基于 CMAC 网络强化学习的电梯群控调度

高 阳,胡景凯,王本年,王冬黎
(南京大学软件新技术国家重点实验室,江苏南京 210093)

摘 要: 电梯群控调度是一类开放、动态、复杂系统的多目标优化问题. 目前应用于群控电梯调度的算法主要有分区算法、基于搜索的算法、基于规则的算法和其他一些自适应的学习算法. 但已有方法在顾客平均等待时间等目标上并不能够达到较好的优化性能. 本文采用强化学习技术应用到电梯群控调度系统中,使用 CMAC 神经网络函数估计模块逼近强化学习的值函数,通过 Q-学习算法来优化值函数,从而获得优化的电梯群控调度策略. 通过仿真实验表明在下行高峰模式下,本文所提出的基于 CMAC 网络强化学习的群控电梯调度算法,能够有效地减少平均等待时间,提高电梯运行效率.

关键词: 电梯群控调度; 强化学习; CMAC 神经网络; 函数估计
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2007) 02-0362-04

Elevator Group Control Using Reinforcement Learning with CMAC

GAO Yang, HU Jing-kai, WANG Ben-nian, WANG Dong-li
(National Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu 210093, China)

Abstract: Elevator group control is a multi-objective optimization problem in an open, complicated and dynamical system. Currently, many algorithms have been applied in elevator group control, such as zoning approaches, search-based approaches, rule-based approaches and other adaptive approaches. However these methods fail of achieving the optimal performance in the average wait time. In this paper, the reinforcement learning technology is applied in the elevator group control system. The CMAC neural network is used to approx the value function of reinforcement learning and Q-learning algorithm is used to optimize the value function, thereby the optimal control policy of the elevator group control is achieved. The simulation experiment shows that the elevator group control using reinforcement learning with CMAC can reduce the average wait time efficiently in the down peak traffic.

Key words: elevator group control; reinforcement learning; CMAC neural network; function approximation

1 引言

早期电梯控制采用单呼梯信号形式,随着计算机控制和智能技术的发展,由计算机统一管理一组电梯的呼叫和指令信号,根据系统设定的优化目标和建筑物中的实际交通状况,产生最优电梯调度策略[1]. 这就是目前常见的电梯群控系统. 其调度的实质是在开放、动态的复杂环境中,对乘客候梯时间、乘客乘梯时间、拥挤度和能耗等多个优化目标进行优化控制. 目前群控电梯调度算法主要有分区算法[2],基于搜索的算法[3]和基于规则的算法[4,5]等等. 随着智能技术的发展,越来越多的研究者采用专家系统、模糊控制、人工神经网络以及遗传算法等技术研究自适应的学习算法. 但由于电梯运行在一个连续时间系统中,其状态空间高维,同时外部状态不能完全感知且随乘客到达率变化而动态改变,因此有效计算电梯群控调度的最优策略仍然是研究界和产业界面临的主要难题之一.

考虑到电梯面临的实际环境是未知的、不确定的,而调度是针对顾客到达模型的在线优化. 因此 Crites 等人提出将强化学习(Reinforcement learning)技术应用到电梯群控调度中,通过仿真实验表明其方法与目前已有算法相比,能够获得较小的顾客平均等待时间[6]. 在 Crites 等人的方法中,采用 BP 神经网络逼近连续状态空间的值函数,但由于 BP 神经网络在增量、在线学习中收敛并不稳定. 而 CMAC 神经网络具有较好的在线增量学习能力,因此在本文中设计了基于 CMAC 神经网络的强化学习算法,对电梯群控调度进行优化. 将基于 CMAC 网络函数估计的强化学习算法应用于在下行高峰模式中,实验表明基于 CMAC 网络函数估计的强化学习算法能有效地减少乘客平均等待时间,提高电梯调度的性能,并相比基于 BP 神经网络的强化学习电梯调度算法具有更优的性能.

