A Reinforcement Learning method based on immune network adapted to semi Markov decision process

Nagahisa Kogawa †
Masanao Obayashi †
Kunikazu Kobayashi †
Takashi Kuremoto †

†Graduate School of Science and Engineering, Yamaguchi University
2-16-1 Tokiwadai Ube, Yamaguchi, 755-8611 Japan
(Tel:0836-85-9500, Fax:0836-85-9501)
(E-mail: kogawa@nn.csse.yamaguchi-u.ac.jp)

Abstract: The immune system attracts attention as new biological information processing type paradigm. It is a large-scale system equipped with the complicated biological defense function. It has functions of memory and learning that use the interaction such as stimulus and suppression between immune cells. In this paper, we propose and construct a reinforcement learning method based on immune network adapted to semi-Markov decision process (SMDP). We reveal the proposed method has a capability to dealing with problem which is modeled as SMDP environments through computer simulation.

Keyword: immune network, reinforcement learning, eligibility, semi-Markov decision process

I Introduction

The immune network has been applied to obstacle avoidance problem[1][2] and function approximator structure and parameter adjustment[3]. The clonal selection theory has been applied to pattern recognition[4].

We are looking forward to building a superior learning system by introducing the immune system with such features. In this study, we propose and construct a reinforcement learning method based on immune network [5] adapted to semi Markov decision process (SMDP) [6]. Through computer simulation, we reveal the proposed method has a capability of dealing with a problem which is modeled as SMDP environments.

II Immune network

An immune network is a dynamic network with interaction between cells, such as stimulus and suppression, inspired from the idiotypic network hypothesis which was introduced by N. K. Jerne. The idiotypic network hypothesis suggests that an antibody reacts with not only an antigen but also the other antibodies in the immune processing mechanism of the living body. So the antibody has properties of antigen against the other antibodies.

Fig.1 shows a structure of immune network. Fig.2 shows a structure of antibody.

A density of antibody in the immune network is expressed by first order differential equations as

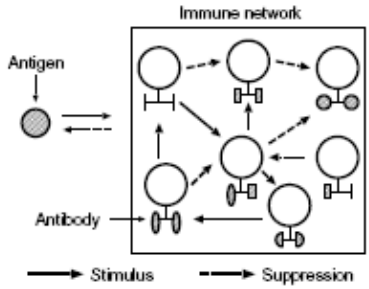

Fig. 1: Structure of immune network

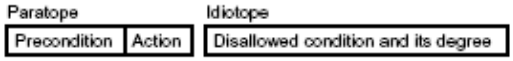

Fig. 2: Structure of antibody

$$\frac{dA_i(t)}{dt} = (\sum_{j=1}^{N} m_{ji} d_j(t) - \sum_{k=1}^{M} m_{ik} d_k(t) + m_i - k_i) d_i(t), \quad (1)$$

$$d_i(t+1) = \frac{1}{1+\exp(0.5 - A_i(t))}, \quad (2)$$

where t is a time to calculate steady states with each antibodies as a result of interactions among antibodies, m_{ji} is stimulus parameter from antibody j

to antibody i, m_{ik} is suppression parameter from antibody k to antibody i, d_i is a density of antibody i, m_i is an affinity between a antigen and an antibody i, k_i is a dissipated coefficient of an antibody i, N, M are the number of antibody that stimulate antibody i and the number of antibody that suppress antibody j, respectively.

III Semi-Markov decision process

SMDP[6] means a model of environment that interval between decision makings isn't constant or a convergence of learning doesn't appear by making decisions frequently at regular intervals. In conventional method, decision making isn't performed in order to adapt to SMDP environments until the time that observed state is changed. However environments are uncertain and incomplete from noise and lack of ability of a sensor for real world problems. These environments are called POMDP[6] environments. In this situation there may be actually a change of state in the whole area even if there isn't it in agent's perceptible area. Therefore it is necessary that agent perform decision making in order to obtain a proper solution even if there isn't a change of state in agent's perceptible area.

In other words, we suggest that this problem is regarded as a problem of the trade-off, which a convergence of learning doesn't appear by making decisions frequently at regular intervals and, a proper solution isn't obtained by making decisions between changes of states. In the proposed method, we perform decision making by using time information moderately and dealing with this problem even if a change of state doesn't appear in the observation.

IV Proposed system

We explain a reinforcement learning method based on immune network adapted to semi Markov decision process in detail. We show a definition of antigen and antibody, a learning method about stimulus and suppression and eligibility, respectively.

1. Definition of antigen and antibody

An antigen consists of perceptible information observed by agent or time information that means time not to observe a change of state in perceptible area. The paratope of antibody is constituted of precondition in order to distinguish the other antibodies and an agent's action. The idiotope of antibody is constituted of disallowed condition (antibody) and its degree.

Under the SMDP environments, there is a case that cannot distinguish a situation only by current perceptible information. we add an antibody distinguishing time information to the immune network in order to adapt to the situation at any time if a situation that a change of state isn't observed appear newly. Decision makings are performed by stimulus and suppression from antibodies distinguishing time information even if the situation that a change of state isn't observed in perceptible area. The antibody distinguishing time information is used to choose an antibody distinguishing perceptible information to show an appropriate action for the agent. And the antibody which distinguishing time information doesn't has the agent's action.

