Research on Control of Move-in-mud Robot Based on Q Learning

Yunming Du[1,a], Bingbing Yan [2,b] and Yongcheng Jiang [2,c]

[1]College of Information and Electronic Technology, Jiamusi University, Jiamusi 154007, China

[2]College of Mechanical Engineering, Jiamusi University, Jiamusi 154007, China

[a]duyunming@126.com, [b]yanbingbing@126.com, [c]jiangyonghcheng@126.com

Keywords: move-in-mud robot; motion control; Q learning; radial basis function; neural network

Abstract. In order to improve the behavior self-control ability of move-in-mud robot in unknown environment, this paper proposes a behavior control algorithm based on radial basis function neural network Q learning. The algorithm enhances the interaction between robot and environment and improves self-learning ability through using the enhanced Q learning method. By employing the radial basis function neural network to approximate the state space and Q function, the learning system has good generalization ability and effectively solves the dimension disaster problem of the state space under complex and continuous environment. Simulation experiment results show that this method not only can make move-in-mud robot have strong motion control ability, but also improve the ability of robot to adapt to the environment.

Introduction

Move-in-mud robot is a new type of special underwater robot, which can complete drilling operation in underwater soil environment. The main purpose of this kind of robot is to salvage the sunken ship [1]. In unknown environment the practice of programming the robot by experience is difficult to deal with all kinds of situations that may be encountered. So it is necessary to introduce the self-learning function that can make the robot realize the independent behavior control after a period of training. Reinforcement learning is widely applied to the control of mobile robot based on behavior [2], and has been proven to be an effective tool to improve the intelligence system through experience. It is remarkable that Q learning is the most commonly used reinforcement learning algorithm [3].

In order to improve the autonomous performance of the mobile robot, it is necessary to introduce reinforcement learning module in the control structure, and design a kind of intelligent control structure. Since the environmental information of robot perception is continuous, the environmental information is required to discrete and the status value needs to be stored in the form of table in the application of reinforcement learning. This could cause a serious problem that great storage space is needed. Neural network has good generalization ability, which can realize the approximation of any nonlinear function. In this paper, the neural network is used to approximate Q function to solve the application problem of reinforcement learning in continuous state.

Structure of Move-in-mud Robot and Environmental Detection

Description of the Structure of Move-in-mud Robot. The move-in-mud robot used in the simulation research is derived from the literature [4], whose body structure mainly includes the attacking clay mechanism and the creeping mechanism. The twist drill is chosen in the attacking clay mechanism, which is mainly used to crush the silt in front of the head. The creeping mechanism mainly consists of an upper supporting joint, a steering joint and a lower supporting joint, whose role is to support the mud around the robot body and drive the robot moving forward. The concrete structure of the move-in-mud robot is shown in Fig. 1.

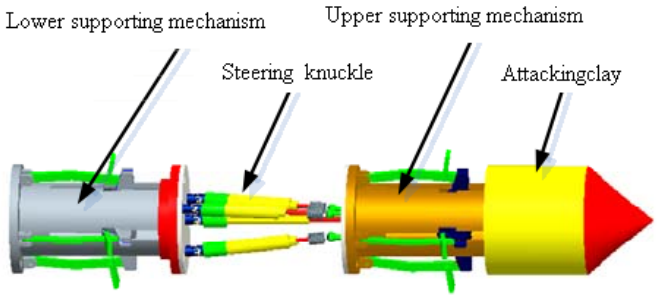

Fig. 1. Schematic diagram of the structure of move-in-mud robot

In the upper and the lower supporting joints, the support is realized by using rolling wheels in the groove and connecting rods. When the slide blocks move in the groove of the barrel wall, the connecting rods move up and down. This will lead to the expansion of the support plates on connecting rods and realize the support. The steering knuckle is a parallel mechanism, which consists of five parts: hydraulic cylinder, universal joint, joint ball bearing, guide shaft, front and back platform. The moving platform and static platform of the steering knuckle are respectively connected with the upper and lower supporting joints through a fixed connection mode. In a movement cycle, the two platforms of the steering joint perform alternate motion to achieve the effect of creep.

Environmental Detection of Move-in-mud Robot. Move-in-mud robot perceives the external environment information through the front-end sensor system and makes decision to implement the corresponding action. Information obtained from the external environment of robot includes the depth of soil, water content, frontal resistance and so on. In addition, the distance between the ship and the robot is also important external environment information, which can be obtained through the soil pressure sensor and ultrasonic sensor. After perceiving the external environment, the agent can complete the walk in the soil by selecting the appropriate actions. In order to make the robot have intelligent control, this paper introduces the reinforcement learning method into the motion control of robot, by which attempts to improve the self-study and control ability of robot.

