PAPER • OPEN ACCESS

Q Value Reinforcement Learning Algorithm Based on Multi Agent System

To cite this article: Xijie Yin and Dongxin Yang 2018 *J. Phys.: Conf. Ser.* **1069** 012094

Q Value Reinforcement Learning Algorithm Based on Multi Agent System

Xijie Yin[1] **and Dongxin Yang**[2]

1 School of Data and Computer Science, Shandong Women's University, No. 2399, University Road, University of science and technology, Changqing, Jinan, Shandong, China.
Email: xijiesd@126.com
2 Dazhong News Group, Shandong media Building, No. 6, Luoyuan Street, Jinan, Shandong,China.
Email: sdxjg@163.com

Abstract. Q-learning algorithm of multi-agent system is studied in this paper. In order to improve the learning efficiency and convergence speed of the Q algorithm which is a typical learning algorithm of multi agent system, this paper proposes an improved reinforcement learning algorithm of multi agent system based on the existing design experience and surrounding environment information. The Q learning algorithm is effectively extended to the multi agent systems by information sharing among multiple agents in the improved algorithm. The effectiveness of the proposed algorithm is verified by simulation.

1. Introduction

Agent technology is a tool that can be a substitute for people to negotiate and make decisions. The intelligence and flexibility of traditional Agent system is low, and its decisions cannot meet the actual needs of people, sometimes even damage the profit of user. Therefore, it is necessary to improve the intelligence and learning ability of Agent system.

With the rapid development of distributed network technology in recent years, the research on theory and application of multi agent system has become a hotspot in the field of artificial intelligence. In the Agent system, each agent possesses a basic behavior and normally it only deals with the local information and goals related to itself. Considering the actual application, multi agent system is required to be real-time, dynamic, random, distributed and etc. Therefore, the agent is required to have the learning ability to autonomously interact with the surrounding environment, analyze the external environment and build the environment model; besides, it is also required to master the learning skills and collaboration manners that mimic the human beings so as to enhance the intelligence of the Agent system. Machine learning is the basic way to solve the above problems at present. In this paper, an improved reinforcement learning algorithm of multi agent system based on the Q reinforcement learning algorithm is proposed, enabling the Agent to perceive the environment effects generated by itself and other Agents. Moreover, the learning status and learning efficiency is further improved with the achievement of information sharing and cooperation among multiple Agents. In addition, this method is validated by simulation.

2. Reinforcement Learning and Q Learning

Reinforcement learning, also known as re-learning or evaluation learning, is an online learning technique which is different from supervised learning and unsupervised learning methods. It is widely used in the fields of intelligent control robots, intelligent analysis and prediction, and etc. For the reinforcement learning, learning is regarded as a process of tentative evaluation, and in this process, reinforcement learning object can perceive the environment state and execute specific action on the environment. When the environment perceives the action, its state will change and a reinforcement signal, usually in the form of reward or punishment, will be generated and fed back to the learning object. Then the reinforcement learning object will execute the next action according to the reinforcement signal and the current environment state. The select principle of the next action is to increase the probability of being rewarded.

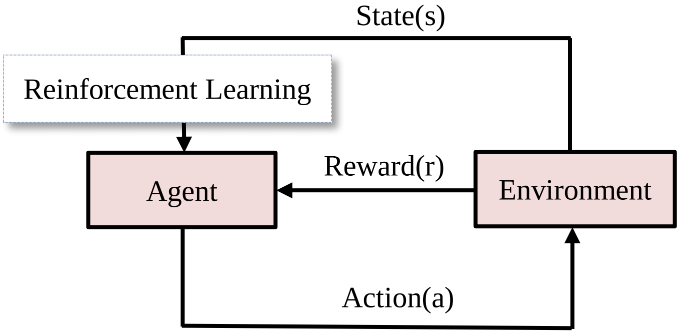

Figure 1. Basic principle of reinforcement learning

Q learning is one of the most important algorithms for reinforcement learning. The key to Q learning is to transform the interaction between environment and learning object into a Markov decision process, namely MDP. That is to say, the learning process is reflected in the selected action and the state, and then a fixed probability distribution of state transition, a next state and an immediate return value are determined by the selected action. The ultimate goal of Q learning is to find a way to maximize the reward by taking samples from the objective world.

For Q learning, each state (s) and action (a) is corresponding to a Q (s, a), and a is selected according to the Q (s, a). From the State s, the relevant a will be executed according to a certain method, and finally an accumulated reward value will be obtained. Q* is the optimal Q value, which represents the sum of discount reward obtained by multiple executions of Action a. Action a is executed by the Agent from State s in accordance with the optimal method. Q* is defined as follows:

$$Q^{*}(s,a)=R(s,a)+\gamma\sum_{s\in S}T(s,a,s')\max_{a'}Q^{*}(s',a'). \tag{1}$$

where s is the state set, a is the action set, T(s, a, s') is the probability that the state will be transformed to s' after Action a is executed in State s, R (s, a) is the reward of executing Action a in State s, Y is discount factor which shows the time interval’s impact on the reward.

