

A Report The Difference Features Of A Multi-agent Using An Overlay Knowledge In The Fire Panic Problem

Yukinobu HOSHINO

School of Systems Engineering
Kochi University of Technology
Kochi, Japan
hoshino.yukinobu@kochi-tech.ac.jp

Abstract—This report of this paper shows a performance of the heterogeneous multi-agent system using the overlay knowledge technique. The overlay knowledge is one kind of layered rule-based system. Basic idea is the tile-coding rule on reinforcement learning. But In the tile-coding rule, all reinforced rules are spread any categorical states of rule table. Proposal style of overlay knowledge is from a tile-coding. We examined the features of the learning principle about the heterogeneous multi-agent. It was proved that the overlay knowledge was useful sufficiently for the heterogeneous multi-agent. The robust characteristic was the cause which outputs absolute behavior. By changing of the robust levels, behaviors of the agent expected a change into the flexible movement.

Keywords—component; multi-agent system ; overlay knowledge ; reinforce learn; fire panic

I. INTRODUCTION

The report on this paper shows a performance of the heterogeneous multi-agent, which have the overlay knowledge. The overlay knowledge is one kind of the layer type rule-based system. Our basic idea is the tile-coding rule of the reinforcement learning. In the tile-coding, all reinforced rules are tiled some categorical states of the rule table. Recently, there are the shaping style rules by using other training method. So we believe the robustness and applicability phenomenon between behavior and rule's amount. Especially, rule style of the overlay knowledge easy to find that phenomenon. We would like to show that in this paper and discuss the learning theorem in Multi-agent system.

II. LEARNING THEORY OF MULTI-AGENT SYSTEM

Learning system and intelligence are related to each other. Learning always has to do about the self-improvement of future behavior based on past experiment, which agent has gone. Learning of the multi-agent system have two principal categories. One is centralized leaning or isolated learning. Another is decentralized leaning or interactive learning. In the multi-agent system, several centralized learning would like to try to obtain same activate at same time by same goal. There are several groups of agents that involved by different decentralized learning on each agent's. Especially, the

heterogeneous multi-agent system has some group agents which have different behavior. Interactive learning process is feature quality. That two learning categories are a general nature and a broad variety forms of learning. Generally, the multi-agent system has six type difference features [9].

- ① The degree of decentralization :
The decentralization of a learning process concerns its distribution and parallelism. One normal case is that a single agent carries out all learning behavior. Other case is that the learning system is distribute and parallelism through all agents of a group.
- ② Interaction-specific features :
There is a number of features that can be applied to classifying the interactions request for realizing a decentralized learning process. Examples are the level interaction, the frequency interaction, pattern interaction, the variable interaction and so on.
- ③ Involvement-specific features :
That feature can be used for characterizing the involvement and role played during involvement. Learning processes of agent are the relevance of involvement and role played during involvement.
- ④ Goal-specific features :
Learning goal of multi-agent are the type of improvement that is tried to be achieved by learning and the compatibility of the learning goals traced by the agents.
- ⑤ The learning method :
The following learning methods or strategies used by an agent are usually distinguished. There are rote learning, learning from instruction and advice taking. Otherwise, learning from examples, practice leaning, analogy learning, learning by discovery and so on.
- ⑥ The learning feedback.

The learning feedback indicates the performance level behavior. This is leading factor to follow classification, success, and goal by some feedbacks, are supervised learning and unsupervised learning. Reinforcement learning is a kind of unsupervised learning but the feedback is only specific signal of the utility, which depends on the actual behavior of learner. And the objective is to maximize for targets.

In the heterogeneous multi-agent system, leaning system has to follow those features. Interaction's reference is key-point and the research point of using the overlay knowledge.

III. Q-LEARNING

Q-learning is a type of unsupervised learning system. The evaluation about selected behaviors and policies, make renew/reinforce by a reward/penalty. The agent selects behaviors according to the observed state. The selected behaviors become the best behavior, because Selective Unit select them according to rules reinforced by $Q(s,a)$ where a is a behavior and s is the observed state.

