第 26 卷第 8 期
2009 年 8 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 26 No. 8
Aug. 2009

文章编号: 1000−8152(2009)08−0884−05

搬运系统作业分配问题的小脑模型关节控制器Q学习算法

唐　昊[1,2], 丁丽洁[1], 程文娟[1], 周　雷[1]

(1. 合肥工业大学 计算机与信息学院, 安徽 合肥 230009; 2. 安全关键工业测控技术教育部工程研究中心, 安徽 合肥 230009)

摘要: 研究两机器人高速搬运系统的作业分配问题. 在系统的Markov决策过程(MDP)模型中, 状态变量具有连续取值和离散取值的混杂性, 状态空间复杂且存在“维数灾”问题, 传统的数值优化难以进行. 根据小脑模型关节控制器(CMAC)具有收敛速度快和适应性强的特点, 运用该结构作为Q值函数的逼近器, 并与Q学习和性能势概念相结合, 给出了一种适用于平均或折扣性能准则的CMAC-Q学习优化算法. 仿真结果说明, 这种神经元动态规划方法比常规的Q学习算法具有节省存储空间, 优化精度高和优化速度快的优势.

关键词: 作业分配; Markov决策过程; Q学习; CMAC

中图分类号: TP202　　**文献标识码**: A

The cerebellar-model-articulation-controller Q-learning for the task assignment of a handling system

TANG Hao[1,2], DING Li-jie[1], CHENG Wen-juan[1], ZHOU Lei[1]

(1. School of Computer and Information, Hefei University of Technology, Hefei Anhui 230009, China;
2. Engineering Research Center of Safety Critical Industrial Measurement and Control Technology, Ministry of Education, Hefei Anhui 230009, China)

Abstract: The task assignment of a high-speed handling system with two robots is studied in this paper. In the underlying Markov decision process(MDP) model, the state variable is composed of both continuous and discrete values, and the state space is complex and suffers from the curse of dimensionality. Therefore, the traditional numerical optimization is prevented from successful application to this system. Since the cerebellar-model-articulation-controller(CMAC) has the advantages of fast convergence and desired adaptability, it is employed to approximate the Q-values in a CMAC-Q learning optimization algorithm for combining the concept of performance potential and Q-learning, and for unifying the average criteria with the discount criteria. Compared with the Q-learning, the proposed neuro-dynamic programming approach requires less memory, but provides higher learning speed and better optimization performance as shown in the simulations.

Key words: task assignment; MDP; Q-learning; CMAC

1　引言(Introduction)

在现代化物流和生产加工等环境中, 存在一种具有多个智能机器人(或手臂)共同工作的搬运系统. 机器人连续和高速的运转会导致其传动装置过热疲劳、发生停机、故障甚至烧毁, 从而影响系统工作的效率、稳定性和安全性. 因此, 在该类作业系统中, 每个机器人需根据当前系统状态, 采取联合行动, 以协同完成搬运任务. 其控制目标就是通过作业合理分担, 避免单个机器人过度疲劳, 使整个系统长时间运行的工件捡取率最高.

理论上, 此类问题可通过Markov决策过程(MDP)来建模和描述, 并采用策略/数值迭代等理论计算方法或其它直接搜索方法来优化求解[1]. 文献[2]将遗传算法与模拟退火相结合, 研究了单机器人搬运系统的行动控制问题. 对于多机器人搬运系统, 状态空间规模较大, 且状态变量可能具有混合性. 因此, 传统方法的应用受到限制. 文献[3,4]采用了不同的强化学习方法对多机器人搬运系统的作业分配问题进行了求解, 但是存在存储空间巨大、学习速度慢等问题. 神经元动态规划(NDP)结合了仿真、逼近、学习和动态规划等技术, 一定程度上可以解决上述问题. 文献[5,6]基于性能势理论, 分别研究了Markov或半Markov系统基于Monte-Carlo方法或即时差分(TD)学习的神经元策略迭代算法, 但策略改进步骤中仍需模型的转移概率知识. 本文运用小脑模型关节控制器(CMAC)作为性能势Q值函数的逼近器, 给出了一种适用于平均和折扣两种性能准则、并且不依赖于系统模型参数的CMAC-Q学

收稿日期: 2008−05−25; 收修改稿日期: 2008−11−16.