2 强化学习函数估计方法

2.1 强化学习

强化学习是一种以环境反馈作为输入的、特殊的、适应环境的机器学习方法. 所谓强化学习是指从环境状态到行为映

收稿日期:2006-01-25;修回日期:2006-04-20
基金项目:国家自然科学基金(No. 60475026);国家杰出青年科学基金(No. 60325207);国家 973 重点基础研究发展规划(No. 2002CB312002)

射的学习,以使系统行为从环境中获得的累计奖赏值最大(即累计代价值最小).该方法不同与传统监督学习技术那样通过正例、反例来告知采取何种行为,而是通过试错(Trial-and-error)的方法来发现最优行为策略[7,8].如果假定环境满足马尔可夫型属性,则顺序型强化学习问题可以通过马氏决策过程(Markov Decision Process)建模.

马氏决策过程　由四元组$\langle S, A, R, P\rangle$定义.包含一个环境状态集 S,系统行为集合 A,代价函数 $R: S\times A\to \mathbf{R}$ 和状态转移函数 $P: S\times A\to PD(S)$.记 R_{ss}^{a} 为系统在状态 s 采用 a 动作使环境状态转移到 s' 获得的瞬时代价值;记 $P_{ss'}^{a}$ 为系统在状态 s 采用 a 动作使环境状态转移到 s' 的概率.

马氏决策问题的目标是发现优化策略(即在每个状态时的动作选择),以使累计的代价和最小.根据 Bellman 最优策略公式,在最优策略 π^* 下系统在 s 状态下的值函数由式(1)定义.式(1)中的 γ 为折扣因子.

$$V^*(s)=\max_{a\in A}E\{r_{t+1}+\gamma V^*(s_{t+1})\mid s_t=s, a_t=a\}$$
$$=\max_{a\in A}\sum_{s'}P_{ss'}^{a}[R_{ss'}^{a}+\gamma V^*(s')] \quad (1)$$

在动态规划技术中,在已知状态转移概率函数 P 和代价函数 R 的环境模型知识前提下,从任意设定的策略 π_0 出发,可以采用策略迭代或值迭代的方法逼近最优的 V^* 和 π^*.但由于在强化学习所应用问题中,P 函数和 R 函数未知,系统无法直接通过动态规划技术进行值函数计算.因此在学习过程中常采用逼近的方法进行值函数的估计,其中最主要的方法之一是先通过 Monte Carlo 采样,然后结合动态规划技术,设计强化学习中的值函数迭代公式.分别如式(2)和式(3).其中式(2)中 R_t 是指当系统采用某种策略 π,从 s_t 状态出发获得的真实的累计折扣代价值;R_t 在式(3)中由当前代价值和下一状态值函数估计;α 为学习率.

$$V(s_t)\leftarrow V(s_t)+\alpha[R_t-V(s_t)] \quad (2)$$
$$V(s_t)\leftarrow V(s_t)+\alpha[r_{t+1}+\gamma V(s_{t+1})-V(s_t)] \quad (3)$$

2.2　基于神经网络函数估计的强化学习

但对于大规模 MDP 或连续状态 MDP 问题中,强化学习过程不可能遍历所有状态.因此要求强化学习的值函数具有一定泛化能力.强化学习函数估计的本质就是用参数化的函数逼近状态-值函数的映射关系.由于应用神经网络进行强化学习中的函数估计,可以有效避免维数灾难的问题[9],因此在本文中只考察基于神经网络函数估计的强化学习方法.强化学习函数估计框架如图 1.

