Neurocomputing 194 (2016) 176–182

Contents lists available at ScienceDirect

Neurocomputing

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

Optimal tracking control for completely unknown nonlinear discrete-time Markov jump systems using data-based reinforcement learning method

He Jiang, Huaguang Zhang*, Yanhong Luo, Junyi Wang

College of Information Science and Engineering, Northeastern University, Box 134, 110819 Shenyang, PR China

ARTICLE INFO

Article history:
Received 2 November 2015
Received in revised form
28 December 2015
Accepted 2 February 2016
Communicated by Z. Wang
Available online 23 February 2016

Keywords:
Optimal tracking control
Markov jump systems
Data-based
Reinforcement learning
Adaptive dynamic programming
Neural networks

ABSTRACT

In this paper, we develop a novel optimal tracking control scheme for a class of nonlinear discrete-time Markov jump systems (MJSs) by utilizing a data-based reinforcement learning method. It is not practical to obtain accurate system models of the real-world MJSs due to the existence of abrupt variations in their system structures. Consequently, most traditional model-based methods for MJSs are invalid for the practical engineering applications. In order to overcome the difficulties without any identification scheme which would cause estimation errors, a model-free adaptive dynamic programming (ADP) algorithm will be designed by using system data rather than accurate system functions. Firstly, we combine the tracking error dynamics and reference system dynamics to form an augmented system. Then, based on the augmented system, a new performance index function with discount factor is formulated for the optimal tracking control problem via Markov chain and weighted sum technique. Neural networks are employed to implement the on-line ADP learning algorithm. Finally, a simulation example is given to demonstrate the effectiveness of our proposed approach.

1. Introduction

Research on Markov jump systems (MJSs) has received significant attention due to its extensive applications among the network control systems, manufacturing systems and power systems [1,2]. For these systems, abrupt variations usually occur in their system structures when there exist subsystem interconnection variations and sudden environmental disturbances. Due to the powerful modeling capability of MJSs, the models of these systems could be expressed and formulated by MJSs [3]. Although there have been diversities of studies on MJSs, such as synchronization [4,5], robust control [6,7] and stability analysis [8,9], there are still few works concerning with the optimal control issue, which can be generally divided into two categories: one is optimal regulation and the other one is optimal tracking [10]. In addition, most existing results depend on the accurate system functions in the design of controllers. However, it is infeasible to obtain the complete knowledge of system models at each time step in the real-world applications owing to the existence of the inherent stochastic property in MJSs. Therefore, it would be interesting and challenging to find out a method to investigate the optimal issue for MJSs under mode-free conditions.

Reinforcement learning (RL) [11–15], which is studied from computational intelligence and machine learning, has the ability to let an agent learn an optimal control policy by utilizing the responses from unknown environment. As a core branch of RL methods, adaptive/approximation dynamic programming (ADP) has provided a successful way to achieve optimal control in a stochastic process [16], which motives our research. ADP brings the advantages of adaptive and optimal control together to address different optimal control issues, such as optimal tracking control [10,17–19], optimal control with constrained control input [20–23], optimal control with time-delays [24–26], optimal control for zero-sum [27,28] and non-zero-sum games [29,30], and optimal control applied on unknown nonlinear systems [31–34] and multi-agent systems [35–39].

This paper proposes a data-based ADP method to solve the optimal tracking control problem (OTCP) for nonlinear discrete-time (DT) MJSs with completely unknown dynamics. Firstly, an augmented system composed of tracking error dynamics and reference system dynamics is introduced for reformulating the issue. According to Markov chain, we transform the multiple-objective optimal problem for each subsystem of MJS into a single-objective one for the whole MJS by utilizing the weighted sum technique. Then, we introduce the policy iteration algorithm and discuss the value of discount factor in the stability analysis. Finally, based on the temporal

* Corresponding author.
E-mail addresses: jianghescholar@163.com (H. Jiang), hgzhang@ieee.org (H. Zhang), neuluo@gmail.com (Y. Luo), wjyi168@126.com (J. Wang).

http://dx.doi.org/10.1016/j.neucom.2016.02.029
0925-2312/

difference learning method, a gradient-descent on-line algorithm is designed only requiring the knowledge of system data. Two neural networks (NNs), critic NN and actor NN, are employed to implement the proposed algorithm and approximate the performance index and control policy, respectively. Note that, ADP can be classified into five groups in general, namely, heuristic dynamic programming (HDP), dual heuristic dynamic programming (DHP), globalized dual heuristic dynamic programming (GDHP), action dependent heuristic dynamic programming (ADHDP) and action dependent dual heuristic dynamic programming (ADDHP). Our proposed approach is most relevant to ADHDP. As a result, the main contributions of this paper can be summarized as follows: Firstly, based on the augmented system which contains tracking error dynamics and reference system dynamics, the optimal tracking control problem for DT MJSs is reformulated by using Markov chain and weighted sum technique. To the best of our knowledge, a novel data-based ADP algorithm is proposed to solve this issue with completely unknown system dynamics for the first time. Secondly, neural networks are used to facilitate the implementation of the on-line ADP algorithm. The proposed RL algorithm iteratively computes the control policy and performance index function through the reinforcement signal from the completely unknown external environment and tries to reinforce the control policy to minimize the future performance index. During the learning procedure, only the system data is required rather than the accurate system functions, whereas other previous works [4–9] investigated the issue with known system dynamics. Thirdly, unlike the traditional identification approaches [31–33], the ADP algorithm is an adaptive learning process, which can still find out the optimal control policy adaptively even when the system parameters change and successfully circumvent the identification errors.

