A Method of Role Differentiation Using a State Space Filter with a Waveform Changing Parameter in Multi-agent Reinforcement Learning

Masato Nagayoshi, Simon Elderton
Niigata College of Nursing, 240 shinnan-cho
Joetsu, Niigata 943-0147, Japan
E-mail: nagayosi@niigata-cn.ac.jp, elderton@niigata-cn.ac.jp

Hisashi Tamaki
Kobe University, 1-1 Rokkodai-cho, Nada-ku,
Kobe, Hyogo 657-8501, Japan
E-mail: tamaki@al.cs.kobe-u.ac.jp

Abstract

Recently, there have been many studies on the multi-agent reinforcement learning (MARL) in which each autonomous agent obtains its own control rule by RL. Here, it is considered that different agents having individuality is more effective than uniform agents in terms of role differentiation in MARL. Then, we have proposed a promoting method of role differentiation using a waveform changing parameter in MARL. In this paper, we confirm the effectiveness of role differentiation by introducing the waveform changing parameter into a state space filter through computational experiments using "Pursuit Game" as one of multi-agent tasks.

Keywords: reinforcement learning, role differentiation, meta-parameter, waveform changing, state space filter

1. Introduction

Engineers and researchers are paying more attention to reinforcement learning (RL) [1] as a key technique for realizing computational intelligence such as adaptive and autonomous decentralized systems. Recently, there have been many studies on multi-agent reinforcement learning (MARL) in which each autonomous agent obtains its own control rule by RL. Then, we hypothesize that different agents having individuality is more effective than uniform agents in terms of role differentiation in MARL. Here, we define "individuality" in this paper as being able to be externally observed, but not a difference that we are incapable observing, such as a difference of internal construction.

We consider that differences in interpretations of experiences in the early stages of learning have a great effect on the creation of individuality of autonomous agents. In order to produce differences in interpretations of the agents' experiences, we utilized Beck's "Cognitive distortions" [2], which is a cognitive therapy.

Then we have proposed a "fluctuation parameter" which is a wave-form changing meta-parameter in order to realize "Disqualifying the positive" * which is one of the "Cognitive distortions", and a promoting method of role differentiation using the fluctuation parameter in MARL [3].

In this paper, we introduce the "fluctuation parameter" into a state space filter [4] in order to realize "Overgeneralizing" † which is one of the "Cognitive distortions", and confirm the effectiveness of role differentiation by introducing the fluctuation parameter into the state space filter through computational experiments using "Pursuit Game" as one of multi-agent tasks.

2. Q-learning

In this section, we introduce Q-learning (QL) [5] which is one of the most popular RL methods. QL works by calculating the quality of a state-action combination, namely the Q-value, that gives the expected utility of performing a given action in a given state. By performing an action $a \in \boldsymbol{A}_{\mathrm{Q}}$, where $\boldsymbol{A}_{\mathrm{Q}} \subset \boldsymbol{A}$ is the set of available actions in QL and $\boldsymbol{A}$ is the action space of the RL agent, the agent can move from state to state. Each state provides the agent with a reward r. The goal of the agent is to maximize its total reward.

The Q-value is updated according to the following formula, when the agent is provided with the reward:

$$Q\big(s(t-1), a(t-1)\big) \leftarrow Q\big(s(t-1), a(t-1)\big) + \alpha_{\mathrm{Q}}\{r(t-1) + \gamma \max_{b \in \boldsymbol{A}_{\mathrm{Q}}} Q(s(t), b) - Q(s(t-1), a(t-1))\} \quad (2)$$

where $Q\big(s(t-1), a(t-1)\big)$ is the Q-value for the state and the action at the time step $t-1$, $\alpha_{\mathrm{Q}} \in [0,1]$ is the learning rate of QL, $\gamma \in [0,1]$ is the discount factor.