Reinforcement Learning Based on Neural Network

Enhanced Q Learning. Reinforcement learning refers to the mapping learning from environmental state to behavior, which makes the system behavior obtain the cumulative reward value from the environment and finds the optimal strategy by trial-and-error. Reinforcement learning system accepts the input s of the environment status and outputs the corresponding action a according to the internal reasoning mechanism. Environment changes to the new state s' under the system action a. When the system accepts the new input of environment state, it also receives the instantaneous reward and punishment feedback r. For the reinforcement learning system, the goal is to learn a behavior strategy π: $s \rightarrow a$, which enables the system to select the optimal action to obtain the maximum cumulative reward value of the environment. In other words, the system is to maximize Eq. 1, where γ is a discount factor.

$$V^{\pi}(s_t) \equiv r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots. \tag{1}$$

Q learning is a widely used reinforcement learning, which does not directly apply the value function above, but uses a similar Q function. Its expression is as follows:

$$\mathrm{Q}(s_t, a_t) \leftarrow r_t + \gamma V(s_{t+1}). \tag{2}$$

where a_t is a selected action from the action set A at time t. The aim of the system is to make the total reward value maximum, therefore using $\max_{a \in A} \mathrm{Q}(s_{t+1}, a_t)$ instead of $V(s_{t+1})$ in the above formula we can get the following equation:

$$Q(s_t, a_t) \leftarrow r_t + \gamma \max_{a \in A} Q(s_{t+1}, a_t). \quad (3)$$

At time t, the agent selects an action *a* based on the current state, and then updates the Q value according to the following expression:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_{b \in A} Q(s_{t+1}, b) - Q(s_t, a_t)]. \quad (4)$$

where *b* is the action selected under the conditions at time t+1.

Q Learning Based on Neural Network. Reinforcement learning is applied in the case of discrete state information, and the corresponding state-action pair is stored in the form of table. This not only takes up a lot of memory space, but slows down the speed of learning convergence. More importantly, when the state information is continuous, it will not be achieved. So this paper proposes Q learning based on radial basis function neural network, which uses the radial basis function neural network to approximate the Q function. It can effectively overcome the defect caused by the Q value stored in the table [5-7]. In practical application a feedforward neural network with three layers is used [8], whose input is the state information of robot and output is the corresponding Q value of each action. The learning structure of the system is shown in Fig. 2.

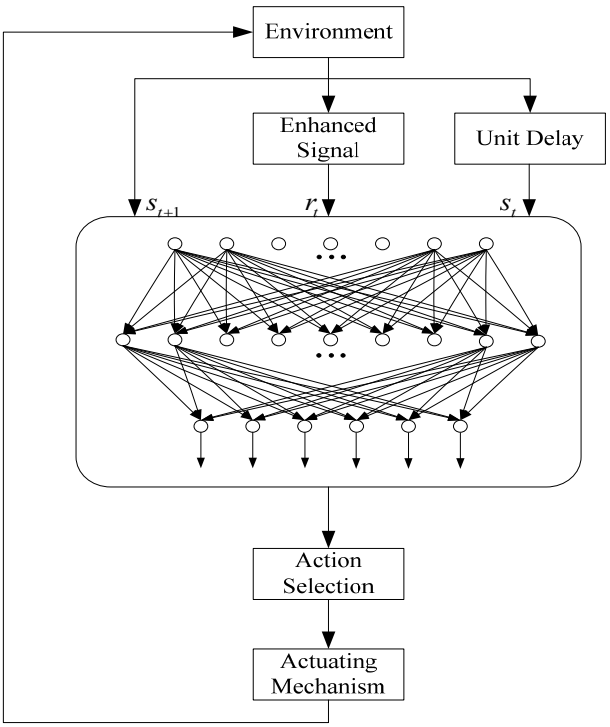

Fig. 2. Q learning based on radial basis function neural network.

The specific implementation steps are as follows:

Step 1: Initialize the neural network and the parameters used in operation;

Step 2: Initialize the state of the move-in-mud robot;

Step 3: Get the robot's current status information s_t;

Step 4: Enter state information into neural network and selecte actions according to Q value;

Step 5: Execute actions to make the robot into a new state s_{t+1}, and obtain the enhanced signal value;

Step 6: Train neural network according to the radial basis function learning algorithm;

Step 7: Repeat step 3-6 until the learning is completed.