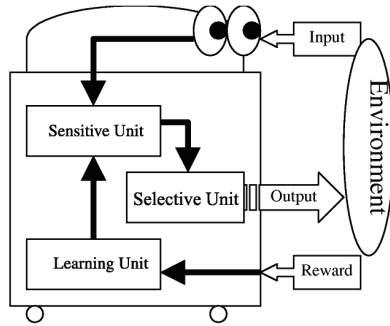


Fig.1 Framework of Q-learning

If there are no fitting rules, a new rule is created on the rule database; also rule base. Q-learning is able to learn from uncertain information. The learning systems are composed three units, including a sensitive unit, a learning unit, and a selection unit, as shown in Fig.1. If agents have received a positive reward, Q-value is renewed a positive direction. So, using rewards and penalties, the learning unit is able to change Q-value with the rules used. The renewal function f_i is given in Eq.(1). $Q(s_t, a_t)$ is an evaluation to make renewal, shown in Eq.(2). In these cases, γ is the discount rate of $Q_{\max}$, d is the discount rate of a reward and α is the learning rate. s_t is observed state at the current time t . a_t is the selectable action at the current time t . Those parameters are in the range $[0,1]$. T is whole time of one episode. t is current time in one episode time T . Also, $Q_{\max}$ is the max evaluation of the best policy at the next state $S_{(t+1)}$. r is +1 as a reward or -1 as a penalty. Eq.(2) explains that Q-value is renewed by the reward.

$$f_t = \gamma Q_{\max} - Q(s_t, a_t) + d^{T-t} r \quad (1)$$

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha f_t \quad (2)$$

An agent selects one behavior based on the view it has of the surrounding environment. The selection unit always uses probability based on Boltzman Distribution, as given Eq. (3). $Q(s,a)$ is Q-value about a target state s and a candidate action a . T is "Temperatures parameter" and p is probability to select a candidate action a .

$$p(a | s) = \frac{\exp(\frac{Q(s,a)}{T})}{\sum_{b \in action} \exp(\frac{Q(s,b)}{T})} \quad (3)$$

IV. THE FIRE PANIC PROBLEM

We show the fire panic problem on fig.2. In this problem, a fire exists in a field and it spreads throughout the space. The purpose of the agent is to reach an established goal without touching the fire that is in the environment. In order for the agents to avoid the fire spreading in the environment, they must acquire behaviors that allow them to reach a goal before the fire reaches them. The fire heats during three steps. After this three steps, the fire jumps to the randomly one direction (up, down, left or right) at one step. In the fire panic problem, all agents have to follow features ① ~ ④ about using the overlay knowledge. Please check a section II. LEARNING THEORY OF MULTI-AGENT SYSTEM

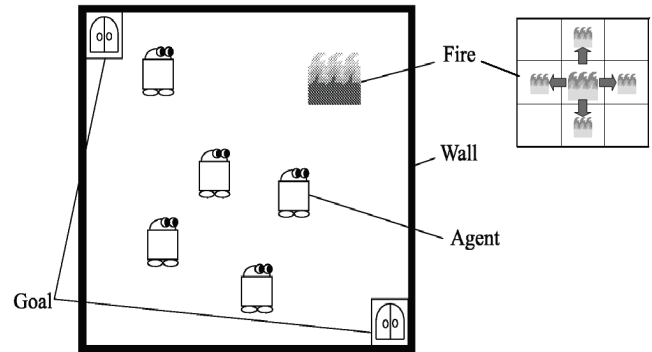


Fig.2 The fire panic problem

In this problem, we use two types of agents, A1 and A2. These agents are heterogeneous multi-agents with differing areas of view. The view area of A1 is 3x3. The view area of A2 is 5x5. Also, all agents can recognize all of these objects in the space: Fire, Wall, Agent, and Goal. All agents can select from these behaviors: Up, Down, Left, Right and Stop.

