (2)$$

The framework map of Q reinforcement learning system model, as shown in Figure.1.

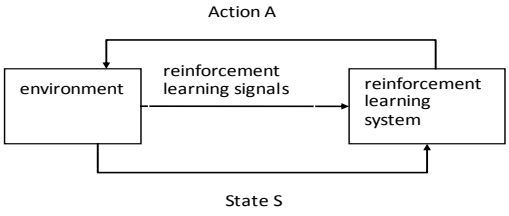

Figure 1. Q Reinforcement Learning System Model Diagram

Each learning of intelligent agent can be seen as the beginning of a random state, the policy can be used to search all the actions of unknown environments, selected action, and given some reward, updated the corresponding of $Q(s_t, a_t)$. As the traditional Q learning algorithm has some limitations on selection policy, so it is of slow convergence and poor locality *etc*. characteristics.

3. Dynamic Greedy Strategy

As the saying goes, "Failure is the mother of success", when the robot in the unknown environment, need to choice every action of their own, robot must learn in their own environment, summed up his experience in the learning process, namely when encountered obstacles, need to punish the action, when this action can be selected, need to reward the action

Therefore, dynamic selection strategy is the main content of reinforcement learning algorithm, which determines the behavior of the learning system to select actions. Namely start from the initial environment states in mobile robot learning system, after repeated the selection of state and the operation, achieve the ultimate goal, in this process, mobile robot need to make each operation, need to use the action selection strategy π . Mainly has greedy strategy, ε_Greedy strategy and random strategy [7] in common action selection strategy.

In the greedy strategy, each step choose the greatest reward of action, that is, when the state of s_t steering s_{t+1} , need to through greedy strategy $a_{t+1} = \arg\max Q(s_{t+1}, a_{t+1})$ to select the greatest action, conduct the function $Q(s_t, a_t)$ steering $Q(s_{t+1}, a_{t+1})$, However, this method is easy appear localized optimum.

For the shortcomings of greedy strategy, many researchers often use ε_Greedy strategy, ε_Greedy strategy randomly choose action adopted ε as the probability, choose the best action for the probability $1-\varepsilon$. This method as long as sufficient learning and attempts, namely countless times learning each action in the unknown environment, so it can find the global optimal action, but the method of downside is that search strategies need to several tests to find a suitable parameter ε before exploring, and intelligent agent is of a certain blindness is and itself is not targeted for action in unknown environment.

Q learning algorithm use two different greedy strategies, namely the use and exploration, by the two strategies, search the optimal path. In the traditional Q learning algorithm, generally selected greedy strategy, with ε represents the proportion between the search and the use of knowledge, When ε is too small, the system is more important than the use while blindly exploration, but make the system easy to fall into local optimum, namely early maturity, maybe could not get the optimal solution; when ε is too large, the system is more important than exploration, enabling long term improve system performance , finally maybe get the optimal solution, but the slow convergence speed of the system. Thus, both contradict each other.

Therefore, this paper based on ε value for the importance of algorithm, use the properties of the exponential function, improved the algorithm. In the learning process, first of all, explore the environment, accumulate more experience, and finally gradually moving to use, choose better actions. Namely the number of learning of mobile robot more, the more familiar to the unknown environment, it will not learning for some actions. Through the number of learning changing, so that s value constantly changes, namely the number of learning as independent variables function, when the number of learning becomes larger, s value declining, so make the algorithm from exploration transform into use. The function is:

$$p(a_i \mid s_k) = \frac{e^{Q(s_k, a_i)\varepsilon(k)}}{\sum_{n=1}^{N} e^{Q(s_k, a_m)\varepsilon(k)}} \tag{3}$$

Among them

$$\varepsilon(k) = 1 - 0.49 \times (2^{k/N}) \tag{4}$$

Where N is the number of learning, k is independent variables of the number of learning, early learning, ε is relatively large, intelligent agent mainly explore the environment, constantly aware of the unknown environment, as k increases, ε decreases, intelligent agent for most priority action's demanded increasingly high, intelligent agent transition from use to exploration, choose the best action. When m greater, the more learning time of intelligent agent for the unknown environment, the more familiar for environment, learning late, intelligent agent for the global optimal action more seriously. Namely get the global optimum action.

