Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 161 (2010) 578–595
FUZZY sets and systems
www.elsevier.com/locate/fss

Exploration and exploitation balance management in fuzzy reinforcement learning

Vali Derhami[a], Vahid Johari Majd[a,*], Majid Nili Ahmadabadi[b,c]

[a]*Intelligent Control Systems Laboratory, School of Electrical Engineering, Tarbiat Modares University, P.O. Box 14115-143, Tehran, Iran*
[b]*Control and Intelligent Processing Center of Excellence, University of Tehran, Tehran, Iran*
[c]*School of Cognitive Science, Institute for Research in Fundamental Sciences, Tehran, Iran*

Received 17 March 2008; received in revised form 13 May 2009; accepted 13 May 2009
Available online 28 May 2009

Abstract

This paper offers a fuzzy balance management scheme between exploration and exploitation, which can be implemented in any critic-only fuzzy reinforcement learning method. The paper, however, focuses on a newly developed continuous reinforcement learning method, called fuzzy Sarsa learning (FSL) due to its advantages. Establishing balance greatly depends on the accuracy of action value function approximation. At first, the overfitting problem in approximating action value function in continuous reinforcement learning algorithms is discussed, and a new adaptive learning rate is proposed to prevent this problem. By relating the learning rate to the inverse of "fuzzy visit value" of the current state, the training data set is forced to have uniform effect on the weight parameters of the approximator and hence overfitting is resolved. Then, a fuzzy balancer is introduced to balance exploration vs. exploitation by generating a suitable temperature factor for the Softmax formula. Finally, an enhanced FSL (EFSL) is offered by integrating the proposed adaptive learning rate and the fuzzy balancer into FSL. Simulation results show that EFSL eliminates overfitting, well manages balance, and outperforms FSL in terms of learning speed and action quality.

Keywords: Reinforcement learning; Decision analysis; Fuzzy control; Exploration; Exploitation

1. Introduction

Reinforcement learning (RL) is a method that uses a scalar reinforcement signal or reward to train an agent in a complex and non-deterministic environment without supervisor [2,18]. To achieve better action quality, agents should generate actions such that they explore environment suitably, and yet exploit their experiences to avoid punishment. Because of these conflicting objectives, balancing exploration and exploitation is an important issue in RL [6,18,24].

Depending on the type of the problem and the aim of learning, different authors assessed the quality of balance in the terms of the number of successes [16,22,23], the learning time-period [22], and the number of failures [16]. However, they did not offer any comprehensive balance measure that includes the effective parameters in balance.

In continuous fuzzy reinforcement learning (FRL), fuzzy inference system is used to obtain an approximate model for the value function in continuous state space and to generate continuous actions [2,10,24]. Critic-only architecture

* Corresponding author. Tel./fax: +98 21 8288 3353.
E-mail addresses: vahid_majd@yahoo.com, majd@modares.ac.ir (V.J. Majd)
URL: http://www.modares.ac.ir/eng/majd.

0165-0114/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2009.05.003

is one widely used architectures in continuous FRL, whose adjustable parameters are updated at each time-step, which allows to have more effective on-line learning. A critic-only based FRL uses only a fuzzy system for approximating action value function (AVF), and generates action with probability proportional to this function; hence, such an action selection strategy allows possible implementation of balance.

Some authors have implemented Q-learning method with linear function approximators using fuzzy systems [5,8,15]. This critic-only FRL algorithm, called fuzzy Q-learning (FQL), was applied to some problems with continuous state and action spaces [4,5,8]. However, FQL is a heuristic method, lacks mathematical analysis, and has the possibility of divergence [1,21]. In the FQL presented in [8], a mixed strategy by combination of Softmax formula and a direct strategy were used for action selection in each rule. The authors in [5,15] have used Softmax formula or ε-greedy method for action selection in each rule, where the temperature factor or ε gradually decreases as a function of episode number. Such methods do not generate final actions according to a continuous probability function; hence small changes in the value of action may cause substantial changes in the control behavior, which may have an impact on the convergence in continuous RL [14]. In fact, in [4,5,8,15], the usual action selection methods in discrete (standard) RL have been used for the continuous case, and no method was offered to improve balance in continuous case.

Fuzzy Sarsa learning (FSL) was first introduced in [3] based on linear Sarsa, and the existence of stationary points was established for it. FSL is the first critic-only FRL with mathematical analysis. This algorithm tunes the parameters of conclusion parts of the fuzzy system online. The experimental results in [3] signify higher learning speed and action quality for FSL compared to FQL. In each rule of FSL, actions are selected according to a modified Softmax formula so that the final inferred action selection probability becomes equivalent to the standard Softmax formula with continuous distribution. As we will show in this paper, FSL algorithm has a suitable potential for balance, and thus we focus on this type of critic-only FRL method.

In discrete RL, the number of states and actions is finite and countable, and the values of states (or state–action pairs) are saved in a value table whose elements are adjusted independently [19]; whereas in continuous RL, the number of states and actions is infinite, and function approximators are used to approximate value function. In this case, changing an approximator parameter may cause changes in the approximate values of the entire space. Considering these differences, the available balance management methods in discrete RL [12,16,22] cannot be directly used or do not improve balance in continuous case. The questions that may arise in this issue are: How to obtain the number of visits of each state in continuous case? How to count the number of selections of each action in continuous case? Moreover, since changes in an approximator parameter may cause changes in the approximate values of the entire space, how can one consider this matter in balance management?

A suitable strategy in RL algorithms is to have higher exploration and lower exploitation at the early stage of learning, and then to decrease exploration and increase exploitation gradually. Before exploiting, however, adequate exploration and accurate estimate of AVF should have been achieved. Detecting adequate exploration and achieving an accurate estimation of AVF are two serious challenges in FRL. Although in discrete RL, having longer exploration leads to more accurate AVF, it may not be the case in continuous RL due to previously mentioned differences between their algorithms. In fact, in continuous RL, the approximation accuracy of AVF depends on the distribution of data.

In off-line learning approaches, the training data set is selected from the entire problem space with uniform distribution. However, since RL is an on-line learning method, data are acquired as the agent interacts with the environment, and thus, data distribution depends on the states that the agent visits. If the training data for the approximator does not have a uniform distribution in the problem space, the weight parameters of the approximator may overfit for a part of problem space and produce large errors for other parts. Overfitting results in an inaccurate AVF, and consequently, it does not allow a proper balance management.

As a practical matter in RL problems, learning rate is gradually decreased as a function of time [15,16,18]. In the FRL given in [20], for every rule, a separate adaptive learning rate is provided to adjust the parameters of the rule. The learning rate for each rule is calculated only based on the firing strength of that rule and independent of other rules. These methods [15,16,18,20] cannot prevent overfitting, because they do not have any mechanism to decrease the influence of the training data of a part of problem space that the agent has visited more frequently than other parts.

In this paper, we will focus on balance in continuous RL, will extend available concepts for balance from discrete RL to continuous case, will discuss the interaction between approximate AVF and action selection, and will define suitable signals that can be used as inputs of a balancer in continuous case. We will propose a new adaptive learning rate to eliminate overfitting problem, and to increase the accuracy of AVF approximation. To the best of our knowledge, this is first time that overfitting is comprehensively discussed and a solution to prevent it, is provided in continuous RL.