Adaptive Q-learning Algorithm for AUV Route Planning

Yingying Zhu\textsuperscript{1}, Changan Dai\textsuperscript{2}, Jiahao Xia\textsuperscript{3*}, Haowei Wu\textsuperscript{4}, Li An\textsuperscript{3}

\textsuperscript{1} Basic Teaching and Research Department of Command and Training, Armed Police Special Police Academy, Beijing, 102211, China
\textsuperscript{2} China National Aeronautical Radio Electronics Research Institute Department, Shanghai, 200233, China
\textsuperscript{3} Science and Technology on Underwater Vehicles Laboratory, Harbin Engineering University, Harbin, 150001, China
\textsuperscript{4} Beijing Institute of Astronautically Systems Engineering, Beijing, 100076, China
\textsuperscript{*}Corresponding author’s e-mail: xiajiahao@hrbeu.edu.cn

Abstract. When applying Q-learning algorithm to AUV route planning, it is difficult to strike a balance between exploration and exploitation. Usually, the fixed greedy rate or the greedy rate with fixed change trend is adopted according to experience. However, the value of greedy rate in the above method cannot be matched with the learning environment, and often occurs a problem of falling into a local optimal solution or slow learning convergence. In order to solve the above problems, we proposed an adaptive Q-learning algorithm. It guides the greedy rate with environmental complexity, and the environmental complexity degree of AUV is evaluated by the environmental complexity mathematical model. By setting up the simulation experiment of two-dimensional environment, the fixed greedy rate and the greedy rate with fixed change trend is compared. The simulation results show that the adaptive Q-learning algorithm considering environmental complexity converges faster and is less prone to fall into local optimal solutions.

1. Introduction
To ensure that AUV can complete the assigned task, the task planning system must comprehensively consider the operating environment, obstacle distribution, task requirements and other factors to develop a feasible route for AUV, so as to determine when and where it should execute the corresponding action. The most important part of the system is the route planning[1]. Gao Hu[2] proposed a hierarchical reinforcement learning method to solve the route planning problem of robots in complex environments. Xin Pan et al.[3] proposed a hybrid algorithm based on genetic ants to solve the problem of low efficiency of global route planning in 3d space. Ran-Xiangru[4] proposed to build the hierarchical structure of AUV task based on the global route planning.

The balance between exploration and exploitation is an important part of reinforcement learning, which can be used to solve the problem of action selected strategy. The $\varepsilon$ -greedy strategy is one of the three main search strategies, and the other two are Boltzmann distribution strategy and probability action selected strategy. Many scholars have studied and improved the greedy strategy in recent years. Olivier Caelen[5] improved the original greedy strategy and proposed the $\varepsilon$ -first algorithm. Zhao Yingnan[6] proposed a strategy of dynamically adjusting the greedy rate based on the success rate.
Chen Konghong[7] proposed an action selected strategy based on search entropy to realize that absolute exploitation does not occur when greedy strategy is adopted in a dynamic environment.

In this paper, we proposed the Q-learning algorithm and set the reward and punishment function. AUV can gradually improve the output strategy of the actuator in the process of information interaction with the environment through self-learning. However, in the absence of search strategy analysis tools, it is very difficult to set the greedy rate. This paper improves the traditional greedy strategy, and the fixed greedy rate or the greedy rate with fixed change trend is no longer adopted, and an adaptive changing greedy rate method considering the environmental complexity is proposed, which not only improves the learning efficiency, but also effectively avoids falling into the local optimal solution.

2. Adaptive Q-learning algorithm

2.1. Environment complexity mathematical model

The complexity of the environment is mostly determined by the complexity of the obstacles, which can be divided into two aspects for discussion. One is the number of obstacles. The other is the proportion of obstacles to the whole environment. The evaluation criteria are shown in equation (1).

\[ C = \frac{m \cdot o}{10 \cdot e} \]  

(1)

Where \( C \) is the complexity of the environment and \( m \) is the number of obstacles in the environment; \( o \) represents the total number of obstacles in the environment. And \( e \) denotes the total number of grids in the environment. When the quantized environment complexity is greater than 1, it can be set as 1. If there are no obstacles in the environment, let \( m = 1, o = 1 \).

