基于高斯回归的连续空间多智能体跟踪学习

陈 鑫 [1,2] 魏海军 [1,2] 吴 敏 [1,2] 曹卫华 [1,2]

摘 要 提高适应性、实现连续空间的泛化、降低维度是实现多智能体强化学习 (Multi-agent reinforcement learning, MARL) 在连续系统中应用的几个关键. 针对上述需求, 本文提出连续多智能体系统 (Multi-agent systems, MAS) 环境下基于模型的智能体跟踪式学习机制和算法 (MAS MBRL-CPT). 以学习智能体适应同伴策略为出发点, 通过定义个体期望即时回报, 将智能体对同伴策略的观测融入环境交互效果中, 并运用随机逼近实现个体期望即时回报的在线学习. 定义降维的 Q 函数, 在降低学习空间维度的同时, 建立 MAS 环境下智能体跟踪式学习的 Markov 决策过程 (Markov decision process, MDP). 在运用高斯回归建立状态转移概率模型的基础上, 实现泛化样本集 Q 值函数的在线动态规划求解. 基于离散样本集 Q 函数运用高斯回归建立值函数和策略的泛化模型. MAS MBRL-CPT 在连续空间 Multi-cart-pole 控制系统的仿真实验表明, 算法能够使学习智能体在系统动力学模型和同伴策略未知的条件下, 实现适应性协作策略的学习, 具有学习效率高、泛化能力强等特点.

关键词 连续状态空间, 多智能体系统, 基于模型的强化学习, 高斯回归

引用格式 陈鑫, 魏海军, 吴敏, 曹卫华. 基于高斯回归的连续空间多智能体跟踪学习. 自动化学报, 2013, **39**(12): 2021−2031

DOI 10.3724/SP.J.1004.2013.02021

Tracking Learning Based on Gaussian Regression for Multi-agent Systems in Continuous Space

CHEN Xin[1,2] WEI Hai-Jun[1,2] WU Min[1,2] CAO Wei-Hua[1,2]

Abstract Improving adaption, realizing generalization in continuous space, and reducing dimensions are always viewed as the key issues for the implementation of multi-agent reinforcement learning (MARL) within continuous systems. To tackle them, the paper presents a learning mechanism and algorithm named model-based reinforcement learning with companion′s policy tracking for multi-agent systems (MAS MBRL-CPT). Stemming from the viewpoint to make the best responses to companions, a new expected immediate reward is defined, which merges the observation on companion′s policy into the payoff fed back from the environment, and whose value is estimated online by stochastic approximation. Then a Q value function with dimension reduced is developed to set up Markov decision process (MDP) for strategy learning in multi-agent environment. Based on the model of state transition using Gaussian regression, the Q value functions w.r.t. the state-action samples for generalization are solved by dynamic programming, which then serve as the basic samples to realize the generalization of value functions and learned strategies. In the simulation of multi-cart-pole in continuous space, even if the dynamics and companions′ strategies are unknown in priori, MBRL-CPT entitles the learning agent to learn the tracking strategy to cooperate with its companions. The performance of MBRL-CPT shows its high efficiency and good generalization ability.

Key words Continuous state space, multi-agent systems (MAS), model-based reinforcement learning (MBRL), Gaussian regression (GR)

Citation Chen Xin, Wei Hai-Jun, Wu Min, Cao Wei-Hua. Tracking learning based on Gaussian regression for multi-agent systems in continuous spaces. *Acta Automatica Sinica*, 2013, **39**(12): 2021−2031

多智能体系统 (Multi-agent systems, MAS) 是由同一环境中简单的、自治的、存在交互的多个智能体组成的系统[1], 目前已在机器人编队、移动式智能体、物流管理等领域广泛应用. 通过建立 MAS 在线学习实现智能体在未知或复杂动态环境下最优策略的探索与优化, 从而提高 MAS 的自主决策能力, 已经成为 MAS 研究的热点问题.

强化学习 (Reinforcement learning, RL) 作为一种交互式机器学习方法, 被视为未知或动态环境下智能体学习的重要手段. 它介于监督学习与非监督学习之间, 结合环境反馈, 通过试错机制不断地与

收稿日期 2012-04-17 录用日期 2013-05-13
Manuscript received April 17, 2012; accepted May 13, 2013
国家自然科学基金 (61074058) 资助
Supported by National Natural Science Foundation of China (61074058)
本文责任编委 陈杰
Recommended by Associate Editor GHEN Jie
1. 中南大学信息科学与工程学院 长沙 410083 2. 先进控制与智能自动化湖南省工程实验室 长沙 410083
1. School of Information Science and Engineering, Central South University, Changsha 410083 2. Hunan Engineering Laboratory for Advanced Control and Intelligent Automation, Changsha 410083

环境进行交互学习, 具有良好的收敛性, 并且实施相对简单[2−3]. 其最优解可通过值函数计算或策略迭代获得[4]. 一般来说, 在 MAS 环境下运用 RL 实现策略学习是解决 MAS 策略学习的主要解决思路之一.

但是, 实际 MAS 应用环境, 例如机器人编队、多传感器网络等, 往往呈现连续状态或连续动作空间, 其内在连续性给多智能体强化学习 (Multi-agent reinforcement learning, MARL) 算法的应用带来了极大的考验. 因此, 泛化和策略学习是连续空间中 MARL 研究的两个重要方面[5].

在泛化方法的研究中, 函数逼近法应用最为广泛, 它包括值函数逼近法和策略函数逼近法两类[6]. 如王雪松等[7] 采用协同最小二乘支持向量机实现了值函数的逼近, 以及 Busoniu 等[8] 应用最小二乘 TD 算法实现了在线策略逼近等. 这类方法采用参数化函数逼近器逼近并替代离散形式的映射关系, 以实现部分泛化功能. 但是, 参数化函数逼近器需要已知函数原型结构, 在某些未知环境下, 很难通过参数化模型实现模型和值函数的泛化.

为了实现模型结构未知条件下的模型或值函数泛化, Rasmussen 等[9] 利用非参数化形式的高斯回归 (Gaussian regression, GR) 建立环境模型, 实现状态空间泛化. Jung 等[10] 以及 Deisenroth 等[11] 研究了基于少量观察数据建立环境的 GR 模型的方法, 在连续状态空间实现了单智能体基于模型的学习. Deisenroth 等[12] 提出了单智能体连续状态–动作空间的 GR 建模与策略学习, 但其样本更新方式难以用于实际系统, 而且在 MAS 环境下, 建模的计算和存储开销随智能体个数呈指数增长. 因此, 提高 MAS 环境下状态和动作空间同时泛化的有效性和实用性是目前连续空间 MARL 研究的关键问题之一.

大多数连续空间 RL 中状态–动作空间的泛化通常需要基于离散的观测样本构造泛化的基础样本, 而离散样本集合往往也作为值函数迭代学习的基本集合, 采用离散空间 RL 实现值函数学习. 因此, 离散空间策略学习是实现连续空间策略学习的基础.

一直以来, 离散空间 MARL 的同时学习、“维数灾难”、信度分配、探索–利用平衡等问题被视为 MARL 研究的难点问题[13]. 由于同时存在多个智能体, 所有智能体试图通过改变自己的策略对动态环境作出最佳响应行为, 从而陷入适应性循环. 因此, 实现策略学习的稳定性和适应性成为研究和设计 MARL 所考虑的主要方面. 从侧重方面的不同, 目前主要研究的 MARL 方法可以大致分为两类.

第一类为基于协调均衡的 MARL 算法, 如 Hu 等提出一般和随机对策 (Stochastic games, SG) 框架下的 Nash-Q[14], 以及 Greenwald 等提出的 CE-Q[15] 学习算法等. 该类算法设定智能体的学习目标是获取一组基于特定类均衡的行为策略, 其本质是对均衡解的学习. 不过它存在均衡解的选择问题, 即在出现多个均衡解的情况下, 智能体之间必须对均衡解的选择意向达成一致协议.

