DOI:10.16185/j.jxatu.edu.cn.2007.04.009
第27卷 第4期 西 安 工 业 大 学 学 报 Vol.27 No.4
2007年08月 JOURNAL OF XI'AN TECHNOLOGICAL UNIVERSITY Aug. 2007

文章编号: 1673-9965(2007)04-325-05

一种在线自学习的移动机器人模糊导航方法*

赫东锋[1,2], 孙树栋[2]

(1. 西安工业大学 机电工程学院, 西安 710032; 2. 西北工业大学 机电学院, 西安 710072)

摘 要: 针对现有移动机器人模糊导航对未知不确定环境缺乏自适应性的缺点, 提出了一种具备在线自学习能力的模糊导航方法. 通过设计模糊规则并确定动作先验值, 完成初始模糊导航系统的构建. 利用Q强化学习算法对模糊规则中各行为的值函数进行在线增量学习, 实现模糊决策的逐步求精. 仿真实验表明, 移动机器人导航系统能够在运行过程中不断调整导航策略, 实现对未知不确定环境的自适应. 同时由于导航先验知识的引入, 有效地克服了强化学习初始阶段进行盲目搜索导致的学习速率低、收敛速度慢的缺点, 实现了移动机器人可靠导航.

关键词: 机器人导航; 模糊逻辑; 在线自学习; Q强化学习

中图号: TP242.6 文献标识码: A

移动机器人导航是移动机器人研究领域内的基础问题, 也是机器人学领域的研究热点. 由于已知确定环境下的移动机器人导航技术已经相对成熟, 因此当前的研究重点主要集中于未知不确定环境下的导航技术[1].

模糊控制具有不需要控制对象模型, 易于引入专家知识, 鲁棒性好等优点, 在移动机器人导航控制中也得到了广泛应用[2]. 由于实际应用的复杂性, 在未知环境下要预先根据专家经验总结出完善的控制规则往往非常困难. 另外, 由于模糊变量的隶属函数与模糊规则在运行过程中往往无法修改, 因此它对不确定环境缺乏适应能力. 一些学者试图使用人工神经元网络方法实现模糊系统的自学习与自适应[3], 但需要大量的输入输出数据对作为教师信号, 这往往在机器人的实际应用中很难获得.

强化学习是一种智能体通过试错法(Trial-and-Error)不断与环境交互并通过环境反馈获取经验实现从状态到动作的最佳映射的机器学习方法. 由于强化学习不需要教师数据且易于实现增量式学习, 因此比较适合于未知不确定环境下的机器人导航任务[4]. 但由于强化学习的延迟回报性以及行为探索时的盲目性, 强化学习的效率往往比较低.

文中提出了具有在线自学习能力的移动机器人模糊导航方法. 通过在模糊导航系统中引入强化学习实现导航系统对未知不确定环境的自适应与自学习. 同时, 通过模糊系统中导航专家先验知识的引入, 以加快强化学习的学习速度, 从而达到模糊导航与强化学习之间相互取长补短的目的, 实现移动机器人在未知不确定环境下的可靠导航.

1 机器人模型与导航方法原理

1.1 机器人模型

机器人模型如图1所示. 为简化问题, 假设传感器在理想状态下工作, 即机器人能够实时检测自身以及目标和障碍物的状态信息. 将机器人正前方135°范围内的区域划分为3个分别为45°的子区间: l, f, r, 分别代表左前方、正前方和右前方. 在每个子区间中设置一定数量的超声波传感器组, 且取每组传感器读数的最小值作为避障距离, 分别记作 d_l, d_f, d_r. 另外, 机器人上装有目标传感器, 探索目标点与机器人运动方向之间的夹角为 t_g, 且当目标位于机器人右前方时, 目标定位 t_g 为正, 否则 t_g 为负. 设机器人的控制量分别为机器人的运动速度 V_a 与机器人的转动角 S_a. 其中 $V_a \in \{0, 15\ m/s\}$, $S_a \in \{-30°, 30°\}$, 当机器人右转时, 转动角 S_a 为

* 收稿日期: 2007-06-12

作者简介: 赫东锋(1975-), 男, 西安工业大学讲师, 博士研究生, 主要研究方向为智能机器人与智能机械. E-mail: dongfenghe@163.com.

正,否则为负,下标 a 为机器人行为(action).设 R_{min} 为最小危险距离,R_{max} 为安全距离.

