Information Sciences 286 (2014) 209–227

Contents lists available at ScienceDirect

Information Sciences

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

Reinforcement learning with automatic basis construction based on isometric feature mapping

Zhenhua Huang, Xin Xu *, Lei Zuo

College of Mechatronics and Automation, National University of Defense Technology, Changsha 410073, PR China

ARTICLE INFO

Article history:
Received 23 June 2013
Received in revised form 23 May 2014
Accepted 6 July 2014
Available online 16 July 2014

Keywords:
Reinforcement learning
Isometric feature mapping
Value function approximation
Approximate policy iteration
Learning control

ABSTRACT

Value function approximation (VFA) has been a major research topic in reinforcement learning. Although various reinforcement learning algorithms with VFA have been proposed, the performance of most previous algorithms depends on the predefined structure of the basis functions. To address this problem, this paper presents a novel basis learning method for VFA based on isometric feature mapping (IFM). In the proposed method, basis functions for VFA are automatically generated by constructing the optimal embedding basis of the data in a d-dimensional Euclidean space, which best preserves the estimated intrinsic geometry of the manifold. Furthermore, the IFM-based basis learning method is integrated with approximation policy iteration (API) for learning control in Markov decision problems with large state spaces. A new manifold reinforcement learning framework termed IFM-based API (IFM-API) is presented. Three learning control problems, including a real control system of the Googol single inverted pendulum, were studied to evaluate the performance of the proposed IFM-API algorithm. The simulation and experimental results show that, compared with other basis selection or learning methods, the IFM-based basis learning method can automatically compute an efficient set of basis functions with much fewer predefined parameters and less computational costs. Besides, it is illustrated that the proposed IFM-API algorithm can obtain better learning control policies than other API methods.

1. Introduction

In recent years, many encouraging achievements and applications have been received in the supervised and unsupervised learning domain [1,5,8,14,15,23]. However, it is still difficult for them to deal with sequential decision-making or learning control problems efficiently [7,24]. Reinforcement learning (RL) [30], as a machine learning framework, has been widely studied [7,18,30,34,35,37] in the past decade for solving sequential decision making problems. These problems are usually modeled as Markov Decision Processes (MDPs). In RL, the action policy of a learning agent is modified to maximize its total rewards during the process of its interaction with an environment. Due to the property of RL, it is more suitable for solving learning control problems (which can be modeled as MDPs) than supervised learning methods and mathematical programming methods [7,9,24]. In earlier stages of RL research, most efforts were focused on learning control algorithms for MDPs with discrete state and action spaces. However, MDPs in many real-world applications are usually with continuous or

* Corresponding author.
E-mail addresses: xuxin_mail@263.net, xinxu@nudt.edu.cn (X. Xu).

http://dx.doi.org/10.1016/j.ins.2014.07.008
0020-0255/

large-scale state spaces. If the value functions in continuous or large-scale state spaces are still represented using discrete tabular, the learning rate and convergence rate will be influenced negatively [24].

Aiming at the above problem, hierarchical RL (HRL) and approximate RL methods were popularly studied. As illustrated in [4], existing work in HRL can be grouped into three aspects: the abstraction of the sets of actions [31], the abstraction of states [11] and the decomposition of state space [10]. Some successful applications of HRL have been reported, e.g., robot path tracking [38], web services composition [12] and so on. Approximate RL, which is also called approximate dynamic programming (ADP) [34], has received increasing attention in recent years. Previous research work on approximate RL methods mainly include there perspectives, e.g., policy search [3], value function approximation (VFA) [29], and actor-critic methods [25]. Actor-critic methods can be regarded as a combination of policy search and VFA. It has been illustrated that actor-critic algorithms can obtain better learning efficiency than policy search or VFA in online learning control of MDPs with large state spaces [25]. In actor-critic algorithms, an actor is used for policy learning or policy improvement and a critic is employed for policy evaluation or value function approximation. In recent years, more and more research efforts have been focused on actor-critic approaches. Adaptive critic designs (ACDs) [18] become a significant class of learning control methods for linear or nonlinear dynamic systems.