第二类算法为基于最佳响应策略的 MARL 算法. 与最优策略追求学习的收敛性相比, 追求学习适应性的最佳响应策略更具实用性. 如 Conitzer 等提出的 AWESOME[16] 算法, 当对方策略为平稳策略时, 则学习最佳响应策略, 否则, 仍然采用保守的 Nash 均衡算法. 而 Weinberg 等[17] 提出的非静态、收敛策略 (Non-statonary converging policies, NSCP) 算法能够在对手策略处于有限变化的条件下, 实现最佳响应策略学习. 虽然这类算法在一定程度上有效解决了智能体的策略适用性问题, 但大多数方法仍采用基于联合状态–联合动作的集中式学习方式, 容易引发 MAS “维数灾难” 问题, 难以适应问题规模较大的学习环境. 显然, 保证智能体策略适应性的同时, 降低由于 MAS 环境带来的高维度是提高 MARL 实用性的另一关键.

对于 “维数灾难” 问题, 常用的有效方法是分层强化学习 (Hierarchical reinforcement learning, HRL), 例如静态分层方法 (Hierarchical Q-learning, HQL)[18] 和动态分层方法 (Dynamic HRL model, DHRL-Model)[19] 等. 分层学习通过对状态空间和动作空间的特征提取, 实现子任务或复合动作抽象, 从而降低学习空间的维度或限定学习空间的大小. 一般来说, 静态分层需要预先知道任务或环境特征的先验知识, 而动态分层需要通过统计或辨识方法完成特征提取, 这在 MAS 环境下需要较多的计算开销. 因此, 结合 MAS 特点, 探索除分层学习之外的降维方法, 是缓解 MARL 的 “维数灾难” 问题的另一思路.

另一方面, 信度分配也是 MAS 应用的难题之一. 实际应用中多智能体观测到的环境反馈大多表征团队行为的效果, 很难分离个体贡献的大小. 因此, 在复杂或未知环境下, 基于预先知识的信度分配很难保证合理性.

上述讨论的连续空间 MARL 几个关键问题是紧密相关的, 信度分配的合理性决定了降维的方式, 进而基于降维的空间表达构造分布式学习的架构, 为同时学习提供可能的解决方案.

因此, 本文综合考虑上述几方面, 从 “AI agenda”[20] 的角度, 提出一种连续空间 MAS 环境下的跟踪式学习算法, 基于利己型智能体期望合作后个体收益最大化的原则, 定义连续空间多智能体学习环境下的降维值函数. 基于该值函数所学习的

策略强调对其他智能体行为的适应性. 在线建立环境状态转移模型, 采用基于模型的学习方式提高学习效率, 从空间规模方面缓解 “维数灾难”, 同时建立值函数模型, 实现状态 – 动作空间的泛化. 最终使智能体在连续环境下能采用跟踪学习的方式获得合作的最佳策略.

1 基于模型的多智能体跟踪学习

基于最佳响应的 MARL 提出在观测同伴行为的基础上实现对同伴策略的跟踪, 具有较好的适应性. 但基于联合状态 – 联合动作空间的集中式学习方法使 MARL 受到 “维数灾难” 的约束. 而基于个体状态 – 个体动作空间的分布式独立 MARL 存在奖赏分配问题, 即 MAS 中智能体只能观察环境对于智能体联合行为的奖赏值.

本文基于 MAS 的特点, 通过合理定义即时回报形式, 定义降维的值函数和相应的值函数学习算法, 实现在保留 MARL 适应性基础上, 降低学习空间的维度, 提高 MARL 效率的目的, 也为泛化提供规模可控的样本集合. 进而将离散 MARL 和模型及值函数泛化相结合, 在降维连续空间建立 MARL 跟踪式学习算法.

1.1 连续空间降维的跟踪学习值函数

在实际 MAS 系统中, MAS 得到的即时回报往往是针对群体行为的评价, 以往的 MAS 分布独立式学习算法通过信度分配获得个体行为的即时回报, 进而降低动作维度, 实现个体策略的独立学习. 但预先的空间信度分配往往带有主观意愿, 并不能保证分配的合理性. 针对这一问题, 本文基于 “AI agenda” 的思想, 结合策略观测定义一种新的个体期望即时回报, 并以此为基础构造降维值函数, 从而避免信度分配, 实现动作空间降维.

设多智能体共同执行联合动作之后的即时回报为 $r(\boldsymbol{s},\boldsymbol{a})$, 其中, $\boldsymbol{s}$ 和 $\boldsymbol{a}$ 分别表示多智能体的联合状态和联合动作.

定义 1. 定义从状态 $\boldsymbol{s}$ 开始, 执行策略 $\pi^i(\boldsymbol{s})$ 获得的期望累计折扣回报 $V^{\pi^i}(\boldsymbol{s})$ 如式 (1) 所示:

$$V^{\pi^i}(\boldsymbol{s}) = \mathrm{E}\left[\sum_k \gamma^k r_k(\boldsymbol{s}, a^i)\right] \tag{1}$$

这里, $r(\boldsymbol{s}, a^i)$ 表示对应智能体 i 执行动作 a^i 获得的期望即时回报, 即:

$$\begin{aligned} r(\boldsymbol{s}, a^i) = {} & \int_{a^1}\cdots\int_{a^{i-1}}\int_{a^{i+1}}\cdots \\ & \int_{a^N} r(\boldsymbol{s}, a^1, \cdots, a^{i-1}, a^i, a^{i+1}, \cdots, a^N) \\ & \mathrm{d}a^1\cdots\mathrm{d}a^{i-1}\mathrm{d}a^{i+1}\cdots\mathrm{d}a^N \end{aligned} \tag{2}$$

$r(\boldsymbol{s}, a^i)$ 体现了在 MAS 中利己型的智能体认为在状态 $\boldsymbol{s}$ 时, 通过一次合作它能获得的期望回报. 这其实是学习智能体对合作效果的一种认知. 显然, 基于 $\boldsymbol{s}$ 获得的 $V^{\pi^i}(\boldsymbol{s})$ 体现了智能体如何适应同伴的策略 (或者说跟踪同伴策略), 并追求自身利益最大化的目标.

连续空间智能体 i 期望累计折扣回报的 Bellman 方程如式 (3) 所示:

$$\begin{aligned} V_*^{\pi^i}(\boldsymbol{s}) = \max_{a^i}\{ & r(\boldsymbol{s}, a^i) + \\ & \gamma\int p(\boldsymbol{s}'|\boldsymbol{s}, a^i) V_*^{\pi^i}(\boldsymbol{s}')\mathrm{d}\boldsymbol{s}'\Big\} \end{aligned} \tag{3}$$

其中, $p(\boldsymbol{s}'|\boldsymbol{s}, a^i)$ 表示状态转移概率函数.

定义 2. 定义联合状态 - 个体动作空间智能体 i 在状态 $\boldsymbol{s}$ 处, 执行动作 a^i 所获得的期望折扣回报, 即 Q 函数如式 (4) 所示:

$$Q^*\left(\boldsymbol{s}, a^i\right) = r\left(\boldsymbol{s}, a^i\right) + \gamma\int p\left(\boldsymbol{s}'|\boldsymbol{s}, a^i\right) V^*\left(\boldsymbol{s}'\right)\mathrm{d}\boldsymbol{s}' \tag{4}$$

联立式 (3) 和式 (4), 可得式 (4) 中 V 函数的表达式如下:

$$V^*\left(\boldsymbol{s}\right) = \max_{a^i} Q^*\left(\boldsymbol{s}, a^i\right) \tag{5}$$

显然, 与集中式学习的期望折扣奖励相比, 定义 1 和定义 2 中的期望折扣奖励不再显性地与同伴智能体动作相关, 从而实现值函数学习空间的降维.

