第 23 卷第 4 期
2006 年 8 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 23 No. 4
Aug. 2006

文章编号：1000−8152(2006)04−0547−05

随机平稳策略下半Markov决策过程的仿真优化算法

代桂平[1,2], 唐　昊[3], 奚宏生[2]

(1. 北京工业大学 电子信息与控制学院, 北京 100022; 2. 中国科学技术大学 自动化系, 安徽 合肥 230027;
3. 合肥工业大学 计算机系, 安徽 合肥 230009)

摘要: 基于性能势理论和等价Markov过程方法, 研究了一类半Markov决策过程（SMDP）在参数化随机平稳策略下的仿真优化算法, 并简要分析了算法的收敛性. 通过SMDP的等价Markov过程, 定义了一个一致化Markov链, 然后根据该一致化Markov链的单个样本轨道来估计SMDP的平均代价性能指标关于策略参数的梯度, 以寻找最优（或次优）策略. 文中给出的算法是利用神经元网络来逼近参数化随机平稳策略, 以节省计算机内存, 避免了“维数灾”问题, 适合于解决大状态空间系统的性能优化问题. 最后给出了一个仿真实例来说明算法的应用.

关键词: 随机平稳策略; 等价Markov过程; 一致化Markov链; 神经元动态规划; 仿真优化

中图分类号: TP202　　**文献标识码**: A

Simulation optimization algorithm for SMDPs with parameterized randomized stationary policies

DAI Gui-ping[1,2], TANG Hao [3], XI Hong-sheng[2]

(1. College of Electronic and Control Engineering, Beijing University of Technology, Beijing 100022, China;
2. Department of Automation, University of Science and Technology of China, Hefei Anhui 230027, China;
3. Department of Computer, Hefei University of Technology, Hefei Anhui 230009, China)

Abstract: Based on the theory of performance potentials and the method of equivalent Markov process, the performance optimization problem is discussed for a class of semi-Markov decision processes (SMDPs) with parameterized randomized stationary policies and a simulation optimization algorithm is proposed. Firstly, a uniform Markov chain is defined through the equivalent Markov process. Secondly, the gradient of the average cost performance with respect to the policy parameters is then estimated by simulating a single sample path of the uniformized Markov chain, so that an optimal (or suboptimal) randomized stationary policy can be found by iterating the parameters. The derived algorithm can meet the requirements of performance optimization of many different systems with large-scale state space, an artificial neural network is also used to approximate the parameterized randomized stationary policies and avoid the curse of dimensionality. Finally, convergence of the algorithm with probability one on an infinite sample path is considered, and a numerical example is provided to illustrate the application of the algorithm.

Key words: randomized stationary polices; equivalent Markov process; uniformized Markov chain; neuro-dynamic programming; simulation optimization

1　引言(Introduction)

半Markov决策过程是一类受到一系列控制决策驱动的半Markov系统, 其状态转移规律和控制决策所采用的行动相互作用决定了系统的演化, 过程在每个状态的逗留时间是一个服从一般分布的随机变量.当系统模型已知时, 其性能优化问题可利用动态规划方法建立Bellman方程来精确描述和求解, 并已有相应的算法来求解最优平稳策略[1~3].

但对很多模型不完全可知的实际系统或者存在“维数灾”（curse of dimensionality) 的大状态空间SMDPs, 常规动态规划方法不再适用, 需要考虑基于仿真的优化方法. 最近几年, 为了克服“维数灾”问题, 神经元动态规划(neuro-dynamic programming, 或称强化学习–reinforcement learning)方法被引入了离散时间Markov决策过程(DTMDP)的性能优化问题中[4,5], 其基本思想是利用神经元网络等逼近结构来逼近系统的性能值或策略, 以节省计算机内存. 前者即为NDP方法中的Critic模型, 后者即

收稿日期: 2004−10−10; 收修改稿日期: 2005−10−21.
基金项目: 国家自然科学基金资助项目 (60274012); 北京工业大学博士科研启动基金资助项目 (00194).