基金项目: 国家自然科学基金资助项目(60404009); 安徽省自然科学基金资助项目(090412046,070416242); 安徽高校省级自然科学研究重点资助项目(KJ2007A063,KJ2008A058); 教育部留学回国人员科研启动基金资助项目.

习优化算法.

2 搬运系统作业分配问题(Task assignment of handling system)

本文主要研究两机器人双传送带搬运系统, 如图1所示[2]. 机器人R_1和R_2对立分布在两个并列的传送带两旁, 两个传送带上的工件分别按照相同的间隔时间分布随机到达, 其上游装有一个视觉传感器, 用来帮助机器人获取传送带上一定长度范围内的工件信息, 即工件所处的位置. 这里, 一般假设工件的大小、方向及传送带速度固定.

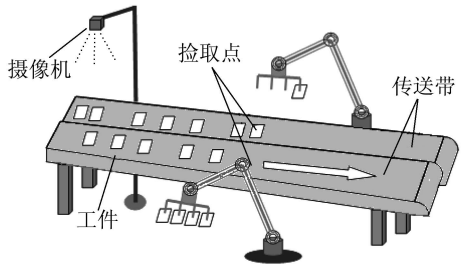

图 1　搬运系统模型

Fig. 1　The physical model of a handing system

每个机器人可执行3种行动: 等待(wait)、捡取(pick)、放置(place). 如图1所示, 两个捡取点并列分布在两个传送带上, 机器人可捡取任一捡取点上的工件(两个机器人不能同时捡取工件, 捡取工件所用时间忽略不计). 即在同一时刻, 机器人一次最多可捡取两个工件. 每个机器人旁都有一个工件箱, 用来存放捡取的工件, 放置工件需一定的时间. 机器人只要持有工件便可执行放置动作, 两个机器人可同时放置工件. 当机器人不在捡取点时, 则一定处在放置工件的过程中, 这段时间机器人不需要进行任何决策. 因此只需讨论机器人在捡取点时的行动模式, 有3种情况: 1) 机器人所持工件个数等于它所能持有的最大数目时, 只能进行放置; 2) 所持工件个数为0, 若捡取点上无工件, 则只能等待, 若有工件, 则可进行捡取或等待; 3) 所持工件个数大于0, 并小于最大持有数目时, 若捡取点上无工件, 可进行等待或放置, 有工件, 则可进行等待、放置或捡取.

机器人高速运转会导致传动装置发热, 其热量累积公式为[3]

$$O(t+\Delta t) = K_1 O(t) + (1-K_1) i^2(t). \quad (1)$$

其中: Δt为离散化采样周期, 表示传送带走过一个工件长度所用的时间, 记$\Delta t = 1$ 步; $O(t)$为t时刻的热量, 且$O(0) = 0$; $K_1 = 0.99917$为机器人传动装置的热量累积参数; $i(t)$为t时刻的电流, 与文献[3]相似, 令机器人在捡取或放置工件时$i(t)$为一常数, 等于5 A, 等待时间为0.

若机器人从传送带上捡取工件, 会得到相应报酬; 若从传送带上流失了工件, 则要付出相应的代价; 若任一个机器人的发热量超出了它所能承受的最大热量, 则整个搬运系统需停机进行维修, 并付出相应的代价. 机器人作业分配的目标就是在决策时刻, 机器人根据当前系统状态, 采取联合行动, 使系统报酬准则函数在无穷时段水平下期望值最大. 系统中, 令M表示一个机器人所能持有的最大工件个数; L表示视觉传感器的视觉范围大小, 单位为一个工件长度; G表示机器人放置工件所需时间, 单位为一步, 即一个step; J表示机器人所能承受的最大热量, 单位为焦耳.

