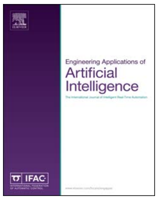
Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Lyapunov theory based stable Markov game fuzzy control for non-linear systems

Rajneesh Sharma

Netaji Subhas Institute of Technology, New Delhi, India

ARTICLE INFO

Article history:
Received 12 February 2016
Received in revised form
4 May 2016
Accepted 17 June 2016

Keywords:
Reinforcement learning
Q learning
Fuzzy Markov games
Lyapunov theory
Inverted pendulum
Two link robotic manipulator
SCARA

ABSTRACT

In this paper we propose a Lyapunov theory based Markov game fuzzy controller which is both safe and stable. We attempt to optimize a reinforcement learning (RL) based controller using Markov games, simultaneously hybridizing it with a Lyapunov theory based control for stability. Proposed technique generates in an RL based game theoretic, adaptive, self learning, optimal fuzzy controller which is both robust and has guaranteed stability. Proposed controller is an "annealed" hybrid of fuzzy Markov games and the Lyapunov theory based control. Fuzzy systems have been employed as generic function approximators for scaling the proposed approach to continuous state-action domains. We test our proposed controller on three benchmark non-linear control problems: (i) inverted pendulum, (ii) trajectory tracking of standard two-link robotic manipulator, and (iii) tracking control of a two link selective compliance assembly robotic arm (SCARA). Simulation results and comparative evaluation against baseline fuzzy Markov game based control showcases superiority and effectiveness of the proposed approach.

1. Introduction

Reinforcement learning (RL) paradigm centers on Markov Decision Processes (MDP) as the underlying model for adaptive optimal control of non-linear systems (Busoniu et al., 2010; Wiering and van Otterlo, 2012). A critical assumption in the MDP based RL technique is the assumption of a stationary environment. However, imposing such a restrictive assumption on the environment may not be feasible, especially when the controller has to deal with disturbances and parametric variations.

Notwithstanding this limitation, RL has been used successfully for controlling a wide variety of non linear systems, e.g., in Kobayashi et al. (2009) a meta-learning method based on temporal difference has been employed for inverted pendulum control (IPC); Kumar et al. (2012) presents a self tuning fuzzy Q controller for IPC; Ju et al. (2014) proposes kernel based approximate dynamic programming approach for inverted pendulum control, and in Liu et al. (2014) an experience replay least squares policy iteration procedure has been proposed for efficient utilization of experiential information. In literature, we can find quite a few variants of the inverted pendulum problem. However, in our work, we have used standard version of the pendulum wherein the pivot point is mounted on a cart which can move horizontally.

Another domain where RL has been applied is robotic manipulator control, which is a highly coupled, non linear and time varying task. The task becomes even more challenging when the controller has to cope with varying payload mass and external disturbances. Both, neural network based RL and fuzzy systems based RL have been employed for robotic manipulator control. In Lin (2009) authors have used an H_∞ reinforcement learning based controller on a fuzzy wavelet network (FWN).They implement an actor-critic RL formulation avoiding complex Ricatti equations for controlling SCARA. An adaptive neural RL control has been proposed in Tang et al. (2014) to counter unknown functions and dead zone inputs in an actor-critic RL configuration, wherein Lyapunov theory has been employed to show boundedness of all closed loop signals. For a comprehensive and in-depth look on controllers employing soft computing techniques, e.g., neural networks, fuzzy systems and evolutionary computation on robotic manipulators; we refer the reader to (Katic and Vukobratovic, 2013).

As stated earlier, all RL based controller design approaches share a basic lacuna that they assume an MDP framework. To make RL controller design process more general and robust, we introduced a Markov game formulation wherein the noise and disturbance are viewed as an "opponent" to the "controller" in a game theoretic setup (Sharma and Gopal, 2008). This formulation helped us in designing RL controllers that are robust in handling disturbances and noise as the controller always tries to optimize against the "worst case" opponent or noise. Markov Game formulation (Sharma and Gopal, 2008) allows broadening of the MDP

E-mail address: rajneesh496@gmail.com

http://dx.doi.org/10.1016/j.engappai.2016.06.008
0952-1976/© 2016 Elsevier Ltd. All rights reserved.

based RL control to encompass multiple adaptive agents with competitive goals. Markov game controller was able to deal with disturbances and parameter variations of the controlled plant. However, both MDP and Markov games based RL approaches failed to address one key concern, namely, stability of the designed controller.

To be specific, there is no guarantee that the controller will remain stable in presence of disturbances and/or parameter variations. Our attempt herein is to design self learning, model free controllers with guaranteed stability. This is sought to be achieved by incorporating a Lyapunov theory based action generation mechanism in the game theoretic RL setup. The controller has all the advantages of game based RL (Markov game control) and has guaranteed stability due to inclusion of a Lyapunov theory based action.

This work is motivated by a need for addressing the stability issue in RL based control by proposing a 'safe and stable' game theoretic controller. The controller is safe as it uses a Markov game framework for optimization; controller always optimises against the worst opponent or plays 'safe' as referred to in the game theory literature (Vrabie and Vamvoudakis, 2013). In the proposed approach Markov game based 'safe' policy is hybridized with a Lyapunov theory based 'stable' policy for generating a 'safe and stable' policy. This hybridization is carried out in an 'annealed' or gradual manner for arriving at a safe and stable game theoretic control.