1.2 跟踪学习架构

连续空间降维的跟踪学习值函数定义式 (4) 说明, 当 MAS 中智能体能够知道其联合状态 - 个体动作的环境状态转移概率函数时, 就能够计算当前联合状态 - 个体动作空间下的 Q 值函数, 进而得到最优执行策略. 显然, 获得计算式 (4) 所需的连续空间状态转移概率 $p(\boldsymbol{s}'|\boldsymbol{s}, a^i)$、值函数 $V^*(\boldsymbol{s})$ 以及联合状态 - 个体动作的即时奖赏函数 $r(\boldsymbol{s}, a^i)$ 成为降维条件下学习跟踪策略的关键.

为此, 引入基于模型的学习方式, 建立如图 1 所示的 MAS 连续空间基于模型的跟踪学习架构 (MAS continuous model-based reinforcement learning with companion′s policy tracking, MAS MBRL-CPT). 使智能体 i 考虑同伴策略平稳的条件下, 建立环境状态转移模型 $p(\boldsymbol{s}'|\boldsymbol{s}, a^i)$, 获得真实的即时奖励函数 $r(\boldsymbol{s}, a^i)$, 以动态规划的方式求解离散 Q 函数, 进而通过建立 Q 函数模型 $Q^*(\boldsymbol{s}, a^i)$、V 函数模型 $V^*(\boldsymbol{s})$ 实现连续空间动作泛化和状态泛化. 基于 Q 函数模型实现离散最优动作决策, 并建立连续空间策略模型, 实现决策的泛化. 由此建立的智能体最优跟踪策略能够基于同伴策略, 最大化自身的收益. 这可以视为学习智能体对同伴策略的一种跟踪行为.

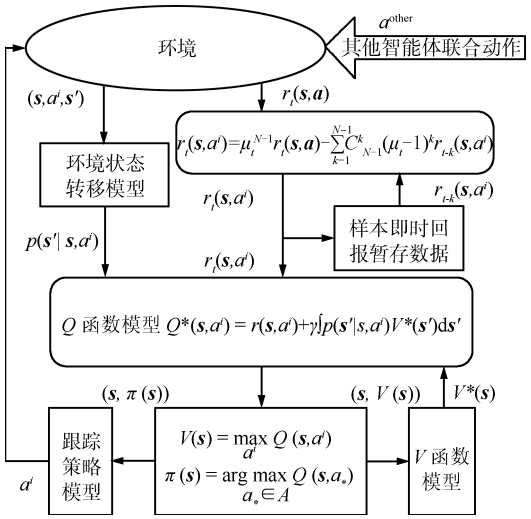


图 1 基于模型跟踪学习框架

Fig. 1 Framework of the tracking learning based on model

显然, 建立状态转移模型是求解离散样本点 Q 函数的关键, Q 函数模型、V 函数模型, 以及策略模型是实现连续空间泛化的核心. 因此, 在定义降维的值函数之后, 考虑 MAS 环境设计实用的建模算法是实现跟踪学习架构的另一关键.

2 基于高斯回归的多智能体系统建模

连续空间 MAS 环境下的建模呈现出模型维度高、函数结构不确定以及模型输出具有一定的概率分布性等特点, 传统的参数化建模方法难以实现在 MAS 环境下的精准建模. GR 作为一种概率建模方法, 不需要事先确定模型对象的函数结构, 对先验知识的依赖小, 其超参数优化也相对容易[21].

GR 中数据集 $\{\boldsymbol{X},\boldsymbol{y}\}$ 的元素由输入向量 $\boldsymbol{x}_i$ 和对应的输出观察值 $y_i=h(\boldsymbol{x}_i)+\varepsilon$, $\varepsilon\sim \mathrm{N}(0,\sigma_e^2)$ 组合而成. 选择平方指数协方差函数与噪声协方差函数的组合为核函数[11], 即:

$$k_h=\alpha^2\exp\left(-\frac{1}{2}(x_p-x_q)^{\mathrm{T}}\Lambda^{-1}(x_p-x_q)\right)+\delta_{pq}\sigma_\varepsilon^2 \tag{6}$$

式中, $\Lambda=\mathrm{diag}\{[l_1^2,\cdots,l_D^2]\}$, l、α^2 以及 σ_ε^2 组成 GR 的超参数 θ, 可通过极大似然法获取最优超参数 θ^*, 实现 GR 的训练, 似然函数的对数形式如式 (7):

$$\begin{aligned}&\log \mathrm{P}\left(\boldsymbol{y}|X,\boldsymbol{\theta}\right)=\\&\log\int p\left(\boldsymbol{y}\,|h\left(X\right),X,\boldsymbol{\theta}\right)p\left(h\left(X\right)|X,\boldsymbol{\theta}\right)\mathrm{d}h=\\&-\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}(K_\theta+\sigma_\varepsilon^2 I)^{-1}\boldsymbol{y}-\frac{1}{2}\log|K_\theta+\sigma_\varepsilon^2 I|-\\&\frac{D}{2}\log\left(2\pi\right)\end{aligned} \tag{7}$$

假设最优超参数 θ^* 已确定, 则对于任意的测试输入 $\boldsymbol{x}_*$, 函数值 $h_*=h(\boldsymbol{x}_*)$ 的预测输出是一个高斯分布, 其均值和协方差的计算分别如式 (8) 和 (9):

$$E_h\left[h_*\right]=k\left(\boldsymbol{x}_*,X\right)\left(K+\sigma_\varepsilon^2 I\right)^{-1}\boldsymbol{y} \tag{8}$$

$$\begin{aligned}\mathrm{var}_h\left[h_*\right]=&\,k\left(\boldsymbol{x}_*,\,\boldsymbol{x}_*\right)-\\&k\left(\boldsymbol{x}_*,X\right)\left(K+\sigma_\varepsilon^2 I\right)^{-1}k\left(X,\boldsymbol{x}_*\right)\end{aligned} \tag{9}$$

式中, K 表示核方差矩阵, 其元素 $k_{ij}=k(\boldsymbol{x}_i,\boldsymbol{x}_j)$.

如果 MAS 环境下智能体的状态、动作可观测, GR 可以实现跟踪学习架构中环境状态转移、值函数、策略的建模和泛化, 从而实现连续空间多智能体策略学习. 显然, 所有建模过程都基于与状态相关的观测样本集, 记为 S_l.

2.1 多智能体环境下的高斯回归模型

采用 GR 实现在线建立环境状态转移模型、值函数模型以及策略模型, 主要问题在于如何在 MAS 环境下实时实现模型样本的采集和回归建模, 并减少高维带来的高复杂度计算.

1) 基于高斯回归的状态转移模型

对于一个状态维度为 D 的环境, 建立环境状态转移模型 (记为 GR_f), 包含 D 个独立 GR 模型 (记为 $GR_f^d, d=1,\cdots,D$), 分别对应状态空间的每个维度, 共同组成联合状态 - 个体动作环境状态转移模型. 所有 GR 模型的输入均为状态 - 动作对 $(\boldsymbol{s},a^i)$, GR_f^d 输出为

$$\Delta s'_d=f_d(\boldsymbol{s},a^i)-s_d,\quad d=1,\cdots,D$$

其中, 算子 f_d 表示状态 - 动作对 $(\boldsymbol{s},a^i)$ 到其后继状态 $\boldsymbol{s}'$ 中第 d 维 s'_d 的映射. 对于任意测试输入 $(\boldsymbol{s}_*,a_*^i)$, 其预测输出值的期望和方差如式 (10) 和 (11) 所示:

$$\mathrm{E}_f\left[f_d(\boldsymbol{s}_*,a_*^i)\,\middle|\boldsymbol{s}_*,a_*^i\right]=s_{*d}+\mathrm{E}_f[\Delta s'_{*d}\,|\boldsymbol{s}_*,a_*^i] \tag{10}$$

$$\mathrm{var}_f\left[f_d(\boldsymbol{s}_*,a_*^i)\,\middle|\boldsymbol{s}_*,a_*^i\right]=\mathrm{var}_f[\Delta s'_{*d}\,|\boldsymbol{s}_*,a_*^i] \tag{11}$$

式中, E_f 和 var_f 分别根据式 (8) 与 (9) 计算获得.