3 数学模型(Mathematical model)

当工件分别按参数为λ的Poisson流(以step为时间单位)到达每个传输带时, 两机器人搬运系统可用一个离散时间MDP模型来近似描述. $X(t) = X(X_P(t), X_R(t))$为t时刻系统的状态, 包括工件信息和机器人信息. t时刻传感器视觉范围内的工件信息记为$X_P(t) = \{x_{p_1}(t), x_{p_2}(t), \cdots, x_{p_L}(t)\}$, $x_{p_1}(t)$为捡取点位置的工件个数, $x_{p_l}(t)$为捡取点向前$l-1$个单位位置上的工件个数, 有$0 \leqslant x_{p_l}(t) \leqslant 2, l = 1, 2, \cdots, L$. $X_R(t) = \{X_{R_1}(t), X_{R_2}(t)\}$为$t$时刻两个机器人$R_1$, R_2的状态. $X_{R_i}(t) = \{x_{c_i}(t), x_{q_i}(t)\}$为$t$时刻第$i$个机器人所持工件个数(或所处位置)以及热量值. 当机器人在放置过程, 无需决策, 它所持有的工件个数也无需考虑, 故可用$x_{c_i}(t) \in \{M+1, M+2, \cdots, M+G-1\}$来表示机器人离开捡取点的位置, $x_{c_i}(t) \in \{0, 1, \cdots, M\}$表示机器人在捡取点位置所持工件的个数, 则$0 \leqslant x_{c_i}(t) \leqslant M+G-1$. 另外, $x_{q_i}(t)$表示热量值. 本文分别用3个整数0, 1, 2来表示等待、放置、捡取3种行动, $a(t) = \{a_1(t), a_2(t)\}$为$t$时刻两个机器人采取的联合行动(共有8种). 令$u(t)$为采取行动$a(t)$后, 机器人捡取的工件数; $l(t)$为采取行动$a(t)$后至下一决策时刻前, 传送带上流失的工件数; r_1和r_2分别表示捡取一个工件所得到的报酬与流失一个工件所付出的代价.

记系统状态空间为Φ, 联合行动集为D, 则平稳策略v表示映射$v: \Phi \to D$, 即$a(t) = v(X(t))$. 状态转移规律满足方程$X(t+1) = \sigma(X(t), a(t))$, 它由系统的转移函数$\sigma(\cdot)$确定, 可用转移概率$P_{X(t)X(t+1)}(a(t))$表示. 记$P^v = [P_{X(t)X(t+1)}(a(t))]$, $f(X(t), a(t)) = r_1 \times u(t) - r_2 \times l(t)$表示系统的报酬函数, 则系统的运行和动态特性可近似用一个MDP$(X(t), \Phi, D, P^v, f^v)$来描述.

定义任意状态 i 的无穷时段折扣报酬准则为

$\eta_\beta^v(i) = (1-\beta)\mathrm{E}[\sum_0^\infty \beta^t f(X(t),a(t))|X(0)=i]$. 这里$0<\beta<1$是折扣因子, $\beta=1$时, 其极限就是平均准则η^v. 系统的优化目标就是选择一个最优策略v^*使得选择的性能准则值最大.

4 CMAC-Q学习算法(CMAC-Q learning algorithm)

4.1 性能势Q学习(Performance potential Q-learning)

Q学习本质上是基于数值迭代的思想[7], 通过仿真学习状态–行动对(X,a)的函数值进行问题的求解. 结合性能势概念[8], 平均和折扣性能准则下统一的即时差分公式为

$$d_t = f(X(t),a(t)) - \beta\eta_t + \beta\max_{a\in D} Q_\beta(X(t+1),a) - Q_\beta(X(t),a(t)). \tag{2}$$

其中

$$\eta_t = (1-\delta_t)\eta_{t-1} + \delta_t f(X(t),a(t)) \tag{3}$$

为平均代价的学习公式, δ_t为学习步长. Q值的迭代学习公式为

$$Q_\beta(X(t),a(t)) := Q_\beta(X(t),a(t)) + \gamma_t d_t. \tag{4}$$

这里γ_t为另一学习步长, 一般比δ_t衰减慢.

Q学习一般用于有限状态行动集的MDP问题, 学习中要为每个状态–行动对建立一个与其一一对应的性能值表格. 最后, 比较给定状态下每个行动的性能值, 找出其最优行动, 构成一个最优或次优策略.