For this learning method, learning ability of the learning object is improved by the optimized learning on a continuous iterative calculated Q function, and the optimized learning is realized by constant reflections. The initial value of Q function can be arbitrarily specified, and after each action and corresponding reward, the Q function will be updated according to Formula (2):

$$Q_{t+1}(s,a)=(1-a)Q_{t}(s,a)+a[r_{t}+\gamma\max_{a'}Q_{t}(s',a')]. \tag{2}$$

where a is the learning rate controlling the convergence. After continuous exploration, the value of Q function gradually approaches Q*, indicating that the Q learning algorithm will inevitably converge on the optimal solution if a can satisfy the certain condition.

The basic algorithm is:

Initialize $Q(s,a)$ arbitrarily

Repeat(for each episode)

```
Initialize $s$
Repeat(for each step of episode)
Chose a from $s$ ushing policy derived from $Q(eg, \varepsilon - greedy)$
Take a a， observer $r, s'$
$Q(s,a) \leftarrow Q(s,a) + a[r + \gamma max_a Q(s', a') - Q(s,a)]$ $s \leftarrow s'$
Until $s$ is terminal
```

3. Improved Algorithm for Q Learning

In general, interactive reinforcement learning is to extend the Markov decision process to multi agent system. The detailed method is: changing the original reward function R: s × a → R to a joint reward function Ri: s ×a (1) ×⋯ × a (n) → R, and changing the original state transfer function P: s × a → PD (s) to P: s ×a (1) ×⋯ × a (n) → PD (s). To complete the above algorithm, each agent who is executing a specific action should be aware of the possible actions executed by other agents, and it is feasible only for a relatively small scale system. If the number n of agent in the system is increased or the possible action of each agent in each state is increased, the state space of the agent will grow exponentially, and in this case, although the above algorithm is theoretically feasible, the possibility of actual realization is minimal considering the computer capabilities now and in the predictable future.

Therefore, it is necessary to improve the existing Q learning algorithm. Agent is set to perceive environment effects generated by itself and other agents; corresponding mark information can also be obtained. In this way, information sharing and cooperation among multiple Agents is achieved.

Each Agent starts to initialize its own state. Q (s, a) is an arbitrary value, typically 1 / N, where N is the execution number of Action a under the State s. Ri (s, a) equals to 0, and Ri is the reward function of Agent i. α is the learning rate and γ is the discount factor.

The information left by Agent in the grid word is called footprint, represented by F. The larger the F is, the deeper the footprint is. The probability that a certain point is the optimal solution is proportional to the number of Agent passing this point. The value of F can be accumulated by Formula (3):

$$\mathrm{F} = F' + f_i^t. \tag{3}$$

where $f_i^t$is the footprint left by Agent(i) at t.

In the traditional Q learning algorithm, the next action is only determined by the value of Q. While in the improved Q learning algorithm, the next action executed by the Agent is determined by both Q and F, as shown in Formula (4), and in this way the Q learning algorithm is improved to a reinforcement one suitable for multi Agent system.

$$\pi_i^{(s)} = \mathrm{argm_a ax} \sum_{s'} \mathrm{p_{ss}^a} + \mathrm{F}_{s'} \mathrm{r} \mathrm{Q}(\mathrm{s}', \mathrm{a}). \tag{4}$$

$p_{ss}^a$is the transform probability from State s (s) to s' (s), $\mathrm{F}_{s'}$ is the footprint depth under the State s'. Information of other Agents can be simultaneously taken into account in the decision-making process by such improvement in the algorithm. The update of Q value complies with Formula (2), which can not only keep its own internal state but also obtain information of other Agents.

The reward function is:

$$\gamma_i^{st} = \begin{matrix} 1, if\ target\ is\ found\ in\ the\ next\ step \\ \gamma r^{s'}, if\ target\ is\ not\ found\ in\ the\ next\ step \end{matrix}. \tag{5}$$

Detailed algorithm is as follows:

1. First, the state set of each agent is initialized; the learning rate and discount are set. N is the execution number of Action a under the State s.

2. Loop the following steps.

2.1 First, observe s and the next possible State s'. Determine whether there is a goal for the next step, if so, start the next learning round.

2.2 According to the above Formula(5), select Action a' and take the specific exploration strategy.

2.3 Execute Action a, getting the reward and the next State s'. The value of Q function will be updated, leaving a footprint f in the grid world, and the learning rate will be decreased progressively. The learning process is over until the learning rate is decreased to zero or the value of Q function is no longer conspicuous changed.

3. Extract the result

When the learning rate is 0 and no exploration strategy is taken, select an Agent to start the search for the goal and record its searching path.

4. Simulation Results

Figure 2 shows the result of searching the goal in a 10 × 10 world with only one Agent. The learning rate α is set to be 0.7, and it will decrease with the learning cycle. The discount factor γ is 0.8. Figure 3 shows the result of searching the goal respectively with two Agents; learning rate α and discount factor γ here are the same as above.

