Effect of action selection on the emergence of one-way communication using Q-learning

Masanobu Nakanishi, Katsunari Shibata
Dept. of Electrical and Electronic Engineering, Oita University
700 Dannoharu Oita 870-1192 Japan

Abstract

In this paper, the effect of action selection in the learning of one-way communication between two agents using Q-learning is examined. The ratio of successful learning becomes larger when the receiver agent's action selection is greedy and the transmitter agent's action selection is not completely greedy but with a small random factor. From the analysis of the learning process, it is known that inappropriate mapping from states to signals in the transmitter agent sometimes breaks the mapping from signals to action severely in the receiver agents. Accordingly, the transmitter agent needs to find an appropriate mapping through exploration, while the receiver agent decides its action after the mappings is fixed in the transmitter. Accordingly, no exploration is necessary in the receiver agent.

1. Introduction

Communication plays a very important role in the supplement of insufficient observation, collision avoidance and cooperative action in multi-robot and multi-agent systems. In order to learn a purposive communication autonomously, evolutionary method[1] or reinforcement learning[2][3] has been used. Autonomous acquisition of one-way communication to supply the receiver's insufficient observation has been examined[1][2]. However, it was not examined what kind of information should be transmitted, and whether the optimal communication can be acquired in any cases. Then, we also focused on the learning of one-way communication that supplies the receiver's insufficient observation. For simple analysis, the number of agents is limited to two, those are a transmitter agent and a receiver agent. We have examined the reason why state confusion occurred in some simulations[4]. In this paper, the effect of action selection in the learning of one way communication between two agents using Q-learning that the authors discovered through the simulations is examined.

2. Learning of one-way communication

2-1 Task description

In this paper, the learning of one-way communication that supplies the receiver's insufficient observation is focused on. The simulation environment was decided referring to [1][2]. Fig.1 shows the image of one-way communication learning. Two agents called "male" and "female" are assumed in a discrete environment. The male can move, but does not have sight. On the other hand, the female can't move, but can transmit some signals to the male. The female's input is the relative position of the male, and its output is a communication signal. The male's input is the communication signal and its output is a physical action. If the male touches the female, a reward is given to the both agents. The meaning of the communication signal is not given to either agent at all beforehand. Therefore, a transmitter agent has to learn what communication signal should be transmitted and a receiver agent have to learn to generate appropriate actions from the signal. If some common language can be built up between the male and the female, the contact can be repeated efficiently.

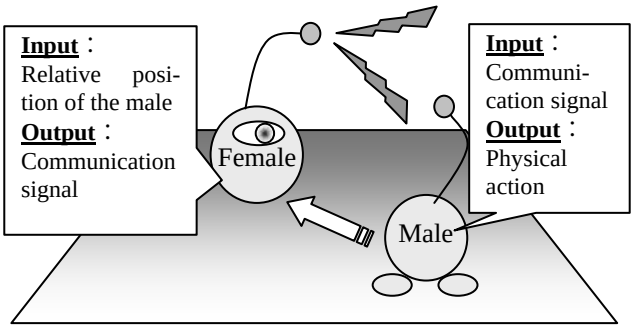


Fig.1 Learning of one-way communication

2-2 Learning method for the both agents

For the learning of the both agents, Q-learning is used. In Q-learning, state-action pairs are evaluated, and the action value is called Q-value. An agent chooses an action with the probability calculated from the Q-values. It is usually applied on a discrete action space.

The algorithm of Q learning is as follows.

(1) The agent observes a state.
(2) The agent selects and executes an action.
(3) The agent observes the state after the transition.
(4) A reward r_{t+1} is received from the environment.
(5) Q-value is modified as

$$Q(s_{t,}a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1},a)\right] \quad (1)$$

where α is a learning rate $(0 < \alpha \leq 1)$
γ is a discount factor $(0 \leq \gamma < 1)$

(6) t -> $t+1$, and the flow returns to the step(2).

For the female, the state s is the male's relative position, and the action a is the communication signal. For the male, the state s is the communication signal, and the action a results in a state transition. α is 0.1, and γ is 0.9 here.

