The Seventeenth International Symposium on Artificial Life and Robotics 2012 (AROB 17th '12),
B-Con Plaza, Beppu, Oita, Japan, January 19-21, 2012

Estimate of current state based on experience in POMDP for Reinforcement Learning

Yoshiki Miyazaki[1], Kentarou Kurashige[1]

[1] Muroran Institute of Technology, Hokkaido, Muroran 050-8585, Japan
(Tel: 81-143-46-5400, Fax: 81-143-46-5499)
(s1823084@mmm.muroran-it.ac.jp, kentarou@csse.muroran-it.ac.jp)

Abstract: Recently, reinforcement learning (RL) attracts attention. Because interaction with environment is important on RL, it is necessary to recognize state of robots more accurate. However, in real environment there is incompleteness on recognition by ability lack and noise of sensors. If recognition has incomplete, there is problem that robots cannot learn appropriately because robots cannot distinguish states that robots should originally distinguish. The model including such incompleteness is known as POMDP and we aim to solve the problem learning does not progress appropriately in POMDP. We pay attention to the experience of robots. When robots cannot identify current state uniquely, robots decide current state using current observation, previous recognition state and action. And, by memorizing this state information as internal state, recognizable state increases. In this way, robots can distinguish states which robots cannot distinguish in conventional method and learn appropriately. We show the effectiveness of proposed technique with simulation.

Keywords: Reinforcement Learning, Partially Observable Markov Decision Process, incomplete recognition

1 INTRODUCTION

Now, RL [1] is known as effective technique as one of technique to adapt to environment. RL also attracts attention as technique which is often used in actual robots. On RL, robots learn by interaction with environment. And sensors are necessary for interaction with environment. Robots recognize a situation as a state that robot faces by using sensors. So the recognition by sensors is important for learning.

In RL, most basically model of environment is Markov Decision Process (MDP) [2]. In this environment, it is assumed that the recognition of environment by robots is complete. In other words, robots have enough recognition ability to achieve tasks in MDP. But in real environment, because the ability of sensors is insufficient or sensors have noise, there is often incompleteness for recognition. So, recognized state does not accord with actual situation. Partially Observable Markov Decision Process (POMDP) [3] is model that deals with such incompleteness. In POMDP, robots may recognize different plural situations in real environment as same state. Because robot recognizes different situations as same state, there is possibility that robot cannot learn appropriately for task. So, there are some studies for environment on POMDP [4],[5].

In this paper, we pay attention to using past state and action. And we propose a system coping with incomplete recognition. In proposed system, robot recognizes a current state by using previous state and action. Not always robot recognizes state by using such technique, but robot recognizes current state by using previous state and action at the time that robot recognizes incomplete recognition. Therefore robot needs to recognize whether current state is incomplete recognition. We use experience of robots for recognition of incomplete recognition. In experience on robot, when there is difference on result of same action between same recognized state, incomplete recognition is likely to have occurred. In the case that there is difference from two results of same action on a recognized state, robot define internal state for the state by using previous state and action. And robot can distinguish a state thought to be incomplete recognition by using internal state. Therefore it is possible to consider POMDP to be MDP by using proposed method. And proposed system has four parts: Recognition part, Judgment of incomplete recognition part, Definition of internal state part, and Accumulation of experience part.

In this paper, at first we describe the summary and problems about recognition in POMDP on RL. Secondly we propose the technique of recognition using previous state and action. We will describe four parts of Recognition part, Judgment of incomplete recognition part, Definition of internal state part and Accumulation of experience part. Thirdly, we perform simulation that shows the effectiveness of proposed system. In simulation, we use the maze problem and adapt Q-learning[1] to agent. Moreover, we compare the proposed technique with normal recognition and complete perception. Finally, we describe conclusion and the future work.

