Reinforced learning by using a learned results of a different form robot

Nobuo Shibata, Hirokazu Matsui

Kurimamachiya 1577 Tsu Mie
(Tel : 81-59-231-9802)
(shibata@eds.elec.mie-u.ac.jp)

Abstract: The reduction on the trial frequency is important for reinforcement learning under an actual environment. In this report, we propose a method that can learn an unknown environment efficiently by applying an already learned Q-table to another new learning. Concretely, an agent (Target agent) learns efficiently by using Q-table that a different from agent (Source agent) has learned at the same environment. The proposed method uses the two Q-tables of Source agent and Target agent and an action-translation-table (AT-table). The Q-table of Source agent has already been learned at the same environment. The AT-table is learned for the suitable mapping of actions between Source and Target agents by information of the stored state transition probabilities when Source agent learned. Concretely, when Target agent takes an action at a situation, the AT-table is learned so which action of Source agent is similar to the action of Target agent. We think that Target agent can learn the suitable action efficiently by using the previous proposed method (for learning efficiently by using dual Q-tables) with Target Q-table and the set of Source Q-table and the AT-table instead of dual Q-tables. We verify that the proposed method is effective in a simulation.

Keywords: Reinforcement learning

I. INTRODUCTION

The reduction on the trial frequency is important for reinforcement learning. We propose an efficient learning method for robots by using knowledge.

When human being walks, he stores them as knowledge, the relations between moving and change of viewing. When he drives a car, he learns efficiently the steering operation by using the stored knowledge in walking. But, he cannot drive a car by only using the knowledge, he must learn new change of viewing, that never happens in walking. We apply an efficient learning method of human being to the proposed learning method for robots. In ordinary research, to reuse the knowledge, a learning task is separated into the smaller tasks[1], or using multi-layered reinforcement learning system[2]. In this report, we propose an efficient learning method that reuses the knowledge learned by another robot without transformation of learning modules. The proposed method is an extended method from our previous proposed method "Reinforcement Learning with Self-Instruction by using dual Q-tables"[3]. We verify that the proposed method is effective in a simulation.

II. LEARNING WITH DUAL Q-TABLES

Here, we describe that the previous proposed method(the dual learning method)[3] can learn the environment, earlier and in more detail by using the dual Q-tables.

1. Algorithm

The dual learning method uses two Q-tables (Fig.1), for experience storing and for knowledge storing. We thought experience and knowledge as followings. An experience is a collection of instances in an environment, so we use the whole space of the environment for the Q-table of experience. A knowledge is a compression of instances in an environment, so we use a compressed space (partial space) of the environment for the Q-table of knowledge. The smaller Q-table (for the knowledge storing) makes the learning finish earlier. The larger Q-table (for the experience storing) makes the learning be in more detail.

The dual learning method consists of repeated steps, the first step is that the action is selected by using the Q-table with lower information entropy, the second is that the two Q-tables are updated at the same time by the action. This research aims at the method that learns an environment earlier and in more detail with using dual Q-tables.

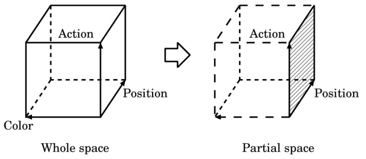


Fig.1. Q-tables for the partial space and the whole space

2. Materialization

Here, we materialize the above-mentioned concept to the procedures, as shown in Fig.2.

(1) Init dual Q-tables

The agent initializes the two Q-tables for the partial and for the whole Q-table.

(2) Select Q-table

The agent selects the Q-table with lower information entropy $H(s)$, out of the partial and the whole Q-table, that is driven by Eq.1.

$$H(s)=\sum_{a\in A} p(a\mid s)\log_2 \frac{1}{p(a\mid s)} \qquad (1)$$

where $p(a\mid s)$ is a probability of selecting action a at state s, that is defined by Eq.2. We consider that the Q-table to be so effective, that the information entropy of Q-table is lower.

