Procedia Computer Science

Volume 88, 2016, Pages 94–101

7th Annual International Conference on Biologically Inspired
Cognitive Architectures, BICA 2016

A Study of an Indirect Reward on Multi-agent Environments

Kazuteru Miyazaki

National Institution for Academic Degrees and Quality Enhancement of Higher Education, Kodaira, Tokyo, Japan. teru@niad.ac.jp

Abstract

In a multi-agent learning where multiple agents are learning, there is a problem about *an indirect reward* that is how to distribute a reward to an agent that does not obtain a reward directly.We have shown the theorem [3] about "negative effect" of an indirect reward. This paper focuses on the "positive effect" of an indirect reward such as an elimination of *the perceptual aliasing problem* [1]. First, we describe the relationship the theorem [3] and the "positive effect" of the indirect reward. Next, we propose a method to eliminate the perceptual aliasing problem and show the effectiveness of the proposed method by numerical examples.

Keywords: Reinforcement Learning, Multi-agent Learning, Indirect Reward, Direct Reward, Perceptual Aliasing Problem, Profit Sharing

1 Introduction

Among machine-learning approaches, reinforcement learning (RL) focuses most on goal-directed learning from interaction [10]. It is very attractive because it can use dynamic programming (DP) to analyze behavior. RL generally treats, rewards and penalties as teaching signals in learning. DP-based RL involves optimizing behavior under reward and penalties signals designed by RL users on the Markov Decision Processes (MDPs).

RL is difficult to design to fit real-world problems because, first, interaction requires too many trial-and-error searches and, second, no guidelines exist on how to design values of reward and penalty signals. While these are essentially neglected in theoretical researches,they become serious issues in real-world applications, e.g., unexpected results arise if inappropriate values are assigned to reward and penalty signals [4].

We are interested in approaches treating reward and penalty signals independently. We also want to reduce the number of trial-and-error searches by strongly enhancing successful experience — a process known as exploitation-oriented learning (XoL) [4].XoL has four features.(1) XoL learns more quickly by strongly tracing successful experiences. (2) XoL treats, rewards and penalties as independent signals, letting these signals be handled more intuitively and easily than the handling of concrete values. (3) XoL does not pursue optimality efficiently, which can

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016
© The Authors. Published by Elsevier B.V.

doi:10.1016/j.procs.2016.07.411

be acquired by multi-start resetting all memory to get a better policy. (4) XoL is strong in the class that exceeds MDPs because it is a Bellman-free method. An example of XoL learning methods for a type of a reward includes Profit Sharing (PS) [2].

In this paper, we focus on a multi-agent learning [7, 3, 9] where multiple agents are learning. In the multi-agent learning, there is a problem about an indirect reward that is how to distribute a reward to an agent that does not obtain a reward directly. We have been shown the theorem about ”negative effect” of an indirect reward in the paper [3] in order to avoid to obtain no reward in the multi-agent system.

This paper focuses on the ”positive effect” of the indirect reward such as an elimination of *the perceptual aliasing problem* [1]. First, we describe the relationship the theorem [3] and the ”positive effect” of the indirect reward. Next, we propose a method to eliminate the perceptual aliasing problem and show the effectiveness of the proposed method by numerical examples.

2 The Domain

Consider an agent in an unknown environment. After perceiving sensory input from the environment, the agent selects and executes an action. Time is discretized by one input-action cycle. *An action* is selected from among the discrete types.Input from the environment is called *a state*. The discrete types of action is called *the number of actions*. The pair consisting of the state and an action selected in a state is called *a rule*. Rewards and penalties based on a series of actions are provided from the environment, and a reward is given to a state or an action causing transition to a state in which our purpose is achieved, whereas a penalty given to a state or corresponding action in which our purpose is not achieved. In this paper, we consider the cast that there is no penalty.

A rule series that begins from a reward/penalty state or an initial state and ends with the next reward/penalty state is called *an episode*. If an episode contains rules of the same state, but paired with different actions, the partial series from one state to the next is called *a detour*. A rule always existing on a detour is called *an irrational rule*, and otherwise called *a rational rule*. A function that maps states to actions is called *a policy*. The policy with a positive amount of reward acquisition expectations is called *a rational policy*. *The optimal policy* is a policy that can maximize the amount of a reward.

We call indistinction of state values a *type 1 confusion*. Furthermore, we call indistinction of rational and irrational rules a *type 2 confusion*. In general, if there is a type 2 confusion in some sensory input, there is a type 1 confusion in it. By these confusions, we can classify environments. *Q-learning* (QL) [10], that guarantees the acquisition of an optimal policy in MDPs, is deceived by the type 1 confusion since it uses state values to make a policy. PS is not deceived by the confusion since it does not use state values. On the other hand, learning systems that use the weight (including QL and PS) are deceived by the type 2 confusion. If the perceptual aliasing problem occurs, the type 2 confusion may occur. Though there are many researches about the perceptual aliasing problem in *Partially Observable Markov Decision Processes* (POMDPs) [8] it is often eliminated by enriching the sensory input in real applications [6].

An environment in which multiple agents are present is referred to as *a multi-agent environment*. The learning of a multi-agent environment is referred to as *a multi-agent learning*.

This paper focuses on a multi-agent environment that only one type of reward is present and assumes *a synchronous* environment in which only one agent is performing in each time as same as the the paper [3]. Though the paper also assume that there is no type 2 confusion, we do

Download English Version:

https://daneshyari.com/en/article/4962250

Download Persian Version:

https://daneshyari.com/article/4962250

Daneshyari.com