2 PROBLEM OF RECOGNITION IN POMDP

2.1 Recognition using sensors on RL

The subjection of our research is a learning agent which has sensors. On RL, at each time period, the agent is in some

state $s \in S$ (S is a set of states). But the agent recognizes a state that agent faces through sensors. On recognition based sensors, the agent decides a state by the combination of value of each sensor. Here, $o \in O$ (O is a set of observations) is an observation that agent recognizes through sensors and v_n are value of each sensor. o is expressed in eq.(1), and n is number of sensors agent has.

$$o = \{v_1, v_2, \ldots, v_n\} \tag{1}$$

In the case of the recognition with sensors, the observation is not usually equals to the state in POMDP. And there is a possibility that the agent recognizes two or more another states as a same observation. There are various factors including performance and property of sensors for this reason to cause these, and we think it is difficult to solve the disagreement in the state and observation using only sensors.

In this paper, in case that there is disagreement in the state and the observation, we call this situation as incomplete recognition. And, in case that the observation is same as the state, we call this situation as complete recognition.

2.2 Problem by the incomplete recognition on RL

On RL, agent learn appropriate action based on the observation. Therefor the observation is important. If agent is in incomplete recognition, agent tries to find appropriate action for an observation includes several states (Fig.1). Agent can not learn appropriately, in such case. But incomplete recognition does not become the problem always. When appropriate actions for states included in an observation are the same, agent can learn without problems. In fact, when the appropriate actions for each state included in an observation are different, it adversely affects for learning. In this way, when agent is in incomplete recognition, it is possibility that agent can not learn appropriately. And it is need to distinguish several states included in an observation for such agent.

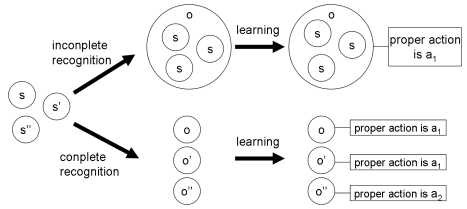

Fig. 1. The relationship of recognition and learning.

3 SUGGESTION OF RECOGNITION TECHNIQUE FOR POMDP

3.1 Estimate of state with previous recognized state and action

When incomplete recognition gives adverse affect for learning, agent should distinguish an recognition state (recognition state is state that agent finally decide) causing incomplete recognition as plural states by subdivision of recognizable state. So, when subdividing an recognition state that is causing the incomplete recognition, we focus on the pair of previous recognition state and action. Specifically when agent is in incomplete recognition at an recognition state, agent subdivides the recognition state by referring to the pair of previous recognition state and action (Fig.2). The subdivided observations are defined new internal state. And this internal state is held as knowledge by addition to X (X is a set of internal states). We define the internal state x in eq.(2). Here, $\hat{o}_t$ is an recognition state at time t, $\hat{o}_{t-1}$ is a recognition state that agent decide at time $t-1$, and a_{t-1} is an action agent takes in $\hat{o}_{t-1}$.

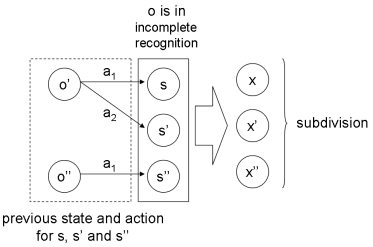

Fig. 2. The conception of estimate of state with previous recognized state and action.

$$x = \{\hat{o}_t, \hat{o}_{t-1}, a_{t-1}\} \tag{2}$$

Agent can recognize current state more particular by referring to previous recognition state and action. However, the number of the internal states that agent should experience increases, if agent subdivides an observation in all situation. It is not desirable that the number of the internal states increases for learning, because it takes much time to learn. In this paper, agent subdivides for the observation that may cause incomplete recognition. Therefore, it is need to judge whether the observation is incomplete recognition for agent. We think the observation includes plural states when a certain observation is incomplete recognition. And in this case, it is thought that different results are shown even if agent takes same action at the observation (Fig.3). Therefore, agent determines an observation that gains different result for same action as situation agent is in incomplete recognition. And we prepare the experience table to compare the results of actions. The proposed system has four original modules:

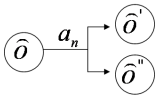

Fig. 3. The problem situation.

Recognition part to decide a internal state by using previous internal state and action, Judgment of incomplete recognition part to determine whether incomplete recognition occurs, Definition of internal state part to subdivide for observation in incomplete recognition, and Accumulation of experience part to saves experience information. And the proposed system also has two conventional modules: Learning part and Choosing action part. We show the construction of proposed system in Fig.4.