2) 基于高斯回归的值函数模型

为了实现状态和动作的泛化, 以离散状态 - 动作对的 Q 值为基础, 利用全概率 GR 模型建立连续空间 $Q^*(\boldsymbol{s},a^i)$ 和 $V^*(\boldsymbol{s})$ 的模型, 分别记为 GR_q 和 GR_v. 训练样本集 S_l 的规模上界可以根据问题合理选择.

状态值函数模型 GR_v 是实现 Q 函数计算的核心模型. 为此, 建立 GR_v 提供整个状态空间 S 下的 V 值分布模型, 任意状态点的 V 值可以通过式 (8) 计算估计均值.

建立 Q 值模型的目的在于使智能体在进行 V 值学习时, 能够根据 Q 值做策略选择, 因此, 只需要

与样本集 S_l 状态对应的 Q 值函数模型, 没有必要建立整个状态 - 动作联合空间 $S \times A$ 上的 Q 函数. 与传统上对整个状态空间建立统一的 Q 函数模型有所不同, 本文算法针对样本集 S_l 中的各个样本状态进行 Q 函数分离建模. 其中, 学习样本的 Q 值根据式 (4) 计算, $V(\boldsymbol{s}')$ 和 $p(\boldsymbol{s}'|\boldsymbol{s}, a^i)$ 分别由 GR_v 和 GR_f 获得. 采用样本空间 Q 值模型分离建模的方式可以减少单个 Q 函数模型的训练样本量, 提高 Q 函数模型的准确性.

3) 基于高斯回归的策略模型

策略可以看成是状态到动作的映射, 为了学习得到整个状态空间下的连续型策略函数, 需要利用 GR 建立策略函数模型 GR_π, 其训练样本的输入是状态样本集 S_l, 输出为 S_l 对应的最佳动作集 $a_l^*(S_l)$. 策略模型 GR_π 的作用是将基于有限状态集的跟踪策略泛化成一个连续型的全局跟踪策略.

3 基于高斯回归建模的多智能体跟踪学习

第 2.1 节采用 GR 建模实现未知环境、先验知识缺乏条件下, 对环境、值函数、策略的观测建模和泛化, 构成基于模型的 MAS 跟踪学习架构的关键模型. 为了完成跟踪学习架构, 仍需要解决两个问题:

1) 从图 1 可以看到, S_l 中 GR_q 的建模需要个体即时回报 $r(\boldsymbol{s}, a^i)$. 在环境未知条件下, 由于无法事先得到联合状态 - 联合动作的 $r(\boldsymbol{s}, \boldsymbol{a})$, 不能使用式 (2) 计算 $r(\boldsymbol{s}, a^i)$. 因此, 需要对 S_l 中的每个状态样本, 观测与之相关的联合即时回报 $r(\boldsymbol{s}, \boldsymbol{a})$, 进而通过逼近学习获得 $r(\boldsymbol{s}, a^i)$.

2) 由于智能体对环境先验知识少, GR 建模过程的状态 - 动作空间样本集 S_l 需要随着智能体的学习不断进行扩充和修改, 使 GR 泛化模型能逼近连续状态 - 动作空间的特征.

3.1 样本空间个体即时回报样本的迭代学习

式 (4) 定义的 Q 值中 $r(\boldsymbol{s}, a^i)$ 是基于联合状态 - 个体动作的个体回报 (2). 在实际 MAS 环境下, 学习智能体大多只能观测联合状态 - 联合动作条件的即时回报 $r(\boldsymbol{s}, \boldsymbol{a})$. $r(\boldsymbol{s}, a^i)$ 描述基于同伴策略分布函数的期望即时回报, 是学习智能体对合作效果的一种认知, 在实际 MAS 环境中无法直接观测. 因此, 需要建立一种迭代逼近算法, 实现基于 $r(\boldsymbol{s}, \boldsymbol{a})$ 的 $r(\boldsymbol{s}, a^i)$ 逼近学习, 从而为样本集 S_l 中 GR_q 的学习提供即时回报函数.

基于 MAS 系统同伴策略学习与即时回报函数学习等效的性质 (这个性质的分析已超出本文讨论范围, 将另文说明), 本文提出单体即时回报迭代公式如式 (12) 所示, 实现 S_l 中所有状态 - 动作对样本所对应的即时回报更新:

$$r_t(\boldsymbol{s}, a^i) = \mu_t^{N-1} r(\boldsymbol{s}, \boldsymbol{a}) - \sum_{k=1}^{N-1} \mathrm{C}_{N-1}^k (\mu_t - 1)^k r_{t-k}(\boldsymbol{s}, a^i) \tag{12}$$

式中, N 表示智能体的个数, $\mu_t \in [0, 1]$ 为学习率, $\mathrm{C}_{N-1}^k = (N-1)!/(k!(N-1-k)!)$.

式 (12) 可以视为从联合回报到即时回报的信度分配, 并且基于多重历史回报的迭代方式能够减少感知误差. 假设迭代次数足够多, 基于式 (12) 最终获得即时回报值能够收敛于期望回报值 $r^*(\boldsymbol{s}, a^i)$.

3.2 动态样本集调整

在未知环境下, 为了提高建模的有效性和效率, 智能体需要在搜索的时候选择具有描述空间特征的样本构成 GR 建模的动态联合状态集 S_l. 因此, 智能体以当前学到策略 GR_π 进行试验, 获得候选状态集 S_{can}, 基于平衡探索 - 利用的判断准则, 建立价值判定函数 (13):

$$U(s_{\mathrm{can}}) = \alpha \mathrm{E}_v \left[V_k^*\left(S_{\mathrm{can}}\right)|S_l\right] + \frac{\beta}{2} \log\left(\mathrm{var}_v\left[V_k^*\left(S_{\mathrm{can}}\right)|S_l\right]\right) \tag{13}$$

式中权系数 α、β 值分别反映了新状态在利用 - 探索平衡方面的取舍. 根据价值判定函数 U 判定 S_{can} 中状态的价值, 选取价值最高的 k 个状态加入样本集 S_l 中. 为了防止新增状态中出现多个状态相似度极高的情况, 采用逐个添加的方式, 即每次只选取 U 值最大的新状态并更新 U 函数中的 S_l 直至添加过程结束.

3.3 算法流程

结合 GR 建模和值函数设计, MAS MBRL-CPT 的执行过程如算法 1 所示. 算法的主体由 L 次迭代学习组成, 动态联合状态集 S_l $(l = 1, \cdots, L)$ 从初始状态集 S_0 开始持续添加新样本以扩充样本集. 每执行一次迭代递归, 动态联合状态集 S_l、环境状态转移函数模型 GR_f、单体即时回报函数 $r(\boldsymbol{s}, a^i)$、Q 函数逼近器 GR_q、V 函数逼近器 GR_v、智能体策略模型 GR_π 均更新一次. 算法最终学到的是一个连续型的全局策略函数 GR_π.