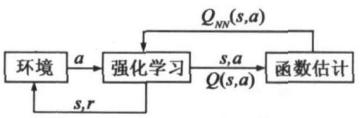


图 1　强化学习函数估计框架

以经典强化学习 Q-学习算法为例,函数估计模块采用神经网络逼近连续状态空间的 Q 函数,设其逼近的 Q 函数为 Q_{NN},则强化学习函数估计中 Q 值的 Bellman 迭代公式可修改为式(4):

$$Q(s,a)\leftarrow Q_{NN}(s,a)+\alpha(r+\gamma\max_{a'\in A}Q_{NN}(s',a')-Q_{NN}(s,a)) \quad (4)$$

对于状态-动作对(s,a)下,函数估计模块的输出为 $Q_{NN}(s,a)$.而强化学习模块期望的输出为 $r+\gamma\max_{a'\in A}Q_{NN}(s',a')$,则通过梯度下降方法,将神经网络权值更新如式(5):

$$\Delta w=\alpha(r+\gamma\max_{a'\in A}Q_{NN}(s',a')-Q_{NN}(s,a))\left(\frac{\partial Q_{NN}(s,a)}{\partial w}\right) \quad (5)$$

在此框架下,应用神经网络来进行函数估计的强化学习算法流程如表 1:

表 1　基于神经网络函数估计的强化学习算法

Step1:观察当前状态 s;
Step2:根据 s 以及函数估计模块,计算 $Q_{NN}(s,a)$;
Step3:使用 ε-greedy 策略选择动作 a,并执行它;
Step4:从环境中获得即时代价 r,并观察下一状态 s';
Step5:根据式(4)学习值函数 $Q(s,a)$;
Step6:根据式(5)更新神经网络权值 w;
Step7:转至 Step1,直到网络稳定或学习达到一定精度.

从表 1 中可知学习过程中同时存在两个收敛过程:一是强化学习的收敛过程;二是神经网络函数逼近的收敛过程.最终基于神经网络函数估计的强化学习算法收敛性和效率与这两个收敛过程有关.因此,神经网络函数估计模块的设计对整个强化学习的性能有着很大的影响.由于 BP 神经网络在增量、在线学习中收敛不稳定,而 CMAC 神经网络具有较好的在线增量学习能力,因此我们设计基于 CMAC 网络函数估计的强化学习算法.下面首先介绍 CMAC 网络结构和学习过程.

Albus 于 1975 年首次提出小脑模型关节控制器(Cerebellar Model Articulation Controller,CMAC),其本质上是一个非线性映射[10].CMAC 网络包含 3 个映射,分别是输入层的非线性映射;中间层的内部映射;以及输出层的线性映射.以下结合图 2 说明 CMAC 网络的运行过程[10].

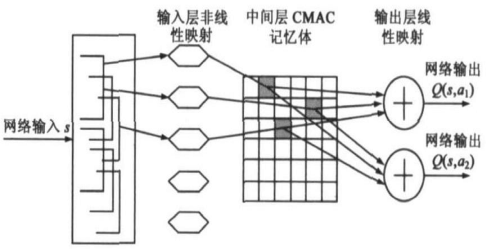


图 2　CMAC 神经网络结构示意图

表 2　*CMAC* 网络学习过程

Step1:用式(6)对输入状态矢量 s 进行规格化;
Step2:根据泛化参数 C 确定激活区域,如式(7);
Step3:通过被激活记忆体的加权求和计算 CMAC 网络输出,如式(8);
Step4:采用 δ 学习规则被激活区域权值进行调整,如式(9).返回 Step1.

$$s=[s_1, s_2, \cdots, s_n]$$
$$s'=[s_1', \cdots, s_n']=\left[\mathrm{int}\left(\frac{s_1}{\Delta_1}\right), \cdots, \left(\frac{s_n}{\Delta_n}\right)\right] \quad (6)$$
$$Addr_i=[s_1'-\mathrm{mod}[(s_1'-i), C], \cdots, s_n'-\mathrm{mod}[(s_n'-i), C]]$$
$$i=1,2,\cdots,C \quad (7)$$

$$Q(s,a)=\frac{1}{C}\sum_{i=1}^{C}W\cdot F(Addr_i) \tag{8}$$

$$\Delta W_i=\delta\cdot(Q(s,a)-Q(s,a))\cdot F(Addr_i) \tag{9}$$

式(8)中 $F(Addr_i)$ 表示在 CMAC 网络在 $Addr_i$ 地址上存储的值. 通过以上步骤可以保证 CMAC 网络的在线增量学习能力,并且其计算效率要远优于其他神经网络[11].