Behavior Selection Strategy. In the initial stage of learning, since the Q value is randomly initialized, it does not have any meaning. In order to explore all possible actions, the Boltzmann distribution is introduced to realize the random selection of action in the initial stage. The probability that an action is selected is expressed as follows:

$$P(a_i / s) = \frac{e^{\mathrm{Q}(s,a_i)/T}}{\sum_{a \in A} e^{\mathrm{Q}(s,a)/T}} . \tag{5}$$

where T is a temperature coefficient, and the T value is proportional to the randomness of selection. As the learning is carried out, the Q value slowly tends toward the desired state-action value. At this time, according to the greedy strategy, the action corresponding to the maximum Q value will be selected.

$$a = \arg\max_{b \in A} \mathrm{Q}(s,b) . \tag{6}$$

Simulation Experiments and Results

In order to verify the feasibility of the method proposed in this paper, experiments are carried out in the simulation environment. Because the upper and lower supporting mechanism of move-in-mud robot are rigidly connected with the steering knuckle, the movement of robot is mainly provided by the steering knuckle. Based on the analysis of the literature [9] and [10], we first decompose the motion of robot into 13 basic actions, which include the moving forward and the 3°, 6° and 9° rotation in the upper, lower, left and right direction. The inputting to neural network information are the distance d between the robot and the hull and the angle θ_{rt} between the moving direction and the target direction of the robot.

The reinforcement signal acquisition is a key problem in Q learning. The convergence rate of learning can be improved by setting the reinforcement signal to the appropriate value. According to the task to be completed, the reinforcement signal is divided into two parts. A part is the enhancement value r_{ro} generated by the distance between robot and hull. The other is the enhancement value r_{rt} generated by the distance between robot and target point. The final reinforcement signal is the sum of the two parts. The values of the two enhancement signals are as follows:

$$r_{\mathrm{ro}} = \begin{cases} \text{-0.5，} & \text{Close to the hull；} \\ \text{+0.5，} & \text{Away from the hull ；} \\ \text{-1.0，} & \text{Collision to the hull；} \\ \text{0，} & \text{Others.} \end{cases} . \tag{7}$$

$$r_{\mathrm{rt}} = \begin{cases} \text{+0.3，} & \text{Close to the target point；} \\ \text{-0.3，} & \text{Away from the target point；.} \\ \text{0，} & \text{Others.} \end{cases} \tag{8}$$

The other parameters used in the experiment are the learning rate α=0.5, the discount factor γ=0.9 and the temperature coefficient T=5. The initial weights of radial basis function neural network are random numbers on the interval [0,1].

When the robot runs in an unknown environment, the robot randomly selects actions at the initial stage. So the robot's path is not smooth and often runs into the hull. After hundreds of times of learning, the robot can successfully reach the target point in the case of avoiding collision with the hull and the running track is relatively smooth. The simulation result is shown in Fig. 3. With the development of learning, the running effect is getting better and better. In the same way, the robot can successfully complete the task with different environments and starting points.

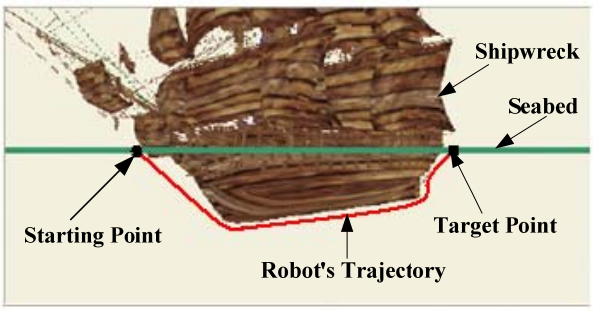

Fig. 3. Running track of move-in-mud robot

The performance of reinforcement learning is mainly measured by the convergence and the convergence rate, which can be investigated by the reinforcement signal and running track. The convergence of reinforcement learning can be mainly displayed by the reinforcement signal, and the increasing average enhanced signal implies that the training effect is becoming better. Conversely, the larger decrease amplitude means that there must be a collision in the training. In order to further test the validity of the algorithm, we take 80 training cycles and count the average step number and the average reward for the robot to reach the target. Experimental results obtained are shown in Fig. 4 and Fig. 5.