算法 1. 算法流程

01 初始化样本集 S_0、环境模型 GR_f、立即回报 $r_0(\boldsymbol{s}, a^i)$
02 $V_0^*(\boldsymbol{s}) = \max\limits_{a^i} r_0(\boldsymbol{s}, a^i) + \omega$, ω 为噪声
03 初始化 V 函数, $V_0^*(S_0) \to GR_v$
04 for $l = 1$ to L do
05 添加 K 个新状态, $S_l \to S_{l+1}$, 更新 GR_f、$r_l(\boldsymbol{s}, a^i)$
06 for all $\boldsymbol{s}_i \in S_l$ do
07 for all $a_j \in A$ do

$$r_l(\boldsymbol{s}_i,a_j)=\mu_l^{N-1}r(\boldsymbol{s}_i,\boldsymbol{a})-\sum_{k=1}^{N-1}C_{N-1}^{k}(\mu_l-1)^k r_{l-k}(\boldsymbol{s}_i,\boldsymbol{a}_j)$$

$$Q_l^*(\boldsymbol{s}_i,a_j)=r_l(\boldsymbol{s}_i,a_j)+\gamma\int p(\boldsymbol{s}'_i|\boldsymbol{s}_i,a_j)V_{l-1}(\boldsymbol{s}_i')\,\mathrm{d}\boldsymbol{s}'_i$$

end for a_j

$Q_l^*(\boldsymbol{s}_i,\cdot)\rightarrow GR_q$

$\pi_l^*(\boldsymbol{s}_i)=\arg\max_{a_*\in A}Q_l^*(\boldsymbol{s}_i,a_*)$

$\mathrm{V}_l^*(\boldsymbol{s}_i)=Q_l^*(\boldsymbol{s}_i,\pi_l^*(\boldsymbol{s}_i))$

end for $\boldsymbol{s}_i$

$V_l^*(\cdot)\rightarrow GR_v$

$\pi_l^*(\cdot)\rightarrow GR_\pi$

end for l

根据 S_L、$\pi_L^*(S_L)$ 建立连续型策略函数逼近器 GR_π

3.4 算法性能分析

本文从算法的稳定性和复杂度两方面分析算法的有效性, 特别是在降低模型参数数量和状态空间维度两方面的效果.

1) 稳定性分析

MAS MBRL-CPT 的核心包括三个部分: 离散样本的个体期望即时回报的迭代更新、离散样本集合的值函数求解, 以及模型与值函数泛化. 三者分别采用随机逼近、动态规划和 Bayesian 推理. 个体期望即时回报的迭代更新只与离散样本集合相关, 与值函数求解和泛化操作无关, 其收敛性不受后两者影响; 离散样本集合的值函数求解与环境模型的泛化建模相关, 但与值函数泛化无关. 因此, MAS MBRL-CPT 的稳定性分析包括 4 个部分:

a) 离散样本的个体期望即时回报的迭代更新

由式 (12) 可知, 期望即时回报的更新过程只与当前 MAS 群体行为的即时回报, 以及历史期望即时回报相关, 与离散样本的值函数无关, 且不显性包含状态转移概率. 如果 MAS 群体行为的即时回报是时不变且可观测的, 式 (12) 满足鞅过程收敛性质. 可以证明其收敛到式 (2), 即从学习智能体角度观测到的个人行为即时回报.

b) 环境状态转移概率模型 GR_f 建模

GR_f 实现对 $p(\boldsymbol{s}'|\boldsymbol{s},a^i)$ 建模. 易知,

$$p(\boldsymbol{s}'|\boldsymbol{s},a^i)=\sum_{a^j\in A^j,j\neq i}\prod_{a^j}p(a^j|\boldsymbol{s})p(\boldsymbol{s}'|\boldsymbol{s},a^i,a^{\mathrm{other}})$$

其中, a^{other} 表示除学习智能体 i 之外其他智能体的联合动作, $p(a^j|\boldsymbol{s})$ 表示同伴在联合状态 $\boldsymbol{s}$ 下选择动作的概率, 即同伴策略. 显然, 若同伴策略平稳, $p(\boldsymbol{s}'|\boldsymbol{s},a^i)$ 时不变, 当样本集规模足够且状态转移记录足够多时, GR 可以对 $p(\boldsymbol{s}'|\boldsymbol{s},a^i)$ 实现逼近.

c) 离散样本集合 Q 值函数求解

显然, 若个体期望即时回报 $r(\boldsymbol{s},a^i)$ 和 GR_f 收敛, 由式 (4) 可知, 智能体 i 的学习过程是 Markov 决策过程 (Markov decision process, MDP), 通过动态规划可以对 Q 函数的 Bellman 方程求解, 得到 $Q(\boldsymbol{s},a^i)$.

d) 连续空间 Q 值函数、V 值函数以及策略函数的 GR 建模

三者的 GR 建模基于离散样本集合 S_l 中样本的 Q 值函数, 若离散样本集合 Q 值函数稳定, 则 GR 可以实现对离散值函数和策略的连续空间泛化.

由以上分析可知, 个体期望回报的学习实现了 MAS 环境下个体行为回报的观测, 结合环境建模, 从学习智能体角度构造 MDP; 通过动态规划实现值函数的在线求解, 构造离散样本集合的值函数样本; 通过策略泛化获得连续空间个体策略. 上述三个过程的交替进行, 实现了智能体在 MAS 环境下跟踪策略的学习.

2) 时间复杂度

训练一个拥有 n 个样本的标准 GR 模型需要 $\mathrm{O}(n^3)$ 次运算, GR 预测均值和方差分别需要 $\mathrm{O}(n)$、$\mathrm{O}(n^2)$ 次运算[11]. 训练比预测的时间复杂度至少大一个数量级.

基于高斯回归的 MAS 跟踪学习的主要计算量来自 GR 在线训练模型. 假设联合环境维度为 d_s, 个体动作维度为 d_a, 在当前第 l 次迭代递归中, 训练 GR_f、GR_q、GR_π、GR_v 分别需要 $\mathrm{O}(d_{\mathrm{s}}|S_l|^3|A|^3)$、$\mathrm{O}(|S_l||A|^3)$、$\mathrm{O}(d_a|S_l|^3)$、$\mathrm{O}(|S_l|^3)$ 运算操作, 因此, 单次迭代的时间复杂度为 $O_l=\mathrm{O}((d_{\mathrm{s}}|A|^3+d_a+1)|S_l|^3+|S_l||A|^3)$. 可见算法时间复杂度取决于状态样本集中样本的数量 $|S_l|$、动作规模 $|A|$、递归次数 L, 其中, 环境模型与状态值函数模型在线更新的计算量决定了总时间复杂度的数量级.

3) 空间复杂度

标准 GR 模型的超参数个数为 $d+2$, 其中, d 表示训练样本输入的维度. 由于 GR_f、GR_q、GR_π、GR_v 分别需要 d_{s}、$|S_L|$、d_a、1 个独立的标准 GR, 相应参数个数为 $d_s(d_s+d_a+2)$、$|S_L|(d_a+2)$、$(d_s+2)d_a$、d_s+2. 因此, MAS MBRL-CPT 共有模型参数 $M_1=|S_L|(d_a+2)+(d_s+2)(d_s+d_a+1)+d_sd_a$ 个. 同样条件下的离散空间无模型学习需要存储的参数为 $M_2=|S|^2|A|+|S|(|A|+2)$ 个. 显然, 在大规模问题中, $|S|^2|A|\gg|S_L|(d_a+2)$、$|S|(|A|+2)\gg(d_s+2)(d_s+d_a+1)+d_sd_a$, 并且随着问题规模的增大, M_2 呈指数增长趋势, M_1 呈线性增长趋势. 因此, 相比离散空间无模型学习算法, MAS MBRL-CPT 极大地降低了对存储空间的需求.

