Journal of Robotics, Networking and Artificial Life
Vol. 6(4); March (2020), pp. 221–224
DOI: https://doi.org/10.2991/jrnal.k.200222.003; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

A Promoting Method of Role Differentiation using a Learning Rate that has a Periodically Negative Value in Multi-agent Reinforcement Learning

Masato Nagayoshi[1,*], Simon J. H. Elderton[1], Hisashi Tamaki[2]

[1]Department of Nursing, Niigata College of Nursing, 240 Shinnan-cho, Joetsu, Niigata 943-0147, Japan
[2]Department of Computer Science and System Engineering, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan

ARTICLE INFO

Article History
Received 31 October 2019
Accepted 16 December 2019

Keywords
Reinforcement learning
multi-agent
negative learning rate
role differentiation

ABSTRACT

There have been many studies on Multi-Agent Reinforcement Learning (MARL) in which each autonomous agent obtains its own control rule by Reinforcement Learning (RL). Here, we hypothesize that different agents having individuality is more effective than uniform agents in terms of role differentiation in MARL. In this paper, we propose a promoting method of role differentiation using a wave-form changing parameter in MARL. Then we confirm the effectiveness of role differentiation by the learning rate that has a periodically negative value through computational experiments.

© 2020 *The Authors*. Published by Atlantis Press SARL.
This is an open access article distributed under the CC BY-NC 4.0 license (http://creativecommons.org/licenses/by-nc/4.0/).

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.

In this paper, we propose 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.

We then confirm the effectiveness of role differentiation by introducing the fluctuation parameter into the learning rate, especially having a periodically negative value, through computational experiments using "Pursuit Game" as one of the multi-agent tasks.

*Corresponding author. Email: nagayosi@niigata-cn.ac.jp

2. Q-LEARNING

In this section, we introduce Q-learning (QL) [3] 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 $a \in \mathbf{A}_Q$, where $\mathbf{A}_Q \subset \mathbf{A}$ is the set of available actions in QL and $\mathbf{A}$ 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 Equation (1), when the agent is provided with the reward:

$$\begin{aligned} &Q(s(t-1), a(t-1)) \leftarrow Q(s(t-1), a(t-1)) \\ &+ \alpha_Q \left\{ r(t-1) + \gamma \max_{b \in \mathbf{A}_Q} Q(s(t), b) - Q(s(t-1), a(t-1)) \right\} \end{aligned} \tag{1}$$

where $Q(s(t-1), a(t-1))$ is the Q-value for the state and the action at the time step $t-1$, $\alpha_Q \in [0,1]$ 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

$$\pi(a|s) = \frac{\exp\left(\frac{Q(s,a)}{\tau}\right)}{\sum_{b \in \mathbf{A}_Q} \exp\left(\frac{Q(s,b)}{\tau}\right)} \tag{2}$$

where τ is a positive parameter labeled temperature.

3. FLUCTUATION PARAMETER

Reinforcement learning 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:

$$\kappa(t_p) = \begin{cases} \kappa + A\cos\left(2\pi\left(\frac{t_p}{\lambda}\right) + \phi\right) & (t_{pa} < t_{ps}) \\ \kappa + A\cos\left(2\pi\left(\frac{t_p}{\lambda}\right) + \phi\right) \times \frac{t_{ps}}{t_{pa}} & (\text{otherwise}) \end{cases} \quad (3)$$

where A, t_p, t_{pa}, t_{ps}, λ 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.

4. COMPUTATIONAL EXAMPLES

4.1. Pursuit Game

The effectiveness of the proposed approach is investigated in this section. It is applied to the so-called "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 five 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 as shown in Figure 1.

The agent has a field of view, and the depth of view set at 3 as shown in Figure 2. 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

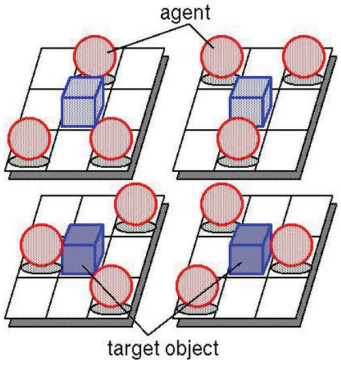

Figure 1 | Capture positions.

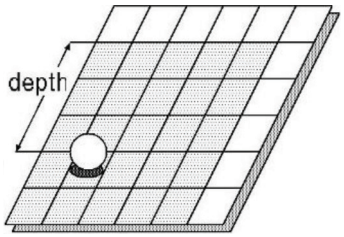

Figure 2 | Range of the agent view (indicated by grayed area).

Table 1 | Parameters for Q-learning

Parameter	Value
α_Q	0.1
γ	0.9
τ	0.1

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.

4.2. 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 one-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.

4.3. Example (A): Same Amplitude

The effectiveness of role differentiation by introducing four fluctuation parameters, in which the initial phase $\phi = 0$, $\lambda = 500$ [step], and the amplitude $A = \{0.1, 0.12, 0.15, 0.17\}$, into the learning rate of QL (hereafter called "0.1", "0.12", "0.15", and "0.17", respectively) are investigated in comparison with an ordinary QL without fluctuation parameter (hereafter called "constant"). Here, the fluctuation parameters of all agents take the same value. The unit of the phase t_p is set [step] which is the same as the wavelength λ, the unit of the damped phase t_{pa} is set [episode], and the initial phase of damping is set at $t_{ps} = 1000$ [episode]. The range of values which the fluctuation parameter for $\alpha_Q = 0.9$ can take e.g. [0.0, 0.2], [−0.05, 0.25] on the condition of $A = 0.1$ and 0.15, respectively. If the unit of the phase t_p is set [episode] and the learning rate is negative, then the control rule which the agent obtains is random. This is the result of negative learning at each step. Therefore, the unit of the phase t_p is set [step].

