Journal of Process Control 69 (2018) 16–29

Contents lists available at ScienceDirect

Journal of Process Control

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

Control of a bioreactor using a new partially supervised reinforcement learning algorithm

B. Jaganatha Pandian, Mathew Mithra Noel*

School of Electrical Engineering, VIT University, India

ARTICLE INFO

Article history:
Received 21 January 2018
Received in revised form 13 July 2018
Accepted 19 July 2018

Keywords:
Machine learning
Reinforcement learning
Neural networks
Nonlinear control
Bioreactor control
Interacting multiple tank control

ABSTRACT

In recent years, researchers have explored the application of Reinforcement Learning (RL) and Artificial Neural Networks (ANNs) to the control of complex nonlinear and time varying industrial processes. However RL algorithms use exploratory actions to learn an optimal control policy and converge slowly while popular inverse model ANN based control strategies require extensive training data to learn the inverse model of complex nonlinear systems. In this paper a novel approach that avoids the need for extensive training data to construct an exact inverse model in the inverse ANN approach, the need for an exact and stable inverse to exist and the need for exhaustive and costly exploration in pure RL based strategies is proposed. In this approach an initial approximate control policy learnt by an artificial neural network is refined using a reinforcement learning strategy. This Partially Supervised Reinforcement Learning (PSRL) strategy is applied to the economically important problem of control of a semi-continuous batch-fed bioreactor used for yeast fermentation. The bioreactor control problem is formulated as a Markov Decision Process (MDP) and solved using pure RL and PSRL algorithms. Model based and model-free RL control experiments and simulations are used to demonstrate the superior performance of the PSRL strategy compared to the pure RL and inverse model ANN based control strategies on a variety of performance metrics.

1. Introduction

Control of bioreactors is a challenging problem of great economic significance. The inherent nonlinearity of biological processes precludes the use of traditional linear control strategies. Thus exploration of alternate nonlinear and optimal strategies is of interest. In recent years Reinforcement Learning (RL) based control strategies have been successfully applied to the control of outstanding nonlinear control problems not amenable to the techniques of classical control [1,2]. Recent years have also witnessed a synergetic growth of the fields of machine learning and nonlinear control leading to the application of traditional artificial intelligence approaches to a wide variety of control engineering problems. RL is inspired by the ability of animals to learn optimal actions to achieve complex long term goals by maximizing cumulative environmental rewards. In the RL framework the control problem is viewed as an optimal sequential decision problem referred to formally as a Markov Decision Process (MDP). The solution to an MDP is learned by interacting with the environment by taking exploratory actions and observing environmental rewards. The learner referred to as an agent takes actions in each state and transitions to the next state after observing a scalar environmental reward signal. The reward signal represents the feedback from the environment regarding the desirability of taking a particular action in a given state. The reward function and the state transition function may me unknown in the general case. RL attempts to choose actions to maximize the expected cumulative discounted reward. Future rewards are discounted to favour quicker progress to the goal state.

A fundamental idea in RL is the concept of a Value function $V : S \rightarrow \mathbb{R}$ which indicates the desirability of each state by assigning a real number to each state (desirable states leading to higher rewards are assigned higher values). Given the Value function, the optimal action in a certain state $\boldsymbol{s}$ is the action that moves the system to the next state $\boldsymbol{s'}$ with the largest value $V(\boldsymbol{s'})$. The control policy is a function $\pi : S \rightarrow A$ that specifies the action to be taken in each state. The ultimate goal of RL is to compute an optimal control policy π^* that specifies the best action to be taken in each state. Since the optimal action in a certain state is the action that moves the system towards the next state with the maximal value, the Value function must first be computed. Thus the problem of com-

* Corresponding author.
E-mail addresses: jaganathapandian@vit.ac.in (B.J. Pandian), mathew.m@vit.ac.in (M.M. Noel).

https://doi.org/10.1016/j.jprocont.2018.07.013
0959-1524/

Nomenclature

s	state vector
a	vector of control actions
$P_{sa}(\mathbf{s'})$	state transition probabilities
R(**s**)	reward for taking action **a** in state **s**
$\pi(\mathbf{s})$	Policy function that determines action taken in state 's'
$V^{\pi}(\mathbf{s})$	cumulative discounted reward for following policy π starting from state **s**
π^{*}	optimal Policy function
$V^{*}(\mathbf{s})$	optimal Value function
$\hat{V}(\mathbf{s})$	estimate of the optimal Value function
x	$[x_1\ x_2]^T$, state vector for the bioreactor system
x'	next state after control action in current state **x**
U	$[u_1\ u_2]^T$, action vector for the bioreactor system
γ	discount factor to favour immediate rewards
L_i	number of discretization levels used for state variable x_i
L_u	number of discretization levels used for feed substrate concentration u_2
θ_i	process parameters, i = 1, 2, 3, 4
h	$[h_1\ h_2\ h_3\ h_4]^T$, state vector for the quadruple tank process
f	$[f_1\ f_2]^T$, action vector for the quadruple tank process
Q(**s**,**a**)	Q value of a state-action pair
ε	exploration probability
α	learning rate
MSE	$\frac{1}{T}\int_0^T e(t)^2 dt$
ITAE	$\frac{1}{T}\int_0^T t\left\|e(t)\right\| dt$
ITSE	$\frac{1}{T}\int_0^T te(t)^2 dt$
IAU	$\frac{1}{T}\int_0^T \left\|u(t)\right\| dt$