4.2 基于CMAC网络的Q学习(Q-learning based on CMAC networks)

CMAC网络是Albus于1975年提出的一种局部逼近神经网络[9], 适合在线学习. 网络的输入可以是多维连续的矢量, 每个输入矢量在网络内被量化, 并激活记忆空间中C个单元(C称为网络的泛化参数), 被激活的单元输出为1, 未被激活的单元输出为0, 网络输出为所有单元输出的加权和. 网络的学习过程就是不断地根据输入数据对记忆空间中的权值进行调整, 使权值的分布能够反映所要求的非线性关系. 其中, 权值调整采用的是有监督的δ学习算法. CMAC网络权值的调整公式为

$$\Delta W_x = \zeta(y_\mathrm{d} - y)/C. \tag{5}$$

W_x为输入样本X对应的C个记忆单元的权值向量, 记$W_x = \{w_{x,1}, w_{x,2}, \cdots, w_{x,C}\}$, ζ为网络自身的学习步长, y_d为期望输出, y为网络的实际输出. 对于输入空间较大的情况, 需记忆单元的数量也较大. 为节省存储空间, Albus提出了hash编码, 将权值存储于数量大大少于记忆单元的hash单元中, 记忆单元中只存储hash单元的散列地址编码[10].

CMAC网络应用到Q学习, 能够解决状态混杂、大规模状态空间等情形下的强化学习问题. 本文为每个联合行动设置一个网络, 用来存储对应的Q值表. 在t时刻, 行动$a(t)$的网络中, 记输入$X(t)$对应的权值向量为$W_{X(t),a(t)}(t)$, 其输出为$Q_\beta(X(t),a(t),W_{X(t),a(t)}(t))$, 后者近似表示$t$时刻状态–行动对$(X(t),a(t))$的性能势学习值. 本文考虑的系统中, 状态$X(t)$可用一个$L+2\times 2$维向量表示, 前$L$维表示从视觉传感器得到的$L$个工件长度传送带上的工件位置及个数, 后$2\times 2$维表示系统中机器人各自所持有的工件个数和热量. 于是, 每个联合行动要设置一个$(L+2\times 2)$维输入单输出的CMAC网络. 本系统中, 由于行动集规模不大, 仅为8个行动, 因此这种设置是合理的.

根据公式(2), 采用CMAC逼近结构时, Q学习的统一即时差分公式为

$$d'_t = f(X(t),a(t)) - \beta\eta_t + \beta\max_{a\in D} Q_\beta(X(t+1),a, W_{X(t+1),a}(t)) - Q_\beta(X(t),a(t),W_{X(t),a(t)}(t)). \tag{6}$$

其中: η_t和δ_t的物理意义同公式(3)中的一致, d'_t可视为公式(5)中的网络期望输出与实际输出的差值$(y_\mathrm{d}-y)$. 因此, CMAC网络的权值调整公式为

$$\Delta W_{X(t),a(t)} = \zeta_t d'_t/C. \tag{7}$$

CMAC-Q学习算法描述如下:

Step 1 初始化每个行动所对应的CMAC网络的权值向量, 设置学习次数I, 令$t=0$, $\eta_t=0$;

Step 2 初始化每次学习的步数K;

Step 3 观察t时刻的状态$X(t)$, 由CMAC网络计算出$X(t)$对应的所有可选行动的性能值;

Step 4 从状态$X(t)$对应的可选行动集中按ε-greedy策略选择行动$a(t)$;

Step 5 执行行动$a(t)$, 仿真或观测实际系统得到下一状态$X(t+1)$;

Step 6 选择学习步长δ_t和ζ_t, 按式(3)(6)(7)调整行动$a(t)$对应的CMAC网络的权值;

Step 7 $t:=t+1$, $K:=K-1$, 若$K>0$, 转step 3; 否则, 令$I:=I-1$, 若$I>0$, 转step 2, 否则学习结束.

5 仿真结果(Simulation results)

令$M=4$, $L=6$, $G=3$, $r_1=1$, $r_2=1$. 根据文献[3], 令$J=17.86$, 系统维修一次时间为180步.