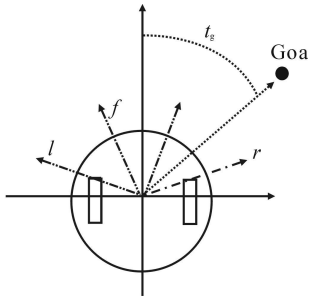


图 1 机器人模型

Fig. 1 Model of mobile robot

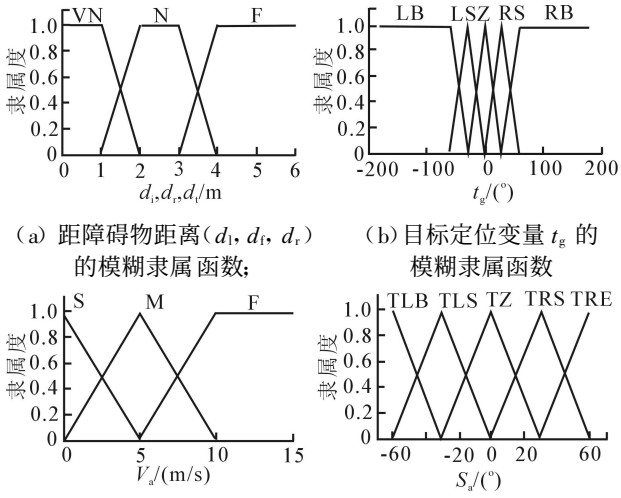
1.2 导航方法原理

导航方法由模糊导航规则和模糊导航规则的 Q 强化学习两部分组成.模糊导航规则在系统的构建阶段通过收集导航专家的经验知识来获得.模糊导航规则的 Q 强化学习则是在系统运行过程中,机器人根据所处状态通过一定策略选择导航动作并执行,然后依靠环境反馈(即强化信号)对动作执行的结果进行评估,根据评估结果反过来对动作选择策略进行改进,从而完成在线自学习过程.导航先验知识与 Q 强化学习通过模糊规则中的动作估计值联系在一起:导航先验知识决定动作估计值的初始值(称为动作先验估计值),Q 强化学习通过环境反馈不断更新动作估计值.动作估计值决定了对应动作被选择的概率,它与模糊规则一起决定了动作选择的策略,因此,动作估计值的更新意味着机器人不断根据环境变化学习动作策略的过程.

2 导航模糊规则的建立

2.1 输入输出变量隶属函数的确定

定义输入变量 d_l, d_f, d_r 的模糊语言变量为{VN, N, F}={“非常近”,“近”,“远”},论域范围为 0~6 m,其隶属函数如图 2(a)所示.目标定位变量 t_g 的模糊语言变量为{LB, LS, Z, RS, RB}={“左大”,“左小”,“零”,“右小”,“右大”},论域范围为(−180°~+180°),其隶属函数如图 2(b)所示.输出速度 V_a 的模糊语言变量为{S, M, F}={“慢速”,“中等”,“快速”},论域范围为(0~15 m/s),其隶属函数如图 2(c)所示.机器人转动角 S_a 的模糊语言变量为{TLB, TLS, TZ, TRS, TRB}={“左大”,“左小”,“零”,“右小”,“右大”},论域范围为(−60°~+60°),其隶属函数如图 2(d)所示.



(a) 距障碍物距离(d_l, d_f, d_r)的模糊隶属函数;

(b) 目标定位变量 t_g 的模糊隶属函数

(c) 输出速度 V_a 的模糊隶属函数;

(d) 转动角 S_a 的模糊隶属函数

图 2 输入-输出变量的隶属函数

Fig. 2 Membership function of input-output variable

2.2 具有强化学习能力的模糊规则

模糊规则可以看作是状态集 $S=\{S_i \mid S_i \in S\}$ 到动作集 $A=(a_i \mid a_i \in A)$ 的映射.而在强化学习中正是通过对这种映射关系强弱估计的逼近实现学习.为了使模糊规则即能体现状态-动作之间的映射关系,又具有 Q 强化学习能力,采用如下的模糊规则形式[5]

$$\begin{aligned} R_i&: \text{IF } S \text{ is } S_i, \text{ then } A \text{ is } a_{i1} \text{ with } q(i,1)\\ &\text{or } A \text{ is } a_{i2} \text{ with } q(i,2)\\ &\text{or } \cdots \text{or } A \text{ is } a_{iJ} \text{ with } q(i,J) \end{aligned} \tag{1}$$