Despite of the above advances, it is well known that VFA is a central problem of all successful applications of RL, where a variety of non-linear and linear approximation architectures have been studied. Particularly, VFA approaches with linear approximation architecture have been widely used due to the advantage of convergence and stability properties. Recent advances in VFA with linear function approximators include linear SARSA-learning [29], and least-squares policy iteration (LSPI) [17], etc. Although better approximation ability of nonlinear VFA is exhibited than that of linear VFA, few conclusions were reported on rigorous theoretical analysis of RL applications using nonlinear VFA. In recent years, many research efforts have been put on RL algorithms using the linear VFA architecture. For example, LSPI has been popularly studied as an efficient RL algorithm with linear basis functions [9,17]. However, the basis functions in LSPI need to be manually selected. In [36], a kernel-based least squares policy iteration (KLSPI) algorithm was proposed by replacing the hand-coded basis functions with kernel-based features. Although the kernel-based features can be generated in a data-driven style, the kernel functions still need to be selected carefully by the designer. Therefore, one common drawback of previous work in VFA is that the basis functions or kernel functions are usually hand-coded by human experts, rather than automatically constructed from the geometry of the underlying state space.

In pattern recognition and machine learning, one of the key problems is to develop appropriate and effective low-dimensional representations for complex high-dimensional data, namely the *dimensionality reduction* problem. A prevalent approach to solve the problem is based on the notion of *manifold*. It has been shown that a set of high-dimensional data can usually be described as a set of vectors in a low-dimensional nonlinear manifold [19,27]. Given a set of data points $x = [x_1, x_2, \ldots, x_n]$ in a low-dimensional space R_{d_L}, let $f : x_i \to R_{d_H}$ be a smooth embedding, for some $d_H > d_L$. Manifold learning aims to recover x and f based on a given set of observed data $\{y = f(x)\}$ in R_{d_H}. Until now, different manifold learning algorithms have been developed, e.g., locally linear embedding (LLE) [26], isometric mapping (ISOMAP) [32], Laplacian eigenmaps (LE) [6], etc. A general graph embedding framework was presented in [19] to unify various dimensionality reduction methods.

Although manifold learning has been widely studied in regression and classification tasks, there are few works on the integration of manifold learning into reinforcement learning problems. Recently, based on the principle of Laplacian eigenmaps, an algorithm called representation policy iteration (RPI) for value function approximation in RL was proposed [21]. Basis functions in RPI can be automatically constructed using the spectral analysis of a self-adjoint Laplacian operator, instead of using a hand-coded parametric architecture. However, in RPI, except for the numbers of nearest neighbors and basis functions, some other parameters also need to be predefined, e.g., the Laplacian type and the width of the Gaussian distance. Besides, it is still difficult to learn or construct a "representative" graph by trajectory sampling in continuous or large-scale state spaces.

Compared with other manifold learning methods, ISOMAP [2,32] requires only one parameter to be determined and seeks to preserve the intrinsic geometry of the data. The key problem in ISOMAP is how to estimate the geodesic distance between faraway points only using input-space distances. As illustrated in [2,32], the geodesic distance for neighboring points can be well approximated by the input-space distance. The geodesic distance for faraway points can be approximated by computing the total length of a series of shortcut paths between neighboring points. After constructing a graph that connects neighboring data points as edges, one can efficiently approximate the geodesic distance by searching the shortest paths in the graph. Inspired by the above idea, this paper presents a novel basis learning method based on isometric feature mapping (IFM). Using this method, basis functions can be constructed automatically by preserving the intrinsic geometry of the collected samples. Furthermore, the IFM-based basis learning method is integrated with approximation policy iteration (API) for learning control in MDPs with continuous or large-scale state spaces. A new manifold reinforcement learning framework called IFM-based API (IFM-API) is proposed. Three learning control problems, including a real system control of the Googol single inverted pendulum, were studied to evaluate the performance of the proposed IFM-API algorithm. The simulation and experimental results show that, compared with some other basis selection or learning methods, the IFM-based basis learning method can automatically compute an efficient set of basis functions with fewer predefined parameters and smaller computational costs. Besides, the clustering-based sub-sampling method [13], which is a better choice than the trajectory-based sub-sampling method [20,21], is used in IFM-API. It is illustrated that the proposed IFM-API algorithm can obtain better learning control policies than some other API methods, e.g., KLSPI [36], LSPI [17] and RPI [21].