Multi-Agent Reinforcement Learning Collaborative Environment

著者	李 嘉誠
出版者	法政大学大学院理工学・工学研究科
雑誌名	法政大学大学院紀要．理工学・工学研究科編
巻	57
ページ	1-3
発行年	2016-03-24
URL	http://hdl.handle.net/10114/12424

Multi-Agent Reinforcement Learning Collaborative Environment

李 嘉誠

Li Jiacheng

指導教員氏名　李　磊

法政大学大学院理工学研究科応用情報工学専攻修士課程

Abstract

As in the multi-agent system, access to new knowledge and adapt to the new environment is one of the essential characteristics of agent.So Agent need to learn to adapt to the dynamic changes in the environment, in order to obtain optimal or sub-optimal behavior strategy. This paper studies multi-robot systems,proposes a further reinforcement learning algorithm and applied to multi-robot system. The multi-robot path trace planning issues were discussed.We designed a reinforcement function, the design method proposed reinforcement piecewise function. The simulation results show the effectiveness of this method.

Key Words *: multi-agent, reinforcement function, unknown environment*

1. Introduction

Machine learning is a smart body to improve intelligence, coordination, adaptability basic approach.

Footprints algorithm by sharing information between Agent effectively extends the Q-learning algorithm is applied to the multi-Agent system.

The basic algorithm：　Q-learing

The paper will solve:Solve the problem of information exchange between Agent and improve the learning speed.

2. Q-learning algorithm

```
Initialize Q(s， a)arbitrarily
Repeat(for each episode)
Initialize s
Repeat(for each step of episode)
chose aform s ushing policy derived from.Q(eg $\varepsilon$-greedy)
Take action a.observer r，s′
Q(s，a) ← Q(s，a)+a[r+ $\gamma$ max Q(s,′ a′ ) ← Q(s a)
s ← s′
Until is terrninal
```

3. Markov model

In this chapter, we assume that the robot interacts with the environment completely observable, then reinforcement learning can be built on model of Markov decision processes, which is defined as follows

1. About the environment finite state space S;
2. Finite action set A;
3. Immediate reward function R:S×A→ R ;To implement decisions in A state S gained instant rewards;
4. State transition equation P: S × A → π (S); (S)is a probability distribution on S.Here, P(s′ | s,a)is probability of system transferred to state s when performing behavior a in that state of s

Its meaning is as follows

1. The state transition probability distribution of the current operation state of the environment and Agent equations, and the previous state and actions independent.
2. Payoff function of its current state of action based solely on reinforcement signal is given.

4. The improved Q learning

Each Agent in the square in the world of information is called a ″ footprint″ F, ″ footprints″ deeper (F value), the greater the representative through the Agent here, the more the more likely is the optimal solution. The F value can be accumulated by the formula : The Agent t time i left by the ″ footprint″ . Q - learning algorithm generally only through the Q value decide the next move, the Q - learning algorithm is improved, and the Agent at the same time by reference to the Q value and F value using the formula (1) the choice of action, thus achieved the Q - learning into suitable for reinforcement learning algorithm of multiple Agent system:

$$\pi_{\mathrm{i}}(s) = \arg\max_{\alpha} \sum_{i'} p_{ss'}^{\alpha} + F_{s'} + \gamma Q(s', a) \qquad (1)$$

To perform an action in the $P_{ss'}^{a}$ said state s to s probability; F, said state s the footprints of the depth, such doing can at the time of decision, the other Agent information into consideration at the same time. Q value updates can use formula , both to maintain its internal state, and achieve the purpose of the other Agent action information

$$\gamma_i^{st} = \begin{cases} 1 & \text{The next step is to find the target} \\ \gamma^{r} & \text{The next step is not to find the target} \end{cases}$$

1. The initial set of the state of each agent, vector set, discount, order, where N for state s executable actions under a number.
2. Repeat the following steps