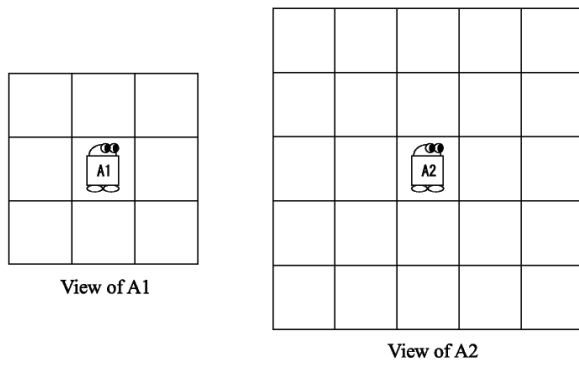


Fig. 3 Fields of view in agents A1 and A2

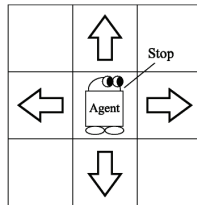


Fig. 4 Behavior of Agents

Agent is not able to enter any spaces that already contain Wall or Agent. In the fire panic problem, each action is one step. Each trial have 100 steps, all agents are able to move during 100 steps in a field. Agents learn behaviors using Q-learning or the our overlay knowledge system.

- If the agent reached a goal during the trial, the agent receives the reward $r=+1$.
- When a trial finished, agents on the field that are not touched by fire receive no-reward $r=0$.

If the agent touches the fire during the trial, the agent dies and receive the penalty $r=-1$.

V. APPLING Q-LEARNING

At first, we implemented Q-learning system for A2 agents in 30,000 trials. Then, after these trials, we introduce A1 agents, with smaller fields of view, into the space and we conducted 30,000 further trials in which A1 and A2 agents learn together in the same environment. Fig.5 shows the environments of the first and second trials. Fig.6 and Fig.7 show the number of times an agent reached a goal for every 1,000 trials.

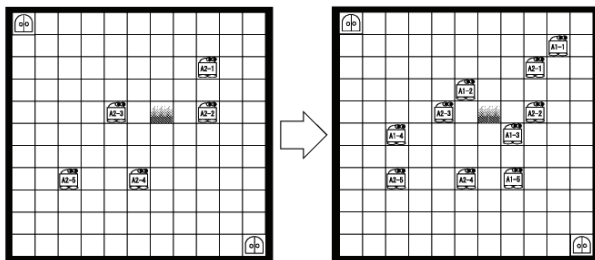


Fig. 5 First and second trials on the experiments

We could observe that A1 got the best behavior to reach a goal, show Fig.6. Then the number of goal on A2 has decreased at Fig.7. Those are results of using normal Q-Learning. About those curves, a vertical axis is a number of goal arrivals for every 1,000 trial. A horizontal axis is a number of trials. The number of goal has never increased again about A2 agents, given as Fig.7. We have kept experiment 60,000 trials, but the number of goal has maintained a low level. See Fig.8 (A), the best action of A2 is selecting a left move. During first term, A2 agents have never seen A1 type agents and finished a leaning to take the best action. But rule base is recorded view and Q-value with all action. See Fig.8 (B), if A2 agent has sensed A1 at first time, then A1 agent is not able to find an applicable rule. Because the rule base system seek a completely fitting rule. Basically, if there is not fitting rule in the rule base, then whole agents has no choice the best action. Thus, A2 agents cannot help choosing the random action, like a confusing behavior. Finally, A2 should learn same action on those views as a different situation.