The average numbers of steps required to capture the target object were observed during learning over 20 simulations with various amplitude parameters in the learning rate, as described in Figure 3.

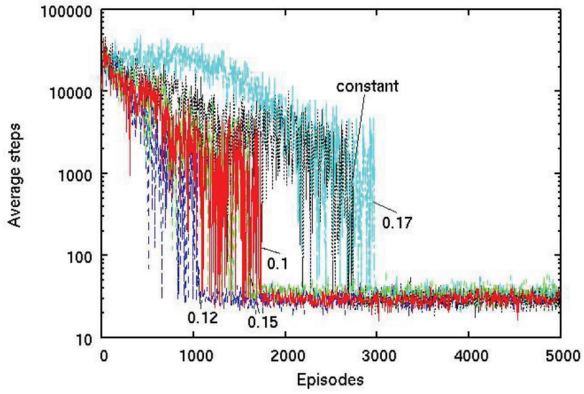

Figure 3 | Required steps of various amplitude parameters in the learning rate.

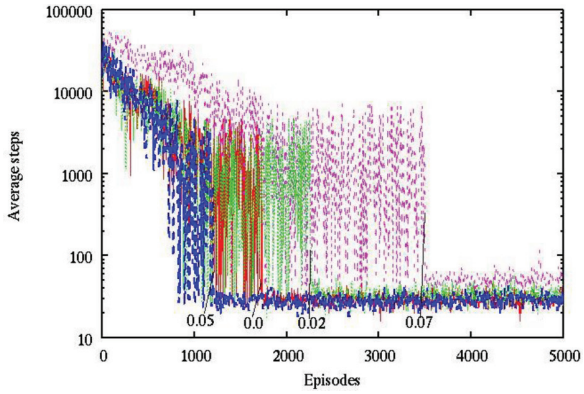

Figure 4 | Required steps of various shifting size of amplitude in the learning rate.

It can be seen from Figure 3 that, (1) "0.12" shows a better performance than any other methods and with regard to promoting role differentiation. (2) "0.17" shows a worse performance than "constant" with regard to promoting role differentiation.

Thus, the learning rate that has a periodically a negative value only slightly shows a better performance than the learning rate that has any non-negative value. It could be considered that this is the result of preventing over-fitting by having periodically a negative value.

4.4. Example (B): Various Amplitude

In this section, the effectiveness of role differentiation by introducing four fluctuation parameters, in which the initial phase $\phi = 0$, $\lambda = 500$ [step], and the shifting size of the amplitude: ±0.0, ±0.02, ±0.05, ±0.07 around $A = 0.1$, into the learning rate of QL (hereafter called "0.0", "0.02", "0.05", and "0.07", respectively) are investigated. For example, the amplitudes of three agents are 0.05, 0.1, 0.15 on the condition of the shifting size of the amplitude: ±0.05 around $A = 0.1$. The same as Example (A), the unit of the phase t_p is set [step] which is same as the wavelength λ, the unit of the damped phase t_{pa} is set [episode], and the initial phase of damping is set at $t_{ps} = 1000$ [episode].

The average numbers of steps required to capture the target object were observed during learning over 20 simulations with various shifting size of amplitude in the learning rate, as described in Figure 4.

It can be seen from Figure 4 that, (1) "0.05" shows a better performance than any other methods with regard to promoting role differentiation. (2) "0.02" shows a worse performance than "0.1". (3) "0.07" shows a worse performance than any other method.

Thus, the moderate shifting size of amplitude in the learning rate among the agents shows a better performance than the in the case of same amplitude in the learning rate among the agent.

5. CONCLUSION

In this paper, we 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.

Through computational experiment using the "Pursuit Game", we confirmed the effectiveness of role differentiation by introducing the fluctuation parameter into the learning rate, especially having a periodically negative value.

Our future projects include evaluating the effectiveness of promoting role differentiation using our proposed fluctuation parameter in order to realize "Jumping to conclusions", "Making "must" or "should" statements", and "Overgeneralizing" with a state space filter [4].

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP19K04906.

REFERENCES

[1] R.S. Sutton, A.G. Barto, Reinforcement learning: An introduction, A Bradford Book, MIT Press, Cambridge, 1998.
[2] A.T. Beck, Cognitive therapy and the emotional disorders, International University Press, New York, 1976.
[3] C.J.C.H. Watkins, P. Dayan, Technical note: Q-learning, Mach. Learn. 8 (1992), 279–292.
[4] M. Nagayoshi, H. Murao, H. Tamaki, A state space filter for reinforcement learning in POMIDPs: application to a continuous state space, 2006 SICE-ICASE International Joint Conference, Busan, South Korea, IEEE, 2006, pp. 6037–6042.

AUTHORS INTRODUCTION

Dr. Masato Nagayoshi

He is an Associate Professor of Niigata College of Nursing. He graduated from Kobe University in 2002, and received Master of Engineering from Kobe University in 2004 and Doctor of Engineering from Kobe University in 2007. He is a IEEJ, SICE, ISCIE member.

Dr. Hisashi Tamaki

He is a Professor of Graduate School of Engineering, Kobe University. He graduated from Kyoto University in 1985, and received Master of Engineering from Kyoto University in 1987 and Doctor of Engineering from Kyoto University in 1993. He is a ISCIE, IEEJ, SICE, ISIJ member.

Mr. Simon J. H. Elderton

He is an Associate Professor of Niigata College of Nursing. He graduated from University of Auckland with an Honours Masters degree in Teaching English to Speakers of Other Languages in 2010. He is a JALT, Jpn. Soc. Genet. Nurs., JACC member.