Acquisition of Cooperative Behaviour among Heterogeneous Agents using Step-up Reinforcement Learning

Wataru Sato
Department of Informatics
Graduate School of Engineering
Kogakuin University
Shinjuku-ku, Tokyo, Japan

Kanta Tachibana
Faculty of Informatics
Kogakuin University
Shinjuku-ku, Tokyo, Japan
kanta@cc.kogakuin.ac.jp

***Abstract*—This paper discusses acquisition of cooperative behaviour among heterogeneous agents and proposes two methods to promote cooperative behaviour: phased learning and selective recognition. For complicated scenarios such as multi-agent tasks, we propose phased learning, in which agents first learn in a simpler environment before learning in the target environment. For heterogeneous multi-agent tasks, we propose selective recognition, in which an agent recognizes a partner, with whom it can cooperate to earn rewards, selectively. By means of simulations in which two types of agents cooperated to capture prey, we verified that, using our proposed methods, agents are able to differentiate agents they should cooperate with from those with whom they should not.**

Keywords—reinforcement learning; multi-agent system; heterogeneous agent;

I. INTRODUCTION

As work in various fields become increasingly automated, societal expectations for the widespread availability of more versatile robots are on the rise. In this situation, multi-agent systems [1] that facilitate the execution of complex tasks by cooperative action among multiple autonomous agents have attracted increased attention. In particular, Multi-Agent Reinforcement Learning (MARL) [2] is viewed as being effective in complex environments such as the real world because of its excellent versatility.

In a multi-agent environment, an agent learns its action regarding other agents as a part of the environment. Other agents also act according to learning rules. Thus, in the view of an agent, environmental changes are affected by the progress of the learning of other agents. Ito and Kanabuchi [3] proposed a learning scheme that reduces learning time significantly. In the first stage of their proposed learning scheme, some agents have a limited perception of the environment, whereas other agents have a full perception. Subsequently, after all agents have become intelligent, they all fully perceive the environment. Not only is the learning duration reduced, but in after-learning solution convergence, the performance of their scheme is almost equivalent to the performance of fully perceiving agents.

In reinforcement learning, if an agent earns no reward over a long period of time, reinforced values become similar to each other/moderate and the agent selects actions randomly, i.e., it forgets good behaviour. More than ever, MARL is likely to frequently encounter this kind of situation. As a solution, this paper proposes a phased learning method in which agents first learn in a simpler environment before learning in the target environment.

Previous studies of MARL reported on acquisition of cooperative behaviour among homogeneous agents. However, in the real world, situations that require cooperative behaviour among heterogeneous agents often arise. Zhang et al. [4] proposed a scheme in which an agent utilizes the reinforced value of others at a rate corresponding to their perceived reliability. This scheme enables an agent to automatically distinguish superior or inferior partners with whom to cooperate. However, this method of using a partner's reinforced value can be applied only in situations in which that agent and the partner's optimal behaviour are the same or similar. Consequently, this method is inappropriate in cases where cooperative behaviour is required among heterogeneous agents.

This paper also proposes a selective recognition method in which an agent recognizes a partner with whom it can cooperate to earn rewards, selectively. Further, scenarios in which the agent has not determined which partner it ought to cooperate with are considered. We also verify via simulation that, using our proposed method, agents are able to distinguish between other agents that they should and should not cooperate with in order to earn rewards.

II. EXPERIMENTAL TASKS

Fig. 1 displays an example of our simulation setup for the prey capturing task. The simulation map has a torus topology. Thirty hunters made up of two different species and 30 preys, a total of 90 agents, exist in the map. A hunter agent of one species was rewarded if it caught prey in cooperation with an agent from the other species. Conversely, it did not get a reward if it did so with an agent from its own species. Agents were able to see the nearest 20 × 20 grids. Preys tried to escape from the nearest hunter in sight. All agents stayed where they were or moved one grid per time step in either of eight

directions. An episode was concluded when all preys were caught or 3000 time steps had passed.

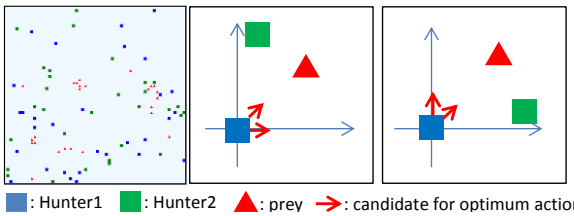


Fig. 1 Simulation map Fig. 2 Partner and prey in the same directions

III. LEARNING RULE OF THE HUNTER

We utilized Q-learning for the learning rules of the hunter, with softmax action selection.