为NDP方法中的Actor模型. 文献[6]将NDP方法推广到了Markov报酬过程中. 文献[7]讨论了NDP方法在连续时间Markov决策过程(CTMDP)中的应用. 这里, 将文献[7]的方法推广到半Markov决策过程, 讨论其在Actor模型下的优化算法, 利用等价Markov过程的方法, 在参数化随机平稳策略范围内, 将其转化成一致化Markov链, 然后通过仿真这个Markov链的单一样本轨道来估计半Markov过程的性能势和平均准则关于策略参数的梯度, 并以此来迭代求解最优(或次最优)随机平稳策略.

2 问题描述(Problem description)

2.1 半Markov决策过程(Semi-Markov decision processes)

考虑一个半Markov过程$Y=\{Y_t;t\geqslant 0\}$, 它有一个有限的状态空间$\Phi=\{1,2,\cdots,K\}$和有限的行动空间D. 对$\forall a\in D$, Y的半Markov核是$Q(a,t)=[A(i,j,a,t)]$, 系统在状态i选择行动a的代价率$f(i,a)$是定义在Φ和D上的实值函数. 现假设任意策略参数向量$\theta=(\theta_1,\theta_2,\cdots,\theta_K)\in\Theta\subseteqq\mathbb{R}^K$都决定了唯一一个随机平稳策略$\mu(\theta)$, 这里$\Theta$为全体策略参数向量构成的集合, 称为策略参数空间, 记全体随机平稳策略集为$\Pi=\{\mu(\theta)|\theta\in\Theta\}$. 一个随机平稳策略$\mu(\theta)$就是把状态$i\in\Phi$映射到$D$上的概率分布的一个函数, 记系统在状态$i$选择行动$a$的概率为$\mu_a(i,\theta)$, 这种由可变参数表示的策略称为参数化随机平稳策略. 这里只考虑Y是不可约、非周期和正常返的, 则根据文献[8]中的定理10.5.22可知, 在策略$\mu(\theta)$下, Y存在唯一的稳态分布$p(i,\theta)>0$, $i\in\Phi$, 用$p(\theta)=(p(1,\theta),p(2,\theta),\cdots,p(K,\theta))$表示$Y$的稳态概率向量. 记$\bar{X}$为$Y$的嵌入Markov链, 则其也存在唯一的稳态分布$\pi(\theta)=(\pi(1,\theta),\cdots,\pi(2,\theta))$,且满足$\pi(i,\theta)>0,\pi(\theta)(p(\theta)-I)=0,\pi(\theta)e=1$, 这里$p(\theta)=[p(i,j,\theta)]_{K\times K}$为策略$\mu(\theta)$下$\bar{X}$的一步转移概率矩阵.

在策略$\mu(\theta)$作用下, 定义矩阵$Q(\theta)=[Q(i,j,\theta,t)]_{K\times K}$和向量$f(\theta)=(f(1,\theta),\cdots,f(K,\theta))^{\mathrm{T}}$如下:

$$\begin{cases} Q(i,j,\theta,t)=\sum\limits_{a\in D}\mu_a(i,\theta)Q(i,j,a,t),\\ f(i,\theta)=\sum\limits_{a\in D}\mu_a(i,\theta)f(i,a), \end{cases}\quad\forall i,j\in\Phi. \tag{1}$$

显然$Q(\theta)$有定义, 它是使用策略$\mu(\theta)$时系统的半Markov核. 称$Y(\theta)=\{Y_t,\Phi,D,Q(\theta),f(\theta)\}$为约束在$\Theta$上的半Markov决策过程. $Y(\theta)$关于参数向量θ的平均代价性能指标为

$$\eta(\theta)=\min_{T\to\infty}\mathrm{E}\{\frac{1}{T}\int_0^T f(Y_t,\theta)\mathrm{d}t\}.$$

