Adaptive Immunity Based Reinforcement Learning

Jungo Ito, Kazunori Sakurama and Kazushi Nakano

Dept. of Electronic Eng., The University of Electro-Communications
1-5-1 Chofu-ga-oka, Chofu, Tokyo 182-8585, Japan

Abstract

Recently much attention has been paid to intelligent systems which can adapt themselves to dynamic and/or unknown environments by use of learning methods. However, traditional learning methods have the disadvantage that learning requires enormously long amounts of time with the degree of complexity of systems and environments to be considered. We thus propose a novel reinforcement learning method based on the adaptive immunity. Our proposed method can provide a near-optimal solution with less learning time by self-learning using the concept of the adaptive immunity. The validity of our method is demonstrated through two simulations for Sutton's maze problem.

Keywords: reinforcement learning, adaptive immunity

1 Introduction

Much attention has recently been paid to intelligent systems which can adapt themselves to dynamic and/or unknown environments by use of learning and memory function. However, traditional learning methods have the disadvantage that learning requires enormously long amounts of time with the degree of complexity of the systems and environments.

On the other hand, there exist a lot of approaches for modeling the various functions of a living creature and the evolution of a creature and then applying them to optimal solution search methods and learning methods. Especially among the approaches, the immune system has gotten much attention in research[1]. As for a typical immune system, Farmer, et al.[2] have proposed an engineering model based on Jerne's idiotypic network hypothesis[3]. But, Ref.[4] does not have a learning mechanism, whose performance depends on the designer's experience. Ref.[5] has a learning mechanism, by which self-optimization can be performed autonomously. But, it has the disadvantage that learning needs an exponentially increasing amount of time with the degree of complexity of environments.

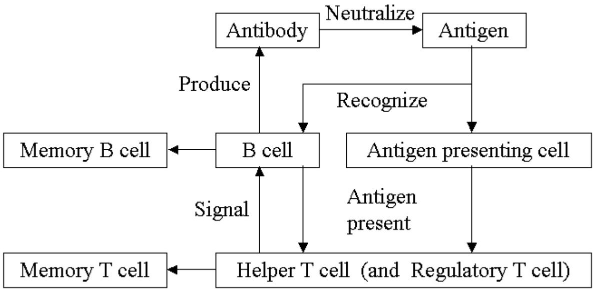

Fig. 1: Adaptive immunity

As well known, "reaction" which is called the adaptive immunity among immune systems makes it possible to detect and eliminate pathogens in coordination with cells having multiple different roles and makes self-optimization possible in a learning mechanism (Fig. 1). This paper proposes a novel reinforcement learning method based on the adaptive immunity. The proposed method can select a suitable action based on a rule selection using the concept of adaptive immunity, and can provide a near-optimal solution with less learning time by introducing the concept of immune system to learning and memory. Lastly, the validity of our method is demonstrated by two simulations for Sutton's maze problem.

2 Proposed Method

2.1 Summary of adaptive immunity

This chapter considers the mechanism of the adaptive immune system eliminating pathogens which invades the human body. The pathogenic is called antigen. The antigen is recognized by the antigen presenting cells. The antigen presenting cells includes the B cells. And the information of the antigen is