(3) Select Action (for both Q-tables)

The agent selects an action by the Boltzmann selection used generally in Q-learning. The selection probability of the action a is shown by Eq.2.

$$p(a\mid s_k)=\exp\left(\frac{Q(s_k,a)}{T}\right)\Big/\sum\exp\left(\frac{Q(s_k,a')}{T}\right) \qquad (2)$$

where $p(a\mid s_k)$ is probability of selecting action a on state s_k, k is times, T is temperature.

(4) Update Q-table (for both Q-tables)

Each Q-value is updated by Eq.3. The equation is used generally when updating Q-value.

$$Q(s_k,a_k)\leftarrow\left(Q(s_k,a_k)+\alpha(r+\gamma\max_a Q(s_{k+1},a)-Q(s_k,a_k)\right) \qquad (3)$$

where, k is time, s_k is state at time k. $Q(s,a)$ is Q-value at state s and action a. $r(s,a)$ is reward decided by each pair of s and a. α is learning rate and γ is discount rate.

3. Result of simulation

We show results of average of the 1000 trials simulation in Fig.3. Fig.3(a) is a result in the case that partial Q-table is 100% effective, In other words, the states component of partial Q-table can distinguish the needed states by 100%. The case of Fig.3(b) and the (c) are 50% and 0% effective, respectively. The ordinary method uses only the partial Q-table or only the whole Q-table. As the results of (a)(b)(c) three cases, the dual learning method is not less effective than the ordinaries in any case.

III. PROPOSED METHOD

We propose a learning method that reuses the already stored Q-table and stored Environment table for another agent with different structure by extending the dual learning method. We call the previous learning agent, Source agent, and the current learning agent, Target agent. For the proposed method, we assume that the agent(Target) can access a result, that another agent(Source) with a different structure, learned in the same environment for the same goal.

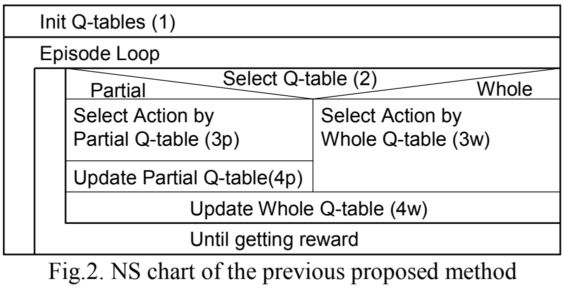
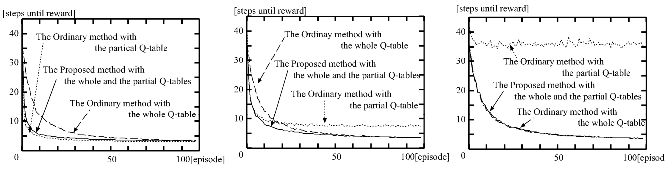

Init Q-tables (1)
Episode Loop
Select Q-table (2)
Partial
Whole
Select Action by Partial Q-table (3p)
Select Action by Whole Q-table (3w)
Update Partial Q-table(4p)
Update Whole Q-table (4w)
Until getting reward

Fig.2. NS chart of the previous proposed method



(a)100% effective (b)50% effective (c)0%effective

Fig.3. Results of computer simulation

1. Algorithm

In the proposed method, Target agent has four tables, two information tables and two learning tables. One information table is the already stored Q-table for Source agent. The other is the already stored Environment table(Env-table), that is the probability table of state transition by the state-action of Source agent. One learning table is an action-translation-table(AT-Table) for associating the actions of Source agent with the actions of Target agent. The other is a new Q-table for state and action of Target agent.

In the proposed method, the action of Target agent by a state is selected in two ways. In one way, the action for Source by the state is selected by using already stored Q-table for Source agent, and then the action for Source is translated to the action for Target by using AT-Table. By these steps, the action for Target agent by the state is selected. In this way, the learning space cannot be expressed perfectly for Target, but it is small. In the other way, the action by the state is selected directly by using a Q-table only for Target agent.