2.2. Adaptive Q-learning algorithm

\( \varepsilon \)-greedy[8] achieves a balance between exploration and exploitation based on probability. Each attempt is explored with the probability of \( \varepsilon \) and utilized with the probability of \( 1-\varepsilon \). According to the research experience of Zhou Zhi-hua, \( \varepsilon \) is usually a smaller constant, such as 0.1 or 0.01[9]. And when the number of explorations is large, \( \varepsilon \) can be set as a function of the number of rounds, and \( \varepsilon = 1/(t)^{1/2} \). In order to slow down the decline rate of the greedy rate, by considering the influence of environmental complexity on the greedy rate, an improved changing greedy rate strategy is proposed, as shown in equation (2).

\[ \varepsilon = (t)^{-1/3} \cdot \frac{m \cdot o}{10 \cdot e} \]  

(2)

Where \( \varepsilon \) is the greedy rate and \( t \) is the number of learning rounds. \( m \) represents the number of obstacles in the environment and \( o \) represents the total number of obstacles in the environment. \( e \) denotes the total number of grids in the environment.

Suppose the robot is in state \( s_i \) , and \( s_j \in S \). Before training, the complexity of the training environment is determined according to the environment complexity mathematical model, and the action selected strategy adopted in the first learning round is shown in equation (3). In a learning round \( t \), the action selected strategy of that turn is shown in equation (4).

\[ \pi(s,a) = \begin{cases} \text{argmax}_{s,a} Q(s,a) & \delta > 1 - C \\ \text{random selection} A(s_i) & \delta \leq C \end{cases} \]  

(3)

\[ \pi(s,a) = \begin{cases} \text{argmax}_{s,a} Q(s,a) & \delta > 1 - C/(t)^{1/3} \\ \text{random selection} A(s_i) & \delta \leq C/(t)^{1/3} \end{cases} \]  

(4)
3. Simulation result
In this section, a two-dimensional learning environment and long island map were set respectively. The adaptive Q-learning algorithm with environmental complexity considering were simulated.

3.1. Two-dimensional environment simulation
The 2d environment took an environment of 3000m×3000m and turn it into a grid map of 30×30. Import 7 concentrated obstacles. Environmental modeling was shown in figure 1, and relevant simulation parameters were shown in table 1. The number of learning rounds of the fixed greedy rate Q-learning algorithm was 20,000, and the rest were shown in table 1.

![Figure 1. Layout of 2d environment](image1)

![Figure 2. Relationship curve between rounds and steps](image2)

As can be seen from figure 1, the total number of obstacles in the environment \( m \) was 7. By calculation, the total number of grids in the environment was 900, and 289 of the grids were occupied by obstacles. Thus, the environment complexity evaluation result under this environment can be obtained as \( C = \frac{7}{10^2} \frac{289}{900} = 0.225 \). We set four different greedy rates below.

1. fixed greedy rate Q-learning: \( \epsilon_1 = 0.1 \)
2. fixed greedy rate Q-learning: \( \epsilon_2 = 0.01 \)
3. general greedy rate Q-learning: \( \epsilon_3 = \frac{1}{(t)^{3/2}} \)
4. adaptive greedy rate Q-learning: \( \epsilon_4 = 0.225/\sqrt[3]{t} \)

| round | Learning rate | discount rate | reward | punishment | One step energy cost |
|-------|---------------|---------------|--------|------------|----------------------|
| 10000 | 0.01          | 0.9           | 100    | -10        | -0.1                 |

AUV started from the upper left grid (0,0), and after continuous exploration of the environment, finally reached the end grid (24,26) safely and with less consumption. The curve of the relationship between the number of rounds and the number of steps in the learning process of the four models was shown in figure 2. According to the experimental results of different greedy rate values, the convergence rounds and experimental planning steps of the algorithm were compared, as shown in table 2.
It can be seen from Table 2 that in the experiment, the convergence rounds obtained by the adaptive greedy rate Q-learning algorithm were less than those with fixed values. At the same time, from Figure 2, we learned that the adaptive greedy rate Q-learning considering environmental complexity had the fastest convergence and the least number of convergence rounds, which was 69.47% less than that in the case of $\epsilon=0.1$, 6.37% less than that in the case of $\epsilon=0.01$, and 5.45% less than that in the case of general greedy rate.