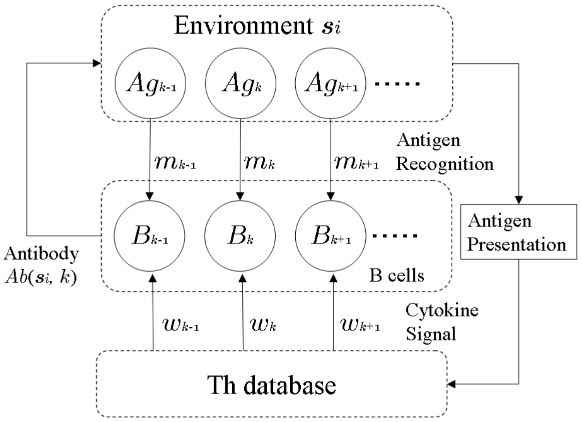

Fig. 2: Engineering model of adaptive immunity

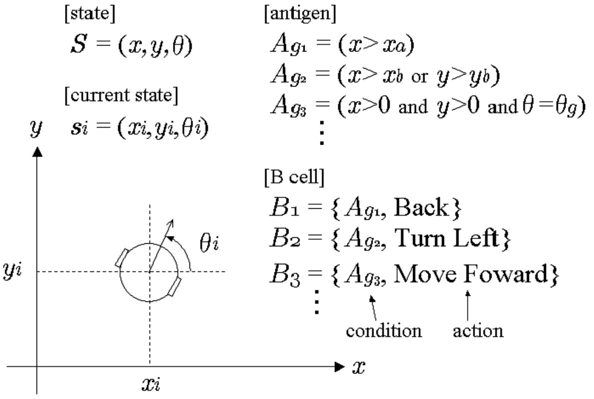
[state]
$S = (x, y, \theta)$
[current state]
$s_i = (x_i, y_i, \theta_i)$
[antigen]
$Ag_1 = (x > x_a)$
$Ag_2 = (x > x_b$ or $y > y_b)$
$Ag_3 = (x > 0$ and $y > 0$ and $\theta = \theta_g)$
[B cell]
$B_1 = \{Ag_1$, Back$\}$
$B_2 = \{Ag_2$, Turn Left$\}$
$B_3 = \{Ag_3$, Move Foward$\}$
condition
action
y
y_i
θ_i
x
x_i

Fig. 3: Example of state, antigens and B cells

presented to T cells. The T cells which the information is presented release cytokines, and send the signal to the B cells for activation. The activated B cells then produce the antibody to neutralize the antigen. Therefore, the invaded antigen can be eliminated. The T cells playing the above role are called Helper T cells (Th cells). The relationships of B cells - antigens and B cells - Th cells are specific. Generally, T cells and B cells die after eliminating the antigen. But, some activated T cells and B cells have a long lifetime, circulate through around the human body and survive as memory cells. As a result, the adaptive immunity becomes able to respond quickly and eliminate effectively the same type of antigen.

2.2 Modeling of adaptive immunity

The dynamics of adaptive immunity is modeled as shown in Fig. 2. First of all, the set of all the states where the agent can exist is defined as $\boldsymbol{S}$. The state where the agent exists is defined as $\boldsymbol{s}_i(\in \boldsymbol{S})$. These two correspond to all the states which the human body can take and the current state of the human body. On the other side, the B cells are considered as a pair of conditions and actions. The condition means the state where the B cell should be activated. This state corresponds to the antigen Ag. The action means that the agent performs it if the B cell is selected. This condition-action pair is the rule.

Now let us consider an action of a mobile robot, the relation between the state, antigen information and B cell is shown in Fig. 3. The state consisting of the x- and y-coordinates and the direction, is expressed as (x, y, θ). The current state $\boldsymbol{s}_i$ is expressed as (x_i, y_i, θ_i). For this, the antigen information is expressed as the conditions for the elements of $\boldsymbol{S}$. And the B cell is expressed as a pair of the antigen and the action where the robot should do.

When the agent exists in $\boldsymbol{s}_i$, kth B cell B_k in the agent is stimulated if the $\boldsymbol{s}_i$ satisfies this condition Ag_k of B_k. If $\boldsymbol{s}_i$ does not satisfy it, B_k is not stimulated. In Fig. 2, m_k means the degree of stimulation in B_k. If $\boldsymbol{s}_i$ contains Ag_k, m_k becomes 1. Otherwise, m_k becomes 0.

On the other hand, the Th cells is considered as the agent's memory where the value of B cells evaluated for each state is recorded. Th database is a group of Th cells. The database receives the information of current state $\boldsymbol{s}_i$ from the antigen presenting cell. After then, a Th cell corresponding to $\boldsymbol{s}_i$ releases the evaluation of each B cell $\boldsymbol{w}= (w_1, w_2, \cdots)$ as a cytokine signal.