puting the optimal control policy can be reduced to the problem of computing the optimal Value function. Thus the difficult problem of learning a control policy is reduced to the easier problem of learning a Value function in RL. The Value function corresponding to a particular policy π is defined as the expected value of cumulative discounted rewards (1).

$$V^{\pi}(\mathbf{s}) = E[R(\mathbf{s_0}) + \gamma R(\mathbf{s_1}) + \gamma^2 R(\mathbf{s_2}) \ldots | \mathbf{s} = \mathbf{s_0}, \pi] \quad (1)$$

The expected value is taken in the Value function since state transitions are probabilistic in general. A deterministic system is a special case where all state transition probabilities except one are zero. The RL problem can be mathematically formulated as a Markov Decision Process (MDP).

Formally a MDP is a 5-tuple $(S, A, P_{sa}, \gamma, R)$:

S – set of all possible states $\mathbf{s}$, $\mathbf{s} \in R^n$ (where n is the dimension of the state vector)

A – set of all possible actions $\mathbf{a}$, $\mathbf{a} \in R^m$ (where m is the dimension of the action vector)

$P_{sa}(\mathbf{s}')$ – the probability of transitioning to state $\boldsymbol{s'}$ by taking an action $\boldsymbol{a}$ in state $\boldsymbol{s}$

$\gamma \in [0, 1)$ - discount factor

$R : S \times A \rightarrow \mathbb{R}$ - reward function

The Policy function π maps the current state to the action to be taken by the controller. A policy that maximizes the total payoff is an optimal policy $\pi*$ and is therefore given by (2):

$$\pi*(\mathbf{s}) = \max_{\pi} V^{\pi}(\mathbf{s}) \quad (2)$$

The principal advantage of the reinforcement learning approach is its general applicability to difficult and poorly understood control problems [1–4] since it requires only the availability of a scalar reward signal. Over the past few years RL based controllers were proposed for controlling highly nonlinear processes [5–9]. The Q-learning approach [10,11] is a class of learning strategies commonly applied to model free learning problems. Data- driven models were also used for RL approaches to consider the actual process dynamics in the working environment [12,13]. Since the RL framework assumes finite state and action spaces, a major challenge in applying reinforcement learning to process control problems is the need for discretization of continuous variables resulting in exponential growth of the number of discretized values. A wide variety of function approximation approaches have been proposed to alleviate the problems associated with discretizing continuous state and action spaces [9,14–18]. Function approximation methods are usually applied to value function or policy function to reduce quantization error introduced by the discretization of continuous state spaces. Since RL operates by taking exploratory actions and receiving environmental rewards convergence of RL algorithms can be improved with directed search [19–22]. Exploratory action are necessary to find a good control policy however random exploratory actions may result in poor rewards; thus a trade-off between exploration and exploitation is unavoidable in RL. Since exploratory actions are in general costly RL approaches that reduce the need for extensive random exploration are of interest. One possible approach is to obtain supervisory input for the RL learning problems from an approximate inverse model neural controller.

In this paper an approach that uses an approximate inverse model ANN controller [23–25] to reduce the need for exploratory actions and speedup convergence of RL algorithms is proposed. In this partially supervised approach an approximate control policy learnt by an inverse model ANN controller is refined using reinforcement learning. Existing approaches that combine RL and ANNs employ ANNs as function approximators to learn the Value or Q functions or to generalize the Policy function learnt using RL to unseen states. The approach proposed in this paper significantly differs from these approaches in using an inverse ANN scheme to learn an approximate control policy which is refined using a computationally efficient RL based approach. A pure inverse ANN approach requires extensive training data to learn a control policy for complex nonlinear systems. In many cases extensive training data is unavailable and an exact inverse model of the plant may not even exist or be stable theoretically. Thus the proposed approach avoids the need for extensive training data to construct an exact inverse model in the inverse ANN approach, the need for an exact and stable inverse to exist and the need for exhaustive and costly exploration in pure RL based strategies. In the PSRL approach proposed in this paper, Bellman's equation which relates the Value function and Policy functions is used to compute an approximate Value function using the approximate control policy learnt with the inverse ANN approach. This approximate Value function is then refined using Value iteration. This approach for initializing RL is computationally cheap because Bellman's equation can be solved efficiently as they are linear in the unknowns V(**s**). The Partially Supervised Reinforcement Learning (PSRL) algorithm proposed in this paper is tested by applying it to the challenging and econom-