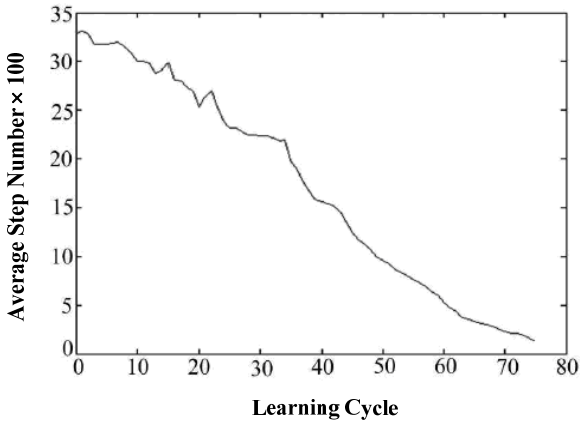

Fig. 4. The learning cycle and average step number

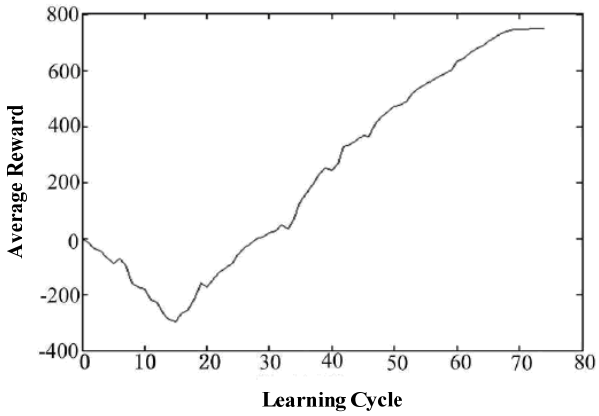

Fig. 5. The learning cycle and average reward

As can be seen from Fig. 4, the average step number to reach the target point is decreasing in the learning process. In addition, according to Fig. 5 we can also see that the average reinforcement signal shows an increasing trend after 15 learning cycles and the collision with ship hull is significantly reduced after 30 learning cycles. This shows that the control strategy used has been continuously optimized.

Conclusion

The Q learning based on radial basis function neural network is applied to the motion control of move-in-mud robot in this paper. Under the condition of unknown environment and lack of prior knowledge, the effective control of move-in-mud robot is realized by using reinforcement learning with the characteristics of self-learning and online learning. In order to avoid the dimension disaster of reinforcement learning state space, the author makes full use of the function mapping and generalization ability of neural network to adjust the Q value table. Using this algorithm, the move-in-mud robot has not only the strong motion control ability, but also the outstanding ability to adapt to the environment. Future work focuses on embedding priori knowledge into the learning system to accelerate the learning process of the system.

Acknowledgments

This work was financially supported by Natural Science Foundation of Heilongjiang Province (E201254) and the Key Scientific Research Project of Jiamusi University (12Z1201519).

References

[1] B.B. Yan: *System Design Move-in-mud Robot and its Virtual Prototype* (Dissertation Harbin University of Science and Technology 2008) (In Chinese).

[2] L.P. Kaelbling, M.L. Littman and A.W. Moore: Artificial Intelligence Research Vol. 4(3) (1996), p. 237-285.

[3] C. Watkins and P. Dayan: Machine Learning Vol. 8(3) (1992), p. 279-292.

[4] M.Q. Xie: *Research on Key Technology of Virtual Prototype of Bionic Move-in-mud Robot* (Dissertation Jiamusi University 2012) (In Chinese).

[5] L. Jun: *Learning Reactive Behaviors with Constructive Neural Networkin Mobile Robotics* (Dissertation Orebro Studies in Technology 2006).

[6] A.R. Webb and S. Shannon: IEEE Transactions on Neural Networks Vol. 9(6) (1998), p. 1155-1166.

[7] X.J. Yang: *Artificial Neural Network and Blink Signal Processing* (Tsinghua University Press, China 2003) (In Chinese).

[8] J. Moody, C. Darken: *Learning with Localized Receptive Fields*. (Carnegie Mellon University Press, US 1988).

[9] B.B. Yan, M.Q. Xie and H.Q. Sun: Science & Technology Review Vol. 29(3) (2011), p. 44-47 (In Chinese).

[10] B.B. Yan, X.M, Liu, M.Q. Xie and B.L. Yin: Modular Machine Tool & Automatic Manufacturing Technique No. 6 (2012), p. 38-41 (In Chinese).