由参数设置, 系统的状态需用$L+2\times 2=10$维的向量表示. 在CMAC-Q学习算法中, 可采用8个结构

相同的10输入单输出的CMAC网络来分别保存8个行动所对应的Q值. CMAC网络的泛化参数$C = 2$, 量化级数$q = (3, 3, 3, 3, 3, 3, 7, 10, 7, 10)$. 于是, 系统状态数为$3^6 \times 7^2 \times 10^2 = 3572100$, 直接使用Q学习算法, 需要28576800个存储单元. 使用CMAC网络存储性能值时, 由于可通过hash函数对存储空间进行压缩, CMAC-Q学习算法中每个网络一般只需存储大约8000个权值, 即总共只需约64000多个存储单元就可以逼近得到所有状态行动对的性能值.

Q学习中, 对热量值进行离散化, 取热量划分等级数$H = 10$. 算法参数 $I = 200$, $K = 50$万, $\delta_t = 1/(t+1)$, $\varepsilon = 0.1$, ζ_t和γ_t都初始化为0.25, 此后每学习1000万步, ζ_t和γ_t均衰减一半. 学习结束后, 对每个给定状态, 比较所有行动的Q值, 找出最优行动, 从而构成一个策略. 然后根据该策略, 模拟实际系统的运行过程(运行步数100万), 计算相关结果, 进行测试比较. 图2∼4为平均准则下, 不同工件到达率时的统计实验结果(5次独立实验平均值).

图2∼4中, 工件处理率为两机器人总共捡取的工件数与传送带上总流过的工件数之比, 反映了系统的工作效率; 机器维修率为仿真期间系统维修所用时间步数与总仿真步数之比, 反映了系统的稳定性; 系统中两个机器人R_1, R_2各自的工件处理率, 反映了他们分摊的任务, 数值越接近则系统的均衡性越好. 显然, 工件到达率越大、处理率越低、维修率越高. 这3个图还分别说明, CMAC-Q算法得到的工件处理率、系统稳定性和均衡性均略优于Q学习算法. 这是因为Q学习需对机器人的发热量进行离散化处理, 导致一定的离散化误差, 并且, 学习过程中没有经历的状态–行动对的Q值不会发生变化. 而CMAC网络能够处理连续输入变量, 尽管仿真中其内部的量化级数与Q学习时所划分的等级数相等, 但该网络本身具有信息的分布式存储功能, 具有一定的泛化能力, 因而学习效果要略好一些. 实验结果也显示出CMAC-Q算法在处理具有连续状态变量和大状态空间问题时所具有的优势.

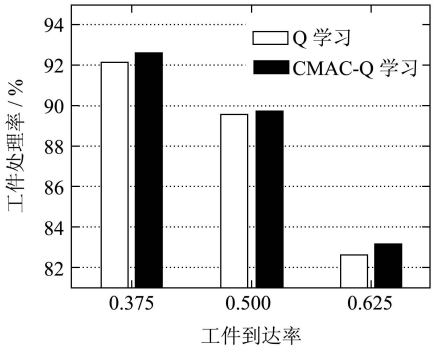

图 2　系统工件处理率($\beta = 1$)

Fig. 2　The processing rates of the system ($\beta = 1$)

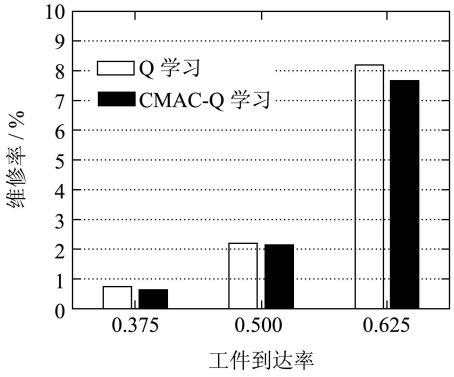

图 3　系统维修率($\beta = 1$)

Fig. 3　The maintenance rates of the system ($\beta = 1$)