Finally, the adaptive immune system selects a B cell based on $\boldsymbol{m}= (m_1, m_2, \cdots)$ and $\boldsymbol{w}$. B cell selecting method shall be explained in the next section. If B_k is selected, the agent executes the action which has been specified by B_k. Just then, the information set composing of a current state $\boldsymbol{s}_i$ and a selected B cell number k is produced and released as antibody $Ab(\boldsymbol{s}_i,k)$.

By repeating B cell selection with the above method, the agent aims toward the target state.

2.3 Algorithm for B cell selection

An algorithm for B cell selection with using the Th database is presented as follows:

1. The agent exists in $\boldsymbol{s}_i$, Th database releases a

cytokine signal $w_k(\boldsymbol{s}_i)$ according to the state. On the other side, B cells present the degree of stimulation m_k according to the current state.

2. After calculating $v_k = m_k \times w_k(\boldsymbol{s}_i)$, a B cell is selected through the roulette selection with using v_k as the selection probability for B_k.

3. Antibody $Ab(\boldsymbol{s}_i,k)$ is produced by B_k. The antibody has the parameter called concentration which means the antibody's lifetime. When the antibody is produced, its concentration is set to 1 ($Ab(\boldsymbol{s}_i,k)$= 1). If the same antibody has already been produced, or if the same B cell has already been selected in the past same state, a new antibody is not produced, and the existing antibody's concentration is reset to 1.

4. The concentrations of other antibodies produced in the past are updated with the following equation:

$$Ab \leftarrow \quad \times Ab \tag{1}$$

where $\quad (0 < \quad < 1)$ is the discount rate.

By performing the above process, the agent decides the B cell which it should select. When the agent receives a reward from environment after it executed an action, the Th database is updated. It means that each $w_k(\boldsymbol{s}_i)$ is updated.

2.4 Update of Th database

When the agent receives a reward from its environment, each $w_k(\boldsymbol{s}_i)$ is updated as follows:

$$w_k(\boldsymbol{s}_i) \leftarrow w_k(\boldsymbol{s}_i) + \quad (r_k(\boldsymbol{s}_i) - w_k(\boldsymbol{s}_i)) \tag{2}$$

$$r_k(\boldsymbol{s}_i) = \begin{cases} Ab(\boldsymbol{s}_i,k) \times R : Ab(\boldsymbol{s}_i,k) \text{ was produced} \\ 0 \qquad\qquad\quad : \text{otherwise} \end{cases} \tag{3}$$

where R is the reward which the agent receives from its environment, and $\quad (0 < \quad < 1)$ is the learning rate. This update is performed for all w. After updating, all antibodies are erased.

The agent becomes able to select an appropriate rule for its environment by repeating the learning with the above process of rule selection.

3 Simulation

This chapter shows two simulation results of our proposed method applying to Sutton's maze problem. This problem is one of the standard ones for reinforcement learning.

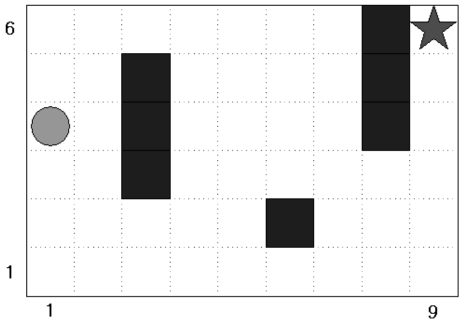

Fig. 4: Sutton's maze

3.1 Simulation setting

We set a problem that the agent located at ○ aims to reach the destination ☆ as shown in Fig. 4. The agent can move to one of four adjacent areas of "front", "back", "right" and "left" by one-time rule selection. We set $m_k = 1$ in all the states of this simulation. Obstacles are shown as ■ in the Fig. 4. When selecting the obstructed movement direction, the agent continues to stay at the existing area. The circumference of in Fig. 4 should be considered as an obstacle. When arriving at the target area through the optimal path in this environment, the agent needs 14 times in rule selection.