3 基于 CMAC 网络强化学习的群控电梯调度算法

对于行驶在楼层间的电梯,如果有乘客希望在下一层出电梯的情况,电梯被强令选择"下一层停靠";否则必须做出"下一层停靠"或"继续行驶"的抉择. 以下算法只给出用强化学习进行"下一层停靠"或"继续行驶"抉择的算法流程:

表 3 基于 CMAC 网络强化学习的电梯群控算法

Step1:在 t_x 时刻电梯 i 到达一个决策点,观察得到状态为 x,根据 CAMC 网络计算 $Q(x,run)$ 和 $Q(x,stop)$. 这里 $Q(x,run)$ 为在 x 状态下继续运行的 Q 值函数,而 $Q(x,stop)$ 为停靠的值函数.

Step2:根据建立在 Q 值上的 Boltzmann 分布选择动作 a,如式(10).

Step3:令电梯 i 的下一个决策点发生在 t_y 时刻,其相应的状态为 y. 根据式(11),更新所有电梯的获得 $R[i]$ 值.

Step4:然后,电梯 i 根据式(12)调节其 $Q(s,a)$ 的估值.

Step5:根据式(13)更新 CMAC 网络权值.

Step6:将 $x\leftarrow y$, $t_x\leftarrow t_y$. 转至 Step1.

建立在 Q 值上的 Boltzmann 分布动作选择公式为:

$$\Pr(stop)=\frac{e^{Q(x,run)/T}}{e^{Q(x,stop)/T}+e^{Q(x,run)/T}} \tag{10}$$

这里 $T>0$ 为温度参数,T 值的大小决定动作选择的随机程度. 在训练学习过程中 T 值将逐渐衰减. 学习初期由于 Q 值不精确,故 T 赋值为较大的数值以确保对每个动作而言都有近似相同的尝试机会. 随着充分学习,Q 值将越来越精确,T 将赋值为较小的数值保证以较大的概率选择当前认为较好的动作.

对电梯控制器而言,设 $R[i]$ 为第 i 部电梯从其上一次决策时间点 $d[i]$ 时开始累计的总折扣强化值. 当在每个事件发生时,对 $R[i]$ 进行以下计算:令 t_0 为上一事件发生的时间,t_1 为当前事件发生的时间. 对于每个在 t_0 和 t_1 之间有效的电梯呼叫键 b 而言,令 $w_0(b)$ 和 $w_1(b)$ 分别为 t_0 和 t_1 时刻按钮 b 按下后逝去的时间. 式(11)中 β 为顾客平方等待时间的指数衰减速率,λ 为顾客的泊松到达率[6].

$$\begin{aligned}\Delta R[i]=\ & e^{-\beta(t_0-d[i])}\sum_b\left\{\frac{2\lambda_b(1-e^{-\beta(t_1-t_0)})}{\beta^3}+\left(\frac{2}{\beta^3}+\frac{2w_0(b)}{\beta^2}+\frac{w_0^2(b)}{\beta}\right)\right.\\ & -e^{-\beta(t_1-t_0)}\left(\frac{2}{\beta^3}+\frac{2w_1(b)}{\beta^2}+\frac{w_1^2(b)}{\beta}\right)\\ & +\lambda_b\left[\left(\frac{2w_0(b)}{\beta^2}+\frac{w_0^2(b)}{\beta}+\frac{w_0^3(b)}{3\beta}\right)\right.\\ & \left.\left.-e^{-\beta(t_1-t_0)}\left(\frac{2w_1(b)}{\beta^2}+\frac{w_1^2(b)}{\beta}+\frac{w_1^3(b)}{3\beta}\right)\right]\right\}\end{aligned}$$

$$R[i]\leftarrow R[i]+\Delta R[i] \tag{11}$$

强化学习的 Q 值学习和 CMAC 网络权值调整公式分别如式(12)和式(13).

