基于高斯过程分类器的连续空间强化学习

王雪松[1,2],张依阳[1],程玉虎[1]
(1. 中国矿业大学信息与电气工程学院,江苏徐州 221116;2. 中国科学院自动化研究所,北京 100190)

摘 要: 如何将强化学习方法推广到大规模或连续空间,是决定强化学习方法能否得到广泛应用的关键. 不同于已有的值函数逼近法,把强化学习构建为一个简单的二分类问题,利用分类算法来得到强化学习中的策略,提出一种基于高斯过程分类器的连续状态和连续动作空间强化学习方法. 首先将连续动作空间离散化为确定数目的离散动作,然后利用高斯分类器对系统的连续状态 - 离散动作对进行正负分类,对判定为正类的离散动作按其概率值进行加权求和,进而得到实际作用于系统的连续动作. 小船靠岸问题的仿真结果表明所提方法能够有效解决强化学习的连续空间表示问题.

关键词: 高斯过程; 分类器; 连续空间; 强化学习; 小船靠岸问题
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2009) 06-1153-06

Reinforcement Learning for Continuous Spaces Based on Gaussian Process Classifier

WANG Xue-song[1,2], ZHANG Yi-yang[1], CHENG Yu-hu[1]
(1. *School of Information and Electrical Engineering, China University of Mining & Technology, Xuzhou, Jiangsu* 221116, *China*;
2. *Institute of Automation, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: The generalization of reinforcement learning methods to large-scale or continuous spaces has become a major focus in the research field of reinforcement learning. Unlike the present reinforcement learning methods for continuous spaces based on a value-function approximation method, the reinforcement learning is constructed as a simple binary-class problem. A kind of reinforcement learning method for continuous state and action spaces based on a Gaussian process classifier is proposed using a classification algorithm to obtain a control policy. At first, a continuous action space is discretized into discrete actions with definite number, and the Gaussian process classifier is used to predict the probability of class for a continuous-state-discrete-action pair. Then a continuous action is generated based on a weighted operation of the positive actions with their probability values. Computer simulations involving a boat problem illustrate the validity of the proposed reinforcement learning method.

Key words: Gaussian process; classifier; continuous space; reinforcement learning; boat problem

1 引言

强化学习基于动物学习心理学的有关原理,采用人类和动物学习中的试错法机制,强调在与环境的交互中学习,可以不需要系统模型而实现无导师的在线学习. 对于很多领域的实际问题,从人工智能的博弈问题、调度优化、智能机器人到实际的工业过程控制都可以描述为强化学习问题,因而强化学习具有广阔的应用前景. 但是,实际系统的空间往往是大规模或连续的,强化学习不可避免的存在状态变量的空间复杂度问题,即维数灾难. 因此,与强化学习问题的理论模型相比,实际的应用问题要复杂得多,这导致了强化学习理论在实际应用中的困难.

在强化学习领域,解决连续空间的表示问题主要有三类方法:离散化方法、参数化函数逼近法以及非参数化函数逼近法. 离散化方法的核心概念是任务分解,将连续的空间量化为若干个离散的区域,在同一区域的状态认为其值函数相等,于是一个连续或较大规模的马尔可夫决策问题(Markov Decision Problem,MDP)被离散化为规模较小的 MDP 问题. 离散化的方法主要有 BOX 方法,模糊划分以及聚类方法等. 采用离散化方法的强化学习已经被证明是收敛的,但其并不一定收敛到原问题的最优解上. 要使收敛的值函数达到一定的精度,离散化的区域不能太少. 因此,对于大规模的 MDP 问题,它

收稿日期:2008-05-16;修回日期:2009-02-16
基金项目:教育部新世纪优秀人才支持计划(No. NCET-08-0836);国家自然科学基金(No. 60804022);江苏省自然科学基金(No. BK2008126);高等学校博士学科点专项科研基金(No. 20070290537,200802901506);国家博士后科学基金(No. 20070411064)

仍然面临着维数灾难的困难,进而在学习时间和存储空间两方面也将降低强化学习控制系统的性能.

