ELSEVIER

Expert Systems with Applications 31 (2006) 351–359

Expert Systems with Applications

www.elsevier.com/locate/eswa

Maximum reward reinforcement learning: A non-cumulative reward criterion

K.H. Quah, Chai Quek *

Centre for Computational Intelligence, School of Computer Engineering, Nanyang Technological University, Blk N4 #2A-32, Nanyang Avenue, Singapore 639798

Abstract

Existing reinforcement learning paradigms proposed in the literature are guided by two performance criteria; namely: the expected cumulative-reward, and the average reward criteria. Both of these criteria assume an inherently present cumulative or additivity of the rewards. However, such inherent cumulative of the rewards is not a definite necessity in some contexts. Two possible scenarios are presented in this paper, and are summarized as follows. The first concerns with learning of an optimal policy that is further away in the existence of a sub-optimal policy that is nearer. The cumulative rewards paradigms suffer from slower convergence due to the influence of accumulating the lower rewards, and take time to fade away the effect of the sub-optimal policy. The second scenario concerns with approximating the supremum values of the payoffs of an optimal stopping problem. The payoffs are non-cumulative in nature, and thus the cumulative rewards paradigm is not applicable to resolve this. Hence, a non-cumulative reward reinforcement-learning paradigm is needed in these application contexts. A maximum reward criterion is proposed in this paper, and the resulting reinforcement-learning model with this learning criterion is termed the *maximum reward reinforcement learning*. The maximum reward reinforcement learning considers the learning of non-cumulative rewards problem, where the agent exhibits a maximum reward-oriented behavior towards the largest rewards in the state-space. Intermediate lower rewards that lead to sub-optimal policies are ignored in this learning paradigm. The maximum reward reinforcement learning is subsequently modeled with the FITSK-RL model. Finally, the model is applied to an optimal stopping problem with a nature of non-cumulative rewards, and its performance is encouraging when benchmarked against other model.

Keywords: Maximum reward; Reinforcement learning; Non-cumulative reward; FITSK-RL; Optimal stopping problem; Financial derivative pricing; Two-cycle task; Maximum reward-oriented behaviour

1. Introduction

Reinforcement learning is a trial and error learning method, where the agent tries to explore and maximize the numerical state values from the feedback it receives over the long run. The process that the agent follows is modeled as the *Markov decision process* (MDP). In the MDP context, the two most well-studied optimality criteria are the expected cumulative reward and the average reward criteria (Puterman, 1994). The expected cumulative reward criterion has been adopted as the de facto objective measure in reinforcement learning research (Bertsekas, & Tsitsiklis, 1996; Kaelbling, Littman & Moore, 1996; Rummery & Niranjan, 1994; Sutton, 1984, 1988; Sutton & Barto, 1998; Watkins, 1989). This criterion employs a technique to exponentially discount the future rewards, such that the near-future rewards are more valuable than far-future rewards. However, some research has demonstrated that the expected cumulative reward criterion is unsuitable for some problems, particularly in the undiscounted reward domain (Schwartz, 1993). Some reinforcement learning methods based on the average reward criterion have been proposed to resolve the undiscounted reward reinforcement learning problem (Mahadevan, 1996; Schwartz, 1993; Tadepalli & Ok 1998). In the essence of this, both classes of methods employ some form of cumulative rewards criteria; the former with a reward discounting approach such that their infinite cumulative rewards are finite, and the latter with a reward averaging approach such that both short and long-term rewards are indistinguishable.

However, a central issue is a problem of whether such a cumulative rewards formulation is a definite necessary requirement in all-learning contexts. Is there any problem that the cumulative rewards formulation is unnecessary, and unjustifiable? This paper discusses two possible scenarios

* Corresponding author. Tel.: +65 6790 4926; fax: +65 6792 6559.
E-mail address: ashcquek@ntu.edu.sg (C. Quek).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.054