因为Y是遍历的, 故有

$$\eta(\theta)=\sum_{i\in\Phi}p(i,\theta)f(i,\theta)=p(\theta)f(\theta). \tag{2}$$

本文的目的是要选择一控制决策方案, 使过程在平均代价性能准则下达到最优的运行效果, 即寻找一θ^*, 使得

$$\eta(\theta^*)\in\min_{\theta\to\Theta}\eta(\theta).$$

2.2 等价Markov过程(Equivalent Markov processes)

记

$$Q(i,\theta,t)=\sum_{j\in\Phi}Q(i,j,\theta,t).$$

令$S(i,\theta)$为在策略$\mu(\theta)$下过程在状态i的平均逗留时间, 即

$$S(i,\theta)=\int_0^\infty t\mathrm{d}Q(i,\theta,t). \tag{3}$$

令$\Lambda(\theta)=\mathrm{diag}(S^{-1}(1,\theta),S^{-1}(2,\theta),\cdots,S^{-1}(K,\theta))$, $\lambda=\sup\limits_{i\in\Phi,\theta\in\Theta}\{S^{-1}(i,\theta)\}$, 定义

$$A(\theta)=\Lambda(\theta)(p(\theta)-I). \tag{4}$$

则显然, $A(\theta)e=0$. 由文献[8]中定理10.5.22可知

$$p(i,\theta)=\frac{\pi(i,\theta)S(i,\theta)}{\sum\limits_{j\in\Phi}\pi(j,\theta)S(j,\theta)}, \tag{5}$$

或用矩阵表示为

$$p(\theta)\Lambda(\theta)=\frac{1}{\sum\limits_{j\in\Phi}\pi(j,\theta)S(j,\theta)}\pi(\theta)I. \tag{6}$$

故有

$$p(\theta)A(\theta)=p(\theta)\Lambda(\theta)(p(\theta)-I)=\frac{1}{\sum\limits_{j\in\Phi}\pi(j,\theta)S(j,\theta)}\pi(\theta)I(p(\theta)-I)=0,$$

因此, $A(\theta)=\Lambda(\theta)(p(\theta)-I)$可以作为一个等价Markov过程的无穷小矩阵.

考虑一个不可约的Markov过程$X=\{X_t;t\geqslant 0\}$, 具有状态空间Φ, 无穷小矩阵为$A(\theta)=\Lambda(\theta)(p(\theta)-I)$, 则这个Markov过程$X$具有唯一的稳态分布, 且由上面的讨论可知, 这个稳态分布就是原半Markov过程Y的稳态分布, 故对相同的性能函数, Markov过程X和半Markov过程X在平均代价性能准则下是等价的, 因此可将半Markov过程Y在平均代价准则下的优化问题转化为与之等价的Markov过程X的优化问题.

对任意给定的$\theta \in \Theta$, 定义CTMDP$X(\theta) = \{X_t, \Phi, D, A(\theta), f(\theta)\}$的Poisson方程[9]为

$$(-A(\theta) + ep(\theta))g(\theta) = f(\theta). \tag{7}$$

因为$-A(\theta) + ep(\theta)$可逆, 则上面方程存在唯一解向量

$$g(\theta) = (-A(\theta) + ep(\theta))^{-1} f(\theta). \tag{8}$$

有下列假设:

假设 1

1) 对$\forall i \in \Phi, a \in D$,函数$\mu_a(i,\theta)$关于$\theta$两次可微, 且一阶和二阶导数有界;

2) 对$\forall i \in \Phi, a \in D, \theta \in \Theta$, 都存在一个有界的函数向量$L_a(i,\theta)$, 使得

$$\nabla \mu_a(i,\theta) = \mu_a(i,\theta) L_a(i,\theta). \tag{9}$$

这里∇表示关于参数向量θ的梯度.