2-3 Action selection

2-3-1 Boltzman selection

When the state is x, the probability of the action a is calculated as,

$$p(a \mid x) = \frac{\exp(Q(x,a)/T)}{\sum_{i \in A} \exp(Q(x,i)/T)} \qquad (2)$$

where A is a set of actions, and T is a temperature coefficient. An action is selected almost randomly when T is large. As opposite to it, when T is close to 0, a little difference of Q-value has a great influence on the action selection, in other words, the action selection is almost greedy. In the following simulations, the initial value of T is 1.0, and it is gradually decreased exponentially to 0.005. In the rest of trials, it was fixed at 0.005. The reason is that when it becomes smaller than 0.005, the computation on our computer becomes impossible.

2-3-2 Greedy selection

In greedy selection, there is no probabilistic factor and the action with the maximum Q-value is always selected.

2-4 Flow of the learning

The agents act in accordance with the following cycle.

(1) The female detects the male's state.
(2) The female's Q-value at t-1 is modified.
(3) The female transmits a signal to the male.
(4) The male receives the female's signal.
(5) The male's Q-value at t-1 is modified.
(6) The male makes an action.
(7) If the both agents touch each other, the trial finishes, and they get a reward. In that case, they learn their Q-values at t according to Eq.(1) with $\max Q(s_{t+1}, a) = 0$ and the flow returns to (1). If the trial finished, t=0, otherwise t -> t +1.

When t =0, the step(2)or(5) is not executed.

3. Simulation

A simulation environment is shown in Fig.1. In this environment, the number of states, signals and actions are decided to be the same to match the condition of the both agents. All the initial Q-values are 1.0 here. When all the initial Q-values are set to be high, which is called Optimistic initial value, the effect of exploration can be realized even in greedy selection[5].

In this simulation, the number of the trials until the temperature coefficient reaches the minimum is varied for each agent in the case of Boltzmann selection, and successful learning ratio was observed.

When the temperature coefficient T reaches the minimum value at the N-th trial, the temperature at the k-th trial is calculated as

$$T(k) = 0.005^{\frac{k}{N}} \quad if\ (k < N) \qquad (3)$$
$$= 0.005 \quad otherwise$$

N is varied from 200 to 1000 with the interval of 200 in each agent. The total number of trials is 1000. Furthermore, the greedy selection is also employed. The average successful ratio over 1000 simulation runs for each combination of exploration ways of two agents is shown in Table 1.

From this figure, it is known that the successful ratio is high when the male's action selection is greedy, and the female's one is Boltzmann selection with the temperature decreased fast. On the other hands, the successful ratio is low when the female's action selection is greedy.

Next, in order to examine the effect of the small exploration factor remaining due to the lower bound of the temperature, greedy selection was employed when the temperature reaches the minimum value 0.005. In the previous simulation, the action which does not have the maximum Q-value sometimes selected because of the probabilistic selection after the temperature coefficient reaches the minimum. On the other hand, in this simulation, the action that has the maximum Q-value is always selected after the temperature coefficient reaches the minimum.

The results are shown in Table 2. The successful ratio is higher than in Table 1 when the male's action selection is greedy, and the female's selection is Boltzmann selection. However, when the female's action selection is greedy, and the male's action selection is Boltzmann selection, the successful ratio is lower than in Table 1. The reason why the successful ratio is high when the male temperature coefficient is decreased faster than the female can be thought that the male's action selection becomes greedy in the early stage of the total trials.

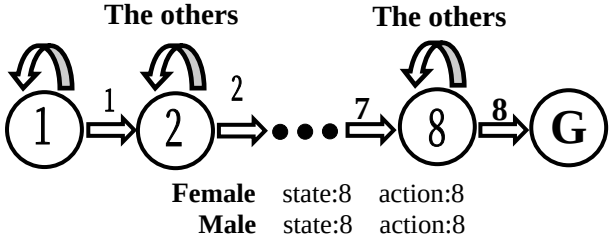


Fig.2 Simulation environment

Table 1: Success ratio according to the number of trials when the temperature reached 0.005

The female's number of trials until T=1000	The male's number of trials until T=0.005					
	greedy	200	400	600	800	1000
Greedy	87.1	44.7	38.7	30.2	19.8	9.8
200	100.0	82.0	77.0	61.1	32.2	7.6
400	99.9	80.1	78.8	61.0	36.0	9.9
600	99.1	64.8	62.4	57.4	34.9	9.0
800	81.9	41.0	36.4	32.1	21.8	7.6
1000	62.8	6.1	6.9	5.6	2.7	0.4

Table 2: Success ratio according to the number of trials when the action was changed to greedy selection after the temperature reached 0.005.