此外, MAS 降维 Q 值函数定义的维度为联合

状态维度与个体动作维度之和, 相比于基于联合状态-联合动作空间 Q 值函数, 在维度上少了其他智能体的联合动作维度. 当智能体数量较多或动作维度较高的情况下, 降维效果尤其明显.

4　连续系统控制仿真及分析

Cart-pole 平衡控制系统是一种强非线性平衡控制问题, 它广泛地被用作验证强化学习效果的基准问题[22]. 增加 Cart-pole 系统中小车的数量使其成为 MAS 环境下的跟踪式平衡控制问题, 改进后得到 Multi-cart-pole 控制系统如图 2 所示, 系统动力学方程如式 (14) ∼ (17):

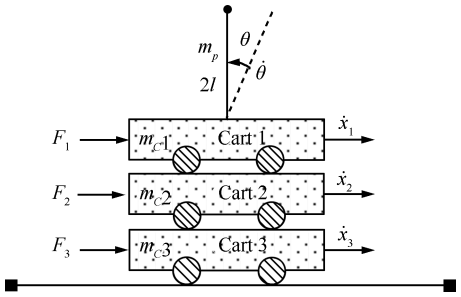


图 2　Multi-cart-pole 控制系统

Fig. 2　The multi-cart-pole control system

$$\ddot{\theta}=\frac{g\sin\theta+\cos\theta\left(\dfrac{-F_1-m_{c1}\ddot{x}_2-m_pl\dot{\theta}^2\sin\theta}{m_{c1}+m_p}\right)}{l\left(\dfrac{4}{3}-\dfrac{m_p\cos\theta^2}{m_{c1}+m_p}\right)} \tag{14}$$

$$\ddot{x}_1=\frac{F_1+m_{c1}\ddot{x}_2+m_pl\left(\dot{\theta}^2\sin\theta-\ddot{\theta}\cos\theta\right)}{m_{c1}+m_p} \tag{15}$$

$$\ddot{x}_2=\frac{F_2+m_{c2}\ddot{x}_3}{m_{c2}} \tag{16}$$

$$\ddot{x}_3=\frac{F_3}{m_{c3}} \tag{17}$$

其中, θ、$\dot{\theta}$ 表示杆的倾斜角度和角速度, x_1, x_2, x_3 分别表示三个小车的水平位移, m_p, m_{c1}, m_{c2}, m_{c3} 分别是杆、三个小车的质量, $2l$ 为杆的长度, g 为重力加速度.

定义 Multi-cart-pole 平衡控制系统中 Cart 1 为学习智能体, $\boldsymbol{s}=\{\theta,\dot{\theta},x_1,\dot{x}_1,x_2,\dot{x}_2,x_3,\dot{x}_3\}$ 为 MAS 联合状态, 所有状态变量均为连续变量, 作用力 F_1、F_2、F_3 分别表示 Cart 1、Cart 2、Cart 3 的动作, Cart 1 的学习目标是生成能够适应 Cart 2, Cart 3 策略的跟踪策略, 与 Cart 2、Cart 3 协作完成倒立杆的平衡控制目标.

依据实际物理系统, 仿真实验中, 取物理参数 $m_p=0.1\,\mathrm{kg}$, $m_{c1}=1.0\,\mathrm{kg}$, $m_{c2}=2.0\,\mathrm{kg}$, $m_{c3}=2.0\,\mathrm{kg}$, $2l=1.0\,\mathrm{m}$, 取仿真采样周期 $T=0.2\,\mathrm{s}$. 设计 Multi-cart-pole 控制系统中 Cart 2、Cart 3 均执行策略 $F_2\sim\mathrm{N}(-\sin\theta,0.1)$、$F_3\sim\mathrm{N}(-0.2\theta,0.1)$. Cart 1 采用 MAS MBRL-CPT 算法, 并且无任何关于系统动力学和 Cart 2、Cart 3 策略的先验知识. 环境对系统行为的立即奖赏函数如式 (18) 和 (19) 所示:

$$o^2=2h^2\left(1-\cos\theta\right) \tag{18}$$

$$r=-1+\exp\left(-0.5c\cdot o^2\right) \tag{19}$$

其中, h 表示倒立杆的长度, c 为常数 (本文取 25).

1) MAS MBRL-CPT 的动态过程和学习效果

图 3 记录了 Cart 1 学习过程中在 4 个时刻的样本空间 S_l. 其中, 十字标志反映倒立杆角度与角速度两维坐标下的样本分布. 学习之前的初始样本空间只包含初始状态和目标状态, 学习过程中采用动态样本集调整方法在线增添新样本. 对比 4 个时间点的样本空间可以发现, 本文采用的样本集调整方法可以根据潜在样本的信息量有效地搜索状态空间, 动态扩展样本集 S_l, 新增样本过程呈现出从初始状态向目标状态逐渐靠拢的趋势, 并且目标状态附近的样本数量明显较多. 显示目标状态附近的信息含量高.

为验证环境状态转移概率模型 GR_f 的建模过程, 实验中随机选择 10 组状态-动作对, 在每次 GR_f 更新后, 以这些状态-动作对作为环境模型的输入, 作一步后继状态的预测. 表 1 给出了 5 次预测中 10 个预测值与真实值之间误差的均值. 容易发现, 随着 GR_f 学习中持续增加环境模型样本, 预测误差越来越小, 说明算法具备在缺乏先验知识条件下, 通过自主选择样本的方式, 实现连续空间 MAS 环境下的建模.

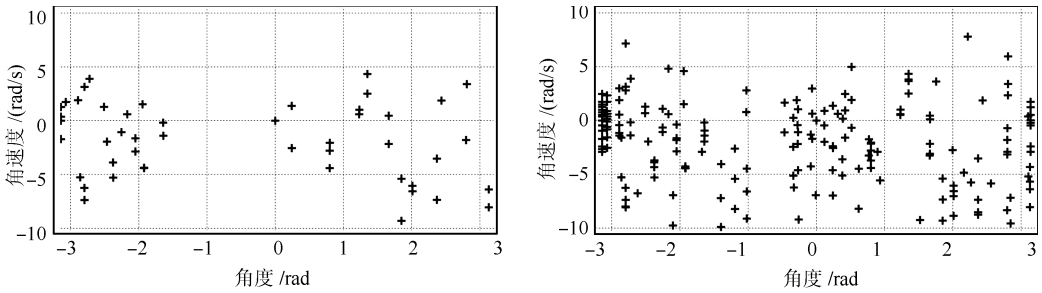
图 4 以极坐标下的倒立摆运动轨迹反映 4 个不同时刻所获得的控制策略效果, 极径表示时间长度, 极角表示倒立摆角度. 随着学习程度的加深 (探索试验次数增加和样本集更新), 控制效果越来越好.

图 5 记录 Multi-cart-pole 控制系统中倒立摆和三个移动平台的 4 个速度变化曲线. 3 秒之前系统处于摇起阶段, 速度剧烈波动, 随后, Pole、Cart 2、Cart 3 相继到达静止状态, 而 Cart 1 仍然在做低速运动, 直到第 6 秒运动结束, 整个 Multi-cart-pole 系统进入准静止状态 (由于 F_2、F_3 存在一定概率的不确定性, 系统不可能到达绝对静止状态).