$$Q(x,a)\leftarrow R[i]+e^{-\beta(t_y-t_x)}\min_{a'\in\{stop,cont\}}Q_{cmac}(y,a') \tag{12}$$

$$\begin{aligned}\Delta W=\ & \alpha[R[i]+e^{-\beta(t_y-t_x)}\min_{a'\in\{stop,cont\}}Q_{cmac}(y,a',W)\\ & -Q_{cmac}(x,a,W)]\nabla_W Q_{cmac}(x,a,W)\end{aligned} \tag{13}$$

4 实验及分析

4.1 电梯系统模型

电梯控制系统通常包括上行高峰、下行高峰、层间交通等几种不同模式[12]. 在不同的交通模式下,系统应采用不同的群控策略来提高服务性能. 显然,客流量高的上行高峰和下行高峰是对调度算法最有挑战的任务.

表 4 给出实验的主要配置. 此配置与 Crites 等人的强化学习电梯调度实验是一致的[6],其中 CMAC 网络权值被统一初始化为-1 到 +1 之间的离散值.

表 4 电梯群控调度仿真实验参数表

	参数描述	参数设置
静态参数	电梯数目	4
	楼层数目	10
	梯内呼叫键数目	10
	额定载客量	20
动态参数	层间运行时间	1.45s
	电梯停止时间	7.19s
	电梯转向时间	1s
	乘客进出电梯时间	均值为 1s 的厄兰分布,范围从 0.6 到 6.0s
顾客到达模型	时间 00	到达率 1
	时间 05	到达率 2
	时间 10	到达率 4
	时间 15	到达率 4
	时间 20	到达率 18
	时间 25	到达率 12
	时间 30	到达率 8
	时间 35	到达率 7
	时间 40	到达率 18
	时间 45	到达率 5
	时间 50	到达率 3
	时间 55	到达率 2
CMAC 网络参数	输入节点	47 个
	输出节点	2 个
	泛化参数	3
强化学习参数	指数衰减速率 β	0.01
	学习率 α	满足 $\sum\alpha=\infty$, $\sum\alpha^2<C<\infty$

4.2 实验结果对比分析

我们将所提出的基于 CMAC 网络的强化学习电梯调度算法与 Crites 等人提出的基于 BP 神经网络的强化学习调度算法以及应用于实际电梯调度的 SECTOR 算法作比较,结果见表 5～7. 其中 SECTOR 算法结果参照 Crites 等人 1998 的实验数据. 由于参数调整的原因,本实验没有能够重复获得 Crites 等人使用 BP 网络进行电梯调度所获得的性能[6]. 但我们认为,在相同条件设置下,不影响对不同电梯群控调度算法之间性能的比较. 基于 BP 和 CMAC 网络强化学习的电梯群控调度实验结果均为 20 个小时模拟后的平均结果,实验分别测试了

下行高峰模式下的三种不同的交通模式:仅下行交通、含上行交通以及两倍上行交通.

表 5　仅含下行交通模式的对比实验结果

算法	AvgWait	SquaredWait	Percent > 60s
SECTOR	21.4	674	1.12
RL-BP	21.2	569	0.09
RL-CMAC	19.7	529	0.07

表 6　含上行交通模式的对比实验结果

算法	AvgWait	SquaredWait	Percent > 60s
SECTOR	27.3	1252	9.24
RL-BP	24.3	1140	9.90
RL-CMAC	21.8	1048	9.14

表 7　两倍上行交通的对比实验结果

算法	AvgWait	SquaredWait	Percent > 60s
SECTOR	30.3	1643	13.50
RL-BP	27.8	1698	8.74
RL-CMAC	23.4	1562	8.20

表 5～7 给出三种交通模式下,三种方法在平均等待时间,平方等待时间和 Percent > 60s 三个指标下的对比结果.从表中可以看出,本文所提出的方法在平均等待时间上缩短 2 秒左右,在平方等待时间上缩短 100 秒左右.表中 Percent > 60s 是指等待顾客中等待时间超过 1 分钟的顾客比,显然本文所提方法也具有更优的性能.