The female's number of trials until T=1000	The male's number of trials until T=0.005					
	greedy	200	400	600	800	1000
greedy	87.1	84.0	67.2	40.5	20.3	9.8
200	90.5	86.9	75.9	47.9	22.4	10.4
400	90.5	90.9	83.4	65.2	28.7	11.9
600	90.4	87.9	85.9	77.8	53.0	14.4
800	83.8	84.4	85.8	79.3	54.5	18.6
1000	62.8	61.5	56.8	36.7	6.8	0.4

Then, in order to investigate the reason of the difference of successful ratio depends on the action selection of the both agents, the change of Q-values in the learning process is observed. The change of Q-values when the female's action selection is greedy from the beginning is shown in Fig.3. The change when Boltzmann selection was employed and the temperature coefficient was fixed at 0.005 from the beginning is shown in Fig.4. Fig.(a) shows the change of female's Q-values when the male is in the state 8 that is located 1 step before the goal, and Fig.(b) shows the change of male's Q-values for the action 8 that takes the agent from the state 8 to the goal, but is not rational for the other states.

From Fig.3,4, it is known that only one Q-value is higher after learning progressed in some degree, and the signal with the highest Q-value in the female corresponds to the signals with the highest Q-value in the male. It is also known that the signal with the maximum Q-value was switched several times. Finally the female decides to select a signal in the state 8, and the male decides to select the proper action from the signal. The both corresponding Q-values converged to 1.0. In Fig.3, the male's maximum Q-value became large in the early stage of learning, but it decreased drastically. On the other hand, in Fig.4, the male's Q-learning is slightly oscillating around a comparatively small value. It can be said that, in greedy selection, reconstruction of the signal-action pair occurs often and drastically, but in the Boltzmann selection, the learning is more stable.

Then, the reason why the difference of successful ratio and the change of Q-value due to the female's exploration factor in action selection is examined. An example to show the influence of one agent's learning to the other agent's learning is shown in Fig.5. The failure of the female's learning means that the female selects the same signal in two or more states. For example, when the female transmitted the signal 1 in the both state 1 and 8, the male continues to select the action 8 as long as the signal 1 is received and the action for the signal 1 is not switched by the greedy selection of the both agents. Accordingly, signal-action mapping is broken not only in the state 1 but in the state 8, and learning has to be done again from scratch.

On the other hand, in the case of the failure of the male's learning, the male selects the same action for the different signals even though the female assigned signals appropriately. For example, although the female transmits the signal 1 at the state 1 and transmits the signal 2 at the state 8, the male selects the action 8 in either case of the signal 1 or the signal 2. Then, for the male, Q-value of the action 8 on the signal 1 decreases, and for the female, Q-value of the signal 1 on the state 1 also decreases. However, in this case, it gives no influence on the Q-values at the state 8. So, insufficient learning of the female breaks the generation of the right action also in other states in the male, while insufficient learning of the male breaks the signal only at the state in the female. That is supposed to be the reason why the successful ratio becomes larger when the male's action selection is greedy, and the female's action selection is not completely greedy, but with a small random factor.

Finally, in order to investigate how the learning progressed in the both agents, the number of steps since which the action with the maximum Q-value was not switched was observed. The result shows that the action with the maximum Q-value began to be fixed from the state that is close to the goal, and then, the range where the action was fixed spreads to the far states from the goal gradually. Furthermore, the time when the female fixed its signal is slightly earlier than the number of steps when the male fixed its action. That must be also because the female's insufficient learning sometimes breaks the male's proper mapping from the signal space to the action space.