2. Learning method

We explain a learning method with stimulus and suppression . Equations of learning method as

$$\begin{aligned} T_r^{Ab_i}(t) &= T_r^{Ab_i}(t\text{-}1)+\beta e_t|r_t| \quad (if\ r_t>0),\\ T_p^{Ab_i}(t) &= T_p^{Ab_i}(t\text{-}1)+\beta e_t|r_t| \quad (if\ r_t<0),\\ T_{Ab_j}^{Ab_i}(t) &= T_{Ab_j}^{Ab_i}(t\text{-}1)+\beta e_t|r_t|, \end{aligned} \tag{3}$$

$$m_{ji}=\frac{T_r^{Ab_i}+T_p^{Ab_j}}{\alpha+T_{Ab_j}^{Ab_i}}, \tag{4}$$

where i, j are the number of antibody. t is a time. Ab_i is antibody which was chosen when a reward or penalty was got. Ab_j is antibody which wasn't chosen but activated. $T_r^{Ab_i}$ is the sum of reward when Ab_i was selected, $T_p^{Ab_i}$ is the sum of penalty when Ab_i was selected, $T_{Ab_j}^{Ab_i}$ is the sum of reward and penalty when Ab_i, Ab_j are activated, r_t is reward or penalty of time t, α is constant, β is learning rate. e_t shows eligibility[7].

3. eligibility

It is necessary to choose different actions under POMDP environments and under the influence of incomplete perception problem. Therefore we equalize an eligibility of each action when agent chose a multiple action in the same perceptible information per each episode. We expect that the antibody which doesn't contribute to reward is not chosen, and contribute to reward is chosen stochastically by learning antibodies distinguishing perceptible information with the eligibility. Fig. 3 shows an average eligibility that is used to learning of antibodies distinguishing perceptible information.

Furthermore, under the SMDP environments, agent performs decision making by using antibodies distinguishing time information when a change of state isn't observed. We use eligibility before the equalization for learning of stimulus and suppression from an antibody distinguishing time information to antibody dis-

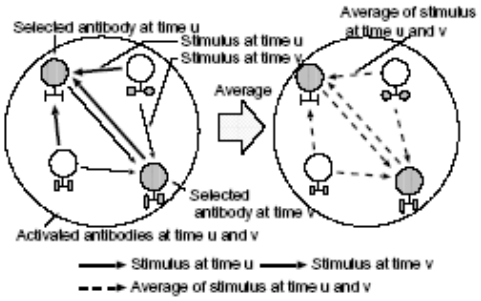

Fig. 3: Stimulus and suppression with eligibility

tinguishing perceptible information. Fig.4 shows stimulus and suppression between antibodies with perceptible information and antibodies with time information.

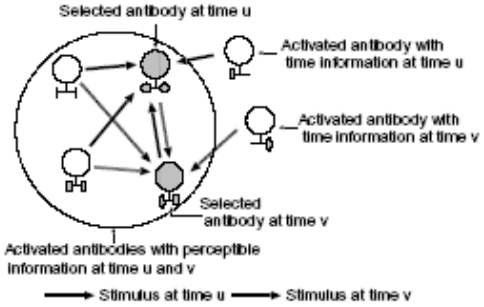

Fig. 4: Stimulus and suppression between antibodies with perceptible information and antibodies with time information

4. Algorithms

Algorithm of the proposed system is as follows.

1) an antigen is perceptible information observed by agent, an antibody is action corresponding to perceptible information.
2) Observe perceptible information and antigen is got.
3) Calculate the density of antibodies by equation(1) ,(2) by interaction between antigen and antibodies.
4) If the density is over threshold Tac_1 , an antibody is selected stochastically in proportion to the density.
5) Decide limited time to continue the same action randomly. agent does action until state isn't changed or limited time is over.
6) Increase time information if state isn't changed. Add an antibody distinguishing time information if time information is new.
7) Repeat procedure 2)~6).

V Computer simulation

We verify that performance of the proposed method by solving a problem of maze shown as Fig. 5. White squares are way, black squares are wall, S shows an episode start position, G shows an episode end position in Fig.5.

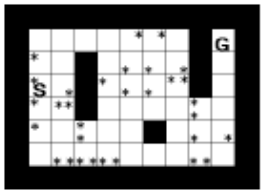

Fig. 5: Small-sized maze

Fig. 6 describes description of antibodies. Agent's actions are up, down, left and right. Perceptible area are information of surrounding agent. Table 1 and Table 2 show parameters of simulation.

A mark $*$ shows positions that have the same perceptible information in Fig.5. This maze has a lot of the same environments. A best solution in maze shown as Fig.5 is 14 steps.