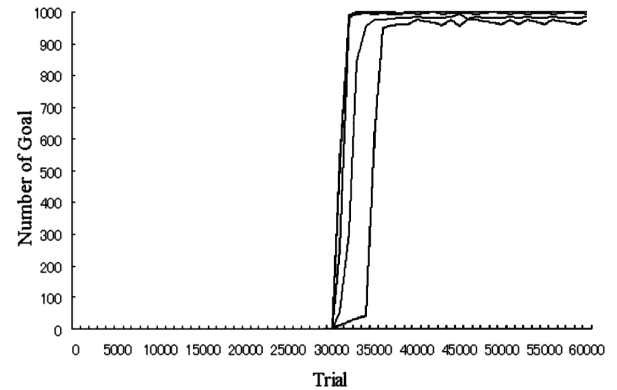


Fig. 6 Number of goal arrivals for A1 agents on Exp.1

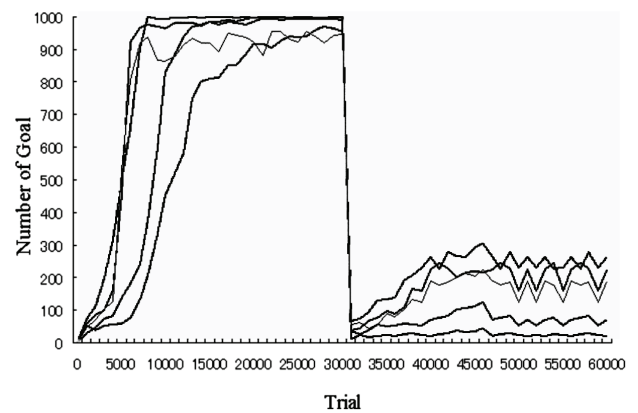


Fig.7 Number of Goal arrivals, taken A2 on Exp.1

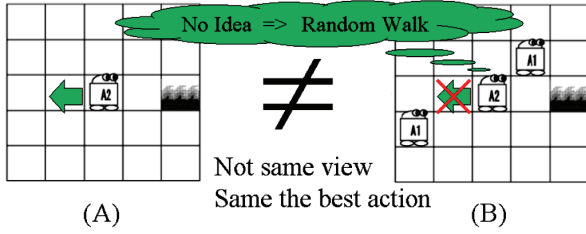


Fig.8 Process of the confusing the best action

VI. REINFORCEMENT LEARNING USING THE OVERLAY KNOWLEDGE

An applicable behavior is selected from a probability with the Boltzman distribution with Q-value. Proposal method is introduced in A2, in which the field of view is 5x5, and learning is determined in accordance with Q-learning. The proposed concept for our method used is shown in Fig. 9. The system renews Q-value with the learning layer. So this system was easy system.

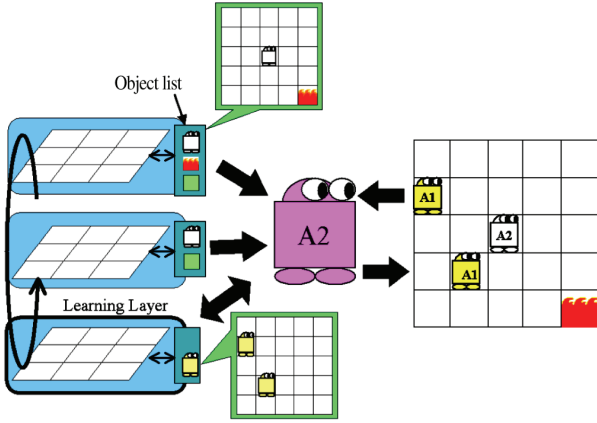


Fig.9 Framework of the overlay knowledge Method

In our learning system, we built the object list to switch layers. About an input status, the system searches object list of a current layer. Some objects are registered to the object list, if a target agent experienced objects in past times. In this fire panic problem, all agents have a possible to find A1 agent, A2 agent, fire, and wall as an object. If designer selected any items for target problem, then it is easy to add/record the object list, which would support any cases on the target problem. When the system learns by a reward, the system renews Q-value $q_{j(x,y)}$ on a current learning layer j . However before that process, the system check that observed objects exists in the object list of the learning layer j . During starting trials in the experiment, agents found new objects at all time, so adding object to the object list is anytime and quickly. When this system had some layers, the system was not able to follow the switching layers quickly. Because the system needs some layer know ledges for overlay Q-values. Then the system keep switch layers quickly, an approximating is impossible.

If new observed objects, which do not exist on the object list of the learning layer, then the system switch a current layer and create a new object list from new observed objects.