The rest of this paper is organized as follows. In Section 2, an augmented system and a new performance index with discount factor are established, and then the OTCP for DT MJSs is reformulated. In Section 3, the policy iteration algorithm is introduced and the value of discount factor is discussed. A novel data-based on-line algorithm is designed through gradient-based methods and implemented by utilizing two NNs in Section 4. Section 5 presents a simulation example to show the effectiveness of our proposed approach. Finally, conclusions are drawn in Section 6.

2. Problem formulation

Consider a class of DT MJSs as below:

$$x(k+1)=f_i(x(k))+g_i(x(k))u(k) \tag{1}$$

where $x(k)\in R^n$ denotes the state; $u(k)\in R^m$ is the control input; the system drift dynamics $f_i(x(k))\in R^n$ and the system input dynamics $g_i(x(k))\in R^{n\times m}$ are both considered to be unknown; i represents the notation of the discrete-time Markov chain $\{\sigma(k)\}$, which denotes the active mode at the kth time step and belongs to a finite set $\rho=\{1,2,\cdots,l\}$, where l is the total number of the jumping modes.

Assumption 1. $f_i(0)=0$ and $f_i(x(k))+g_i(x(k))u(k)$ is Lipschitz continuous on a compact set Ω containing the origin.

Assumption 2. The system (1) is controllable, namely, there exists a continuous control on Ω which stabilizes the system.

The transition probability matrix for the DT MJS can be defined as

$$P=\begin{bmatrix}\pi_{11} & \cdots & \pi_{1l}\\ \vdots & \ddots & \vdots\\ \pi_{l1} & \cdots & \pi_{ll}\end{bmatrix}\in R^{l\times l} \tag{2}$$

where $\pi_{ab}=\Pr\{\sigma(k+1)=b|\sigma(k)=a\}$ for $\forall a,b\in\rho$ and $\sum_{b=1}^{l}\pi_{ab}=1$ for $\forall a\in\rho$.

Define the reference system which generates the desired tracking trajectory as follows:

$$x_d(k+1)=h(x_d(k)) \tag{3}$$

where $x_d(k)\in R^n$ is the tracking objective state and $h(x_d)\in R^n$ denotes the command generator function with $h(0)=0$. Therefore, the tracking error can be given by

$$e_d(k)=x(k)-x_d(k). \tag{4}$$

Remark 1. The existing works concerning about the optimal tracking control problems all require the information of system dynamics [10,17–19]. Especially for MJSs which are essentially stochastic systems, it is infeasible to obtain the accurate system functions at each time step. Thus, for the completely unknown systems, these results are invalid. In this paper, in order to solve the OTCP for MJSs without the requirement of system dynamics, we present a new formulation for this issue by utilizing the idea of augmented system and discounted performance function in [10].

Combining (1) and (3) yields the tracking error dynamics

$$e_d(k+1)=f_i(x(k))-h(x_d(k))+g_i(x(k))u(k). \tag{5}$$

Let $X(k)=\left[e_d^T(k)\ x_d^T(k)\right]^T$ be the state of the augmented system. Then, bringing together (3) and (5) yields the dynamics of augmented system

$$X(k+1)=F_i(X(k))+G_i(X(k))u(k) \tag{6}$$

where $F_i(X(k))=\begin{bmatrix}f_i(e_d(k)+x_d(k))-h(x_d(k))\\ h(x_d(k))\end{bmatrix}$ and $G_i(X(k))=\begin{bmatrix}g_i(e_d(k)+x_d(k))\\ 0\end{bmatrix}$.

The augmented system (6) can be viewed as a new formulated MJS, and we define the performance index function for each subsystem mode as follows:

$$\begin{aligned}J_i(X(k))&=\sum_{t=k}^{\infty}\alpha^{t-k}\left[X^T(t)Q_iX(t)+u^T(t)R_iu(t)\right]\\&=X^T(k)Q_iX(k)+u^T(k)R_iu(k)+\alpha J_i(X(k+1))\end{aligned} \tag{7}$$

where $0<\alpha\le 1$ is a discount factor; $Q_i=\begin{bmatrix}H_i & 0\\ 0 & 0\end{bmatrix}$ with $H_i>0$ and $R_i>0$ being the weight matrices for $\forall i=1,2,\ldots,l$.

Remark 2. It is essential to employ a discounted performance function when dealing with the optimal tracking problems. This is mainly because the desired tracking trajectory generally does not go to zero in most real-world applications, then the performance index function would be infinite without a discount factor as the control policy contains a part which depends on the desired tracking trajectory, i.e., $u^T(t)R_iu(t)$ would not go to zero as time goes to infinity.

According to the transition probability matrix (2), we can bring together the performance index functions (7) of all the subsystem modes to act as that of the whole MJS as below:

$$\begin{cases}V_1(X(k))=\pi_{11}J_1(X(k))+\pi_{12}J_2(X(k))+\cdots+\pi_{1l}J_l(X(k))\\ V_2(X(k))=\pi_{21}J_1(X(k))+\pi_{22}J_2(X(k))+\cdots+\pi_{2l}J_l(X(k))\\ \vdots\\ V_l(X(k))=\pi_{l1}J_1(X(k))+\pi_{l2}J_2(X(k))+\cdots+\pi_{ll}J_l(X(k))\end{cases} \tag{8}$$

From (8), it can be observed that the optimal tracking problem for MJS is converted into a multi-objective optimization problem. In order to simplify the complexity of the problem, we employ the weighted sum technique and transform the multi-objective optimization problem into the single-objective one by introducing the weight vector $\lambda=[\lambda_1\ \lambda_2\cdots\lambda_l]^T$. Hence, the performance index