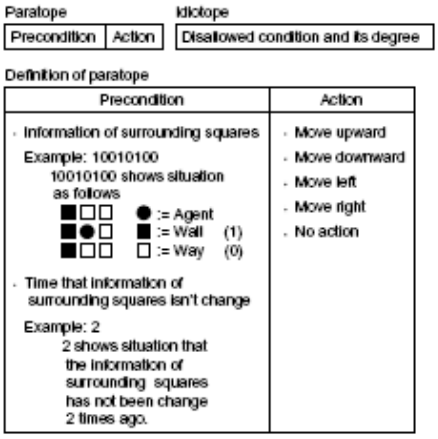

Fig. 6: Description of antibody

Fig.7, Fig.8 show results of computer simulation. 1 trial contains 1000 episodes. 20 trials were performed in the experiment. Fig.7 shows a result of transition of steps on average. Fig.7 illustrates it is enable to learn a near best solution by the proposed method. Fig.8 shows a result of transition of steps of a best trail. Fig.8 confirmed that steps converge to the best solution. This phenomena are shown in 8 trials. It is appeared that the best solution or the near best solution

Table 1: Parameters of simulation

Reward at goal	1.0
Reward at wall	−1.0
Reward at space	−0.01
α	1.0
k_i (all antibodies)	0.5
Tac_1	0.8

Table 2: Parameters of simulation of antibody

	antibody with perceptible information	antibody with time information
β	0.1	0.01
m_i (activated)	1.0	1.5
m_i (not activated)	0.0	−1.5

are selected stochastically if antibody distinguishing time information isn't used. Because antibody distinguishing time information operates on convergence.

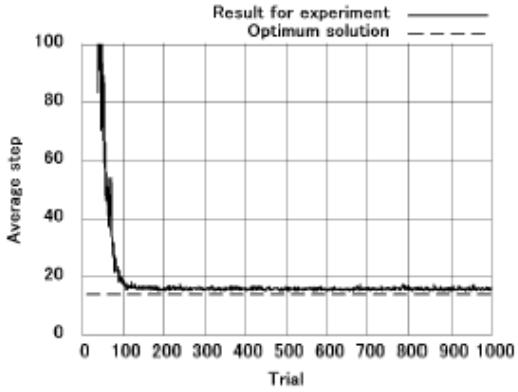

Fig. 7: Result for experiment(average step)

VI Conclusion

We proposed a reinforcement learning method based on immune network adapted to semi Markov decision process. Through computer simulation, we revealed the proposed method has the capability of dealing with the problem which is modeled as POMDP environments. It was verified that effective action is selected by using antibody distinguishing time information. Future problem is to verify a capability for SMDP environments in detail. In computer simulation, the proposed method was used to a small-sized maze. And time information is expressed simply. However it is necessary that description of time information is more flexible to apply a problem of real

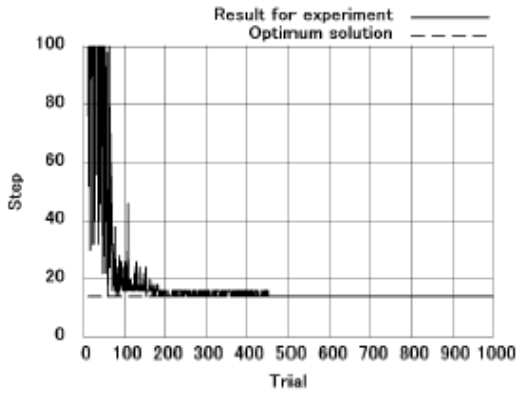

Fig. 8: Result for experiment(best step)

world. The immune network is a dynamic network. And state is expressed more flexible by using this network. It is expected that more complex problem is solved by applying capability of immune network.

References

[1] A. Ishiguro, Y. Watanabe, T. Kondo, Y. Uchikawa, "Construction of a Decentralized Consensus-Making Network Based on the Immune System-Application to an Action Arbitration for an Autonomous Mobile Robot-", *Transactions of the Society of Instrument and Control Engineers*, Vol. 33 No. 6, pp. 524-532, 1997 (in Japanese)

[2] A. Ishiguro, T. Kondo, Y. Watanabe, Y. Shirai, Y. Uchikawa, "An Evolutionary Construction of Immune Network-Based Behavior Arbitration Mechanism Mobile Robot", *IEEJ Transactions on Electronics, Information and Systems*, Vol. 117-C No. 7, pp. 865-873, 1997 (in Japanese)

[3] Yixin Diao, Kevin M. Passino, "Immunity-based hybrid learning methods for approximator structure and parameter adjustment", *Engineering Applications of Artificial Intelligence*, Vol. 15, pp. 587-600, 2002

[4] Kemal Polat, Salih Güneş, Sülayman Tosum, "Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted pre-processing", *Pattern Recognition*, Vol. 39, pp. 2186-2193, 2006

[5] Y. Ishida at el.,*Immunity-Based Systems and Its Applications -Intelligent Systems by Artificial Immune Systems* -, Corona Publishing, 1998 (in Japanese)

[6] H. Kimura, K. Miyazaki, S. 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 (in Japanese)

[7] Richard S. Sutton, Andrew G. Barto, Reinforcement Learning: An Introduction(Adaptive Computation and Machine Learning), *The MIT Press*, 1998