3) 记P为集合$\{p(\theta)|\theta \in \Theta\}$的闭包, 相应于$p$中的每一个元素的半Markov过程均是遍历的.

3 基于单个样本轨道的仿真优化算法(Simulation optimization algorithm based on single sample path)

3.1 一致化Markov链(Uniformized Markov chain)

定义

$$\tilde{p}(\theta) = I + \lambda^{-1} A(\theta). \tag{10}$$

易知$\tilde{p}(\theta)$为随机矩阵[10]. 定义一个DTMDP $\tilde{X}(\theta) = \{X_n, \Phi, D, \tilde{p}(\theta), f(\theta)\}$, 称$\tilde{X}(\theta)$为$X(\theta)$的一致化Markov链, λ为一致化参数. 根据文献[2], DTMDP$\tilde{X}(\theta)$的实现因子可由下式给出:

$$\tilde{d}_{ij}(\theta) = \mathrm{E}\{\sum_{n=0}^{N_{ij}(\theta)-1} [f(X_n,\theta) - \eta(\theta)] | X_0 = i\},$$
$$\forall i, j \in \Phi.$$

其中$N_{ij}(\theta) = \min\{n : n > 0, X_n = j, X_0 = i\}$且$\mathrm{E}\lfloor N_{ij}(\theta)\rfloor < \infty$.$\tilde{X}(\theta)$的性能势可定义为[2]

$$\begin{cases} \tilde{g}(j,\theta) = \tilde{d}_{i,j}(\theta) + c = \\ \mathrm{E}\{\sum\limits_{n=0}^{N_{ij}(\theta)-1} [f(X_n,\theta) - \eta(\theta)] | X_0 = j\} + c, \\ \forall i, j \in \Phi. \end{cases} \tag{11}$$

类似于文献[7], 有下列定理:

定理 1 对固定的参数向量θ, 前面定义的CTMDP$X(\theta)$和DTMDP$\tilde{X}(\theta)$具有相同的平稳分布$p(\theta)$,其平均代价准则函数在随机平稳策略范围内相等, 且有

$$\nabla \eta(\theta) = p(\theta)(\nabla f(\theta) + \nabla \tilde{p}(\theta) \tilde{g}(\theta)). \tag{12}$$

该定理说明, 对一个CTMDP$X(\theta)$的性能优化问题, 可把它转换成一致化Markov链$\tilde{X}(\theta)$来实现, 而不改变其稳态性能. 在理论上, 可通过求解式(8)和(12)获得精确的梯度方向, 然后用最陡下降法进行参数寻优, 以寻找最优(或次最优)策略参数. 这种基于理论计算的方法, 在大规模SMDP中, 由于系统的状态空间很大, 矩阵求逆将占用较多计算机内存, 甚至造成内存溢出而不可行, 因此考虑基于样本轨道的方法, 在$\tilde{X}(\theta)$上, 用计算机仿真一条样本轨道, 并根据样本信息来估计势和梯度, 然后进行参数更新. 由于神经元网络具有较强的函数逼近能力, 一般可用一个多层感知器或径向基函数网络的输出来构造随机策略, 状态作为网络的输入, 参数向量θ即为相应的网络权系数, 由于逼近结构的参数数目比系统的状态数要少, 从而起到节省计算机内存的目的. 这就是神经元动态规划优化方法中的Actor模型. 下面进行详细分析.

3.2 仿真优化算法(Simulation optimization algorithm)

对任意给定的参数θ(即给定随机策略$\mu(\theta)$), 仿真由随机矩阵$\tilde{p}(\theta)$决定的DTMDP$\tilde{X}(\theta)$得到一条样本轨道$(X_{u_0}, X_{u_0+1}, \cdots, X_{u_1}, X_{u_1+1})$并且定义

$$u_0 = 0, \ X_{u_0} = i^*,$$
$$u_{k+1} = \min\{n : n > u_k, X_n = i^*\}, \ k \geqslant 0.$$

显然, u_k为第k次回到初始状态i^*的时刻, 且$u_1 - u_0, u_2 - u_1, \cdots$为独立同分布随机变量.

