Neurocomputing 166 (2015) 244–255

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Data-driven heuristic dynamic programming with virtual reality

Xiao Fang [a,b], Dezhong Zheng [a], Haibo He [b,*], Zhen Ni [b]

[a] Institute of Electrical Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China
[b] Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881, USA

ARTICLE INFO

Article history:
Received 15 August 2013
Received in revised form
9 April 2015
Accepted 9 April 2015
Communicated by T. Heskes
Available online 20 April 2015

Keywords:
Machine learning
Reinforcement learning
Adaptive critic
Goal representation heuristic dynamic programming
Virtual reality

ABSTRACT

In this paper, we propose a virtual reality (VR) platform as a case study of machine learning, in this case applied to the goal representation heuristic dynamic programming (GrHDP) approach. In general, a VR platform normally includes a physical module, a control/learning module, and a VR module. It facilitates machine learning research, where scientists and engineers can participate in the simulation process to analyze dynamic experiments. The internal structure of the VR platform can be replaced according to different research targets, so the platform can be extended to other applications. In this paper, we present the detailed VR design strategy, with a number of applications, including a triple-link inverted pendulum balancing problem, a maze navigation problem, and a robot navigation with obstacle avoidance.

1. Introduction

Recent research on intelligent systems focuses on developing self-adaptive systems to imitate the signal processing mechanism of biological brain organisms for closing the gap between human behavior and computer decisions [1–4]. Generally, an intelligent system should be designed with the following abilities:

- to acquire and retain knowledge through the perception of environments;
- to analyze and predict actions from prior experiences;
- to produce and adjust the actions to obtain consistent and reliable results over time [3–6].

There are two important issues to be considered regarding intelligent systems. The first is how to develop biologically inspired general purpose learning models that can dynamically learn how to achieve a goal [4,7–9], and the second is how to represent data information to support the decision-making process, in an efficient and effective way [10–13].

In the literature, adaptive dynamic programming (ADP) [14–17] provides a powerful learning mechanism [5], which has been successfully applied in many fields, such as dynamic control systems [8,16], power systems [18,19], missile systems [20], and communication systems [21] as well as partial observable Markov decision process [22]. The ADP method can be categorized as heuristic dynamic programming (HDP) [16], dual heuristic dynamic programming (DHP), and globalized dual heuristic dynamic programming (GDHP) [23]. The foundation of ADP can be traced back to the classic Bellman's principle of optimality [24]. It is related to the reinforcement learning (RL) while using the adaptive-critic (AC) design framework [25–27]. For example, in model-free direct HDP structure, two neural networks are used in order to approach on-line learning. The action network interacts with the external environment in order to produce actions according to the state vector observation. The critic network evaluates the performance by means of an external reinforcement signal. The weights of both networks are randomly initialized. The optimal solution is obtained by training the weights using back-propagation to approximate the optimal cost-to-go function [8,28]. Such model-free technique has also been demonstrated in DHP design successfully in the community [29].

Based on the direct HDP design [16], an enhanced approach, called goal representation heuristic dynamic programming (GrHDP), was proposed in [4,7]. In the GrHDP model, an additional network, called goal network, was integrated in the framework in order to interact with the critic network [7,30]. The motivation of the goal network is to provide an internal reinforcement representation to help on-line learning and optimization [8,31]. The goal network learns from an external reinforcement signal $r(t)$ and adaptively generates an internal reinforcement signal $s(t)$ to guide system behavior [8,30,31]. The internal reinforcement signal is also used to feed into the critic network. Since a continuous value is

* Corresponding author.
E-mail addresses: xfang@ele.uri.edu (X. Fang), qhdzdz@sina.com (D. Zheng), he@ele.uri.edu (H. He), ni@ele.uri.edu (Z. Ni).

http://dx.doi.org/10.1016/j.neucom.2015.04.014
0925-2312/

provided, the internal reinforcement signal can be considered more informative. This additional input will contribute to the fine-tuning of the critic network. Also, this internal reinforcement signal can be automatically and efficiently adjusted in order to provide a learning performance improvement in the goal network. Meanwhile, the GrHDP approach also saves the previous cost-to-go value in order to enable on-line learning, association, and optimization over time [7]. More recently, the corresponding goal representation dual heuristic programming (GrDHP) design is also investigated and developed on several control examples with promising results [32,33].

During the simulation design, virtual reality (VR) provides a straightforward way for the construction of this solution [34,35]. In general, VR is a powerful technique, which might be implemented through a set of complex hardware devices, such as a mainframe computer, a head-mounted display (HMD), motion trackers, sensor gloves, or either a computer automatic virtual environment (CAVE) [35,36]. In fact, the key of VR is not related just to the hardware devices, but also to the human–computer interaction (HCI) involving the people participating and exploring a virtual environment (VE). Recently, many researchers focused on the design of visualization/interaction applications instead of the improvement of VR rendering algorithms or hardware devices [37]. In psychology, VE is used to treat patients suffering from anxiety or phobias [38,39]. In education, a virtual learning environment was developed to help students learn their courses [40,41]. In urban traffic research, VE has been used to analyze the interactive driving behaviors under different traffic scenarios [42,43]. In military research, VE simulations have been developed to improve the soldiers' skills on decision-making [44] and the cross-cultural communication under different situations [45]. In surgery, a VR simulator was proposed as a cost-effective training platform [46,47]. In physical rehabilitation, a VE has been used to help the patients to train their movement patterns and enhance their physical rehabilitation [48,49].