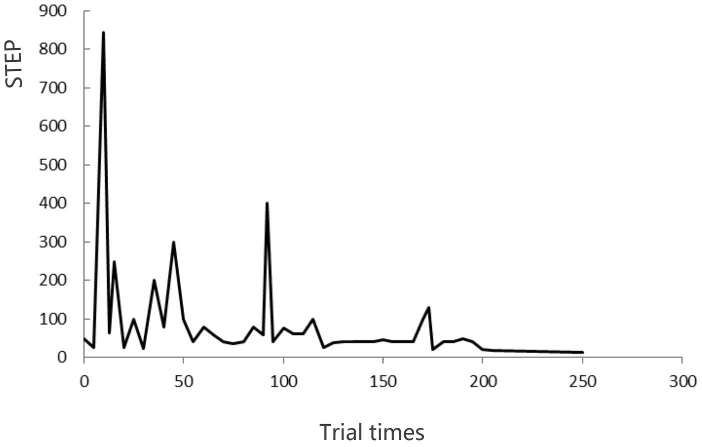
   2.1 observe state s and the next step may state s, whether the next target exists, if have, will start to a new study

   2.2 according to the above formula action selection a, at the same time to take certain exploration strategy

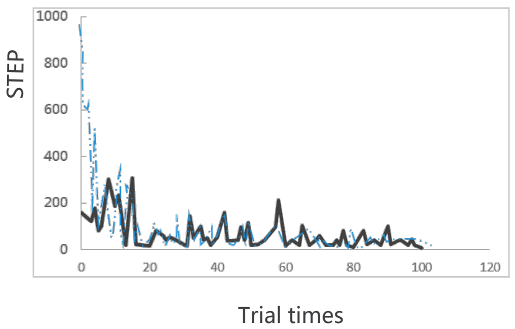
   2.3 to perform an action a. Rewards, and the next state. Update the Q value at the same time corresponding footprints left in the world, f vector. Vector until reduced to 0 or sound Q value is not big changes, says end of learning
3. The extraction results: By an Agent in the vector for O, don′ t take the exploration strategy, looking for it again Goals, and write down the path

5. The simulation results

There is only one Agent in 20 x 20 square target search results in the world, including vector instrument = 0.7, decrease with learning cycle, the discount factor: 0.8. Figure 2 for two under the condition of the Agent, the respective target search results, vec-tor and discount factor gamma ditto. By two figure can be seen, there is only one Agent, the learning cycle is long, and there is no advantage in comparison to the original Q learning method, but when the Agent number increased to more than two, learning cycle significantly shortened, and the learning effect is better. The advantage of this learning method is very good to solve the problem of the exchange of information between the Agent, to a certain extent, to avoid the problem of the index of state space disaster, for cooperation of multi Agent system can quickly find the problem of the optimal solution or approximate optimal solution. Defect is when each state s optional action a number of very much, its operation efficiency still needs to be improved.



Figure：1 There is only one learning Agent



Figure：2 Two learning situation of the Agent

Table： 1 There is only one learning Agent

Learning cycle	step
10	844
25	100
50	99
75	35
100	76
125	38
150	45
175	20
200	20
250	13

Table：2 There are two learning situation of the Agent

Learning cycle	step
5	82
10	87
20	17
30	23
40	54
50	20
70	60
80	15
90	10
100	8

参考文献

[1] Li Qiang. The complicated continuous system reinforcement learning system: algorithm design and application [D]. Shanghai: Shanghai Jiao Tong University,2000

[2] Zeghal K . A Comparison of Different Approaches Based on Force Fields for Coordination among Multiple Mobiles[C]. Proc. IEEE Int. Conf. Intell.

[3] Azarm K , Schmidt G . A Decentralized Approach for theConflict-free Motion of Multiple Mobile Robots, IEEEIROS96, 1996: 1667- 1675.

[4] Mataric M . Interaction and Intelligent Behavior[D]. MIT,1994.

[5] Zhijun Gao. Experimental study of some key technologies based on MAS collaborative intelligent multi manipulator system [D]. Shanghai: Shanghai Jiao Tong University, 2002