在已提出的参数化函数逼近方法中,按照函数逼近器的类型,可以分为基于线性值函数和非线性值函数逼近的连续空间强化学习.前者的具体做法是,首先假定强化学习的值函数是一些给定线性基函数的加权组合,然后采用最小二乘或递归最小二乘方法对权值进行估计,进而得到关于值函数的估计值.该方法的缺陷是,若假定不合适,则估计的偏差会很大.基于非线性值函数逼近方法的思想是,利用神经网络的并行计算、容错性和非线性函数逼近能力,对连续空间下强化学习的值函数进行回归估计.但是,基于神经网络的值函数逼近法存在网络结构不易确定、参数调整过程比较复杂、易于陷入局部极小等缺点.

由 Vapnik 依据结构风险最小化原则提出的支持向量机具有坚实的理论基础,良好的泛化性能,可以在一定程度上解决神经网络中的局部极小、网络结构难以确定以及泛化能力问题[1].近年来,一些学者采用支持向量机来解决强化学习的连续空间表示问题,并成为强化学习领域研究的热点方向之一.该方法的思想是,类似于神经网络值函数逼近,首先将强化学习问题构造为能用支持向量机求解的数学描述形式,然后采用经典的支持向量机及其各种改进形式,如最小二乘支持向量机、岭回归等方法对状态的值函数、状态-动作对的值函数或回报函数进行回归估计计算[2,3],应用实例包括资源限制排程问题[4]、3 维现场可编程门阵列的布局与布线[5]等.

为了将强化学习方法推广到连续空间,上述函数逼近法均是将强化学习构建为值函数或回报函数的回归估计问题,通过计算系统的值函数来组织对最优策略的搜索.不同于已有的值函数逼近法,本文把强化学习构建为一个简单的二分类问题,利用分类算法来得到强化学习中的策略,提出一种基于高斯过程分类器的连续空间强化学习方法.该方法的主要思想是,首先将连续动作空间离散化为确定数目的离散动作,然后利用高斯过程分类器对系统的连续状态-离散动作对进行正负分类,对判定为正类的离散动作进行加权求和,进而得到实际作用于系统的连续动作.小船靠岸问题的仿真结果表明所提方法能够有效解决强化学习的连续空间表示问题.

2 基于高斯过程分类器的强化学习

强化学习以马尔可夫决策过程为基础,通过试错机制来获得最优行为策略[6].一个有限的 MDP 可由一个 5 元组表示:$\{S, A, p(s_t, a_t, s_{t+1}), r(s_t, a_t), Q; s_t, s_{t+1} \in S, a_t \in A\}$,其中 S 为状态空间,A 为动作空间,$p(s_t, a_t, s_{t+1})$ 为系统处于状态 s_t 时,执行决策动作 a_t 后转移到下一状态 s_{t+1} 的转移概率,$r(s_t, a_t)$ 为在状态 s_t 下执行动作 a_t 获得的立即回报,Q 为值函数,按下式进行迭代计算:

$$Q_{t+1}(s, a) = (1 - \eta) Q_t(s, a) + \eta[r_t + \gamma \max_{a_{t+1}} Q_t(s_{t+1}, a_{t+1})] \quad (1)$$

其中,学习率 $0 \le \eta \le 1$ 控制学习的速度,学习率越大,收敛越快,但容易产生振荡;学习率越小,收敛越慢.折扣因子 $0 \le \gamma \le 1$ 表示学习系统的远视程度,如果取值比较小,则表示系统更关注最近的动作的影响;如果比较大,则对比较长的时间内的动作都很关注.一般来说,η 取得较小,γ 取得较大.状态转移中两相邻状态值函数的时间差分定义为 TD 误差:

$$\delta_t = r_t + \gamma Q(s_{t+1}) - Q(s_t) \quad (2)$$

强化学习的基本思想为:若某一动作获得环境正的奖赏,那么系统以后产生这个动作的趋势便会加强;否则,系统产生这个动作的趋势便减弱.因此,TD 误差实际上反映了所选动作的优劣程度.如果把分类器和强化学习结合起来,必须转换观点:在学习过程中,若 TD 误差呈减小趋势,则将当前所选动作定义为"正类";反之,将其定义为"负类",这样就可将整个状态-动作空间粗略地划分为两类.高斯过程是一种概率意义上的核机器,主要优点体现在:它是一种非参数概率模型,不仅能对未知输入做输出预测,而且同时给出该预测的精度参数;可以以先验概率的形式表示过程的先验知识,从而提高过程模型性能;与神经网络、支持向量机等方法相比,高斯过程模型参数明显减少,因而参数优化相对容易,且更易收敛[7].因此,可以利用高斯过程分类器来得到强化学习中的策略.