Fig. 2 and Fig. 3 show that the learning period is longer if there is only one Agent, and the improved algorithm has no obvious advantage than the traditional Q learning algorithm. While when the number of agents increases to two, the learning period is obviously shortened which illustrates a more effective learning.

The advantage of the improved algorithm is the effective solution of information exchange problem among multiple agents and the avoidance of the exponential disaster of state space to a certain extent. The optimal solution or the approximate one can be found faster through the cooperation of multi agent system. The disadvantage of this algorithm is the large number of possible Action a under each State s, with the operation efficiency still to be improved.

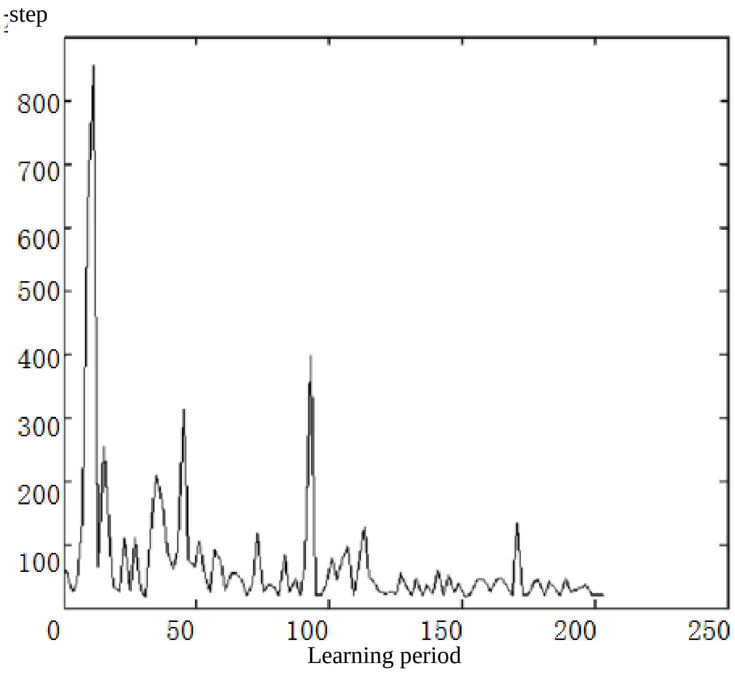

Figure 2. Learning with only one Agent

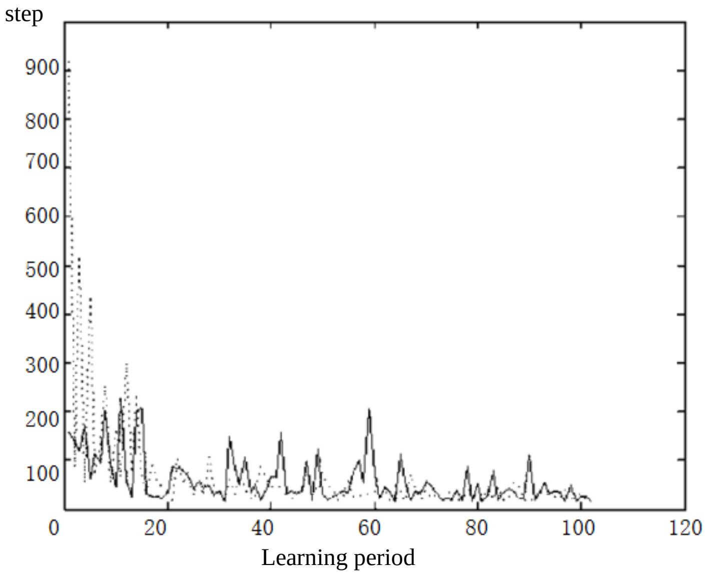

Figure 3. Learning with two Agents

5. Conclusion

Machine learning is a key point of the research on multi agent system, and also a hot spot of research on complex intelligent system. This paper proposes an improved reinforcement learning algorithm for Agent system, solving the problem that the traditional Q learning algorithm for a single Agent cannot be directly applied to a multi agent system. With the improved algorithm, the problem of information exchange between Agents is solved, the exponential disaster of state space is avoided to a certain extent, and the learning speed is improved.

6. Acknowledgments

Research and innovation team of data mining in Shangdong Women’s University. 2014, the key issue of statistical research in Shandong Province, 86.

7. References

[1] Zheng Xiao. Research on Cooperation and Coordination Mechanism among Multi Agent System [D]. Fudan University, 2009.

[2] Bingqiang Huang. Research on Reinforcement Learning Methods and Their Applications [D]. Shanghai Jiaotong University, 2007.

[3] Zhongli Zhan, Qiang Wang, Peixia Wang. Research on Q Learning Algorithm in Multi Agent System [J]. Journal of Liaoning Agricultural Vocational College, 2008, 05: 48-50.

[4] Xiaohu Yin. Research on Reinforcement Learning Methods Based on Multi Agent Cooperation [D]. National University of Defense Technology, 2003.

[5] Jingliang Yu. Research and Implementation of Agent Intelligent Decision Based on Q Learning [D]. Hefei University of Technology, 2005.