PERGAMON

Neural Networks 15 (2002) 665–687

Neural Networks

www.elsevier.com/locate/neunet

2002 Special issue

Control of exploitation–exploration meta-parameter in reinforcement learning

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

[a]*Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0101, Japan*
[b]*CREST, Japan Science and Technology Corporation, Japan*

Received 10 October 2001; accepted 16 April 2002

Abstract

In reinforcement learning (RL), the duality between exploitation and exploration has long been an important issue. This paper presents a new method that controls the balance between exploitation and exploration. Our learning scheme is based on model-based RL, in which the Bayes inference with forgetting effect estimates the state-transition probability of the environment. The balance parameter, which corresponds to the randomness in action selection, is controlled based on variation of action results and perception of environmental change. When applied to maze tasks, our method successfully obtains good controls by adapting to environmental changes. Recently, Usher et al. [Science 283 (1999) 549] has suggested that noradrenergic neurons in the locus coeruleus may control the exploitation–exploration balance in a real brain and that the balance may correspond to the level of animal's selective attention. According to this scenario, we also discuss a possible implementation in the brain. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords: Reinforcement learning; Exploitation–exploration problem; Neuromodulator; Attention; Partially observable Markov decision process

1. Introduction

Reinforcement learning (RL) (Sutton & Barto, 1998) is a learning framework in order to adapt to an environment based on trial and error. This paper discusses an RL scheme for dynamic environments, i.e. environments that change with time. Conventional RL schemes are formulated in terms of Markov decision process (MDP), that is, a decision-making problem or an optimal control problem in a stochastic but static environment. Since an optimal control problem in a dynamic environment can approximately be formulated as an MDP when RL is faster than the environmental change, this study also adopts that approximation. In addition, we also use a formulation of partially observable Markov decision process (POMDP). A POMDP assumes that the environment involves unobservable information, typically, unobservable state variables.

Although RL is a machine learning framework, recent studies (Schultz, Dayan, & Montague, 1997; Waelti, Dickinson, & Schultz, 2001) showed that in a real brain a dopaminergic system including the basal ganglia and the frontal cortex seems to realize a similar learning scheme. Doya (2000b) has suggested that parameters used in RL, which are called 'meta-parameters', may correspond to neuromodulators such as serotonin, noradrenaline and acetylcholine. Thus, the motivation of our study is not only on the machine learning but also on the brain learning.

In RL, an agent is provided by the environment with a scalar reward corresponding to a behavior (action) for each sensory state. The reward indicates instantaneous goodness of the action at the state. The objective of the agent is to maximize the rewards accumulated toward the future, and the maximization is done by improving its strategy to select an action for each state. Such a strategy is called a policy. The estimation and prediction of the accumulated rewards are important for improving the policy. Therefore, a standard RL scheme estimates the reward accumulation which is called the value function.

In order to make a good prediction, it is important to know the dynamics of the environment, i.e. how the current state changes by an action. Model-free RL methods like the actor–critic learning (Barto, Sutton, & Anderson, 1983) and the Q-learning (Watkins & Dayan, 1992) require no model of the environmental dynamics; instead, they try to directly estimate the value function. In contrast, model-based RL methods (Dayan & Sejnowski, 1996; Dearden, Friedman, & Andre, 1999; Doya, 2000a; Matsuno, Yamazaki, Matsuda,

* Corresponding author. Tel.: +81-743-72-5980; fax: +81-743-72-5989. *E-mail address:* ishii@is.aist-nara.ac.jp (S. Ishii).

PII: S0893-6080(02)00056-4

& Ishii, 2001; Moore & Atkeson, 1993; Sutton, 1990) try to model the environmental dynamics and the value function is approximated using the model. Especially when the environment is complicated, e.g. partially observable, a model-based RL has an advantage, because the environmental model can explicitly deal with the complexity. A model-based RL learns faster than a model-free alternate. Our study presents a model-based RL method using the Bayes inference.