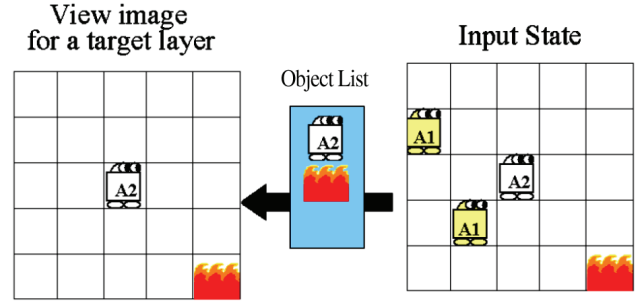


Fig.10 Object list and the masked view

The learning system adds the new rule and renews Q-value on the new learning layer. If a current layer is a last layer, then a current layer is switched to a first layer by the system; turn over to a first layer. This learning system can support several objects by the rule base on other layers. Proposal system has five layers with an object list. The input system masks out the unnecessary objects by checking the object list. And the system applies some Q-value from other masked layers, if the learning layer has no applicable rule. See Fig.10. Those preexisting Q-values on layers are possible to support an action about new situation, which the agent did not experience in past times. Additionally, proposal system has weight parameter w_i for each layer i . In a proposal system, the evaluation value $Q(s,a)$ is calculated by $q_i(s,a)$, as given Eq.(4). $q_i(s,a)$ is a output Q-value from a layer i . In this case, all q_i has evenly reach to $Q(s,a)$. The learning system searches a fitting $q_i(s,a)$ from all recorded q values with each layer i . The observed state keep changing dynamically, so the best policy will change by those situations. Weights show a possibility the applicable behavior for $q_i(s,a)$. We have tried some simulation experiment about weight parameters.

$$Q(s,a) = \sum_{i=1}^m w_i \times q_i(s,a) \quad (4)$$

VII. MULTI LYAERD QL METHOD FOR FIRE PANIC WITH A1 AND A2

Proposal system make $Q(s,a)$ for select behavior by Eq.(3)(4). And we implement our overlay knowledge system for A2 agents, has **three layers**. A1 agents have normal Q-learning system. When an agent finished one trial, Q-value of used rules are renewed by the reinforce function as Eq.1 and Eq.2. And $Q_{\max}$ value is selected from a Q-value record in an experiment. Thus, this learn processes are executed when an agent finish one episode. See Fig.11. A current layer has 50.0% weight and a second layer and a last layer have 25.0% and 25.0%.

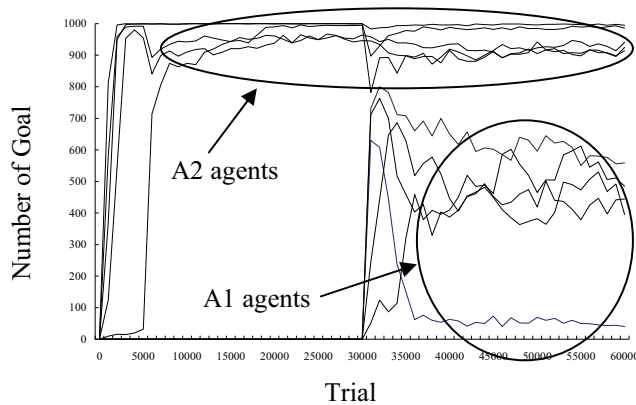


Fig.11 The curve of three layers and Result of weights are 50.0% 25.0% 25.0%

In this case, we took a good result normally, a goal number of A2 agent is decrease little bit when A1 agents came in. And the number of goal came back and kept high values. In this case, preexisting rules could work and support for new states with A1 agents coming. As a result A2 agents did not confuse about coming A1 agents in those views. Here, agent 2 supports a new condition without being confused by the show in the robust-ness of the overlaid knowledge. In this way, the overlaid knowledge is acting on the valid product. However, we can not judge applicability about behavior of the agent. Therefore, we discuss the possibility to realize applicable behavior by the weight parameter change and the algorithm change. Therefore, we change a parameter.