式中:R_i 为第 i 条规则;S 为当前状态;S_i 为状态的模糊集,它是模糊变量 d_l, d_f, d_r, t_g 取值的组合,共有 3×3×3×5=135 种可能组合.$A=\{a_{ij} \mid j=1, \cdots J\}$ 表示动作集合,其中 a_{ij} 表示第 i 种状态时对应的第 j 个动作,它是模糊输出变量 V_a 与 S_a 取值的组合,共有 3×5=15 种可能组合,即 $J=15$. $q(i,j)$ 表示在状态 S_i 下,它代表了该动作被执行的概率.各动作的归一化估计值,归一化的目的是为了保证先验估计与后续学习估计在量值上的统一,即 $\sum_{j=1}^{J} q(1,j)=1$.

2.3 动作先验估计值的确定

通过动作先验估计将先验知识引入强化学习从而大大提高强化学习的学习速度是本文的特色之一.动作先验估计通过模糊综合决策方法来获得[6],具体步骤是:①针对每一种状态,根据专家驾驶经验,分别对输出变量 V_a 与 S_a 的各模糊语言变

量进行打分.②根据 d_l, d_f, d_r 的最小值所属的模糊语言变量,分别确定输出变量 V_a 与 S_a 的权重系数.③对输出变量进行加权平均并进行归一化处理,即可得到动作先验估计.例如,对于状态 S_i: d_r is FAR and d_f is FAR and d_l is NEAR and t_g is LS 的情形,其输出变量专家打分、输出变量权值以及各动作先验估计值分别见表 1~表 4.其他状态的动作先验估计值可用类似方法得到.

表 1　S_i 时输出变量 V_a 得分

Tab. 1　Score of output variable V_a with S_i

V_a	S	M	F
得分	40	50	10

表 2　S_i 时输出变量 S_a 得分

Tab. 2　Score of output variable S_a with S_i

S_a	TLB	TLS	TZ	TRS	TRB
得分	10	0	20	60	10

表 3　S_i 时输出变量权值

Tab. 3　Power of output variable with S_i

最小距离的模糊语言变量	VN	NEAR	FAR
V_a 权重	0.9	0.4	0.1
S_a 权重	0.1	0.6	0.9

表 4　S_i 时各动作先验估计值

Tab. 4　Transcendent estimate of action with S_i

动作(V_a, S_a)	(S, TLB)	(S, TLS)	(S, TZ)	(S, TRS)
估计值 $q(i, j)$	0.0579	0.0421	0.0737	0.1368
动作(V_a, S_a)	(S, TRB)	(M, TLB)	(M, TLS)	(M, TZ)
估计值 $q(i, j)$	0.0579	0.0684	0.0526	0.0842
动作(V_a, S_a)	(M, TRS)	(M, TRB)	(F, TLB)	(F, TLS)
估计值 $q(i, j)$	0.1474	0.0684	0.0263	0.0105
动作(V_a, S_a)	(F, TZ)	(F, TRS)	(F, TRB)	
估计值 $q(i, j)$	0.0421	0.1053	0.0263	

3　模糊导航控制的 Q 强化学习

3.1　系统动作输出的确定

机器人系统动作输出分两步进行,即判断自身所处状态和动作选择.

机器人自身所处状态的判断由模糊规则的前件匹配完成.用最小最大算子为当前状态,即

$$\begin{cases} \partial_i(s) = \min[\mu_i^1(s^1), \mu_i^2(s^2), \cdots, \mu_i^k(s^k)] \\ S_c = \arg\max[\partial_1(s), \partial_2(s), \cdots, \partial_i(s)] \end{cases} \tag{2}$$

式中:$\partial_i(s)$ 为第 i 条规则的适应度;$\mu_i^k(s_k)$ 为状态分量 s_k 的隶属度;S_c 为机器人当前所处状态.

机器人的动作选择由选定的模糊规则的后件决定.后件中每个动作的估计 $q(i, j)$ 值即代表了该动作被选择的概率.