4. Multi-target Search

Inside the traditional Q learning algorithm, algorithms are the ultimate goal of for the purpose that always looking for the whole unknown environment, which makes the algorithm has some limitations and blindness, each study must search the final destination, so that intelligent agent add sport steps and time of learning. Therefore, this paper put forward increase the number of targets, namely conducting the search around the true target and the radius is R of the area, looking for the action that could reach the true goal, these actions called pseudo targets. Such makes the algorithm to search to find the true goal and pseudo target of learning, make the mobile robot more directional and goal in the environment.

Improving single target Algorithm, firstly, study and search the ultimate goal, it do not returned to the initial position, however, find a pseudo target around the target and radius R of the area, then correspondingly incentives pseudo target, prove that the pseudo target could reach the target place, Namely in the latter part of the study, as long as search to pseudo target or true goal, then abort the study. Thereby, increase the efficiency of that the mobile robot searching in unknown complex environment. In the latter study, eliminate the pseudo target, so let intelligent agent to consider the global movements, find the optimal action.

When searching the surrounding environment in the target, proposed another reward method, namely

$$Q(s^{'},a^{'}) = \delta \times r \times Q(s_g,a_g) \quad 0 < \delta \le R \tag{5}$$

$Q(s^{'},a^{'})$ is value function of pseudo target, is value function of true target. $Q(s_g,a_g)$ is searching radius, $0 < \mathrm{r} \le 1$ is the reward, namely the farther away from the true goal, its $Q(s^{'},a^{'})$ smaller.

5. Improved Q Learning Algorithm is as Follows

Step1: Initialize Q(s,a)=0;

Step2: According twofold dynamic greedy strategy, select an action from the initial unknown.

Step3: Given immediately return r for the selected action;

Step4: When face with an obstacle, don't restart from the initial position, find a viable path action in original place, until search to the end of point, before returning to the starting point.

Step6: Online update Q value function, calculate the cumulative return and update the Q value

$Q(\mathrm{s}_t, a_t)=\mathrm{Q}(\mathrm{s}_t,a_t)+\alpha(r+\gamma \max(Q(s_{t+1},a_{t+1})\text{-}Q(s_t,a_t))$

Step7: When the first time o search the true goal, searching for the pseudo target around the true goal and radius R of the area, and given reward to the pseudo target r. Updated Q value function of the pseudo target.

$Q(s^{'},a^{'}) = \delta \times r \times Q(s_g,a_g)\ 0 < \delta \le R$

Step8: Conducting second study, looking pseudo target or true target, repeat steps 2-6, until the end of the study.

6. Simulation

In order to facilitate compare performance of improved Q-learning algorithm with traditional Q learning algorithm, this time to compare with traditional Q learning algorithm, the traditional Q learning algorithm and improved Q-learning algorithm has the same factor: the learning rate α=0.1, the discount rate γ=0.6; traditional Q-learning algorithm has ε_Greedy strategy, and ε=0.8 through different learning times. Through compare the convergence rate of both. Each cell is an action of mobile robot, when confronted obstacles, does not return to

Starting point, while go back and continue to look for the true goals and pseudo target, when looked for true target or pseudo target, recording complete a learning cycle, the mobile robot return to starting point, continue the next round of learning. Various grids as environment, as shown in the Figure below, each grid represents a state of the mobile robot, the coordinate of S point as the start point, G point as the end point (namely true goal), black grid as obstruction. Mobile robot is unknown environment for this, mobile robot has around four optional actions, the results of simulation experiments as follows.

Figure. 2 and Figure. 3is the number of steps needed to reach the target for each search in the 10 * 10 grid:

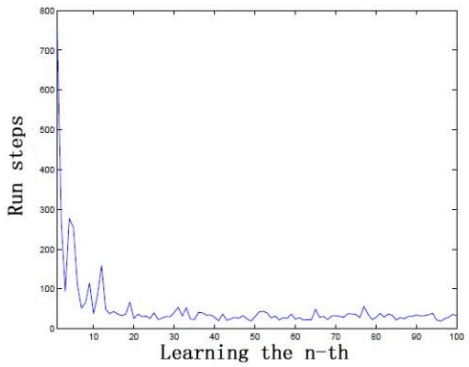

Figure 2. Traditional Q Learning Algorithm

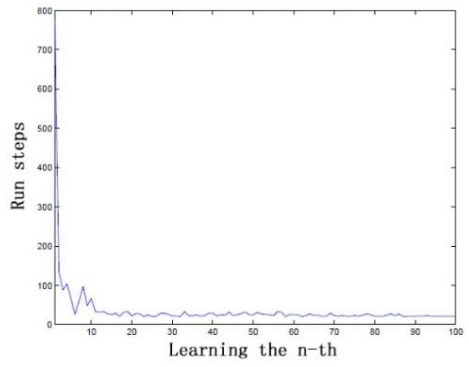

Figure 3. Improved Q-Learning Algorithm

Figure 4 and Figure 5 is the number of steps needed to reach the target for each search in the 20 * 20 grid:

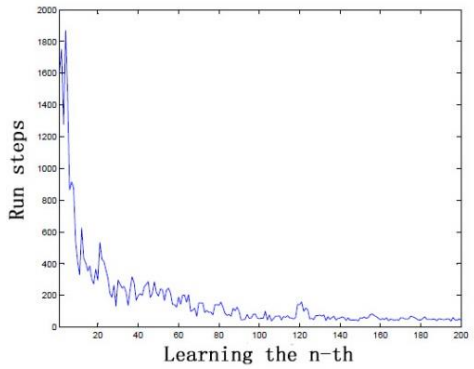

Figure 4. Traditional Q Learning Algorithm

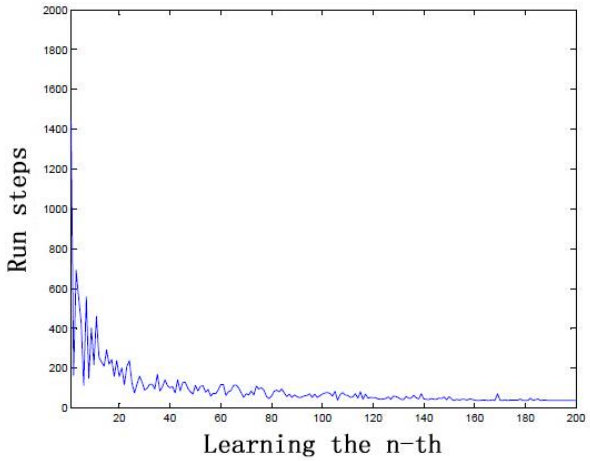

Figure 5. Improved Q-Learning Algorithm

Figure 6 and Figure 7 is the number of steps needed to reach the target for each search in the 30 * 30 grid:

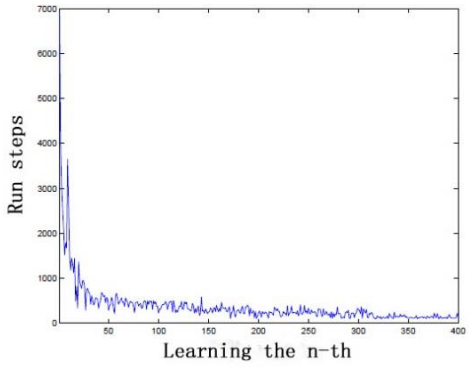

Figure 6 Traditional Q Learning Algorithm

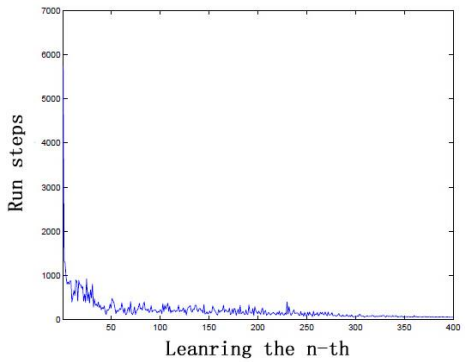

Figure 7 Improved Q-Learning Algorithm

Figure 8 and Figure 9 is motion trail diagram of 10 * 10 search results:

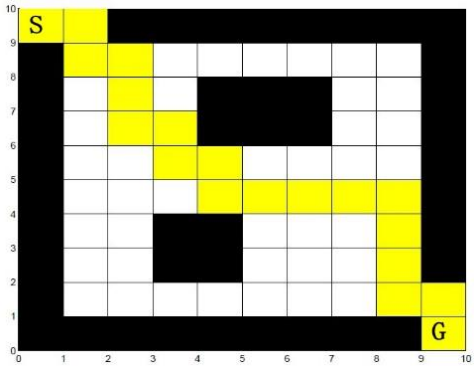

Figure 8 Traditional Q Learning Algorithm

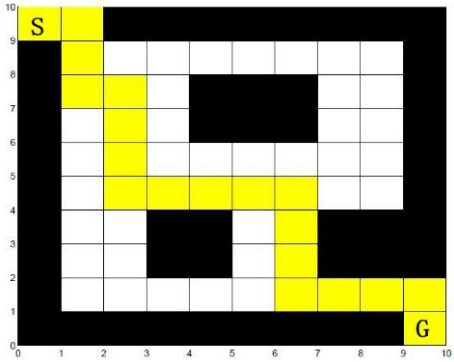

Figure 9 Traditional Q Learning Algorithm

Figure 10 and Figure 11 is motion trail diagram of 20 * 20 search results:

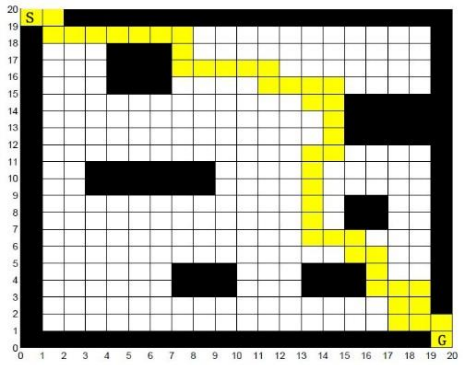

Figure 10. Traditional Q Learning Algorithm

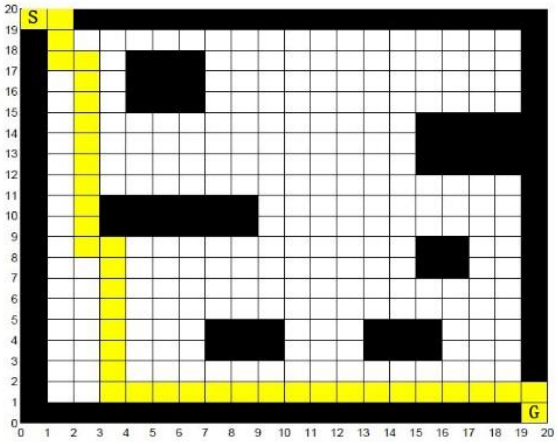

Figure 11. Improved Q-Learning Algorithm

Figure 12 and Figure 13 is motion trail diagram of 30 * 30 search results:

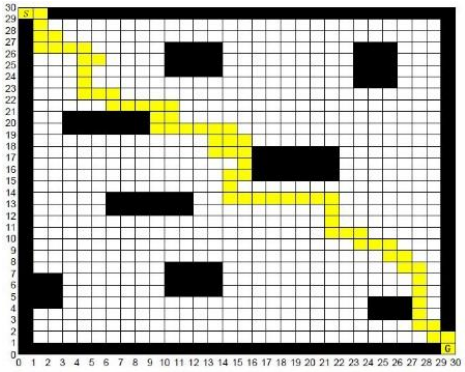

Figure 12 Traditional Q Learning Algorithm

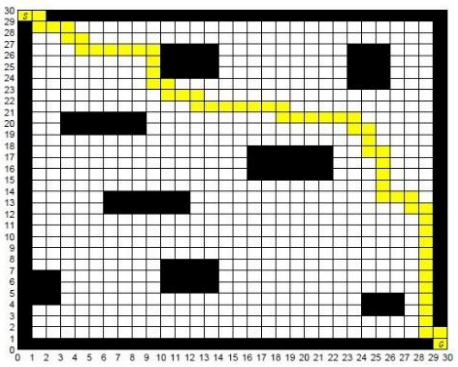

Figure 13 Improved Q-Learning Algorithm

Table 1. Is a Comparison of the Number of Steps of the Two Algorithms Running

	search the maximum number of steps			search results of steps			the average number of steps		
map	10*10 grid environment	20*20 grid environment	30*30 grid environment	10*10 grid environment	20*20 grid environment	30*30 grid environment	10*10 grid environment	20*20 grid environment	30*30 grid environment
traditional Q learning algorithm	756	1832	6906	19	44	70	51.24	121.25	373.24
improved Q-learning algorithm	780	1445	5474	19	39	58	36.79	97.92	213.60
raise the percentage	3.1%	21.1%	20.7%	0%	11.3%	17.1%	22.5%	19.2%	42.3%

