Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neurocomputing 63 (2005) 253–269
NEUROCOMPUTING
www.elsevier.com/locate/neucom

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

Wako Yoshida[a,b,*], Shin Ishii[a,b]

[a]*Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan*
[b]*CREST, Japan Science and Technology Agency, Japan*

Received 4 November 2003; received in revised form 27 February 2004; accepted 23 April 2004
Available online 20 August 2004

Abstract

In this paper, we discuss an optimal decision-making problem in an unknown environment on the bases of both machine learning and brain learning. We present a model-based reinforcement learning (RL) in which the environment is directly estimated. Our RL performs action selection according to the detection of environmental changes and the current value function. In a partially-observable situation, in which the environment includes unobservable state variables, our RL incorporates estimation of unobservable variables. We propose a possible functional model of our RL, focusing on the prefrontal cortex and the anterior cingulate cortex. To test the model, we conducted a human imaging study during a sequential learning task, and found significant activations in the dorsolateral prefrontal cortex and the anterior cingulate cortex during RL. From a comparison of the mean activations in the earlier and later learning phases, we suggest that the dorsolateral prefrontal cortex maintains and manipulates the environmental model, while the anterior cingulate cortex is related to the uncertainty of action selection. These experimental results are consistent with our model.

Keywords: Reinforcement learning; Partially-observable Markov decision process (POMDP); Prefrontal cortex; Functional MRI

*Corresponding author. Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan. Tel.: +81-743-72-5985; fax: +81-743-72-5989.
E-mail address: wako-y@is.naist.jp (W. Yoshida).

0925-2312/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2004.04.012

1. Introduction

Although the environment around us is constantly changing, humans can learn the features of their current environment and determine optimal behaviors. Assuming optimality is defined as rewards received from the environment, an adaptation to the environment can be formulated as an optimal decision-making problem with an on-line identification of the current environment. We discuss here a possible reward-based decision-making method that is based on both machine learning and brain learning. To understand brain functions, it is important to integrate findings from various research fields. The aim of our study is to discuss brain functions based on a theoretical model and to evaluate it by means of neuropsychological experiments.

In the machine learning field, an optimal decision-making problem in a known or unknown environment is often formulated as a Markov decision process (MDP). If an MDP includes the direct identification of an unknown environment, the problem can be solved by a model-based reinforcement learning (RL) method [7–9,15,26]. In RL, the objective of an agent is to maximize rewards accumulated for the future, which is achieved by improving the agent's action selection. To improve the strategy for receiving rewards (i.e., policy), a standard RL scheme then estimates expected reward accumulation with respect to current policy, which is the value function. A model-based RL [7–9,15,26] tries to identify current environment directly, and the value function is approximated from the resulting environmental model. However, human environments often include unobservable state variables. If we consider decision-making by a card player, for example, cards held by the other players are not directly observed and hence are unobservable variables. In a previous paper [11], we presented a model-based RL method for a partially-observable Markov decision process (POMDP) to deal with such a realistic problem. The POMDP assumes that the environment contains unobservable variables. In this paper, we briefly re-introduce our model-based RL method, and then propose a functional model for the brain, in which our RL method is realized within a POMDP environment.

Because RL has provided a good model of animal behaviors (conditioning), with recent neurophysiological and neuroimaging studies suggesting that RL algorithms are associated with neural processing in the brain [1,23,30], the components of RL could correspond to brain functions. We assume that the major parts of our RL method (the environmental model, the value function and the estimation of unobservable state variables) are involved in the prefrontal cortex (PFC): the dorsolateral prefrontal cortex (DLPF), the orbitofrontal cortex (OFC) and the anterior prefrontal cortex (APF). Our RL method also requires that action selection is dependent on estimation of the current environmental model and the value function. We consider that this operation occurs in the anterior cingulate cortex (ACC).

To test our functional model, a human imaging study using functional magnetic resonance imaging (fMRI) was conducted in this study. Although POMDP is discussed in the computational section, the experimental study does not assume a partially-observable situation. Our RL method within a POMDP, an extended version of MDP, can be formulated by adding a hidden state estimation process to