基于高斯过程分类器的强化学习如图 1 所示.图中,$s_t \in R^n$ 表示 t 时刻 n 维系统状态,待选动作集 $A = \{a_k \in R \mid k = 1, 2, \cdots, m\}$,$m$ 为待选动作的个数.将系统的状态与 m 个待选动作分别配对,构成状态-动作对 (s_t, a_k) 顺序输入给高斯过程分类器,高斯过程分类器的输出 π_k^* 为 (s_t, a_k) 属于正类的预测概率值.然后,对判定为正类的离散动作($\pi_k^* > 0.5$)按其概率值进行加权求和,即可得到实际作用于系统的连续动作 a_t,具体操作如式(3)和(4)所示.环境在动作 a_t 的作用下,得到立即回报 r_t,由式(2)计算系统的 TD 误差.根据 TD 误差判断 (s_t, a_k) 的类别标签 y_t,进而得到高斯过程分类器新的训练样本 $\{(s_t, a_t), y_t\}$.

$$\pi_k^* = \begin{cases} \pi_k^*, & \pi_k^* > 0.5 \\ 0, & \pi_k^* \le 0.5 \end{cases} \quad (3)$$

$$a_t = \sum_{k=1}^{m} a_k \pi_k^* \Big/ \sum_{k=1}^{m} \pi_k^* \tag{4}$$

由于强化学习强调在与环境的交互中学习,可以不需要环境模型而实现无导师的在线学习. 因此,高斯过程模型的训练样本需要通过强化学习系统不断地与环境交互而顺序生成,如果将新增样本与已有样本合并后处理,一方面会增加学习的难度,另一方面也因样本集过大而消耗过多的时间和存储空间. 为此,引入滚动时间窗机制实现高斯过程模型的在线学习,即在强化学习系统学习的同时获取样本数据并进行高斯过程模型的训练. 建立一个随时间窗滚动的建模数据区间[8],并保持该区间长度不变,随着新数据$((s_t, a_t), y_t)$的不断加入,旧数据则从建模区间滚动出去.

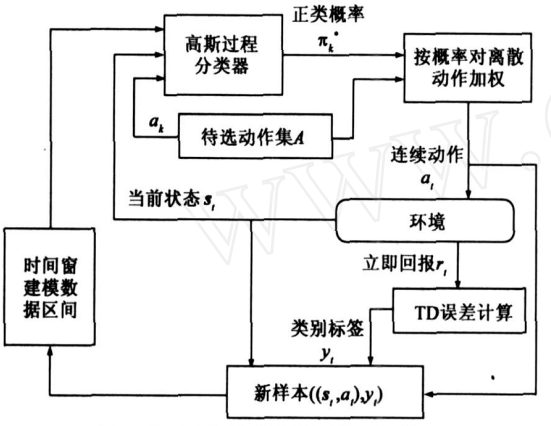

图1　基于高斯过程分类器的强化学习

3　在线高斯过程分类器学习

设图 1 中的时间窗宽度为 L,则当前 t 时刻高斯过程模型的学习训练样本集由过去 L 组数据构成 $D = \{(x_i, y_i) \mid i = t-L, t-L+1, \cdots, t-1\}$,其中样本输入数据 $x_{t-1} = (s_{t-1}, a_{t-1})^{\mathrm{T}} \in X = R^{n+1}$ 表示由 $(t-1)$ 时刻 n 维系统状态 s_{t-1} 和 1 维动作 a_{t-1} 构成的状态-动作对,样本输出数据 $y_i \in Y = \{-1, +1\}$ 为类别标签,其中 +1 和 -1 分别对应正负类.

假定高斯过程模型的训练集是按 $X \times Y$ 上的某个概率 $p(x, y)$ 选取的独立同分布的样本点,样本 x_i 属于类别 y_i 的概率为:

$$p(y_i \mid f(x_i)) = \Phi(y_i f(x_i)) \tag{5}$$

式中,$f(x_i) = f_i$ 为隐函数,$\Phi(\cdot)$ 为典型的概率单位函数或对数函数.