We apply the dual learning method mathematically to the learning of the AT-table and Target Q-table as smaller Q-table and larger Q-table. We expect as similar to the dual learning, that only by the first way Target agent learn more roughly and earlier, only by the second way it learn later and in more detail. But applying the previous, Target agent can learn earlier in more detail. In the proposed method, in order to learn more earlier, the agent can be evaluated by each action step, adding an evaluation by the Env-table to the first way.

2. Materialization

Here, we materialize the concept in the above-mentioned to the procedures of the proposed method, as shown in Fig.5.

(a) Leaning of the Source Agent (the previous learning)

Source agent learns a suitable action at the environment and stores the state transition probabilities to Env-table at each combination of action, current state and next state. The Env-table is shown in Fig.6(a).

(b) Learning of the Target Agent(the main learning)

(1) Init Target Q-table and Load Source Q-table and Env-table

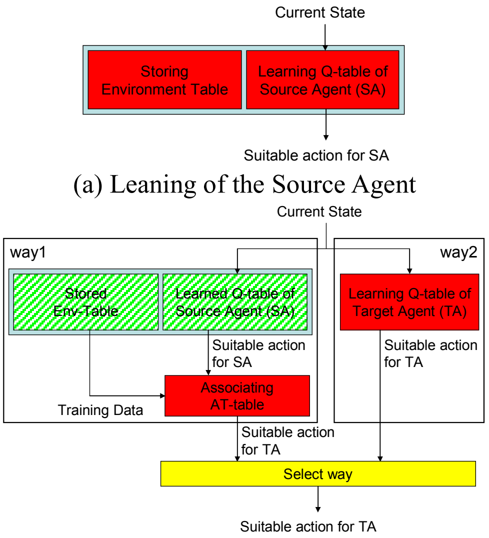
Target agent initializes the Q-table of Target agent to set each Q-value to an init value and load the AT-table and the Env-table from the previous learned data. The values of the AT-table and the Env-table are fixed (shown by the hatching areas in Fig.4(b)).

(2) Select way

Target agent selects the way with lower information entropy H_{way1} or H_{way2} , out of the two ways, that are driven by Eq.4, Eq.5, Eq.6, Eq.7 and Eq.8.

$$H_{t\arg et-Q}(s)=\sum_{a_t\in A_t} p(a_t\,|\,s)\log_2\frac{1}{p(a_t\,|\,s)} \tag{4}$$

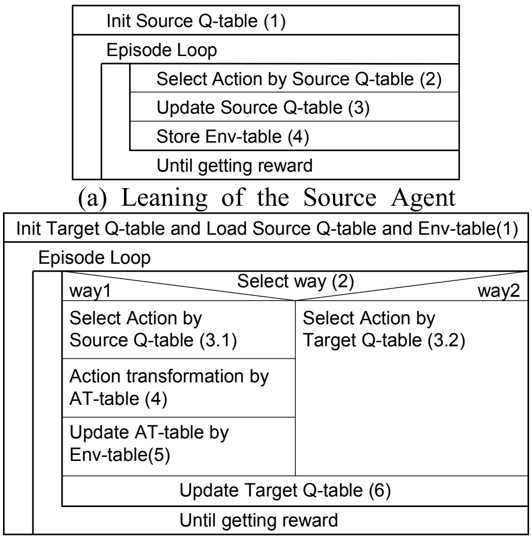
$$H_{source-Q}(s)=\sum_{a_s\in A_s} p(a_s\,|\,s)\log_2\frac{1}{p(a_s\,|\,s)} \tag{5}$$


Current State
Storing Environment Table
Learning Q-table of Source Agent (SA)
Suitable action for SA

(a) Leaning of the Source Agent

(b) Learning of the Target Agent

Fig.4. Block diagram of the proposed method


Init Source Q-table (1)
Episode Loop
Select Action by Source Q-table (2)
Update Source Q-table (3)
Store Env-table (4)
Until getting reward

