ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 313-318

A model-based reinforcement learning: a computational model and an fMRI study

Wako Yoshida* and Shin Ishii

Nara Institute of Science and Technology
CREST, Japan Science and Technology Corporation

Abstract. In this article, we discuss an optimal decision making problem in a dynamic environment on the bases of both machine learning and brain learning. We present a model-based reinforcement learning (RL) in which the model of environment is directly estimated. Our RL makes the action selection depending on the detection of environmental changes and the current value function. We suggest a possible functional model of our RL with a focus on the prefrontal cortex and the anterior cingulate cortex. In order to examine our model, an imaging study is conducted.

1 Introduction

Although natural environments surrounding humans change with time, humans learn the features of the current environment and determine their optimal behaviors. In this article, we discuss a reward-based decision making method on the bases of both machine learning and brain learning. In the machine learning field, an optimal decision making problem is termed Markov decision process (MDP). If an MDP involves the direct identification of an unknown environment, it is solved by a model-based reinforcement learning (RL) method.

In RL, the objective of an agent is to maximize the rewards accumulated toward the future, and it is achieved by improving its action selection. A standard RL scheme then estimates the expected reward accumulation, that is the value function. Model-based RL [10], however, tries to identify the current environment directly and the value function is approximated using the model. In our previous paper [3], we presented a model-based RL method in which the environmental model is estimated based on a Bayes inference.

In this article, we propose a possible functional model in the brain, which realizes our model-based RL method. We assume that the estimation of reward-based environmental models is involved in functions of the dorsolateral prefrontal cortex (DLPF) and the anterior prefrontal cortex (APF). Our RL method also needs the action selection depending on that estimation. We consider this operation is done within the anterior cingulate cortex (ACC).

In order to examine our functional model, a human imaging study using functional magnetic resonance imaging (fMRI) is conducted in this study.

*Correspondence to: Tel.(fax):+81-743-72-5986(5989); E-mail: wako-y@is.aist-nara.ac.jp

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 313-318

2 Model-based RL method

In a Markov environment, where $P(s'|s,a)$ gives the probability of reaching state s' by selecting action a at state s, the value function for state s, $V(s)$, should satisfy the following optimal Bellman's equation:

$$V(s) = \max_a Q(s,a) \tag{1a}$$

$$Q(s,a) \equiv r(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s'), \tag{1b}$$

where $r(s,a)$ denotes an immediate reward and $0 \leq \gamma \leq 1$ is a discount constant. The action-value function $Q(s,a)$ represents the expected reward accumulation when the agent takes action a at state s and the optimal actions at the subsequent states. The objective of an RL, often termed MDP, is to obtain an optimal policy, which outputs the action maximizing $Q(s,a)$ for any state s. In many RL problems, the state-transition probability $P(s'|s,a)$ is unknown. The model-based RL [10] tries to model directly the environment by approximating the state-transition probability based on past experiences.

2.1 Partially-observable MDP

A partially-observable MDP (POMDP) [4] deals with a Markov environment with unobservable state variables. Let $s \equiv (y,z)$ be an environmental state, where y and z denote observable and unobservable state variables, respectively. One way to deal with a POMDP is called a belief state MDP, in which the Bellman's equation is modified from that in MDP, (1), by replacing a state s with a belief state b. A belief state is typically a probability distribution of the observable and unobservable variables. Since there is no probabilistic factor for the observable variables, $b = [y, \hat{P}(z)]$, where $\hat{P}(z)$ is estimated from the past observations. We assume that an agent is able to estimate a new belief state $b' = [y', \hat{P}'(z)]$, using the new observation y'. Even in a finite world, where both state and action spaces are discrete and finite, the belief state MDP is hard to solve, because the value function is often intractable. Therefore, we need an approximation. If an RL agent is certain of the estimation of the unobservable variables, $\hat{P}(z)$ is equivalent to $\hat{z}$, i.e., the most probable value of z. If we further assume for simplicity that the reward function does not depend on the unobservable variables, the Bellman's equation is approximated as

$$V([y,\hat{z}]) = \max_a Q([y,\hat{z}],a) \tag{2a}$$

$$Q([y,\hat{z}],a) = r(y,a) + \gamma \sum_{y'} P(y'|[y,\hat{P}(z)],a)V([y',\hat{z}']). \tag{2b}$$

Since this approximation may not be valid when the RL agent is uncertain of the unobservable variables, our previous model [3] introduced an exploration bonus that encourages exploratory behaviors in an uncertain situation.

In our RL method, there are unobservable variables reflecting the stochastic nature of the environment, and distribution $\hat{P}(z)$ is estimated by a Bayes

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 313-318

inference with "forgetting" effect on past experiences and a non-informative prior. Detailed formulations were described in our previous paper [3].