由于训练样本间相互独立,可得联合概率密度函数:

$$p(y \mid f) = \prod_{i=t-L}^{t-1} p(y_i \mid f(x_i)) = \prod_{i=t-L}^{t-1} \Phi(y_i f(x_i)) \tag{6}$$

已知输入样本 $x = [x_{t-L}, \cdots, x_{t-1}]$,隐函数 $f = [f_{t-L}, \cdots, f_{t-1}]$,为计算方便,高斯过程分类器假设 $f(x_i)$ 服从均值为零的高斯过程分布,亦即先验概率 $p(f \mid x)$ 是一个多维高斯密度函数(高斯过程):

$$p(f \mid x) = N(0, K) = \frac{1}{(2\pi)^{L/2} \mid K \mid^{1/2}} \exp\left\{-\frac{1}{2} f^{\mathrm{T}} K^{-1} f\right\} \tag{7}$$

式中,K 是 f 的协方差矩阵,它是 x 的对称正定函数,是一个核函数. 因此,由贝叶斯推断,可得后验概率:

$$p(f \mid x, y) = \frac{p(y \mid f)\, p(f \mid x)}{p(y \mid x)} \tag{8}$$

式中,$p(y \mid f)$ 为似然函数,$p(y \mid x) = \int p(y \mid f)\, p(f \mid x)\, \mathrm{d}f$ 为边缘概率分布.

对高斯过程分类器训练,也就是对后验概率进行估计. 此处采用 Laplace 方法求后验概率 $p(f \mid x, y)$ 的估计值 $q(f \mid x, y)$[9],将 $\log p(f \mid x, y)$ 在最大后验概率处按二阶泰勒级数展开,即可得到高斯估计:

$$q(f \mid x, y) = N(\hat{f}, A^{-1}) \propto \exp\left(-\frac{1}{2}(f - \hat{f})^{\mathrm{T}} A (f - \hat{f})\right) \tag{9}$$

式中,最大后验概率 $\hat{f} = \arg\max_f p(f \mid x, y)$,海森矩阵 $A = -\nabla\nabla \log p(f \mid x, y) \mid_{f=\hat{f}}$.

求解 $\hat{f}$ 时,由于 $p(y \mid x)$ 与 f 相互独立,因此,由式(8)知,若要最大化 $p(f \mid x, y)$,只需最大化下述函数 $\psi(f)$ 即可:

$$\psi(f) = \log p(y \mid f) + \log p(f \mid x) = \log p(y \mid f) - \frac{1}{2} f^{\mathrm{T}} K^{-1} f - \frac{1}{2}\log \mid K \mid - \frac{L}{2}\log 2\pi \tag{10}$$

式(10)分别对 f 求一阶和二阶偏导,可以得到:

$$\nabla\psi(f) = \nabla \log p(y \mid f) - K^{-1} f \tag{11}$$

$$\nabla\nabla\psi(f) = \nabla\nabla \log p(y \mid f) - K^{-1} = -W - K^{-1} \tag{12}$$

式中,$W = -\nabla\nabla \log p(y \mid f)$ 为对角矩阵. 若似然函数 $p(y \mid f)$ 是对数凸的,那么矩阵 W 的对角元素非负,且 $\nabla\nabla \log p(y \mid f)$ 负定,因此,$\psi(f)$ 是凹函数且必定有一个最大值. 当 $\psi(f)$ 取最大值时,由式(11)得:

$$\nabla\psi(f) = 0 \Rightarrow \hat{f} = K(\nabla \log p(y \mid \hat{f})) \tag{13}$$

由于 $\nabla \log p(y \mid \hat{f})$ 是 $\hat{f}$ 的非线性函数,式(13)无法直接求出,因此,可由 Newton-Raphson 法对 $\hat{f}$ 进行迭代计算[9]:

$$\begin{aligned} \hat{f} &= f^{nr} = f - (\nabla\nabla\psi)^{-1} \nabla\psi \\ &= f + (K^{-1} + W)^{-1}(\nabla \log p(y \mid f) - K^{-1} f) \\ &= (K^{-1} + W)^{-1}(Wf + \nabla \log p(y \mid f)) \end{aligned} \tag{14}$$