表 1 环境模型误差分析表

Table 1 The table for prediction error of environment model

L	θ	$\dot{\theta}$	x_1	$\dot{x}_1$	x_2	$\dot{x}_2$	x_3	$\dot{x}_3$
03	−0.0072	−0.0046	−0.0132	−0.0321	−0.0072	−0.0024	0.0006	−0.0037
06	0.0021	0.0042	0.0066	−0.0050	0.0027	0.0031	0.0003	−0.0011
09	0.0015	0.0025	0.0007	−0.0060	−0.0016	0.0021	0.0006	−0.0004
12	0.0005	0.0005	0.0002	−0.0021	−0.0003	0.0005	0.0003	0.0001
15	0.0001	−0.0003	−0.0000	−0.0013	−0.0000	−0.0001	−0.0000	−0.0001



(a) 样本数量为 50

(a) The distribution when the number of sample is 50

(b) 样本数量为 200

(b) The distribution when the number of sample is 200

(c) 样本数量为 400

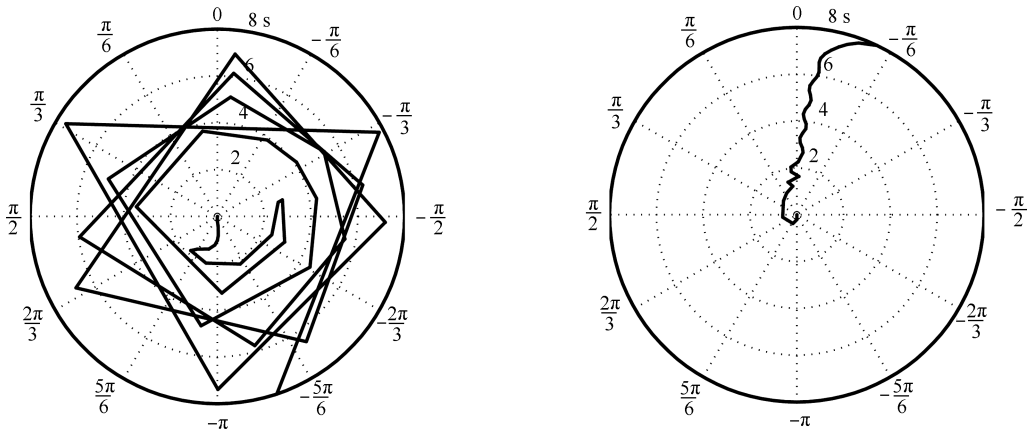
(c) The distribution when the number of sample is 400

(d) 样本数量为 800

(d) The distribution when the number of sample is 800

图 3 学习过程不同时刻状态样本分布

Fig. 3 Sample states distribution during learning

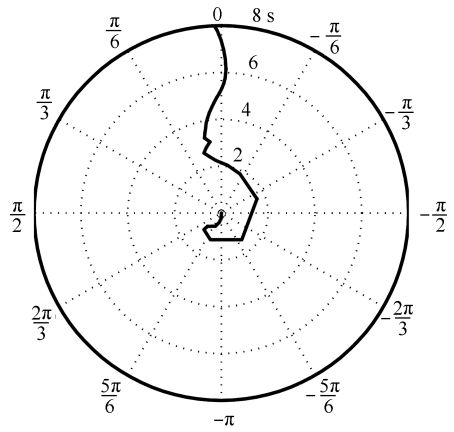


(a) 第 6 次迭代学习

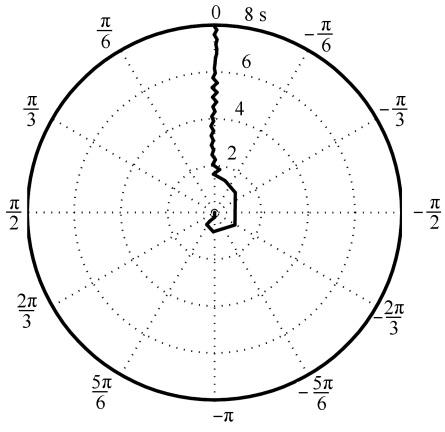
(a) The 6th iteration

(b) 第 9 次迭代学习

(b) The 9th iteration



(c) 第 12 次迭代学习
(c) The 12th iteration



(d) 第 15 次迭代学习
(d) The 15th iteration

图 4　学习过程不同时刻 Cart 1 学习的协作策略控制效果

Fig. 4　Control effect during learning under the cooperative strategy learned by Cart 1

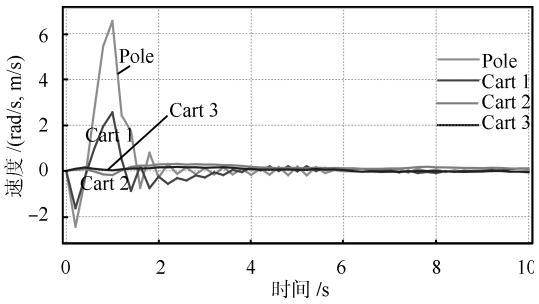


图 5　速度曲线

Fig. 5　Curves of speed

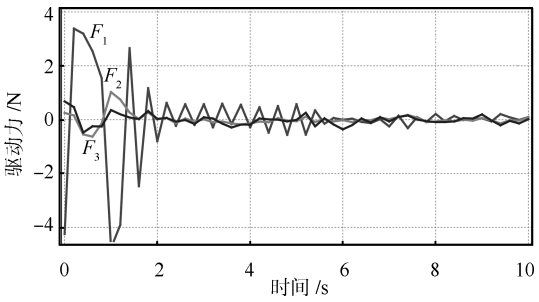


图 6　动作曲线

Fig. 6　Curves of movement

图 6 记录是 Multi-cart-pole 控制系统中的三个移动平台动作 (驱动力) 曲线. 在 3 秒之前的摇起阶段, 三个小车驱动力 F_1、F_2、F_3 均变化剧烈, 直到第 6 秒 F_1、F_2、F_3 趋近零, 系统进入准平衡状态. 从控制过程来看, Cart 1 学习到的合作控制策略使控制过程大致分为两个时段. 第一个控制时段为 0～3 秒, 实现倒立杆的摇起控制, 将倒立杆送达目标区域, 但移动平台仍然处于运动状态; 第二个控制时间段为 3～6 秒, 该控制时段在保持倒立杆基本平衡的前提下, 实现移动平台从运动状态转为准静止状态.

另一方面, 由于真实环境模型未知, 平衡区动力学特性复杂, 基于有限样本集的状态转移泛化模型与真实环境存在差异, 以及有限时间搜索连续最优策略的特点, 都会引起输出动作 (即驱动力) 在平衡区邻域内出现小幅震荡. 这可以通过扩展平衡区域附近的样本集, 延长策略搜索行为来缓解, 但同时也会增加强化学习的存储和计算负担.

2) MAS MBRL-CPT 与分布式独立学习的对比

为了进一步说明本文学习算法的特点, 选择分布式独立学习 (Distributed independent learning, DIL) 算法作为比较对象 (在 DIL 的基础上, 采用 GR 进行值函数逼近, 实现状态空间的泛化). 在仿真环境设置和折扣因子 γ 设置相同时, 重复运行两种算法 20 次求取平均值, 得到图 7 的对比结果, 其中, 横轴表示试验次数, 纵轴表示完成一次试验所需要的时间步数.

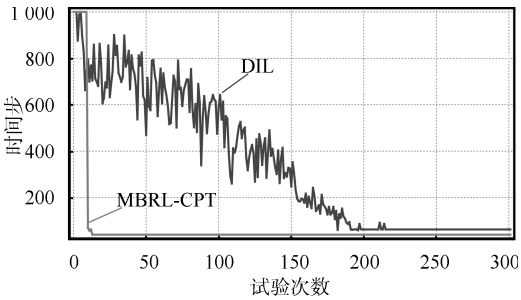


图 7　MBRL-CPT 与 DIL 的对比仿真

Fig. 7　Comparison between MBRL-CPT and DIL

图 7 中的曲线表明, MAS MBRL-CPT 在 15 次探索试验后 ($L = 15$) 可以获得跟踪策略. 相比之下, DIL 则需要大约 190 次探索试验 (Episode ≈

190) 才能实现最佳策略. 可见, 与 DIL 相比, MAS MBRL-CPT 使用较少的探索试验次数就可以获得智能体的协作策略. 而且, 从最终收敛位置来看, 其最终学习到的收敛策略也比 DIL 更优.