在这条样本轨道上, 对不同的$k \geqslant 1$, 把

$$\hat{g}(X_n,\theta) = \begin{cases} 0, & \text{如果} n = u_{k-1}, \\ \sum\limits_{t=n}^{u_k-1} (f(x_t,\theta) - \tilde{\eta}), & \text{其他}. \end{cases} \tag{13}$$

作为势$\tilde{g}(X_n,\theta), u_{k-1} \leqslant n < u_k$的估计, 其中$\tilde{\eta}$为平均代价$\eta(\theta)$的一个在$n$时刻已经得到的估计, 它不依赖于当前时刻的状态$X_n$, 只与$n$时刻以前的历史有关. 类似于文献[7], 式(12)可写为

$$\nabla \eta(\theta) =$$
$$\sum_{j \in \Phi} \mathrm{E}[\mathscr{X}_i(X_n) \nabla f(i,\theta)] + \sum_{i,j \in \Phi} \sum_{a \in A} \mathrm{E}[\mathscr{X}_i(X_n) \cdot$$
$$\mathscr{X}_j(X_{n+1}) \mathscr{X}_a(a_n) L_a(i,\theta) \hat{g}(j,\theta)]. \tag{14}$$

这里: a_n表示系统在状态 X_n 时根据概率 $\mu_a(X_n,\theta)$ 随机选取的行动, $\mathscr{X}_i(\cdot)$ 和$\mathscr{X}_a(\cdot)$分别是状态i和行动a的示性函数. 记

$$F^{(k)}(\theta,\tilde{\eta})=\sum_{n=u_{k-1}}^{u_k-1}\sum_{a\in A}\mu_a(X_n,\theta)(f(X_n,a)-\tilde{\eta})z_n(\theta).$$

其中

$$z_n(\theta)=\begin{cases}L_{a_n}(X_n,\theta), & \text{如果}X_n=i^*,\\ \alpha z_{n-1}(\theta)+L_{a_n}(X_n,\theta), & \text{其他}.\end{cases}$$

α为遗忘因子. 根据式(14), 可把$F^{(k)}(\theta,\tilde{\eta})$当作梯度方向的估计.

于是得到下列更新算法:

$$\begin{cases}\theta_{k+1}=\theta_k-\gamma_k\sum\limits_{n=kT}^{(k+1)T-1}(f(X_n,a_n)-\tilde{\eta}_k)z_n(\theta_k),\\ \tilde{\eta}_{k+1}=\tilde{\eta}_k+\beta\gamma_k\sum\limits_{n=kT}^{(k+1)T-1}(f(X_n,a_n)-\tilde{\eta}_k).\end{cases}\tag{15}$$

这里: $\{\gamma_k\}$是步长序列; β为正标量系数, 以调整θ和$\tilde{\eta}$ 的相对更新速率. 其中参数$\tilde{\eta}$ 的更新相当于一种自适应调整. 具体的算法过程如下:

步骤 1 令$X_0=X_{u_0}=i^*$, 选择合适的整数T和正标量系数β以及步长序列$\{\gamma_k\}$, 令$k=0$, 并初始化参数向量$\theta_0,\tilde{\eta}_0$;

步骤 2 在线观测随机平稳策略$\mu(\theta_k)$作用下的系统, 得到一条样本轨道$(X_{kT},X_{kT+1},\cdots,X_{(k+1)T})$;

步骤 3 按式(15)进行参数更新;

步骤 4 判断算法是否满足终止条件, 若不满足, 令$k:=k+1$, 回到步骤2; 若满足, 则停止.