由 $A = -\nabla\nabla \log p(f \mid x, y) \mid_{f=\hat{f}}$ 和式(12)知,$A = W + K^{-1}$. 因此,后验概率为:

$$p(f \mid x, y) \approx q(f \mid x, y) = N(\hat{f}, (W + K^{-1})^{-1}) \tag{15}$$

检测系统当前状态 s_t,分别与离散动作集合 A 中

的 m 个动作配对，得到 $x^* = \left(x_1^*, \cdots, x_m^*\right)^{\mathrm{T}} = \left((s_t, a_1), \cdots, (s_t, a_m)\right)^{\mathrm{T}}$. 由高斯过程分类器式(15)，预测第 k 个状态-动作对 $x_k^* = (s_t, a_k)$ 属于正类的概率为：

$$\begin{aligned}\pi_k^* &= p(y_k^* = +1 \mid x, y, x_k^*) \\ &= \int \Phi(f_k^*)\, q(f_k^* \mid x, y, x_k^*)\, \mathrm{d}f_k^* \\ &= \int \Phi(f_k^*)\, N(E_q[f_k^* \mid x, y, x_k^*], V_q[f_k^* \mid x, y, x_k^*])\, \mathrm{d}f_k^*\end{aligned} \tag{16}$$

式中，$q(f_k^* \mid x, y, x_k^*)$ 为高斯过程，其均值为 $E_q[f_k^* \mid x, y, x_k^*] = (k_k^*)^{\mathrm{T}} K^{-1} \hat{f} = (k_k^*)^{\mathrm{T}} \nabla \log p(y \mid \hat{f})$，方差为 $V_q[f_k^* \mid x, y, x_k^*] = k(x_k^*, x_k^*) - (k_k^*)^{\mathrm{T}} (K + W^{-1})^{-1} k_k^*$，$k(\cdot,\cdot)$ 为协方差函数，$k_k^* = k(x_k^*) = [k(x_{t-L}, x_k^*), \cdots, k(x_{t-1}, x_k^*)]^{\mathrm{T}}$.

4 仿真研究

为了验证本文所提强化学习方法的有效性，针对如图 2 所示的小船靠岸问题(Boat Problem)进行仿真研究[10]. 系统的状态由两个连续状态变量 x_t 和 y_t 表示，分别表示船头的水平和垂直位移. 小船在控制量(舵角) ω_t 的作用下，由河的左岸任意位置出发，穿过一条宽度为 200 的河流，到达河右岸的指定目标码头位置. 仿真中，系统的动力学特性由如下方程描述：

$$\begin{cases} x_{t+1} = \min(200, \max(0, x_t + v_{t+1}\cos(\theta_{t+1}))) \\ y_{t+1} = \min(200, \max(0, y_t - v_{t+1}\sin(\theta_{t+1}) - E(x_{t+1}))) \end{cases} \tag{17}$$

式中，水流力 $f_c = 1.25$，水流力作用于小船上的推力为 $E(x) = f_c[x/50 - (x/100)^2]$，$\theta_t$ 和 v_t 分别为小船的角度和速度，可由下式计算得到：

$$\begin{cases} \theta_{t+1} = \theta_t + (I\Omega_{t+1}) \\ \Omega_{t+1} = \Omega_t + ((\omega_{t+1} - \Omega_t)(v_{t+1}/\mathrm{MaxSpeed}) \\ v_{t+1} = v_t + (\mathrm{DesiredSpeed} - v_t) I \\ \omega_{t+1} = \min(\max(p(U_{t+1} - \theta_{t+1}), -45^0), 45^0) \end{cases} \tag{18}$$

式中，系统惯性 $I = 0.1$，小船的最大允许速度 MaxSpeed = 2.5，期望速度 DesiredSpeed = 1.75，舵角 ω_t 在 $U_t \in [-100°, 90°]$ 区间内取值，$p = 0.9$ 为比例系数.