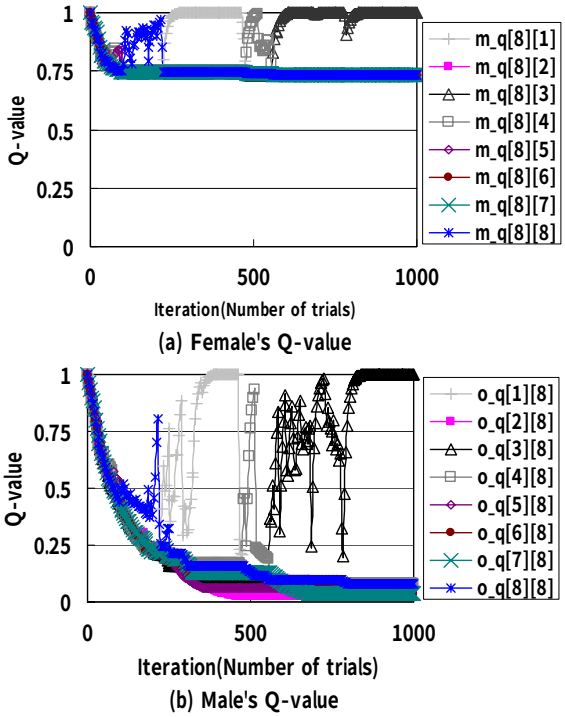


Fig.3 Change of the Q-values when the female's action selection in greedy selection

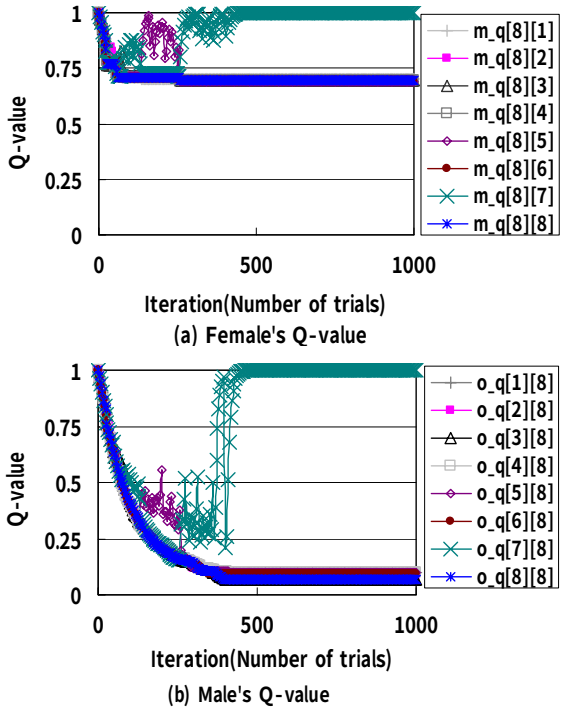
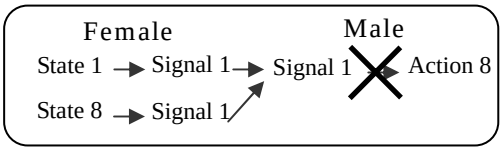
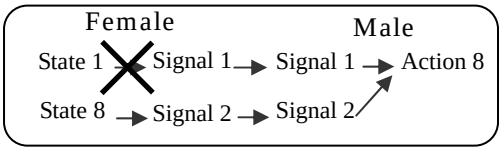


Fig.4 Change of the Q-values when the female's action selection is Boltzmann selection



(a) When the mapping in the female is not appropriate



(b) When the mapping in the male is not appropriate

Fig.5 The influence of one agent's learning to the other's learning

4. Conclusion

In this paper, the effect of action selection in the learning of one-way communication between two agents using Q-learning was examined. The successful ratio became higher when the receiver agent's action selection is greedy, and the transmitter agent's is not completely greedy but has a small random factor.

By observing the change of Q-values and the number of the steps when the both agents determine their actions, it is found that insufficient learning in the female may break learning fatally.

Acknowledgement

This research was partially supported by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Reearch, #14350227 and #15300064.

Reference

[1] G.M. Werner & M. G. Dyer : Evolution of Communication in Artificial Organizing System, Proc.of Artificial life II, 1-47(1991)

[2] N.Ono, T. Ohira, and A. T. Rahmani: Emergent Organization of Interspecies Communication in Q-Learning Artificial Organism,*Advances in Artificial Life*, pp. 396-405 (1995)

[3] K. Shibata and K. Ito: Learning of Communication for Negotiation to Avoid Some Conflicts of Interests-Learning of Dynamic Communication Using a Recurrent Neural Network- Trans. of SICE Vol.35, No.11, 1346-1354 (1999) (in Japanese)

[4] Masanobu Nakanishi, Masanori Sugisaka & Katsunari Shibata: Occurrence of State Confusion in the Learning of Communication Using Q-leaning,Proc. of The 9th AROB (Int'l Sympo. on Artificial Life and Robotics), Vol. 2, pp. 663-666, (2004)

[5] R.S. Sutton and A.G.Barto,: Reinforcement Learning, The MIT Press,1998