2.2 Action selection

We define a stochastic policy π by a conditional probability $P^\pi(a|s)$ [1]. Especially in a finite world, a greedy policy, which maximizes $\int Q(s,a)P^\pi(a|s)da$, will assign probability zero to possible actions except one or several. Then, it becomes difficult for the agent to adapt to the environmental change. In order to preserve the adaptability, free energy is introduced:

$$F(P^\pi) = \int Q(s,a)P^\pi(a|s)da - \frac{1}{\beta}\int P^\pi(a|s)\log P^\pi(a|s)da. \tag{3}$$

The maximization of the first and second terms corresponds to exploitation for obtaining a large reward based on the current value function and exploration for searching for a better policy, respectively. Coefficient β is called inverse-temperature; it controls the exploitation-exploration balance.

Using the variational method, the maximization of $F(P^\pi)$ is achieved by

$$P^\pi(a|s) = \frac{\exp(\beta Q(s,a))}{\int \exp(\beta Q(s,a))da}, \tag{4}$$

which is called the soft-max policy. When the inverse-temperature is small, the soft-max policy randomly selects one of the possible actions. With a large value, it selects a greedy action that maximizes the current action-value function.

We proposed a two-folded control method of the inverse-temperature [3], one based on the variation of the action-value function and the other based on the detection of the environmental change. Although the details are omitted, important is that the action selection is modified such as to depend on the estimation of the environment.

2.3 Working hypothesis

Here, we present a possible brain implementation of our RL method.

DLPF has been mainly studied in terms of a working memory function for goal-directed behaviors. Rao et al. [7] reported sustained activities of DLPF neurons depending on state and/or action. In addition, recent recording studies have revealed that DLPF neurons predict the quality and the quantity of the future reward. Thus, we assume that DLPF represents the estimation of accumulated reward, which depends on state and/or action, i.e., the value function and/or the action-value function in RL.

In the model-based RL, the value function is approximated using the environmental model. According to a recent view, DLPF constructs cascade networks representing transitions of states [11]. A recent study [6] suggested that DLPF is involved in the preparation of forthcoming sequence of actions based on information stored in working memory. The behavioral planning requires

[1] In the following descriptions, $[y, \hat{z}]$ is represented as s.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 313-318

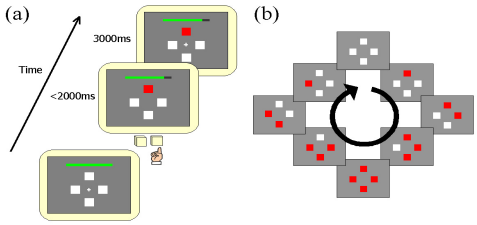

Figure 1: A sequence learning task with visual stimuli and response buttons.

predicting the environmental changes induced by own action. We speculate that the environmental models in RL are expressed in DLPF.

The environmental model of our RL method requires the estimation of unobservable variables. Recent human imaging studies [5, 9] suggested that APF is involved in active switching of behavioral rules without explicit cues. Since such switching is induced by the estimation of environmental change, we consider that APF is possibly related to the estimation of unobservable states.

In our RL, the action selection is based on both the current value function and the detection of environmental change. A recent imaging study [1] observed ACC activations when a subject voluntarily switched his behavioral rule. ACC is also activated by the detection of behavioral error and/or response conflict. We consider that ACC is related to the uncertainty of the action selection, depending on the current environmental model maintained in DLPF.

According to our hypothesis, DLPF maintains and manipulates the environmental model and the reward-based environmental model, i.e., $P(y'|[y, \hat{P}(z)], a)$ and $Q([y, \hat{z}], a)$ in (2). APF estimates the unobservable state variables, $\hat{z}$. These estimations are carried to ACC that executes the action selection (4).

3 fMRI experiment

3.1 Material and methods

Behavioral task Sixteen subjects (13 males and 3 females) performed sequential learning tasks and they were paid in proportion to their task scores. At an initial state, a fixation cross was displayed in the screen center surrounded by four gray squares. A subject was required to press left or right button within 2000 ms. Immediately after the response, the next state was represented (Fig.1(a)). When the subject pressed a correct button, the color of one square changed, while a wrong button resulted in no change. A state was represented by a color pattern of squares, and the color of the squares changed from gray to red in the first round and from red to gray in the second round (Fig.1(b)). Thus, to reach the goal, a subject had to learn an eight-response sequence by feedbacks indicating whether each response was correct or not.