学习系统的目标是在没有任何模型先验知识的前提下，实现小船从左岸任意位置 $\{(x, y) \mid x = 10, y \in [0, 200]\}$ 停靠到右岸码头 $\{(x, y) \mid x = 200, y \in [97.5, 102.5]\}$ 的控制. 上述学习控制问题可以用一个确定性 MDP 来建模，回报函数为：

$$r(x, y) = \begin{cases} +1, & (x, y) \in Z_s \\ R(x, y), & (x, y) \in Z_v \\ -1, & (x, y) \in Z_f \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

式中，河右岸 $Z_b \triangleq \{(x, y) \mid x = 200, y \in [0, 200]\}$，成功区域 $Z_s \triangleq \{(x, y) \mid x = 200, y \in [97.5, 102.5]\}$，可行区域 $Z_v \triangleq \{(x, y) \mid x = 200, y \in [92.5, 97.5]$ or $y \in [102.5, 107.5]\}$，失败区域 $Z_f \triangleq Z_b \setminus Z_s \setminus Z_v$，$R(x, y)$ 的值在 Z_s 和 Z_f 区间内由 +1 至 -1 线性减小，有：

$$R(x, y) = \begin{cases} (2y - 190)/5, & y \in (92.5, 97.5) \\ (210 - 2y)/5, & y \in (102.5, 107.5) \end{cases} \tag{20}$$

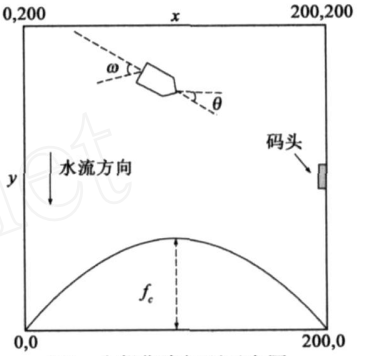

图2 小船靠岸问题示意图

仿真中，每次小船的初始状态为 $\{(x, y) \mid x = 10, y \in [0, 200]\}$，当小船到达河右岸或时间步数超过设定值，则当前航行结束，然后系统状态重新进行初始化，开始下一次航行. 强化学习控制器以及高斯过程分类器的相关参数设置为：$\gamma = 0.9$，$\eta = 0.2$，$L = 30$，采样周期为 1s. 根据定义的不同离散动作个数，分别设计 4 种类型的强化学习控制器：类型 Ⅰ：8 个离散动作 $\{-100°, -75°, -55°, -30°, 0°, 35°, 65°, 90°\}$；类型 Ⅱ：12 个离散动作 $\{-100°, -90°, -75°, -60°, -45°, -30°, -15°, 0°, 15°, 45°, 75°, 90°\}$；类型 Ⅲ：16 个离散动作 $\{-100°, -90°, -80°, -70°, -60°, -45°, -30°, -15°, 0°, 15°, 30°, 40°, 50°, 65°, 80°, 90°\}$ 以及类型 Ⅳ：20 个离散动作 $\{-100°, -90°, -80°, -70°, -60°, -50°, -40°, -30°, -20°, -10°, 0°, 10°, 20°, 30°, 40°, 50°, 60°, 70°, 80°, 90°\}$.

每种类型强化学习控制器的学习过程均由学习阶段和测试阶段构成，其中学习阶段定义为：小船连续 40 次到达非失败区域 Z_f，或尝试(航行)次数超过 5000 次. 学习阶段采用测试阶段前需要花费的尝试次数 n_T 来衡量系统的学习速度[10]，可以看出，$40 \le n_T \le 5000$. 表 1 给出了不同离散动作集合类型的强化学习控制器的学习速度比较. 由于类型 Ⅳ强化学习控制器的离散动作个数为 20 个，对连续动作空间划分得较为细致，因此，类型 Ⅳ控制器的控制精度较高，从而使其到达非失败区域的成功率有所提高，亦即 n_T 值较小. 由文[10]可知，Jouffe 提出一种适用于连续状态和连续动作空间的多步模糊 Q 学习方法 FQL(λ)，小船靠岸仿真结果表明，当离散动作个数取 12 且 $\lambda = 0$ 时，FQL(0) 的 $n_T = 504$. 由表 1 可知，类型 Ⅱ(离散动作个数为 12)强化学习控制器(本文所提强化学习方法均为单步，即 $\lambda = 0$)的 $n_T = 363$. 因此，与 FQL(0) 相比，本文所提基于高斯过程

分类器的强化学习方法具有较快的学习速度.