4 收敛性(Convergence of the algorithm)

首先给出下列假设, 以保证算法的收敛性:

假设 2 步长序列$\{\gamma_k\}$非负、单调不增, 且存在正整数p和正标量B, 使得

$$\sum_{k=0}^{\infty}\gamma_k=\infty,\quad\sum_{k=0}^{\infty}\gamma_k^2<\infty,$$
$$\sum_{k=n}^{n+t}(\gamma_n-\gamma_k)\leqslant Bt^p\gamma_n^2,\ \ \forall n,\ \ t>0.$$

假设 3 对状态i和确定性正整数T, 存在正整数N_0 , 使得对Φ 中的每个状态i和P中的每N_0个矩阵的集合$p_l(\theta)(l=1,2,\cdots,N_0)$, 由式(10)定义的相应矩阵$\tilde{p}_l(\theta)(l=1,2,1\cdots,N_0)$ 满足

$$\sum_{n=1}^{N_0}[\prod_{l=1}^{n}\tilde{P}_l^{\mathrm{T}}]_{ii^*}>0.$$

这样, 有下面的收敛性定理, 其证明可参考文献[6], 此处不再详细叙述.

定 理 2 在假设1∼3下, 对任意正整数T, 记$\{\theta_k\}$为算法产生的参数向量序列, 则$\eta(\theta_k)$ 以概率1收敛, 且$\nabla\eta(\theta_k)$以概率1收敛于零.

5 实例(Example)

考虑一个31个状态的受控半Markov过程, $\Phi=\{1,2,\cdots,31\}, A=\{1,2\}$.在行动$a\in A$ 下, 嵌入Markov链的一步转移概率矩阵为

$$p_{ij}(a)=$$
$$\begin{cases}\exp(-a/j)/(31(1+\exp(-a))), & j\neq i+1,\\ 1-\sum\limits_{j\neq i+1}p_{ij}(a), & j=i+1.\end{cases}$$

代价率为

$$f(i,a)=\ln[(1+i)a]+\sqrt{i}/(2a).$$

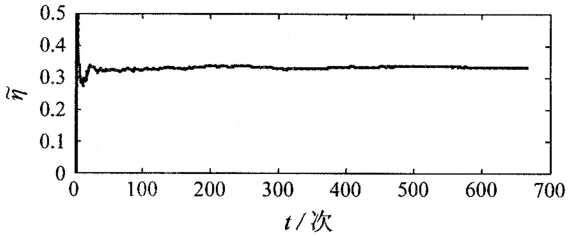
已知过程Y处于状态i下一次转移到状态j时, 在状态i 的逗留时间服从区间$[0,ja]$上的均匀分布, 即分布函数为

$$F(i,j,a,t)=\begin{cases}\dfrac{t}{ja}, & 0\leqslant t\leqslant ja,\\ 1, & t>ja.\end{cases}$$

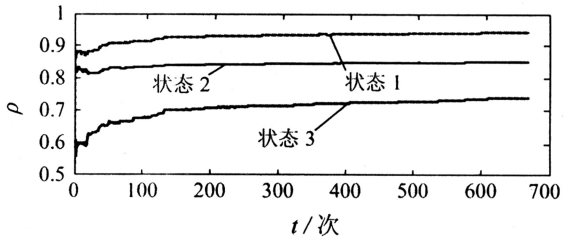
半 Markov 核为$Q(i,j,a,t)=P(i,j,a)F(i,j,a,t)$.

在式(1)中, 定义$f(i,a)=\sum\limits_{j\in\Phi}p_{ij}(a)f(i,a,j)$, 则可直接由相应的Bellman方程解得最优策略为: 在状态1, 2, $\cdots$, 13采取行动1的概率分别为0.95, 0.84, 0.72, 0.64, 0.58, 0.49, 0.37, 0.28, 0.22, 0.17, 0.13, 0.08, 0.03; 其余为0. 对应的最优平均代价为$\eta=0.32$.