When reaching the destination, the agent receives a reward and is returned to the starting area. After repeating 1000 episodes, we examined the change in the number of times of the rule selection which the agent needed to reach the destination. The number of times is defined as step. Here, the parameters used in the proposed method are $\quad = 0.1$, $\quad = 0.1$, initial $w = 1$ and $R = 1$. The following figure plots the average of ten trials.

3.2 Simulation result with deterministic state transition

Fig. 5 shows a simulation result for the environment in Fig. 4. The horizontal axis is the number of episodes, and the vertical axis is the number of steps which the agent needs to reach the destination in each episode. The learning almost converges at about 300 episodes. In addition, the number of steps at 1,000 episodes is 14, that is, the optimal solution 14 is gained for all the ten trials.

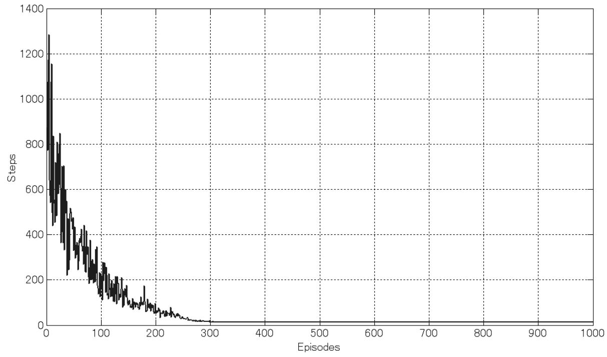

Fig. 5: Simulation result with deterministic state transition

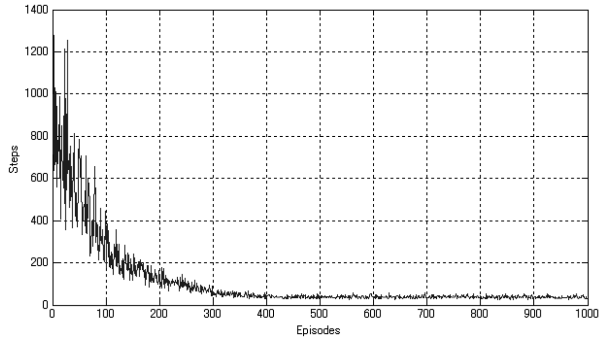

Fig. 6: Simulation result with probabilistic state transition

3.3 Simulation result with probabilistic state transition

Next, we show the other simulation result in the environment with a probabilistic state transition. In this case, once a movement direction is decided, the agent is assumed to move in the direction with the probability of 70% in this simulation. Then, the agent is assumed to move in the other direction with the probability of 10%. Fig. 6 shows the result in this case. The learning almost converges at about 400 episodes. This result shows that the proposed method has a learning capability even in the environment with the probabilistic state transition. But, the number of steps at 1,000 episodes is 32.4. In the environment with the probabilistic state transition, we can see that the agent could not reach the destination through the optimal path even if it always select the optimal action.

4 Conclusion

In this paper, we proposed a novel method for reinforcement learning based on the adaptive immunity. We demonstrated in the simulation that about 400 episodes can produce a near-optimal solution in Sutton's maze problem with deterministic and probabilistic state transitions.

This paper dealt with a static environment problem. However, it should be consider in the near future to develop a method for a dynamic environment and continuous states from the practical point of view.

References

[1] L. N. de Castro and J. Timmis (2002), Artificial Immune Systems: A New Computational Intelligence Approach. Springer

[2] J. D. Farmer and N. H. Packard (1986), The Immune System, Adaptation, and Machine Learning. Physica, 22D:187-204

[3] N. K. Jerne (1984), Idiotypic Networks and Other Preconceived Ideas. Immunological Reviews, No.79, pp.5-24

[4] G. C. Luh, W. W. Liu (2004), Reactive Immune Network Based Mobile Robot Navigation. ICARIS2004, pp.119-132

[5] J. Ito, K. Sakurama, K. Nakano (2005), A Soccer Robot Control Design Based on the Immune System. AROB10th, GS23-4