An experiment consisted of two behavioral conditions. In a memory (MEM) condition, since the state transition was deterministic for each state, a subject

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 313-318

Region	BA	MNI			*t*-value
		x	y	z	
Prefrontal Cortex					
Middle Frontal Gyrus	46/9	48	30	22	6.15
Middle Frontal Gyrus	8	32	18	48	7.11
Parietal Cortex					
Inferior Parietal Cortex	40	38	-50	40	8.85
Anterior Cingulate					
Cingulate Gyrus	32	6	22	38	7.38
Cingulate Gyrus	32	-4	18	46	5.96

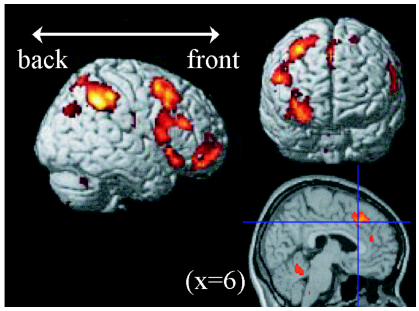

Figure 2: MDP-MEM comparison based on a group random effect analysis.

needed to memorize a fixed sequence of eight correct responses. In an MDP condition, on the other hand, the state transition was first-order Markov, in which a correct response resulted in state transition with 85% or stay at the same state with 15%; these probabilities did not change with time. Thus, especially in the early leaning stage, a subject was required to estimate and maintain the multiple state transitions, i.e., the environmental models $P(s'|s,a)$, and to learn the optimal control $P(a|s)$ using the estimation.

A subject performed two runs, each of which included three MDP sessions of 20 responses and one MEM session. The correct response sequence was the same among three MDP sessions. Before and after a single session, the condition and the number of achieved goals were displayed, respectively.

Procedures Using a 1.5-tesla scanner, functional images were obtained with a T2*-weighted EPIs, with blood oxygenation level-dependent (BOLD) contrast (TE, 55 ms; FA, 90°). The volumes were acquired every 3.0 sec (TR), and contained 28 slices of 5mm thickness.

The data were analyzed with SPM99 software. All EPIs were realigned to first image, registered to anatomical image, and then normalized to MNI reference brain. The normalized EPIs were spatially smoothed with a Gaussian kernel of 8 mm (FWHM). BOLD activations were statistically compared with respect to group random effects with significance $p < 0.005$.

3.2 Results and Discussion

The mean response time (RT) was 3773 ms in the MEM session, and 4082 ms, 3774 ms and 3497 ms in the three MDP sessions; it significantly decreased as the MDP sessions proceeded ($p < 0.05$). In the MDP sessions, we measured the moving average of both behavioral variation by means of entropy, and behavioral correctness by means of overlap with the correct response sequence. The behavioral variation decreased with time through learning (data not shown).

The comparison of MDP trials to MEM trials revealed significant increase in activations in four regions (Fig.2). Although correct automata in both conditions have eight lengths, subjects in the MDP condition were required to preserve some possible sequences due to the stochastic nature of the task. We

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 313-318

consider that the PFC activation is related to the manipulation and maintenance of the automata representing possible environmental models. Actually, when subjects repeated the learned sequence automatically in the MEM conditions, that part was not activated. Activations of both posterior DLPF (BA8) and a cortex in the intraparietal sulcus (IPS) were also found. These areas have been known to reflect the maintenance of working memories without further executive processing [8]. It was suggested that IPS is important for visuospatial attention [2], and this interpretation is possible in our study's case, because error-induced visual attention will occur in the MDP condition.

We also found significant activation increase in ACC during the MDP condition, and the activation significantly decreased as the MDP sessions proceeded. The decrease of behavioral variation through learning is related to the decrease in the activity of ACC. This interpretation is consistent with our hypothesis, in which ACC represents the uncertainty of action selection.

4 Concluding Remarks

In this study, we presented a model-based RL method in which the environment is directly estimated. In order to adapt to changes in the environment, a control method of action selection was introduced.

We proposed a possible functional model of our RL method, in which DLPF maintains and manipulates the environmental models and ACC is related to action selection, and conducted an fMRI experiment. Although we also suggested that the estimation of unobservable states in RL is expressed in APF, its possibility has not yet been examined. This is an issue in our future study.

References

[1] Bush, G., et al. (2002). *PNAS*, **99**, 507-512.

[2] Corbetta, M., et al. (2000). *Nature Neuroscience*, **3**, 284-291.

[3] Ishii, S., et al. (2002). *Neural Networks*, **15**, 665-687.

[4] Kaelbling, L.P., et al. (1998). *Artificial Intelligence*, **101**, 99-134.

[5] Koechlin, E., et al. (2000). *PNAS*, **97**, 7651-7656.

[6] Pochon, JB., et al. (2001). *Cerebral Cortex*, **11**, 260-266.

[7] Rao, S.C., et al. (1997). *Science*, **276**, 821-824.

[8] Rowe, J.B., et al. (2000). *Science*, **288**, 1656-1660.

[9] Strange, B.A., et al. (2001). *Cerebral Cortex*, **11**, 1040-1046.

[10] Sutton, R.S. (1990). In *Machine Learning: Proceeding of the Seventh International Conference*, pp. 216-224.

[11] Tanji, J., et al. (2001). *Current Opinion in Neurobiology*, **11**, 164-170.