表 1　学习阶段系统学习速度比较

强化学习控制器	类型 Ⅰ	类型 Ⅱ	类型 Ⅲ	类型 Ⅳ
学习速度 n_T	422	363	293	237

学习阶段结束后,转测试阶段,将强化学习控制器用于对小船进行航行仿真测试. 图 3 给出了 3 组 12 条曲线,分别表示 4 种类型强化学习控制器从 3 个不同起始点($\{(x,y)\mid x=10,y=40\}$、$\{(x,y)\mid x=10,y=100\}$和$\{(x,y)\mid x=10,y=160\}$)出发的某次航行轨迹曲线. 由图可以看出,这 12 条曲线均比较光滑,这是由于本文所提强化学习控制器的控制舵角输出为连续型控制变量的缘故.

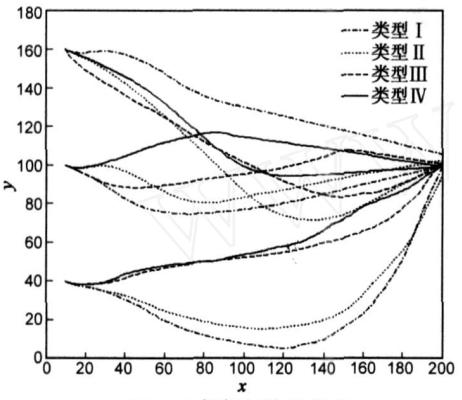

图3　小船航行轨迹曲线

为定量评价强化学习控制器在学习阶段学到的策略,表 2 给出了图 3 中 12 条曲线所对应的时间步和距离测度. 时间步指小船为完成某次航行所需的时间步数,距离测度的定义如式(21)所示,用于衡量控制器的精度[10]. 时间步数越小,且距离测度值越小,表明控制器的性能越好,即小船能以较高精度较快地到达目标码头.

$$d(x,y)=\begin{cases}|y-100|, & x=200\\ 100+(200-x), & \text{otherwise}\end{cases} \tag{21}$$

表 2　测试阶段系统学习性能比较

强化学习控制器	起始点为 $(x=10,y=40)$		起始点为 $(x=10,y=100)$		起始点为 $(x=10,y=160)$	
	距离测度	时间步	距离测度	时间步	距离测度	时间步
类型 Ⅰ	0.088	182	0.362	169	5.218	143
类型 Ⅱ	4.258	175	2.478	153	1.699	144
类型 Ⅲ	2.227	173	0.370	173	0.771	158
类型 Ⅳ	0.841	177	1.032	145	0.907	149

为了消除一次运行结果中诸多随机因素的影响,对每种类型的强化学习控制器独立运行 100 次尝试实验,进行统计分析. 以起始点$\{(x,y)\mid x=10,y=40\}$为例说明,表 3 给出了不同离散动作集合类型下,小船能够成功到达码头的次数、距离测度和时间步数的统计值. 在这 100 次尝试中,在类型 Ⅰ强化学习控制器的作用下,小船有 52 次成功到达码头,平均需要花费 176.7 时间步以精度 7.28 靠近码头;类型 Ⅳ控制器成功的次数为 92 次,大约需要 178.2 时间步到达码头,平均精度为 3.216. 可以看出,4 种类型控制器的平均时间步数相差不大,但是在到达码头的次数和平均距离测度指标方面,类型 Ⅳ控制器的性能要明显好于类型 Ⅰ控制器,这是由于前者对于动作区间的离散划分比较合理的原因所致.

为与文[10]进行对比,对类型 Ⅱ强化学习控制器独立运行 100 次尝试实验,计算平均距离测度. 在这 100 次尝试中,小船的出发点为左岸任一随机位置,即初始状态为$\{(x,y)\mid x=10,\text{random } y\in[0,200]\}$. Jouffe 提出的 FQL(0)的平均距离测度值为 5.96,而本文所提强化学习方法的平均距离测度值为 4.41. 因此,与 FQL(0)相比,本文所提基于高斯过程分类器的强化学习方法具有较高的控制精度.