where the cumulative reward formulation is to be avoided for superior learning result. This is demonstrated in Sections 3 and 4. In Section 3, a two-cycle task from (Mahadevan, 1996) is discussed. One of the cycle leads to a sub-optimal policy and the other cycle leads to an optimal policy, but the rewards of the optimal policy is further away. Therefore, the focus of the learning is on how to bypass the sub-optimal policy in the search for an optimal policy. The classical Q-learning (Watkins, 1989) was demonstrated to suffer from this problem because it was being dragged down by the sub-optimal policy, causing it to be lagging behind from switching to a better policy. In Section 4, an optimal stopping problem arising from financial derivative pricing (Tsitsiklis & Van Roy, 1999) is discussed. This is a case where the cumulative rewards formulation is unjustifiable. On the other hand, this problem is concerned with finding the supremum of the payoff values of each state, which is a function of the expected discounted maximum rewards in a non-cumulative manner. Furthermore, a discounted reward is significant in this domain as the discounting operation represents a compensation of future rewards with respect to the drift rate of financial pricing. This represents a contrary argument against the claim by (Schwartz, 1993) that discounting of rewards is not necessarily important. Having these, a reinforcement learning formulation with non-cumulative *maximum reward* criterion is proposed to accommodate the learning needs of such tasks.

The maximum reward reinforcement-learning agent exhibits a maximum reward-oriented behavior. It attempts to search for the policy that leads to the largest reward; irregardless of intermediate lower rewards along the path. This positions it to have superior immunity to the existence of sub-optimal policies in the problem domain, such as the two cycles problem in Section 3. Furthermore, the maximum reward learning is demonstrated to be effective in approximating the future payoff in a financial pricing context. The learning system exhibits an asymptotic behavior of approximating the supremum value with the maximum reward formulation, as will be discussed in Section 4.

This paper is organized as follows. Section 2 describes the general formulation of the maximum reward reinforcement learning and its update formulae as part of the modular design of the earlier proposed generic reinforcement-learning framework from (Quah & Quek, submitted for publication). Section 3 studies the learning behavior of the maximum reward reinforcement learning using a two-cycle task, and compares against two existing reinforcement-learning models. Section 4 applies the proposed maximum reward reinforcement learning with its generic learning framework to an optimal stopping problem arising from the financial derivative pricing context. Section 5 concludes this paper.

2. Maximum reward reinforcement learning

This section formulates the maximum reward reinforcement learning. Section 2.1 provides some basic definitions from the existing research. Section 2.2 mathematically formulates the maximum reward reinforcement learning, discusses the three existing reinforcement-learning models, and the equivalent form of these three models for the maximum reward reinforcement learning. Section 2.3 solves the modelling problem of the maximum reward reinforcement learning. This is made possible by mapping the maximum reward error update formula to a generic reinforcement-learning framework to inherit the modularity and functionality of the framework.

2.1. Background definitions

Consider a sequence of an infinite horizon *Markov decision process* (MDP) (Puterman, 1994) with discrete time step $t = 0,1,2,3,\ldots,\infty$. At each discrete time step, the agent receives some observable environment's states, $S_t \in S$, where S is the set of possible states. The agent implements a mapping from the state to probabilities of selecting each possible action, $a_t \in A(s_t)$, where $A(s_t)$ is the set of actions available in state s_t. The mapping is called the agent's policy, π_t, and is depicted by Eq. (2.1).

$$\text{Policy mapping;} \quad \pi_t(s,a) = \Pr(a_t = a | s_t = s) \tag{2.1}$$

indicates the probability of choosing action a when the perceived state is s at time t. Due to the action a_t, the agent receives a numerical reward r_{t+1}, and observes a new state s_{t+1} at time $t+1$.

The goal of the learning algorithm is to maximize the successive total reward, or the *return* value as described in Eq. (2.2).

$$\text{Return;} \quad R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2.2}$$

The discount rate, γ, determines the present value of future rewards, a reward received k time steps in the future is worth only γ^{k-1} times what it would be worth if it were received immediately. If $\gamma = 0$, then the agent's actions only influence the immediate reward, r_{t+1}, as in most supervised learning agent. As γ approaches 1, the agent will take future rewards more seriously, or become more farsighted.

2.2. Formulation of maximum reward reinforcement learning

This section defines a possible variation of return value, and extends the definition to various Bellman's optimality formulae. This is followed by the formulation of the maximum reward form of temporal difference learning models in reinforcement learning.

The return value is defined as the cumulative future rewards in Eq. (2.2) where the agent is learning to maximize the return value. However, if the objective of the agent is to identify the largest reward within the value space, then the return value formulae in Eq. (2.2) can be suitably modified into Eq. (2.3).

$$R_t = \max(r_{t+1}, \gamma r_{t+2}, \gamma^2 r_{t+3}, \ldots) = \max_{\forall k \in [0,\infty]} (\gamma^k r_{t+k+1}) \tag{2.3}$$

The maximum reward agent will attempt to discover the shortest path that lead to the highest reward, without