The agent selects an action according to the stochastic policy $\pi(a|s)$, which is based on the Q-value. $\pi(a|s)$ specifies the probabilities of taking each action a in each state s. Boltzmann selection, which is one of the typical action selection methods, is used in this research. Therefore, the policy $\pi(a|s)$ is calculated as

$$\pi(a|s) = \frac{\exp\left(Q(s,a)/\tau\right)}{\sum_{b \in A_{\mathrm{Q}}} \exp\left(Q(s,b)/\tau\right)} \quad (3)$$

where τ is a positive parameter labeled temperature.

3. Reinforcement Learning with a State Space Filter

We have proposed a state space filter based on the entropy which is defined by action selection probability distributions in a state[5].

The entropy of action selection probability distributions using Boltzmann selection in a state $H(s)$ is defined by

$$H(s) = -(1/\log|\boldsymbol{A}|) \sum_{a \in \boldsymbol{A}} \pi(a|s) \log \pi(a|s) \quad (2)$$

where $\pi(a|s)$ specifies probabilities of taking each action a in each state s, $\boldsymbol{A}$ is the action space and $|\boldsymbol{A}|$ is the number of available actions.

The state space filter is adjusted by treating this entropy $H(s)$ as an index of a correctness of state aggregation in the state s. In particular, in case of mapping from the inner state space roughly digitized to the inner state space, a perceptual aliasing problem is happened. That is, the action which an agent should select cannot be identified clearly. Thus, the entropy may not be small in the state space should be divided. In this paper, sufficiency of the number of learning opportunities is judged using a threshold value θ_{L}.

Therefore, if the entropy does not get smaller than a threshold value θ_{H} despite the number of learning opportunities is sufficient, the state space filter is adjusted by dividing the state due to that the perceptual aliasing problem is happened.

Similarly, if the entropy is smaller than θ_{H} in a state s and a different state mapping from a transited input state s', and representative actions in each other's states are same, the state space filter is adjusted by integrating the states due to that the states is too divided.

4. Fluctuation Parameter

RL has meta-parameters κ to determine how RL agents learn control rules. The meta-parameters κ include the learning rate α, the discount factor β, ε of ε-greedy which is one of the action selection methods, and the temperature τ of Boltzmann action selection method.

In this paper, the following fluctuation parameter using damped vibration function is introduced into this κ.

$$\kappa(t_{\mathrm{p}}) = \begin{cases} \kappa + A\cos(2\pi(t_{\mathrm{p}}/\lambda) + \phi) & (t_{\mathrm{pa}} < t_{\mathrm{ps}}) \\ \kappa + A\cos(2\pi(t_{\mathrm{p}}/\lambda) + \phi) \times t_{\mathrm{ps}}/t_{\mathrm{pa}} & (otherwise) \end{cases} \quad (4)$$

where $A, t_{\mathrm{p}}, t_{\mathrm{pa}}, t_{\mathrm{ps}}, \lambda$ and ϕ is the amplitude, the phase, the damped phase, the initial phase of damping, the wavelength, and the initial phase parameter of the fluctuation, respectively. The phase t_{p}, the damped phase t_{pa}, the initial phase of damping t_{ps}, and the wavelength λ are needed to set proper units.

5. Computational Experiments

5.1. *Pursuit Game*

The effectiveness of the proposed approach is investigated in this section. It is applied to the so-called

Table 1. Parameters for Q-learning with a state space filter

Parameter	Value
α_{Q}	0.1
γ	0.9
τ	0.1
θ_{H}	0.3
θ_{L}	1,000

"Pursuit Game" where three RL agents move to capture a randomly moving target object in a discrete 10×10 globular grid space. Two or more agents or an agent and the target object cannot be located at the same cell. At each step, all agents simultaneously take one of the 5 possible actions: moving north, south, east, west or standing still. A target object is captured when all agents are located in cells adjacent to the target object and surrounding the target object in three directions.

The agent has a field of view, and the depth of view set at 3. Therefore, the agent can observe the surrounding $(3 \times 2 + 1)^2 - 1$cells. The agent determines the state by information within the field of view.