(a) Leaning of the Source Agent

Init Target Q-table and Load Source Q-table and Env-table(1)
Episode Loop
Select way (2)
way1
way2
Select Action by Source Q-table (3.1)
Select Action by Target Q-table (3.2)
Action transformation by AT-table (4)
Update AT-table by Env-table(5)
Update Target Q-table (6)
Until getting reward

(b) Learning of the Target Agent

Fig.5. NS chart of the proposed method

$$H_{AT-table}(a_s)=\sum_{a_t\in A_t} p(a_t\,|\,a_s)\log_2\frac{1}{p(a_t\,|\,a_s)} \tag{6}$$

$$H_{way1}(s)=H_{source-Q}(s)+H_{AT-table}(a_s(s)) \tag{7}$$

$$H_{way2}(s)=H_{t\arg et-Q}(s) \tag{8}$$

where a_t is an action of Target agent, a_s is action of Source agent, $p(a\,|\,s)$ is probability of selecting action a at state s, that is defined by Eq.2. $p(a_t\,|\,a_s)$ is probability of selecting action a_t at already selected action a_s, that is stored in the AT-table. We consider that the way to be so effective, that the information entropy of the way is lower. Function $a_s(s)$ in Eq.7 is represented for that a_s is decided by s (selected by using Boltzmann selection with Source Q-table).

(3) Select Action (for both Q-tables)

Target agent selects an action by the Boltzmann selection. The selection probability of action a is shown by Eq.3.

(4) Action Translation by AT-table

Action a_s selected by Source agent Q-table is translated to action a_t by using the AT-table. Action a_t corresponding to action a_s is selected by the Boltzmann selection with AT-table. The selection probability of action a_t is driven by Eq.9.

$$p(a_t\,|\,a_{s_k})=\exp\left(\frac{AT(a_{s_k},a_t)}{T_{AT}}\right)\Bigg/\sum\exp\left(\frac{AT(a_{s_k},a_t)}{T_{AT}}\right) \tag{9}$$

where $p(a_t\,|\,a_{s_k})$ is probability of selecting action a_t at already selected action a_{s_k}, k is times, T_{AT} is temperature.

(5) Update AT-table

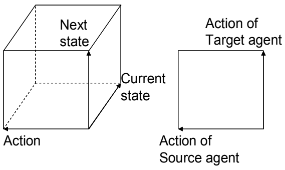
The AT-table is learned for the associating of actions between Source and Target agents by Env-table. AT-table is updated by Eq.10.

$$AT(a_{s_k},a_{t_k})\leftarrow\left(\alpha_{AT}AT(a_{s_k},a_{t_k})+\gamma_{AT}\,p(s_{k+1}\,|\,s_k,a_{s_k})\right) \tag{10}$$

where $p(s_{k+1}\,|\,s_k,a_{s_k})$ is state transition probability to state s_{k+1} by pair of action a_{s_k} and state s_k, α_{AT} and γ_{AT} are such that $0\le\alpha_{AT}<1$ and $0\le\gamma_{AT}<1$.

(6)Update Q-table

Each Q-value is updated by Eq.3.



(a) Env-table (b) AT-table

Fig.6. Env-table and AT-table

IV. SIMULATION

In this chapter, we verify that the proposed method is effective in a simulation.

1. Environment

Experiment environment is shown in Fig.7. It consists of an agent, a goal and an object in a field(100[cm]x50[cm]). In the experiment, the agent learns the task to carry the object to the goal area, given reward at the case that the object carried into the goal. We define an episode to be time until getting reward. The initial pose of the agent, the object and the goal are shown in Fig.7. An initial pose of agent is selected randomly from two pose shown in Fig.7 at each episode. The pose of agent is changed 0.5[cm] and 0.1[deg] per move respectively. We define one step to be time until the current state is changed to next state. The agent keeps taking the same action until the current state is changed.

2. Action set and State set

Action set: The actions of Source and Target agents are shown in Fig.8. Source agent has 4 actions "go forward", "go backward", "pivot turn right" and "pivot turn left" and Target agent has 6 actions "go forward", "go backward", "turn right (forward)", "turn left (forward)", "turn right (backward)" and "turn right (backward)".