综合上述的实验结果说明, 即使环境和同伴策略未知, 学习智能体基于图 1 的跟踪学习框架能够建立连续空间值函数、策略和环境模型, 并运用基于模型的策略学习获得学习智能体的最优跟踪策略, 实现与同伴合作完成任务的目的.

5 结论

针对传统多智能体学习算法的适应性不足、泛化能力弱、学习效率低等问题, 以实现智能体跟踪同伴行为策略为目标, 提出了一种连续空间基于模型的多智能体跟踪学习算法. 通过构造一种跟踪学习架构, 将降维值函数的学习和环境建模结合起来, 在实现学习空间降维的同时, 达到连续空间泛化的目的. 通过设计个体回报的逼近学习算法和样本集动态调整方法, 解决了建模过程中重要的值函数样本计算和样本集动态调整问题. 智能体通过基于模型的学习可以在环境未知条件下获得适应同伴行为模式的最优跟踪策略, 从而有效地与同伴合作完成目标任务.

本文假设同伴策略为平稳策略的目的是保证个体期望折扣回报更新的收敛性, 因此, 算法并不能实现严格意义的分布式学习条件下的 MAS 同时学习. 但是, 跟踪学习实现了 MAS 环境下团队奖励的合理分配, 以及分布式学习的结构. 通过引入协调机制, 使多智能体在状态空间子区域交替采用 MBRL-CPT, 可以实现合作策略的持续优化, 并最终形成合作策略. 这可以在宏观时间上实现类似同时学习的效果, 也是我们的下一步研究方向.

References

1 Busoniu L, Babuska R, De Schutter B. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2008, **38**(2): 156−172

2 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 1996, **4**: 237−285

3 Chen Xue-Song, Yang Yi-Min. Reinforcement learning: survey of recent work. *Application Research of Computers*, 2010, **27**(8): 2834−2838, 2844
(陈学松, 杨宜民. 强化学习研究综述. 计算机应用研究, 2010, **27**(8): 2834−2838, 2844)

4 Cheng Yu-Hu, Feng Huan-Ting, Wang Xue-Song. Policy iteration reinforcement learning based on geodesic Gaussian basis defined on state-action graph. *Acta Automatica Sinica*, 2011, **37**(1): 44−51
(程玉虎, 冯涣婷, 王雪松. 基于状态 - 动作图测地高斯基的策略迭代强化学习. 自动化学报, 2011, **37**(1): 44−51)

5 Xu Xin, Shen Dong, Gao Yan-Qing, Wang Kai. Learning control of dynamical systems based on Markov decision processes: research frontiers and outlooks. *Acta Automatica Sinica*, 2012, **38**(5): 673−687
(徐昕, 沈栋, 高岩青, 王凯. 基于马氏决策过程模型的动态系统学习控制: 研究前沿与展望. 自动化学报, 2012, **38**(5): 673−687)

6 Busoniu L, De Schutter B, Babuška R. Approximate dynamic programming and reinforcement learning. In: Proceedings of the 2010 Interactive Collaborative Information Systems, Studies in Computational Intelligence. Berlin Heidelberg: Springer, 2010, **281**: 3−44

7 Wang Xue-Song, Tian Xi-Lan, Cheng Yu-Hu, Yi Jian-Qiang. *Q*-learning system based on cooperative least squares support vector machine. *Acta Automatica Sinica*, 2009, **35**(2): 214−219
(王雪松, 田西兰, 程玉虎, 易建强. 基于协同最小二乘支持向量机的 Q 学习. 自动化学报, 2009, **35**(2): 214−219)

8 Busoniu L, Ernst D, De Schutter B, Babuska R. Online least-squares policy iteration for reinforcement learning control. In: Proceedings of the 2010 American Control Conference. Baltimore, USA: IEEE, 2010. 486−491

9 Rasmussen C E, Kuss M. Gaussian processes in reinforcement learning. In: Proceedings of the 17th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2003. 751−759

10 Jung T, Stone P. Gaussian processes for sample efficient reinforcement learning with RMAX-like exploration. In: Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases, Part I. Berlin, Heidelberg: Springer-Verlag, 2010. 601−616

11 Deisenroth M P, Rasmussen C E. PILCO: a model-based and data-efficient approach to policy search. In: Proceedings of the 28th International Conference on Machine Learning. Washington, USA, 2011. 465−472

12 Deisenroth M P, Rasmussen C E, Peters J. Gaussian process dynamic programming. *Neurocomputing*, 2009, **72**(7−9): 1508−1524

13 Wu Jun, Xu Xin, Wang Jian, He Han-Gen. Recent advances of reinforcement learning in multi-robot systems: a survey. *Control and Decision*, 2011, **26**(11): 1601−1610, 1615
(吴军, 徐昕, 王健, 贺汉根. 面向多机器人系统的增强学习研究进展综述. 控制与决策, 2011, **26**(11): 1601−1610, 1615)

14 Hu J L, Wellman M P. Nash Q-learning for general-sum stochastic games. *The Journal of Machine Learning Research*, 2003, **4**: 1039−1069

15 Greenwald A, Hall K. Correlated Q-learning. In: Proceedings of the 20th International Conference on Machine Learning. Washington D. C., USA: AAAI Press, 2003. 242−249

16 Conitzer V, Sandholm T. AWESOME: a general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, 2007, **67**(1−2): 23−43

17 Weinberg M, Rosenschein J S, Paul K. Best-response multiagent learning in non-stationary environments. In: Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems. Washington D. C., USA : IEEE, 2004. 506−513

18 Chen C L, Li H X, Dong D Y. Hybrid control for robot navigation: a hierarchical Q-learning algorithm. *IEEE Robotics and Automation Magazine*, 2008, **15**(2): 37−47

19 Dai Zhao-Hui, Yuan Jiao-Hong, Wu Min, Chen Xin. Dynamic hierarchical reinforcement learning based on probability model. *Control Theory and Applications*, 2011, **28**(11): 1595−1600, 1606
(戴朝晖, 袁姣红, 吴敏, 陈鑫. 基于概率模型的动态分层强化学习. 控制理论与应用, 2011, **28**(11): 1595−1600, 1606)

20 Shoham Y, Powers R, Grenager T. Multi-agent Reinforcement Learning: a Critical Survey, Technical Report, Computer Science Department, Stanford University, 2003

21 Rasmussen C E, Williams C K I. *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: The MIT Press, 2006

22 Florian R V. Correct Equations for the Dynamics of the Cart-pole System. Technical Report, Center for Cognitive and Neural Studies, 2007

陈 鑫 中南大学副教授. 主要研究方向为多智能体系统, 智能控制和过程控制.
E-mail: chenxin@csu.edu.cn
(**CHEN Xin** Associate professor at Central South University. His research interest covers multi-agent systems, intelligent control and process control.)

魏海军 中南大学硕士研究生. 主要研究方向为多智能体系统, 强化学习.
E-mail: ecnavy@163.com
(**WEI Hai-Jun** Master student at Central South University. His research interest covers multi-agent systems and reinforcement learning.)

吴 敏 中南大学教授. 主要研究方向为鲁棒控制, 智能控制, 过程控制.
E-mail: min@csu.edu.cn
(**WU Min** Professor at Central South University. His research interest covers robust control, intelligent control, and process control.)

曹卫华 中南大学教授. 主要研究方向为多智能体系统, 智能控制, 过程控制. 本文通信作者.
E-mail: caowh@csu.edu.cn
(**CAO Wei-Hua** Professor at Central South University. His research interest covers multi-agent systems, intelligent control, and process control. Corresponding author of this paper.)