The positive reinforcement signal $r_t = 10$ (reward) is given to all agents only when the target object is captured, and the positive reinforcement signal $r_t = 1$ (sub reward) is given to the agent only when the agent is located in the cell adjacent to the target object and the reinforcement signal $r_t = 0$ at any other steps. The period from when all agents and the target object are randomly located at the start point to when the target object is captured and all agents are given a reward, or when 100, 000 steps have passed is labeled 1 episode. The period is then repeated.

5.2. *RL Agents*

All agents observe the only target object in order to confirm the effectiveness of role differentiation, e.g. moving east of the target object. Therefore, the state space is constructed with a 1 dimensional space.

Computational experiments have been done with parameters as shown in Table 1. In addition, all initial Q-values are set at 5.0 as the optimistic initial values, and θ_{H} was set referring to about 0.288: the maximal value of the entropy when the highest selection probability for one action is 0.9,

5.3. *Experiment (A): Learning Rate*

The effectiveness of role differentiation by introducing 3 the fluctuation parameters, in which the initial phase $\phi = 0$, the amplitude $A = 0.09$, and the wavelength $\lambda = \{50, 100, 500\}$ [episode], into the learning rate of QL with the state space filter (hereafter called "50", "100", and "500", respectively) are investigated in comparison with an ordinary QL with the state space filter without fluctuation parameter (hereafter called "constant"). Here, the fluctuation parameters of all agents take the same value. The unit of the phase $t_{\mathrm{p}}$is set [episode] which is the same as the wavelength λ, the unit of the damped phase $t_{\mathrm{pa}}$is set [episode], and the initial phase of damping is set at $t_{\mathrm{ps}} = 250$ [episode]. The range of values which the fluctuation parameter for $\alpha_{\mathrm{Q}} = 0.1$ can take e.g. [0.01, 0.19] on the condition of $A = 0.09$.

The average numbers of steps and the average size of the state space required to capture the target object were observed during learning over 20 simulations with various wavelength parameters in the learning rate, as described in Figs. 3 and 4, respectively.

It can be seen from Figs. 3, 4 that, (1) "50", "100" and "500" show a better performance than "constant" with regard to the obtained control rule, (2) "50", "100" and "500" are smaller than "constant" with regard to the size of the state space.

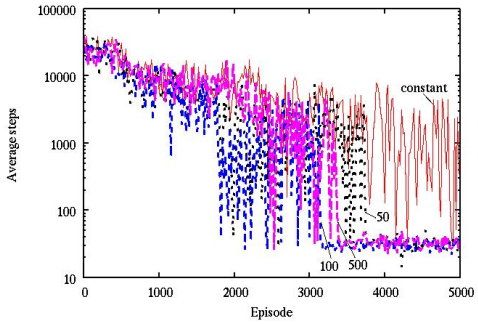

Fig. 3. Required steps of various wavelength parameters in the learning rate (ϕ = 0 [rad]).

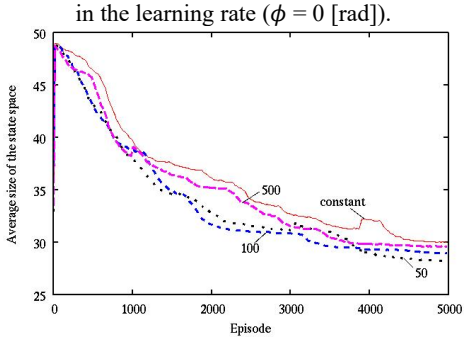

Fig. 4. Required size of the state space filter of various wavelength parameters in the learning rate (ϕ = 0 [rad]).