There also seems to be an increasing need for visualization and VR platforms in the machine intelligence research, with the increasing dimension and volume of data information. To this end, a VE must be developed, using computer graphics (CG) software, such that researchers are able to design a virtual experiment adequate to their research tasks. Second, real data information can be used in the VE, making it a suitable representation for the real world. Third, during the dynamic simulation process, the researchers can interact with the experiment through the VE, analyzing the system behavior and possibly proposing changes to the experiment. For example, an additional virtual disturbance might be designed to verify the stability of a dynamic system.

Motivated by our previous works on ADP designs [7,8,30] and its applications using VR [50–52], the goal of this paper is to propose the integration of VR in the development of machine learning solutions. Specifically, we built an interactive VR platform and applied it to GrHDP experiments in order to illustrate how the use of a VR platform can enhance the development of a machine learning application. We first apply VR to GrHDP in the triple-link inverted pendulum balancing problem [7] and the maze navigation problem [30,31], and then develop a new application benchmark, a robot navigation with obstacle avoidance problem. We hope to demonstrate how the combination of machine learning research and VR improves investigation of many real-world applications.

The rest of this paper is organized as follows. In Section 2, we discuss the structure of the GrHDP approach. The design of a VR interactive platform is described in detail in Section 3. Based on this platform, in Sections 4, 5 and 6, we study, respectively, the design and simulation of experiments for the triple-link inverted pendulum balancing problem, the maze navigation problem, and the robot navigation with obstacle avoidance problem. Finally, the conclusion and discussion is provided in Section 7.

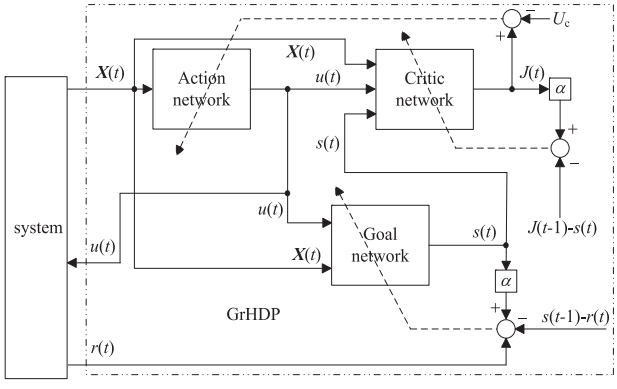

Fig. 1. The structure of GrHDP. An additional network (goal network) is integrated into the direct HDP structure, interacting with both the critic network and the action network.

2. Online learning with the GrHDP approach

The main structure of GrHDP is given in Fig. 1. When compared to other ADP approaches [16], in GrHDP, an additional network (goal network) is integrated into the direct HDP structure, interacting with both the critic network and the action network. All the neural networks are multi-layer perceptrons (MLP) with one hidden layer [50]. The input of goal network and critic network could be defined as $x_g=[X,u]$ and $x_c=[X,u,s]$, respectively. Meanwhile, the input of action network is still the current state vector [30]. The objective of the goal network is to represent an internal reinforcement signal and to approximate the discounted total future reward. The internal reinforcement signal $s(t)$ can be defined as

$$s(t)=r(t+1)+\alpha r(t+2)+\alpha^2 r(t+3)+\cdots \quad (1)$$

where $r(t+1), r(t+2), r(t+3),\ldots$ are the external reinforcement signals at time $t+1, t+2, t+3$, respectively; α is a discount factor ($0<\alpha<1$). In this paper, we used $\alpha=0.95$ in our implementations. Meanwhile, the critic network's purpose is to approximate the discounted total future internal reinforcement signal $s(t)$ to the cost function J. The cost function at time t could be written as

$$J(t)=s(t+1)+\alpha s(t+2)+\alpha^2 s(t+3)+\cdots \quad (2)$$

Comparing with the direct HDP design, the optimization error function and the learning process in the goal network and in the critic network are different: the error function of goal network is related to the primary reinforcement signal $r(t)$, while the error function of the critic network is related to the internal reinforcement signal $s(t)$ [7]. Thus, the error function of the goal network can be written as

$$e_g(t)=\alpha J(t)-[J(t-1)-r(t)] \quad (3)$$

$$E_g(t)=\tfrac{1}{2}e_g^2(t) \quad (4)$$

The error function of the critic network can be defined as

$$e_c(t)=\alpha J(t)-[J(t-1)-s(t)] \quad (5)$$

$$E_c(t)=\tfrac{1}{2}e_c^2(t) \quad (6)$$

The action network is similar to the one in the direct HDP approach. The objective of action network is to indirectly back-propagate the error between the ultimate utility function U_c and the cost function J. Therefore, the error function of action network can be written as

$$e_a(t)=J(t)-U_c(t) \quad (7)$$