Through simulation data in different environments shows, in more complex environments, the rate of convergence of improved Q-learning algorithm is faster, the number of required steps to search results is fewer, learning ability is stronger.

Therefore, in the simulation calculation results, it can be seen that at the beginning of the study, it is unknown that mobile robot for the overall environment, no experience, gain experience only through continuous looking for true goal or pseudo target in the process. Search results motion trail display, the traditional Q-learning algorithm still appears confusion in the end of learning, and don't select action pertinence, the rate of convergence is slow. While for the improved Q learning algorithm, with the mobile robot continuously explore on an unknown environment, and continuously improve their ability to avoid obstacles, which targeted began to select the optimal action in the middle of the learning, the whole learning step is gently, gradually reduce the number of steps of learning, prove its convergence speed fast. the track of improved Q-learning algorithm tends to global optimal path, the path track flat, while the track of traditional Q learning algorithm appears many repetitive movements, path confusion.

7. Main Text

This paper put forward a dynamic strategy and multi-objective search for improved Q-learning algorithm, the convergence rate of traditional Q-learning algorithm is slow, easy to fall into the local optimum *etc*. defects, and defect of ε_Greedy strategy, compare improved Q learning algorithm with traditional Q-learning algorithm, the total number of steps decrease in the late of learning, and the searching result tend to global optimal. Therefore, greatly improve the learning speed, optimization and the ability of adapt to environment through improved Q learning algorithm, overcome the local optimum and slow convergence *etc*. problem. The strategy is simple, effective, and better for use in a variety of environments, Through the use of mobile robot optimization simulation, results show that the improved algorithm is effective, feasible. Therefore, improved Q-learning algorithm has a good prospect.

Acknowledgements

The authors are highly thankful for the Guangxi Natural Science Foundation (ID: 2013GXNSFBA019282), Research Program of science at Universities of Guang Xi Autonomous Region (ID: ZD2014112, 2013YB205), Hechi College special projects (2003ZX-N003), Guangxi Higher outstanding young teachers training project.

References

[1] A. P. Braga, "Robot Navigation in Complex and Initially Unkown Environments," Proc. of the 14 IFAC Work Congress, Beijing, P. R. China, **(1999)**, pp. 179-184.
[2] C. H. Watkins and P. Dayan, "Technical note: Q-learning," Machine Learning, vol. 8 no. 3/4, **(1992)**, pp. 279-292.
[3] Peng J. and Williams R. J., "Incremental Multi—Step Q-Learing," Machine Learning, vol. 22 no. 4, **(1996)**, pp. 283-290.
[4] Wang C., Guo J. and Bao Z. J., "The improved Q learning algorithm in the application of the job shop scheduling," Computer application, vol. 283, **(2008)**, pp. 268-3270.
[5] Yong D. and Xinhe X., "Fuzzy reinforcement Learning and application in robot navigation," IEEE International Conference on Machine Learning and Cybernetics, Guangzhou, china, **(2005)**, pp. 899-904.
[6] Zong X. H., "Intelligent traffic control research in urban areas," Nanchang: Nanchang University, **(2013)**.
[7] Zhao H., "Manipulator trajectory planning research based on Q-learning algorithm," Heilongjiang: Northeast Petroleum University, **(2013)**.

Author

Jiansheng Peng, He received his M.Sc. in Mechanical and Electronic Engineering (2009) and PhD in Nuclear Control Technology (2013) from University. Now he is an associate professor of Physics and Mechanical & Electronic Engineering at Department, University. His current research interests include Development and application of embedded, Multi-robot cooperation. More than 40 papers published in various journals, 2 teaching materials.