A. Definition of state and action

State s_t of an agent at time step t consists of the direction of the nearest prey s_{1t} and that of the nearest partner hunter s_{2t}. The direction is either of four quadrants {the first to the fourth} or four axes $\{+x, +y, -x, -y\}$. To reduce the total number of possible states, the nearest prey and hunter directions are rotated so that the nearest prey is either in the first quadrant or on the $+x$ axis. Optimal action differs depending on the locational relation of prey and partner. As depicted in Fig. 2, s_{2t} is in the first quadrant, s_{2t}' must therefore be partitioned more. States are partitioned into two such that the partner is located on either the right or left side of a line through the hunter and the prey. From the above, $S_{1t}(\ni s_{1t})$, $S_{2t}(\ni s_{2t})$ have three states and 10 states, respectively. A $\ni a_t$ has nine actions (eight directions and stay).

B. State transition probability and entropy

State transition probability is calculated by softmax as outlined below. Entropy is also calculated by the following equation:

$$P(a_t|s_{1t}, s_{2t}) = \frac{\exp[Q(s_{1t}, s_{2t}, a_t)/T]}{\sum_{b\in A}\exp[Q(s_{1t}, s_{2t}, b)/T]}$$

$$I = \frac{1}{N}\sum_{i}^{N}\sum_{s} P_i(s)\sum_{a} P_i(a|s)\log P_i(a|s)$$

$P_i(s)$ is the probability that agent i has entered state s in the episode. $P_i(a|s)$ is the state transition probability, and N is the number of agents. Entropy is the standard value of learning progress calculated at the end of each episode.

C. Learning parameters

The Q-learning parameters are the following, learning parameter, $\alpha = 0.1$, discount parameter, $\gamma = 0.99$, temperature parameter, $T = 0.6$, reward, $R = 10$, and initial Q-values = 0.1. Reward applies to all hunters to cooperate for prey. Update Q-values is the same as in the regular Q-learning method.

IV. PROPOSED METHODS

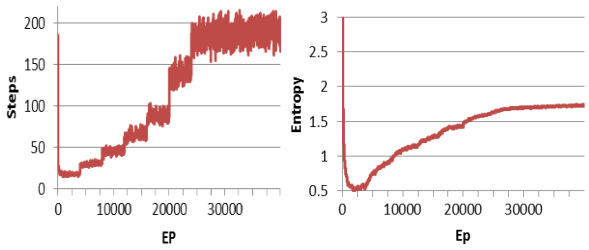
A. Phased learning method

At the beginning of learning, each agent's action is virtually random (according to initial Q-values), so an agent's learning will not proceed. In our simulation, an agent could not earn reward without cooperating with a dissimilar agent to catch prey; thus, agents rarely earned a reward. Further, if a number of agents earned rewards and performed their actions well, none of those agents could get the next reward, and they also had to forget the Q-values learned before. Earning reward is difficult without advanced learning by many agents. We conducted simulations in maps of sizes 50 × 50 and 60 × 60. In both simulations, the hunters could not catch all preys within 3,000 time steps, even after learning in 10,000 episodes. Further, entropy did not converge.

Consequently, we made the agents first learn in a simple environment and applied a moderate superior Q-value to all agents; after which they were made to learn in the target environment. We consider that this method results in a rapid start to learning. Figs. 3 and 4 show that the entropy is now smooth, with moving average of 30 episodes.

Fig. 3 shows the map length and breadth expanded by five grids after every 4,000 episodes. The map size also changed from 30 × 30 to 60 × 60. The total number of episodes began at 40,000 and changed afterwards to 60 × 60 or 16,000 episodes.

Fig. 3 shows that agents could catch all preys in almost all episodes and that the steps converged. In the 60 × 60 map, entropy was likely to converge. The experiment showed that prior learning is effective as a precursor to learning in complex environments in which agents have many states.



(a) Step (b) Entropy

Fig. 3. Result of phased learning method

B. Selective recognition method

In this section, we consider the selective recognition method, which is applicable for heterogeneous agents. In this method, an agent recognizes a partner, with whom it can cooperate to earn rewards, selectively. We conducted a simulation experiment to verify this method. In the simulation, several hunters knew partner hunters with whom they could cooperate and recognized only such partners.

Further, we considered the situation in which an agent does not know which agent it ought to cooperate with.

Consequently, we verified that agents were able to automatically distinguish other agents with whom they ought to cooperate as well as those with whom they should not. Each agent had an array of binary values with length equal to the total number of other hunters. Each element of the array signified whether the corresponding indexed hunter was to be cooperated with. In each episode, the agent either 1) recognized all hunters and updated the values of the array or 2) recognized only hunters with a true value in the array. When an agent applied 1), that agent recognized all agents. If the agent then spotted and captured a prey, the element corresponding to the agent it cooperated with changed to true. On the other hand, if the agents spotted a prey but could not capture it, the element corresponding to the partner agent changed to false. When an agent applied 2), the agent recognized only agents with the value true in the array and acted. In addition, when 2) is applied, the array is not updated. The exploration rate was 1):2) = 2:8 in each episode.

