Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 24 (2013) 126 – 136

Procedia Computer Science

17th Asia Pacific Symposium on Intelligent and Evolutionary Systems, IES2013

Overtaking Method based on Variance of Values: Resolving the Exploration–Exploitation Dilemma

Kento Ochi[a], Moto Kamiura[a,b]*

[a]Tokyo Denki University, Graduate School of Science and Engineering, Ishizaka, Hatoyama-cho, Hiki-gun, Saitama, 350-0394, Japan
[b]Research Institute of Electrical Communication, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan

Abstract

The exploration–exploitation dilemma is an attractive theme in reinforcement learning. Under the tradeoff framework, a reinforcement learning agent must cleverly switch between exploration and exploitation because an action, which is estimated as the best in the current learning state, may not actually be the true best. We demonstrate that an agent can determine the best action under certain conditions even if the agent selects the exploitation phase. Under the conditions, the agent does not need an explicit exploration phase, thereby resolving the exploration–exploitation dilemma. We also propose a value function on actions and how to update this value function. The proposed method, the "overtaking method," can be integrated with existing methods, UCB1 and UCB1-tuned, for the multi-armed bandit problem without compromising features. The integrated models show better results than the original models.

Keywords: Reinforcement learning; Exploration–exploitation dilemma; Variance; Overtaking method

1. Introduction

Reinforcement learning is a type of machine learning that is based on the maximization of total reward[1]. A reinforcement learning agent adapts to an environment through behavioral trial and error, which are determined on

* Corresponding author. Tel.: +81-49-296-1069; fax: +81- 49-296-6185.
E-mail address: kamiura@mail.dendai.ac.jp; moto@goo.jp

doi:10.1016/j.procs.2013.10.035

the basis of the policy of the agent. In supervised learning, a system is given a desired output (i.e., supervisory signals). On the other hand, a reinforcement learning agent can know only a reward for each behavior, which is indirect information about the environment. Therefore, policy against the uncertainty of the environment is essential in reinforcement learning.

The uncertainty of the environment manifests as the exploration–exploitation dilemma, which is a decision tradeoff of the agent, i.e., search for a better action (exploration) or take a temporally selected action as the current optimal solution (exploitation). The exploration–exploitation dilemma has been well researched, and many models that address this dilemma have been proposed[2,3,4].

The dilemma is based on inconsistency between explorative action and exploitative action; taking a temporally selected action (exploitation) and searching for a better action (exploration) are incompatible. However, this dilemma is not absolute. When outputs of a value function for all actions are higher than the true values of averages, exploitative action can be consistent with explorative action.

Here we propose a new method based on the above perspective to resolve the dilemma, which we call the Overtaking method. The Overtaking method can be integrated into existing methods. In addition, we show that, when using the proposed method, integrated models demonstrate a better performance.

2. Reinforcement Learning

2.1. Conditions to accomplish best action in exploitation phase

Here we discuss the conditions required to accomplish the best action in the exploitation phase.

In the value function approach of reinforcement learning, given a state s and an action a, the value $Q^{\pi}(s,a)$ for a policy π is defined. There are many types of policies and value functions in existing methods. Greedy is one such policy, in which an agent selects an action for a state subject to the maximum value of Q.

In the simplest sense, given a state s and an action a, a value function can be defined by a conditional expectation,

$$Q(s,a) := E[R \mid s,a],$$

where R is a random variable that represents a reward from the environment. However, it is non-trivial whether an expectation is suitable for a value for an action. Generally, we can construct a value function that is not equal to an expectation.

To simplify, we assume the multi-armed bandit problem in which we can ignore the state of the agent. An action a indicates the selection of the arm a. Let μ_a be an expected reward for an action a, and $Q_{a,n}$ be a value function, where n is the number of times action a is taken. Here the agent never knows the expected rewards. The value $Q_{a,n}$ is not the same as an expectation and is updated by taking action a. Using this situation, we can prove the following theorem.

Theorem 1. *Assume a multi-armed bandit problem in which there are expected rewards* μ_a *for each action* $a \in S_{action}$*, where* S_{action} *is a set of actions for the problem. If a value function* $Q_{a,n}$ *of a reinforcement learning agent with a greedy policy fulfills the following conditions* (1) *and* (2)*, then the agent can select the best action in the exploitation phase, where* $n \in \mathbb{N}$ *is the number of times action* a *is taken.*

$$Q_{a,n} \geq \mu_a \quad \text{for all } a \in S_{action} \text{ and all } n \in \mathbb{N} \tag{1}$$

$$\lim_{n\to\infty} Q_{a,n} = \mu_a \quad \text{for all } a \in S_{action} \tag{2}$$