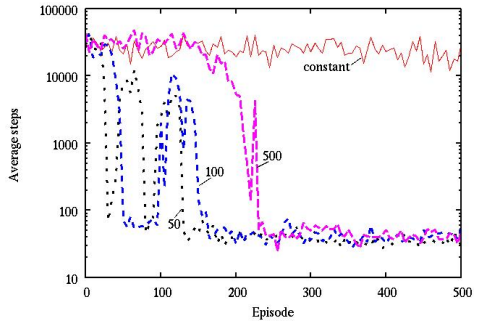

Fig. 5. Required steps of various wavelength parameters in the temperature (ϕ = 0 [rad]).

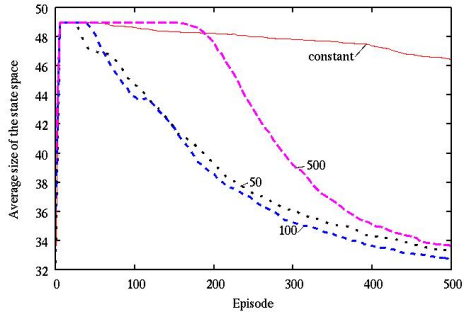

Fig. 6. Required size of the state space filter of various wavelength parameters in the temperature (ϕ = 0 [rad]).

5.4. *Example (B): Temperature*

The effectiveness of role differentiation by introducing 3 the fluctuation parameters, in which the initial phase $\phi = 0$, the amplitude $A = 0.09$, and the wavelength $\lambda = \{50, 100, 500\}$ [episode] , into the temperature of QL with the state space filter (hereafter called "50", "100", and "500", respectively) are investigated in comparison with an ordinary QL with the state space fliter without fluctuation parameter (hereafter called "constant"). Here, the fluctuation parameters of all agents take the same value. The unit of the phase $t_{\mathrm{p}}$is set [episode] which is the same as the wavelength λ, the unit of the damped phase $t_{\mathrm{pa}}$is set [episode], and the initial phase of damping is set at $t_{\mathrm{ps}} = 250$ [episode]. The range of values which the fluctuation parameter for $\tau = 0.1$ can take e.g. [0.01, 0.19] on the condition of $A = 0.09$. If the temperature is zero, then action selection of the agent is greedy and the situations where agents cannot capture the target object occur. Therefore, A is set at 0.09

The average numbers of steps and the average size of the state space required to capture the target object were observed during learning over 20 simulations with various wavelength parameters in the temperature, as described in Figs. 5 and 6, respectively.

It can be seen from Figs. 5, 6 that, (1) "50", "100" and "500" show a better performance than "constant" with regard to the obtained control rule, (2) "50", "100" and "500" are smaller than "constant" with regard to the size of the state space.

Thus, it could be confirmed that the effectiveness of role differentiation by introducing the fluctuation parameter into the state space filter. It could be considered that this is the result of overgeneralizing.

6. Conclusion

In this paper, we introduced a "fluctuation parameter" into a state space filter in order to realize "Overgeneralizing" which is one of the "Cognitive distortions". Through computational experiments, we confirmed the effectiveness of role differentiation by introducing the fluctuation parameter into the state space filter. It could be considered that this is the result of overgeneralizing.

Our future projects include to apply real world problems, e.g. improving a schedule for each nurse.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP19K04906.

References

1. R. S. Sutton and A. G. Barto, Reinforcement Learning (A Bradford Book, MIT Press, Cambridge, 1998).
2. A. T. Beck, Cognitive Therapy and the Emotional Disorders (International University Press, New York, 1976).
3. M. Nagayoshi, S. J. H. Elderton and H. Tamaki, A Promoting Method of Role Differentiation using a Learning Rate that has a Periodically Negative Value in Multi-agent Reinforcement Learning, *Journal of Robotics, Networking and Artificial Life*, 6(4), 2020, pp.221–224.
4. M. Nagayoshi, H. Murao and H. Tamaki, A State Space Filter for Reinforcement Learning, *Proc. of AROB 11th'06,* 2006, pp.615-618.
5. C. J. C. H. Watkins and P. Dayan, Technical note: Q-Learning, *Machine Learning* **8** (1992), 279-292.