5　结语

使用神经网络进行函数估计来解决大空间或连续状态空间强化学习问题,可以初步解决"维数灾难"问题.本文将基于 CMAC 网络函数估计的强化学习算法应用于电梯群控调度中,取得了良好的效果.然而如何调整 CMAC 网络参数从而更好地逼近值函数仍然是值得进一步研究的地方.除此之外,在目前仅能通过仿真实验的方法确定本文提出算法的有效性.因此从理论上证明算法的性能是本文的重要后续工作之一.

参考文献:

[1] 郑延军,张惠侨,叶庆泰,朱昌明.电梯群控系统客流分析与仿真[J].计算机工程与应用,2001.22:139 - 141.
Zhen Yanjun, Zhang Huiqiao, Ye Qingtai, Zhu Changming. Passenger flow analysis and simulation of elevator group control system[J]. Computer Engineering and Applications, 2001. 22: 139 - 141. (in Chinese)

[2] Y Sakai, K Kurosawa. Develop of elevator supervisory group control system with artificial intelligence[J]. Hitachi Review, 1984, 33:25 - 30.

[3] M L Siikonen. Elevator traffic simulation[J]. Simulation, 1993, 61:257 - 267.

[4] H Ujihara, S Tsuji. The revolutionary AI-2100 elevator-group control system and the new intelligent option series[J]. Mitsubishi Electric Advance, 1988, 45:5 - 8.

[5] H Ujihara, M Amano. The latest elevator group-control system[J]. Mitsubishi Electric Advance, 1994, 67:10 - 12.

[6] Crites R H, Barto A G. Elevator group control using multiple reinforcement learning agents[J]. Machine Learning, 1998, 33(2):235 - 262.

[7] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: a survey[J]. Journal of Artificial Intelligence Research, 1996, 4:237 - 285.

[8] R S Sutton and A G Barto. Reinforcement Learning[M]. Cambridge, MA: MIT Press, 1998.

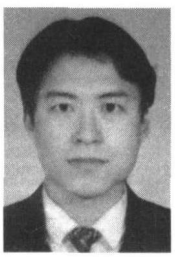
[9] Rich S Sutton. Generalization in reinforcement learning: successful examples using sparse coarse coding[A]. D Touretzky, M Mozer, M Hasselmo, Advances in Neural Information Processing Systems 8[C]. New York: MIT Press, 1996. 1038 - 1044.

[10] Albus J S. A new approach to manipulator control: The cerebellar model articulation controller(cmac)[J]. Journal of Dynamic Systems, Measurement, and Control, 1975, 97(3):220 - 227.

[11] 杨艳丽,曹广忠. CMAC 的无交叠感受域变分辨率学习方法[J].电子学报,2002,30(12A):2153 - 2154.
Yang Yanli, Cao Guangzhong. The CMAC learning algorithms of non-overlapping receptive field with variable resolution[J]. Acta Electronica Sinica. 2002, 30(12A): 2153 - 2154. (in Chinese)

[12] 郑延军,张惠侨,叶庆泰,朱昌明.电梯动态分区算法及其遗传进化研究[J].计算机工程与应用,2001.22:142 - 144.
Zheng Yanjun, Zhang Huiqiao, Ye Qingtai, Zhu Changming. The research on elevator dynamic zoning algorithm and it's genetic evolution[J]. Computer Engineering and Applications, 2001. 22:142 - 144. (in Chinese)

作者简介:



高　阳　男,1972 年生于江苏,2000 年获南京大学计算机科学与技术专业博士学位,现为南京大学计算机系副教授,中国人工智能学会理事,中国机器学习专业委员会常务委员.主要研究方向为强化学习、分布式人工智能和智能系统等. E-mail: gaoy@nju.edu.cn

胡景凯　男,1982 年生于湖北,硕士研究生,研究方向为机器学习和智能信息处理.

王本年　男,1967 年生于安徽,博士研究生,研究方向为机器学习和智能信息处理.