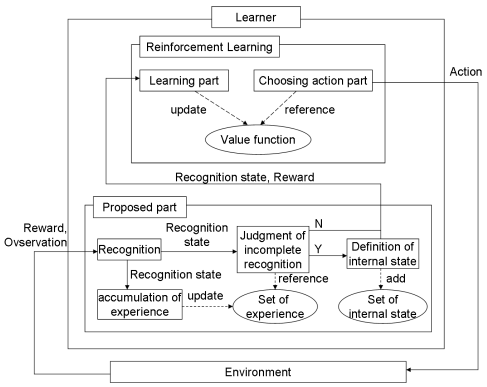

Fig. 4. The construction of proposed system.

3.2 Process of each module

3.2.1 Recognition part

In Recognition, agent decide a recognition state $\hat{o}_t$ by using current observation o_t and previous recognition state $\hat{o}_{t-1}$ and action a_{t-1}. And agent search the internal state x about $\{o_t, \hat{o}_{t-1}, a_{t-1}\}$ in set of internal state X. If agent find the internal state, agent decide x as $\hat{o}_t$. If agent can not fine the internal state, agent decide o_t as $\hat{o}_t$.

3.2.2 Accumulation of experience part

In Accumulation of experience, agent get the experience e as knowledge. This knowledge at time t is expressed in eq.3. And agent add this knowledge to set of experience E.

$$e_t = \{\hat{o}_{t-1}, a_{t-1}, \hat{o}_t\} \tag{3}$$

3.2.3 Judgment of incomplete recognition part

Agent judges whether an recognition state $\hat{o}_t$ is incomplete recognition in following procedures.

- Assume that a_t is an action in $\hat{o}_t$ and $\hat{o}_{t+1}$ is the result of action.
- Search the information about $\{\hat{o}_t, a_t, *\}$ in E
- Assume the found information as $\{\hat{o}_t, a_t, \hat{o}^{'}_{t+1}\}$
- Compare $\hat{o}_{t+1}$ with $\hat{o}^{'}_{t+1}$
- If $\hat{o}_t \neq \hat{o}_{t+1}$, agent judges that $\hat{o}_t$ is incomplete recognition.

3.2.4 Definition of internal state part

When agent subdivides a recognition state $\hat{o}_t$, a new internal state is defined in eq.2. And new internal state defined by eq.2 is add to set of internal state X.

4 SIMULATION USING PROPOSED SYSTEM FOR MAZE PROBLEM

4.1 Outline

We show the proposed system can distinguish the state enough to take task in POMDP through a simulation. We apply proposed system to an agent which recognizes state based on the sensor. And we apply maze problem to the agent with simulation. In the simulation, incomplete recognition happens when an agent decide a state by using only sensors. So the simulation is environment based on POMDP.

In the simulation, we define as one trial to reach a goal, and focus on the number of actions at each trial as the result. In addition, we prepare two agents for comparison. One is an agent deciding a state only with sensors which is same as sensors proposed system has. The other is an agent which has sensor that does not occur incomplete recognition. We unify other conditions for these three agents and apply maze problem and compare the results with these three agents.

4.2 Setting

We show the maze using the simulation in Fig.5. In Fig.5, we assume the coordinate at upper left square as (0,0), and lower right square as (2,2). The start position (•) is (0,2), and goal position ($\star$) is (2,2). In this simulation, agents fall into fatal incomplete recognition in the squares (0,1) and (1,1). Because appropriate action is difference in (1,0) and (1,1), we think it affects learning.

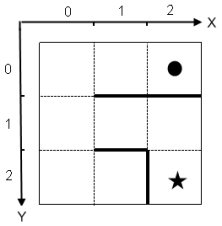

Fig. 5. The maze for simulation.

We prepare three agents (Agent-A, Agent-B, Agent-C). Agent-A and Agent-B has same four sensors. These sensors can recognizes whether there is a wall in front of sensors. Agent-A and Agent-B is set these sensors at vertically and horizontally. In addition Agent-A is applied proposed system. Agent-C own a sensor can know the coordinate in the maze And each agents select an action from four actions :move up, down ,left, right.