If the agent knows the correct optimal value function including the correct estimation of the environmental dynamics, the optimal policy is the one that just selects a 'greedy' action that maximizes the value function at each state. If the estimation and prediction are fairly good, therefore, a good policy is the one that selects a greedy action; this is called exploitation. During the process of trial and error, however, the agent does not know the correct optimal value function. Especially in a POMDP, an approximated value function may be apart from the correct optimal one, due to the uncertain estimation of unobservable state variables. In such a situation, the greedy action is not necessarily optimal. In addition, when the environment changes with time, the value function approximated using the past experiences will not be optimal. In order to know the optimal value function, the agent should execute trial actions, i.e. actions that are not optimal with respect to the current value function; this is called exploration. Since these two strategies, exploitation and exploration, cannot be operated at once, their control has long been an important issue in the control fields (Fe'ldbaum, 1965).

Methods for exploration can roughly be classified into two: undirected exploration methods and directed exploration methods (Thrun, 1992). Undirected exploration methods try to explore the whole state–action space by assigning positive probabilities to all possible actions. For example, semi-uniform (ϵ-greedy) exploration and the Boltzmann exploration (Sutton & Barto, 1998) are undirected methods.

Directed exploration methods use the statistics obtained through the past experiences in order to execute efficient exploration. Kearns and Singh (1998) proposed an exploration algorithm called E^3 algorithm, in which states were classified into known or unknown states based on the visit number. At a known state the agent executes directed exploration under a specific condition, while at an unknown state the agent mainly executes undirected exploration. *R*-max algorithm by Brafman and Tennenholtz (2001) is a modification of the E^3 algorithm so that the agent executes directed 'optimistic' exploration at an unknown state.

Exploration bonus is one popular technique for directed exploration. In Sutton's DYNA system (Sutton, 1990), exploration bonus is added to the immediate reward based on the time period that has passed since the state–action pair was previously experienced. Kaelbling (1993) proposed the interval estimation algorithm using exploration bonus based on the upper bound of the confidence interval for the value function. Moore and Atkeson (1993) also proposed exploration bonus in their learning algorithm called prioritized sweeping. In this method, an unfamiliar state is connected to a fictitious absorbing state with a high value and the agent is encouraged to visit such unfamiliar states. In the method by Dayan and Sejnowski (1996), due to the forgetting effect of the environmental dynamics, the agent comes to try an action that is not optimal with respect to the current estimation of the value function.

We discuss in this paper a new control method of the exploitation–exploration balance. The balance control was also studied by Thrun (1992). Our method is mainly an undirected method, in which the balancing parameter is controlled depending on the current state. Our method also uses exploration bonus. Usher, Cohen, Servan-Schreiber, Rajkowski, and Aston-Jones (1999) has suggested that the exploitation–exploration balance in a real brain may be controlled by noradrenergic neurons in the locus coeruleus (LC) and that the balance may correspond to the level of animal's selective attention. According to this scenario, we will discuss a possible implementation in the brain, which realizes our learning scheme.

Section 2 describes preliminaries to the RL. We propose in Section 3 a Bayes inference method for identifying the current environment. We next propose in Section 4 a control method of the exploitation–exploration balance. An exploration bonus is also introduced in the same section. Section 5 shows computer simulation results. Section 6 discusses a possible implementation in the brain, and Section 7 concludes the paper.

2. Reinforcement learning preliminaries

2.1. Markov decision process

We first consider Markov environments; $P(s'|s,a)$ gives the probability of reaching state s' by selecting action a at state s. If the state-transition probability $P(s'|s,a)$ is known, the value function for state s, $V(s)$, should satisfy the following (optimal) Bellman's equation:

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

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

$Q(s,a)$ is often called the action-value function. The reward function $r(s,a)$ defines the immediate reward for a state–action pair (s,a). The reward function is assumed to be deterministic for simplicity, although the extension to stochastic reward functions is straightforward. $0 \leq \gamma \leq 1$ is a discount constant. The value function defines the summation of the discounted rewards accumulated toward the future. The action-value function $Q(s,a)$ represents the reward accumulation when the agent takes action a at state s and the optimal actions at the subsequent states.