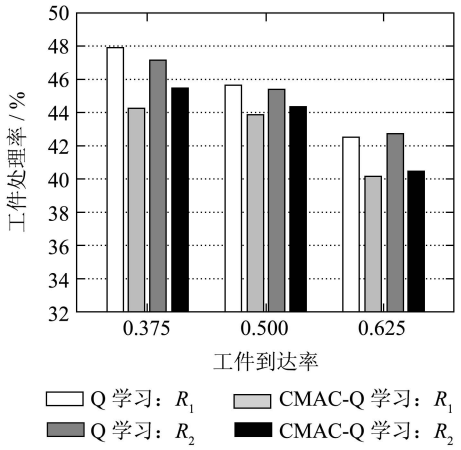

图 4　两机器人各自的工件处理率($\beta = 1$)

Fig. 4　The processing rates of each robot ($\beta = 1$)

本文还把CMAC-Q与一种改进的Q学习算法, 即基于模拟退火Metropolis准则的Q学习(简记SA-Q), 进行了比较. 图5为$\beta = 1$和$\lambda = 0.5$时, 3种算法的平均报酬优化曲线, 其中, 每学习50万步, 产生一个策略, 根据该策略模拟实际系统的运行(100万步), 统计得到相应平均报酬. 显然, SA-Q要优于一般Q学习. 与SA-Q和Q学习相比, CMAC-Q在学习较少的步数后就能得到较好的结果. 原因是CMAC作为一种局部逼近的神经网络, 每次学习修正的权值极少, 学习速度快; 且CMAC网络具有一定的泛化能力, 某一局部点的学习结果可以影响到未被训练的地址单元, 从而能大大提高学习效率.

在INTEL双核CPU1.60 GHz、内存为1 G的PC机上对上述3种算法进行测试比较, 结果见表1. 可见, CMAC-Q的学习值达到相对稳定的学习步数和用时最小, 但每学习一次(50万步)的平均用时最大(因为要进行相对复杂的CMAC网络计算); Q学习的学习步数和用时最多, 但每次学习的平均用时最小, 计算最简单. SA-Q算法处于两者之间.

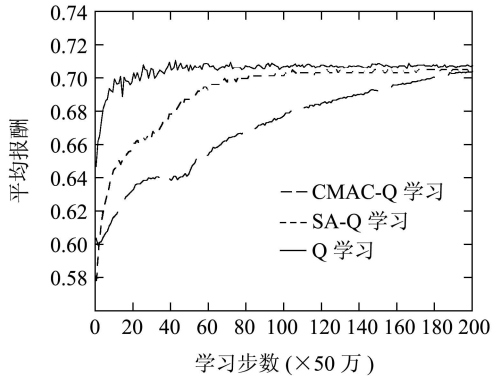

图 5 平均报酬优化曲线($\beta = 1, \lambda = 0.5$)

Fig. 5 The optimization plots of average rewards ($\beta = 1, \lambda = 0.5$)

表 1 3种算法相关学习结果

Table 1 Learning results of three algorithms

算法	平均报酬	学习时间/s	步数/万	存储单元
Q	0.7042	约620	200×50	28576800
SA-Q	0.7048	约300	85×50	28576800
CMAC-Q	0.7076	约225	45×50	约64000

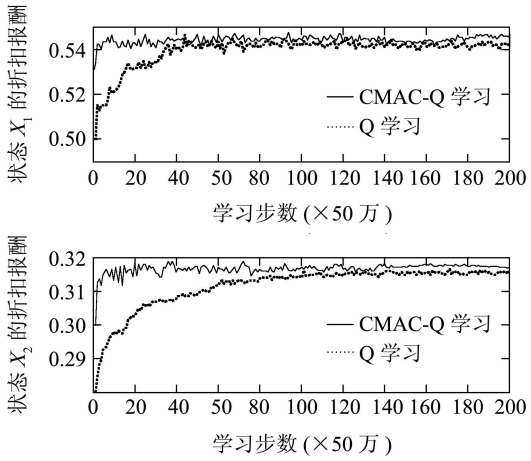

图 6 状态X_1和X_2的折扣报酬优化曲线($\beta = 0.9, \lambda = 0.5$)

Fig. 6 The optimization plots of discounted rewards of X_1 and X_2($\beta = 0.9, \lambda = 0.5$)