In this paper, we apply Q-learning to all agents as learning part. Agents learn based on eq.4 in Q-learning.

$$\begin{aligned} Q(s_t, a_t) &\leftarrow Q(s_t, a_t) + \alpha[r_{t+1} \\ &+ \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \end{aligned} \quad (4)$$

$Q(s_t, a_t)$ is action value of a_t in recognized state s_t at time t. r_{t+1} is reward which agent gains in recognized state s_{t+1} at time $t+1$. α and γ are parameters ($0 \leq \alpha, \gamma \leq 1$). And we apply ε-greedy method to all agents as action selection part. When agent select an action, agent take random action by probability of ε and take greedy action by probability of $1 - \varepsilon$. We show the parameters in Table 1.

Table 1. Setting for simulation

Reward(only a goal)	100
Total of trial	50
Initial value of Q-value	0.001
Size of maze	3×3
α	0.5
γ	0.7
ε	0.05

4.3 Result and Consideration

We show the result of the simulation in Fig.6. Fig.6 show the number of the actions for reaching the goal every trials for each agent. The X-axis expresses the number of the trials, and the Y-axis is the number of the actions that took for reaching a goal.

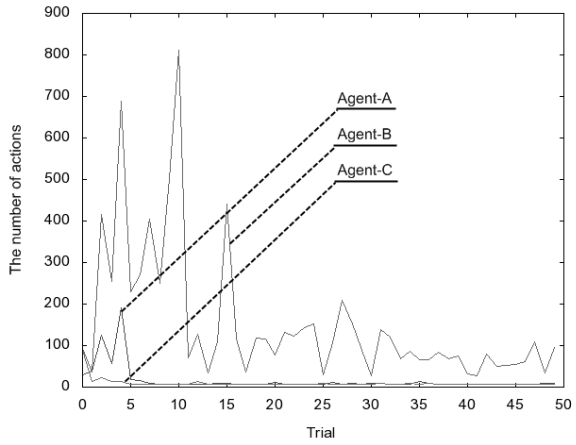

Fig. 6. Comparison of the number of actions each trial about Agent-A,B and C.

In this graph, the number of actions about Agent-B does not converge. Because Agent-B causes incomplete recognition in coordinate of (1,0) and (1,1), the agent can not learn appropriately. And the number of actions about Agent-B is bigger than other agents in all trials. Therefore, we say that incomplete recognition cases adverse affect to learning.

On the other hand, Agent-C does not cause incomplete recognition. Therefore the number of actions about Agent-C is less than Agent-A and B. Agent-C finds the most suitable route to goal by approximately five trials.

We focus on Agent-A. There are trials that there is more number of actions than Agent-C in early period of learning. However, the number of actions about Agent-A and Agent-C is almost same after 7 trials, and we think the proposed system works effectively for incomplete recognition. In proposed system, when agent causes incomplete recognition in coordinate of (1,0) and (1,1), agent distinguishes these state by subdividing the recognized state with previous state and action. Therefor, agent can distinguish enough to learn appropriately.

5 CONCLUSION

In this paper, we focus on the problem of RL in POMDP. There is a problem that agent cases incomplete recognition. And it is possibility that incomplete recognition cases adverse affect for learning. Therefor, agent can not learn appropriately in POMDP. We pay attention to previous recognized state and action for solving this problem. We proposed the recognition technique using not only sensors but also previous recognized state and action. In this way, agent can subdivide the recognized state and learn appropriately. And we proposed the system applied proposed technique to RL. We applied the maze problem to show the usefulness of the proposed system with simulation and we compared with three agents. By this simulation, we showed effectiveness of proposed system for incomplete recognition. In the future, we will apply the proposed system to actual robot.

REFERENCES

[1] Richard S. Sutton and Andrew G. Brato (1998), Reinforcement Learning. The MIT Press

[2] Daniel W. Ttroock (2005), An Introduction to Markov Process, Springer

[3] L.P. Kaelbling, M.L. Littman, A.R. Cassandra (1998), Planning and acting in partially observable stochastic domains. Artificial Intelligence Journal 101: 99-134

[4] E.J. Sondik (1978), The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. Operations Research 26(2):282-304

[5] K.J. Astrom(1965), Optimal control of Markov decision processes with incomplete state estimation. Journal of Mathematical Analysis and Applications 10:174-205