在仿真计算中, 采用一个5×3×1的前向神经网络来逼近随机策略, 网络的输入为状态的编码, 隐节点没有固定偏置输入, 输出层的节点采用恒等函数, 且输出表示采用行动1的概率, 这样, 网络的可调参数为15个, 小于系统的状态数. 选择零初始参数向量, 即初始策略以相等概率选用行动1和2, 其他算法参数为$T=3,\alpha=0.99,\beta=0.2$. 对应2000次状态转移的仿真波形见图1所示, 其中横坐标表示迭代步数t, 上图曲线表示平均代价的估计值$\tilde{\eta}$, 下图3条曲线分别表示在状态1, 2, 3采用行动1的概率ρ.



(a) 平均代价的估计值



(b) 在状态1, 2, 3情况下采用行动1的概率

图 1 仿真结果曲线

Fig. 1 Curve of simulation results

从图中可见, 算法在经过最初的调整后, 在3个状态都以较大概率采用行动1, 相应的平均代价也与理论计算的结果相近.

6 总结(Conclusion)

本文讨论了一类有限状态半Markov决策过程在随机平稳策略范围内的优化问题, 利用神经元网络来逼近随机策略, 无需记录大量信息, 节省了计算机内存, 避免了维数灾问题. 文中给出的算法也可推广到具有可数状态空间和一般行动空间的半Markov决策过程.

参考文献(References):

[1] CAO X R. Semi-Markov decision problems and performance sensitivity analysis[J]. *IEEE Trans on Automation Control*, 2003, 48(5): 758 – 769.

[2] CAO X R, WAN Y W. Algorithms for sensitivity analysis of Markov system through potentials and perturbation realization[J]. *IEEE Trans on Control Systems Technology*, 1998, 6(4): 482 – 494.

[3] BEUTLER F J, ROSS K W. Uniformization for Semi-Markov decision processes under stationary policies[J]. *J of Applied Probability*, 1987, 24(3): 644 – 656.

[4] BERTSEKAS D P, TSITSIKLIS J N. *Neuro-dynamic Programming*[M]. Belmont, MA: Athena Scientific, 1996.

[5] SUTTON R S, BARTO A G. *Reinforcement Learning: An Introduction*[M]. Cambridge, MA: MIT Press, 1998.

[6] MARBACH P, TSITSIKLIS J N. Simulation-based optimization of Markov reward process[J]. *IEEE Trans on Automatic Control*, 2001, 46(2): 191 – 209.

[7] TANG H, XI H Sh. A Simulation optimization algorithm for CT-MDPs based on randomized stationary policies[J]. *Acta Automatica Sinica*, 2004, 30(2): 229 – 234.

[8] CINLAR E. *Introduction to Stochastic Processes*[M]. Engle-wood Cliffs, NJ: Prentice-Hall, 1975.

[9] 奚宏生, 唐昊, 殷保群. 连续时间MCP在紧致行动集上的最优策略[J]. 自动化学报, 2003, 29(2): 206 – 211.
(XI Hongsheng, TANG Hao, Yin Baoqun. Optimal policies for a continuous time MCP with compact action set[J]. *Acta Automatica Sinica*, 2003, 29(3): 206 – 211.)

[10] LIU Z K, TU F S. Single sample path-based sensitivity analysis of Markov processes [J]. *IEEE Trans on Automation Control*, 1999, 44(4): 872 – 875.

作者简介:

代桂平　(1977—), 女, 博士, 研究方向为离散事件动态系统、混杂系统及其应用, E-mail: daigping@bjut.edu.cn;

唐　昊　(1972—), 男, 博士, 副教授, 研究方向为神经元动态规划方法以及系统优化理论, E-mail: tangh@ustc.edu;

奚宏生　(1950—), 男, 教授, 博士生导师, 研究方向为鲁棒控制、离散事件动态系统及其应用等, E-mail: xihs@ustc.edu.cn.