图6为β=0.9, λ=0.5时, 两种算法下状态X_1 = (0 0 0 0 2 1 0 0 0 0)和X_2=(0 0 0 0 2 1 0 0 17.85 17.85)的折扣报酬曲线. 显然, 状态X_2的折扣报酬小于状态X_1的折扣报酬, 与实际情况相符. 这是因为状态X_2时的热量值已接近机器人所能承受的最大热量, 因而从该状态出发产生的折扣报酬统计意义上应小些. 图6还说明, 在折扣性能准则下, CMAC-Q学习与传统Q学习和有关改进Q学习相比, 学习速度也同样有很大的提高, 并且优化结果也略优于Q学习算法产生的结果.

6 结论(Conclusions)

本文针对两机器人双传送带搬运系统, 研究给出的适用于平均或折扣准则的CMAC-Q学习优化算法, 与传统的Q学习相比, 在求解具有连续状态变量和状态空间巨大的Markov系统优化问题时, 具有一定优越性. 一方面, 它能节省大量存储空间, 另一方面, 它能提高学习速度. 在实际的生产过程中, 搬运系统可能包含更多数目或功能不同的机器人, 其作业分配问题将有待深入研究.

参考文献(References):

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

[2] YAMADA A, TAKATA S. Reliability improvement of industrial robots by optimizing operation plans based on deterioration evaluation[J]. *Annals of The International Academy for Production Engineering*, 2002, 51(1): 319 – 322.

[3] KAKAMU K, YAMANOBE N. Task assignment of high-speed handling operations to multiple robots considering robot fatigue[C] *//Design Engineering Workshop, 5th Japan-Korea CAD/CAM Workshop*. Japan: FuJi Technology Press, 2005: 185 – 189.

[4] 丁丽洁, 唐昊, 周雷. 基于对等SAP的Q学习算法在机器人作业分配中的应用[C] //第26届中国控制会议. 北京: 北京航空航天大学出版社, 2007: 536 – 539.
(DING Lijie, TANG Hao, ZHOU Lei. The application of peer to peer SAP-based Q-learning in task assignment to multiple robots[C] *//Proceedings of the 26th Chinese Control Conference*. Beijing: BUAA Press, 2007: 536 – 539.)

[5] TANG H, YUAN J B, LU Y, et al. Performance potential-based neuro-dynamic programming for SMDPs[J]. *Acta Automatic Sinica*, 2005, 31(4): 642 – 645.

[6] 唐昊, 周雷, 袁继彬. 折扣平均准则MDP基于TD(0)学习的统一NDP方法[J]. 控制理论与应用, 2006, 23(2): 292 – 296.
(TANG Hao, ZHOU Lei, YUAN Jibin. Unified NDP method based on TD(0) learning for both average and discounted Markov decision processes[J]. *Control Theory & Applications*, 2006, 23(2): 292 – 296.)

[7] WATKINS C J C H. *Learning from delayed rewards*[D]. UK: King's College, 1989.

[8] CAO X R. *Stochastic Learning and Optimization: A Sensitivity-Based View*[M]. New York: Springer, 2007.

[9] ALBUS J S. A new approach to manipulator control: The Cerebellar model articulation controller(CMAC)[J]. *Journal of Dynamic Systems, Measurement, and Control Transactions of ASME*, 1975, 97(3): 220 – 227.

[10] ALBUS J S. Data storage in the Cerebellar model articulation controller(CMAC)[J]. *Journal of Dynamic Systems, Measurement, and Control Transactions of ASME*, 1975, 97(3): 228 – 233.

作者简介:

唐 昊 (1972—), 男, 教授, 安徽省高校优秀中青年骨干教师, 主要从事离散事件动态系统(DEDS)、强化学习(RL)和神经元动态规划(NDP)及智能优化等理论和应用研究, E-mail: htang@hfut.edu.cn;

丁丽洁 (1981—), 女, 硕士, 主要研究方向为强化学习等;

程文娟 (1970—), 女, 副教授, 研究方向包括智能决策支持系统和DEDS等;

周 雷 (1981—), 男, 助教, 研究方向为DEDS、强化学习以及智能优化方法.