State set: A state set consists of the position states of goal and object. The state set has 160 states that are each position and size of the object and the goal. The object state has 6 states, combinations of positions(left, center, or right) and sizes(large or small). The goal state has 18 states, combination of positions(left, center, or right), sizes(large or small) and directions(left, front, right). In addition to these 78 states, we add states that only the object or only the goal view or lose.

3. Result

The results of the simulation are shown in Table.1 and Fig.9. Each graph is the average of 1000 trials. In these results, all the values of Q-tables, AT-table and Env-table are initialized to 0.0 and the parameters for Q-learning and AT-table mapping are set to following. The reward r is set to 1.0, the learning rate α is set to 0.4, and the discount rate γ is set to 0.9, the temperature T of Boltzmann selection is set to 0.05, α_{AT} is set to 0.9, γ_{AT} is set to 0.35, T_{AT} is set to 0.5.

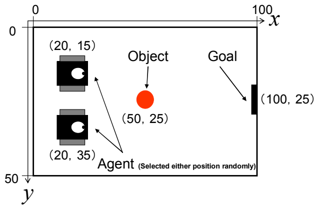
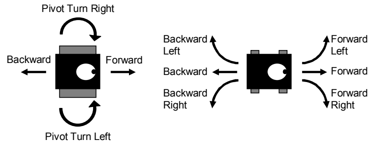


Fig.7. Experiment environment



(a) Source agent (b) Target agent

Fig.8. Actions of the Source and Target agents

In Table.1, high value is represented for correct mapping between Source and Target action pair. For example, "Forward Right" of Target agent allocated to "Pivot turn Right" of Source agent. So in Fig.9, the method with Source agent Q-table and AT-table (way1) can learn earlier than the method with Target agent Q-table (way2). But way1 cannot learn in detail since the pair of Q-table of Source agent and AT-table cannot express all the corresponding action to Target agent. Way2 can learn in detail but needs a lot of time. However, the proposed method (way1 and way2) can learn the environment earlier and in more detail by using Source and Target Q-table and AT-table.

Table 1. Result of AT-table

Source Target	Forward	Backward	Pivot turn Right	Pivot turn Left
Forward	2.573	0.172	0.517	0.150
Backward	0.175	1.761	0.469	0.405
F Right	0.874	0.405	1.758	0.002
F Left	0.652	0.339	0.028	1.292
B Right	0.451	0.645	0.002	1.189
B Left	0.687	0.711	1.848	0.003

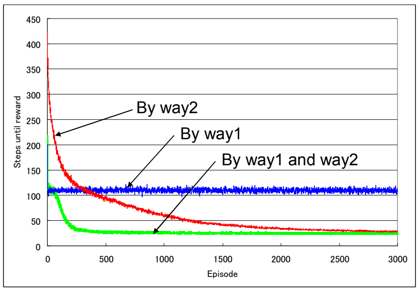


Fig.9. Results of computer simulation

V. CONCLUSION

We verified that the proposed method can learn more efficient than the MTQ and MSQAT in simulation.

In the future, we will apply the proposed method to an actual environment and verify that the proposed method can learn efficiently in an actual environment.

REFERENCES

[1] Akihiko Yamaguchi,Norikazu Sugimoto,Mitsuo Kawato(2009):Reinforcement Learning with Reusing Mechanism of Avoidance Actions and its Application to Learning Whole-Body Motions of Multi-Link Robot(in Japanese), Journal of the Robotics Society of Japan Vol.27 No.2,pp.209-220

[2] Yasutake Takahashi,Minoru Asada(2003):State-Action Space Construction for Multi-Layered Learning System(in Japanese),Journal of the Robotics Society of Japan Vol.21 No.2,pp.164-171

[3] Osamu NISHIMURA,Hirokazu MATSUI,Chieko HIOKI,et al(2006):Reinforcement Learning with Self-Instruction by using dual Q-tables,AROB 11th