Fig. 4 shows the results obtained using this method. "heteroN" signifies that N hunters out of 30 were heterogeneous hunters. Further, "discriminate" signifies all hunters automatically distinguished, as described in the previous paragraph.

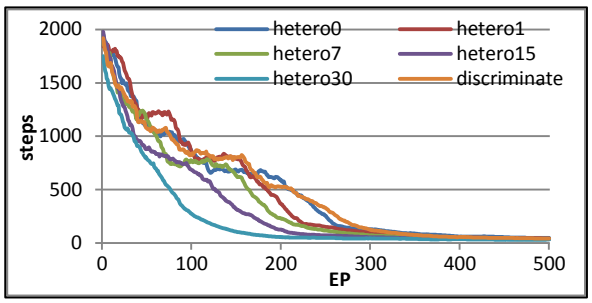
Fig. 4(a) shows that the convergence speed of the time steps was quicker according to the increase in hetero, and Fig. 4(b) shows that the convergence steps decreased according to the increase in hetero. Fig. 4(c) shows that entropy decreased with increase in hetero. In addition, the result shows that the selective recognition method had the faster learning speed. In addition, Fig. 4 shows that discriminate trial was similar to the hetero30 trial and that the performance was superior to hetero0s after converging. Thus, the validity of automatic distinction was verified.

V. CONCLUSION

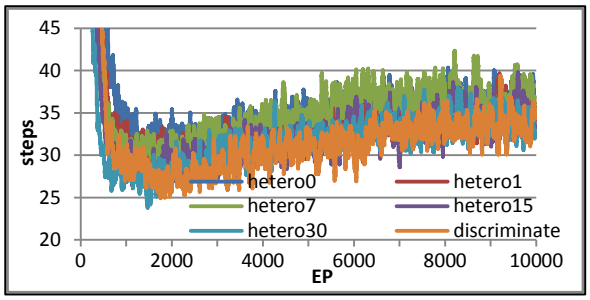
This paper discussed the acquisition of cooperative behaviour among heterogeneous agents through simulations in which two types of agents must cooperate in order to capture a prey.

We proposed a phased learning method in which agents first learn in a simpler environment before learning in the target environment. The resulting experiment conducted showed that the learning solution converged in the early episodes using this learning method. Further, in environments in which learning solutions cannot converge using common reinforcement learning, learning by the phased learning method still resulted in convergence. Thus, this method can be applied in complex environments and environments with a large number of states.

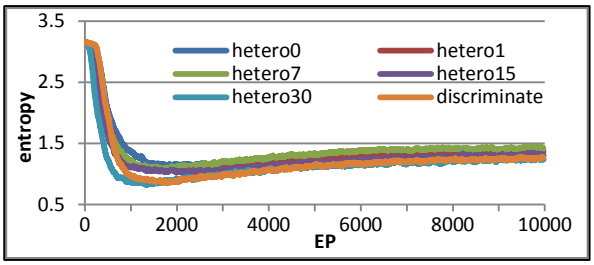
This paper also proposed a selective recognition method in which an agent recognizes a partner, with whom it can cooperate to earn rewards, selectively. The subsequent experiment conducted showed that learning using this method is superior in convergence performance to the learning solution. Further, we verified experimentally that agents were able to differentiate agents they should cooperate with from those with whom they should not, proving that the method is effective.



(a) Steps (to 500 episodes)



(b) Steps



(c) Entropy

Fig. 4. Result of selective recognition method

REFERENCES

[1] Yoichiro Maeda, "Evolutionary simulation for co-operative behavior learning on multi-agent robots," Japan Society for Fuzzy Theory and Systems, vol. 13, no. 3, pp. 57-67, 2001.

[2] Hajime Kimura, Kazuteru Miyazaki, Shigenobu Kobayashi, "A guideline for designing reinforcement learning systems," Journal of the Society of Instrument and Control Engineers, vol. 38, no. 10, pp. 618-623, 1999.

[3] Akira Ito, Mitsuru Kanabuchi, "Speeding up multiagent reinforcement learning by coarse-graining of perception: The hunter game," Electronics and Communications in Japan, vol. 84, no. 12, pp. 37–45, December 2001.

[4] Kun Zhang, Yoichiro Maeda, Yasutake Takahashi, "Learning model considering the interaction among heterogeneous multi-agents," Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, vol. 24, no. 5, pp. 1002-1011, 2012