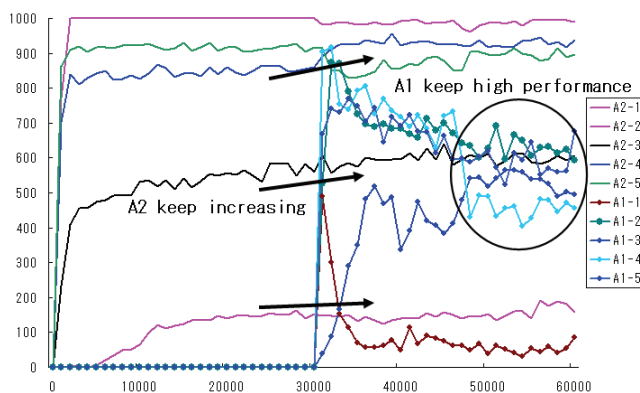


Fig.12 The curve of five layers and Result of weights are 92%, 2%, 2%, 2%, 2%

This system uses the Q value of the all layers. But layer weights, which except current layer, made low and even weight for experiment. On those weights, learning can make robust-ness to the overlay knowledge. So, a flexible action is led from the overlay. The learning uses the overlay knowledge of **five layers**. The weight of a current layer is 92%. The

weights of other 2nd-5th layers are 2%. In fig.12, phenomena of curve show the flexible behavior successfully.

VIII. CONCLUSION

We examined the characteristic of overlay knowledge from the experiment from applying a fire panic problem. All agents have to follow features ①~④ in the fire panic. We examined the nature of the featuring learning principle about the heterogeneous multi-agent which followed features ⑤~⑥. About features ①~⑥, see. II. LEARNING THEORY OF MULTI-AGENT SYSTEM. As a result, it was proved that the overlay knowledge was useful sufficiently for the heterogeneous multi-agent. Also, we have showed the robust characteristic which depends on layer in overlay knowledge. Then, the robust characteristic was the cause which outputs absolute behavior. By the change of the robust levels, behavior of the agent expected a change into the flexible movement. The behavior which is passive about the reward is output. However, the overlay knowledge which was taken by the experimental result and we must analyze it for the rote. This is future work.

REFERENCES

- [1] C.J.C.H. Watkins, P. Dayan : "Technical Note: Q-Learning", Machine Learning, Vol.8, No.3, pp.279-292(1992)
- [2] R.S. Sutton : Integrated architectures for learning, planing, and reacting based on approximate dynamic programming: Proceedings of the 7th International Conference on Machine Learning, pp.216-224 (1990)
- [3] R.S. Sutton and A.G. Barto : Reinforcement Learning: The MIT Press (1998)
- [4] Y. Hoshino and K. Kamei: A proposal of Reinforcement Learning System to Use Knowledge effectively: Proceedings of SICE Annual Conference 2003 in Fukui, Fukui University, Fukui, Japan, CD-ROM Version No.0538 TAI-14-2(4 Pages) (2003)
- [5] Y. Hoshino, A. Sakakura, and K. Kamei: Research of Foresight Knowledge by CMAC based Q-learning in Heterogeneous Multi-Agent System : *Proceedings of 4th International Symposium on Advanced Intelligent System*, pp.280-283 (2003)
- [6] Yukinobu HOSHINO, Akira SAKAKURA and Katsuari KAMEI: A proposal of CMAC with Foresight Knowledge Layer For Heterogeneous Multi-agent, *Proceedings of International Conference on Fuzzy Systems 2004*, Budapest, Hungary, CD-ROM Version 0047-1382.pdf, pp.449-452 (2004)
- [7] J.S. Albus. A new approach to manipulator control: The cerebellar model articulation controller (cmac). *Transactions of the ASME, Journal of Dynamics Systems Measurement and Control*, Vol. 97, pp.220-227, 1975.
- [8] Albus, J.S., Mechanisms of planning and problem solving in the brain (1979) *Mathematical Biosciences*, 45 (3-4), pp. 247-293.
- [9] Gerhard Weiss : Multi agent Systems, The MIT Press
- [10] Yukinobu HOSHINO, Akira SAKAKURA and Katsuari KAMEI: A proposal of the learning system using the recordable multi-layer type rule base and its application for the Fire panic problem, *Proceedings of the Joint International Conference on Cyber Games and Interactive Entertainment 2006 (CGIE2006)*, pp.137-140
- [11] E. Uchibe, and K. Doya.: Competitive-Cooperative-Concurrent Reinforcement Learning with Importance Sampling, In *Proc. of International Conference on Simulation of Adaptive Behavior: From Animals and Animats*, pages 287-296, 2004