Robotic manipulators (Katic and Vukobratovic, 2013) are highly coupled, non linear and time varying uncertain systems. Furthermore, industrial robotic manipulators are employed for picking up and releasing objects or they have to deal a with varying payload mass. This presents a highly challenging and complex task for testing our proposed approach. We test our approach on two degrees of freedom (DOF) robotic manipulators as they capture all the intricacies of a six DOF manipulator and are computationally less expensive. We employ the approach on two robotic arms, i.e., a standard two link robot arm and a SCARA.

Proposed controller belongs to the class of self learning/adaptive systems with roots in Machine Learning (Wiering and van Otterlo, 2012). In contrast to other Artificial Intelligence based and conventional controllers, RL based controllers do not assume access to desired response or trajectory. Proposed controller neither assumes knowledge of desired response nor system model. The controller discovers optimal actions by repeated trial and error interactions with the systems/plant it intends to control. It has access to only a heuristic reinforcement signal emanated by the plant telling the controller whether the action taken by it is "good" or "bad". This makes control task a very challenging one. The advantage is that the designed controller is a self learning, adaptive and is suitable for controlling an unknown system.

The paper is structured as follows: a systematic presentation of the RL approaches that lead to the formulation of the proposed methodology is presented in Section 2. Formulation of Lyapunov theory based stable Markov game fuzzy controller for the three tasks: a) inverted pendulum b) Two link robotic manipulator and c) SCARA, along with simulation models and parameters thereof, have been described in Section 3. Section 4 presents simulation results and comparative evaluation of Lyapunov Markov game fuzzy control against baseline fuzzy Markov game control for the three problems. Section 5 summarises the paper and outlines scope for future work.

2. Lyapunov theory based Markov game fuzzy approach

To facilitate reader understanding of the proposed approach, we give a brief description of some relevant RL approaches.

2.1. Reinforcement learning algorithms

Reinforcement Learning is an online learning paradigm wherein the learning agent's goal is to adapt its behavior to maximize/minimize a cumulative reward/cost obtained from the environment (Busoniu et al., 2010). The key feature that sets RL apart from other Artificial Intelligence based techniques is its extremely goal-oriented nature, and ability to sacrifice short term gains for long term benefits.

There are various ways for designing an RL based controller (Wiering and van Otterlo, 2012). However, in principle, they can be broadly classified as a) model based, and b) model free. In model based RL an explicit model of the system is constructed while in model free RL the model is built impromptu, when the agent attempts to control the system. Herein, we briefly describe the model free RL approach of Q learning (Busoniu et al., 2010). For other RL approaches the reader is referred to (Wiering and van Otterlo, 2012).

2.1.1. Q learning

At every time stage k, an adaptive agent (controller) chooses an action a^k to be applied in current state s^k. The agent then receives a reinforcement signal r^k from the environment and the environment transitions to the next state s^{k+1} under action a^k, as chosen by the agent. Transition from state s^k to s^{k+1} is made as per the underlying state transition probability $p(s^k, s^{k+1})$. Agent's aims to find the optimal policy π^k: $\pi^k(s^k) \rightarrow a^k; a^k \in A(s^k)$ so as to minimize expected sum of discounted cost, i.e., $E\left\{ \sum_{j=0}^{\infty} \lambda^j r^{k+j} \right\}$ where r^{k+j} is the cost incurred j steps into future and λ is the discount factor; $0 \le \lambda < 1$.

In Q learning (Wiering and van Otterlo, 2012), Q-value defines the quality of a state-action pair, and is the total expected discounted cost incurred by a policy, that takes action $a^k \in A(s^k)$ in state $s^k \in S$ and follows the optimal policy in the subsequent states. Q values implicitly contain information regarding transition probabilities. For the state-action pair (s^k, a^k), the Q-value is defined as:

$$Q\left(s^k, a^k\right) = r^k + \lambda \sum_{s^{k+1} \in S} p\left(s^k, a^k, s^{k+1}\right) V\left(s^{k+1}\right) \tag{1}$$

where

- $V(s^{k+1}) = \min_{a \in A(a^{k+1})} Q(s^{k+1}, a)$ is the state value
- $r^k =$ Immediate cost of taking action a^k at state s^k.

Q-values for every state-action pair can be evaluated by considering (1) as an update rule in an iterative manner. In some RL domains where system model is not exactly known, implying that the transition-probabilities $p(s^k, a^k, s^{k+1})$ are unknown, this can't be implemented. Watkins (Wiering and van Otterlo, 2012) generalized the above Eq. (1), doing away with the need of an explicit system model, either in the form of cost structure or transition probabilities:

$$Q\left(s^k, a^k\right) \leftarrow Q\left(s^k, a^k\right) + \alpha \left\{ r^k + \lambda \min_{a \in A\left(s^{k+1}\right)} Q\left(s^{k+1}, a\right) - Q\left(s^k, a^k\right) \right\} \tag{2}$$

where, α is the learning rate parameter, $0 \le \alpha < 1$; it governs the degree to which newly acquired information supersedes the earlier information.

Q-learning is guaranteed to converge to optimal Q-values provided each state-action pair is visited infinitely often and the learning rate parameter is reduced in a gradual manner. Q learning can be extended to continuous state-action space problems by