3.2. Long island environmental simulation
The Long Island is a chain of islands south of Liaodong Peninsula. Its satellite map is shown in Figure 3. In order to verify the reliability of the algorithm, the simulation verification was carried out in the Long Island environment.

As can be seen from Figure 4, the total number of obstacles in the environment $m$ was 6. Through calculation, the total number of grids in the environment is 900, and 191 of the grids were occupied by obstacles.
obstacles. The environment complexity evaluation result of this environment was 
\[ C = \frac{191}{10^{900}} = 0.127 \]. We set four different greedy rates like the two-dimensional environment, while the adaptive greedy rate was set as 
\[ e_\varepsilon = \frac{0.127}{(t)^{1/3}} \]. And the others were the same.

AUV started from the upper left grid (0,0), and finally reached the end yellow grid (25,25) safely and with less consumption. Through simulation experiments, the simulation results of different exploration strategies were compared, as shown in table 4.

| \( \varepsilon \) | Convergence rounds | Planning steps |
|------------------|------------------|----------------|
| 0.1              | 14000            | 30             |
| 0.01             | 5220             | 31             |
| general \( \varepsilon \) | 5311             | 31             |
| self-adaption \( \varepsilon \) | 4494             | 30             |

As can be seen from table 4, when taking \( \varepsilon \) with a fixed greed rate of 0.01 and a general change, the route obtained was not the optimal one, but the voyage cost was higher than the other two cases. The number of convergence rounds in the adaptive variable a model decreased by 67.90% when the greedy rate was 0.1, 13.91% when the greedy rate was 0.01, and 15.38% when the general greedy rate was 0.01. Simulation results show that in the two-dimensional real island environment, adaptive Q-learning model is adopted to minimize the learning cost and achieve the best route planning task.

4. Conclusion

The mathematical model of environmental complexity can better express the environmental model, which play a guiding role in the adjustment of balance between exploration and exploitation in Q-learning algorithm. When using the adaptive Q-learning algorithm considering the complexity of the environment to carry out AUV route planning, the two-dimensional environment have better effects than the fixed greedy rate and the fixed trend greedy rate, which can not only accelerate the convergence rate, but also effectively avoid the local optimal solution caused by insufficient exploration.

Acknowledgments

This work was supported by National Key Research and Development Plan (No. 2017YFC0305700), National Natural Science Foundation of China (No. U1806228, 51879057), The Fundamental Research Funds for the Central Universities (No. HEUCFG201810), Financially Supported by Qingdao National Laboratory for Marine Science and Technology (No. QNLM2016ORP0406) and Research Fund for Science and Technology on Underwater Vehicles Laboratory (No. 6142215180102).

References

[1] Mao, B.h., Tian, S., Chao, A.n. (2015) UAV Mission Planning. National Defense Industry Press, Beijing.
[2] Gao, H. (2016) Mobile Robot Path Planning Based on Reinforcement Learning. Southwest Jiaotong University, Chengdu.
[3] Pan, X., Wu, X.s., Hou, X.g. (2017) Global path planning based on genetic-ant hybrid algorithm for AUV. J. Huazhong Univ. of Sci. & Tech.(Natural Science Edition), 45(05):45-76.
[4] Ran, X.r. (2017) Research on AUV Path Planning Method Based on Hierarchical Reinforcement Learning. Harbin Engineering University, Harbin.
[5] Caelen, O., Bontempi, G. (2007). Improving the exploration strategy in bandit algorithms. In: International Conference on Learning and Intelligent Optimization. Berlin. 56-68.
[6] Zhao, Y.n. (2017) Research of Path Planning Problem Based on Reinforcement Learning. Harbin Institute of Technology, Harbin.

[7] Chen, Z.h. (2016) The Analysis and Research of Exploration Strategies and Algorithms in Reinforcement Learning. Nanjing University, Nanjing.

[8] Castronovo, M., Maes, F., Fonteneau, R., & Ernst, D. (2012). Learning exploration/exploitation strategies for single trajectory reinforcement learning. In: Proceedings of the 10th European Workshop on Reinforcement Learning (EWRL 2012). 1-9.

[9] Zhou, Z.h., (2016) Machine Learning. Tsinghua University Press, Beijing.