表 3　测试阶段系统学习性能统计结果

强化学习控制器	到达码头的次数	距离测度			时间步		
		最大值	最小值	平均值	最大值	最小值	平均值
类型 Ⅰ	52	19.148	0.046	7.280	198	163	176.7
类型 Ⅱ	69	23.092	0.011	6.068	201	169	183.0
类型 Ⅲ	75	17.178	0.045	5.558	188	165	177.6
类型 Ⅳ	92	9.906	0.124	3.216	192	161	178.2

5　结论

提出了一种基于高斯过程分类器的连续空间强化学习方法,该方法的特点主要体现在如下 6 个方面:(1)有效解决强化学习的连续空间表示问题,不仅适用于连续状态空间,而且适用于连续动作空间,进一步拓展了强化学习的应用领域;(2)不同于已有的值函数逼近法,把强化学习构建为一个简单的二分类问题,利用高斯过程分类算法来得到强化学习中的策略;(3)为了描述分类的不确定性和避免简单分类导致的学习精度下降问题,利用高斯过程分类器预测输入类别的概率值,使样本在分类时,不仅具有定性的解释,还具有定量的评价;(4)由于强化学习强调在与环境的交互中学习,高斯过程分类器是一种在线学习方式,即在强化学习系统学习的同时获取样本数据并进行高斯过程模型的训练;(5)与神经网络、支持向量机等方法相比,高斯过程模型参数明显减少,因而参数优化相对容易;(6)离散动作个数的多少影响强化学习方法的性能. 需要进一步研究的问题包括离散动作集合的自适应确定以进一步提高算法的整体性能.

参考文献:

[1] V Vapnik. The Nature of Statistical Learning Theory[M]. New York:Springer Verlag,1995.

[2] R Goto, H Matsuo. State generalization method with support vector machines in reinforcement learning[J]. Systems and Computers in Japan,2006,37(9):77 - 86.

[3] Xuesong Wang, Xilan Tian, Yuhu Cheng. Value approximation with least squares support vector machine in reinforcement learning system[J]. Journal of Computational and Theoretical Nanoscience,2007,4(7/8):1290 - 1294.

[4] G Kai, H Barbara. A reinforcement learning algorithm to improve scheduling search heuristics with the SVM[A]. Proceedings of the IEEE International Conference on Neural Networks[C]. Piscataway:IEEE Inc,2004. 1811 - 1816.

[5] R Manimegalai, E S Soumya, V Muralidharan, et al. Placement and routing for 3D-FPGAs using reinforcement learning and support vector machines[A]. Proceedings of the International Conference on VLSI Design[C]. Piscataway:IEEE Inc,2005. 451 - 456.

[6] 秦斌,吴敏,王欣,等. 基于多智能体强化学习的焦炉集气管压力多级协调控制[J]. 电子学报,2006,34(10):1847 - 1851.
Qin Bin, Wu Min, Wang Xin, et al. Multi-level coordination control based on multi-agent reinforcement learning for the pressure of gas collectors of coke ovens[J]. Acta Electronic Sinica,2006,34(10):1847 - 1851. (in Chinese)

[7] C K I Williams, D Barber. Bayesian classification with Gaussian processes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1998,20(12):1342 - 1351.

[8] M Kyriakos, P Dimitris. Continuous nearest neighbor queries over sliding windows[J]. IEEE Transactions on Knowledge and Data Engineering,2007,19(6):789 - 803.

[9] C E Rasmussen, C K I Williams. Gaussian Processes for Machine Learning[M]. USA:MIT Press,2006.

[10] L Jouffe. Fuzzy inference system learning by reinforcement methods[J]. IEEE Transactions on System, Man and Cybernetics,1998,28(3):338 - 355.

作者简介:

王雪松 女,1974 年生于安徽泗县,2002 年获中国矿业大学控制理论与控制工程专业博士学位,2002 年至 2004 年于北京理工大学控制科学与工程博士后流动站从事博士后工作,现为中国矿业大学信息与电气工程学院副教授. 主要研究方向为机器学习、复杂系统优化与控制、生物信息学等.
Email:wangxuesongcumt@163.com

张依阳 男,1984 年生于河南偃师,2006 年获中国矿业大学学士学位,现为中国矿业大学控制理论与控制工程专业硕士研究生,研究方向为强化学习.
Email:zyiyang@126.com

程玉虎 男,1973 年生于安徽淮南,2005 年获中国科学院自动化研究所控制理论与控制工程专业博士学位,现为中国矿业大学信息与电气工程学院副教授. 主要研究方向为机器学习和智能系统等.
Email:chengyuhu@163.com