3.2　强化信号 r_t 的计算[7]

强化信号是环境对于机器人所执行动作的一种评价.由于机器人导航行为由避障与趋进目标两个子行为构成,因此强化信号必须对这两个子行为进行综合评价.同时,当机器人与障碍物之间的距离不同时,机器人行为的侧重点也不同.本文采用式(3)作为强化信号.

$$r_t = \begin{cases} -10, & d_t < R_{\min} \\ \omega_1(|\alpha_t| - |\alpha_{t+1}|), & d_t > R_{\max} \\ \omega_2(|\alpha_t| - |\alpha_{t+1}|) + \omega_3(|d_{t+1}| - |d_t|), & \text{其他} \end{cases} \tag{3}$$

其中 $d_t = \min(d_l, d_f, d_r)$ 为 t 时刻机器人距离障碍物的最小距离.$\alpha_t = t_g$,表示 t 时刻机器人与目标点之间的夹角.ω_1, ω_2, ω_3 分别为权重系数.当 $d_t < R_{\min}$ 时,机器人与障碍物之间发生碰撞的可能性很大,机器人处于极度危险之中,应给予最大程度的立即惩罚.当 $d_t > R_{\max}$ 时,机器人距离障碍物较远,发生碰撞的可能性较小,所以重点考虑趋向目标行为.当 $R_{\min} \leqslant d_t \leqslant R_{\max}$ 时,机器人与障碍物之间存在碰撞的潜在危险,但情况并不是很紧急,所以应兼顾避障与趋向目标行为,即在避障的同时向目标点靠近.

3.3　动作估计值 q(i, j) 的学习

由于动作估计值 $q(i, j)$ 是动作先验值归一化的结果,与强化信号 r_t 在数量上并不统一,因此不能直接使用 Q 强化学习对动作估计值 $q(i, j)$ 进行估计,必须首先寻找动作估计值 $q(i, j)$ 与状态-动作值函数 $Q(S_i, a_j)$ 之间的关系.

当采用 Boltzmann 方法进行行为探索时,行为选择的概率 $p_{i,j}$ 为

$$p_{i,j} = e^{Q(s_i, a_j)/T} / \sum_{a_j \in A} e^{Q(s_i, a_j)/T} \tag{4}$$

其中 T 为温度系数.如果将动作估计值 $q(i, j)$ 看作是当前状态下动作 a_j 发生的概率,则有 $q(i, j) = p_{i,j}$,即动作估计值 $q(i, j)$ 与状态-动作值函数 $Q(S_i, a_j)$ 之间满足式(4).因此,直接用 Q 强化学习来估计状态-动作值函数 $Q(S_i, a_j)$,然后用式(4)完成状态-动作值函数 $Q(S_i, a_j)$ 到动作估计值 $q(i, j)$ 的转换.

根据 Q 强化学习公式,状态-动作值函数 $Q(S_i, a_j)$ 的估计式为[8]

$$Q(S_i, a_j) = Q(S_i, a_j) + \beta[r_t + \gamma \max Q(S_{i+1}, a_j) - Q(S_i, a_j)] \quad (5)$$

式中:β 为学习率;γ 为折扣系数.

对于由学习得到的动作估计值 $q(i, j)$,采用与原有值进行加权平均的方法进行更新,即

$$q(i, j) = (1 - \zeta)q(i, j) + \zeta p_{i,j} \quad (6)$$

其中 ζ 为更新率系数.

3.4 学习算法小结

整个在线学习算法如图 3 所示.

在起始位置采集传感器的值(d_l, d_f, d_r, t_g).
在运行期间不断循环:

① 根据隶属函数计算各分量的隶属度;由式(2)确定机器人当前状态,根据 对应的规则查找各动作估计值并以此为概率选择输出动作.

② 机器人执行动作,采集当前传感器的值(d_l, d_f, d_r, t_g),根据式(3)计算强化信号.

③ 根据式(5)计算新的状态-动作值函数 $Q(S_i, a_j)$;根据式(4)完成状态-动作值函数 $Q(S_i, a_j)$ 到动作估计值 $q(i, j)$ 的转换;根据式(6)更新动作估计值 $q(i, j)$.

④ 返回①,开始下一学习周期.

图 3 在线学习算法简图

Fig. 3 On-line learning algorithm

4 仿真实验及其结果

对于上述算法,采用 Mobotsim 软件进行仿真实验并与单纯基于模糊规则方法以及 Q 强化学习方法的导航系统进行了对比实验. 图 4 为仿真环境,图中线条为经过 200 次训练后机器人的运动轨迹. 图 5 为本文所用算法与单纯基于模糊规则算法和单纯基于 Q 强化学习算法的性能对比曲线,所有方法均采用每步平均回报作为性能指标,实线表示具有在线自学习能力的模糊推理导航方法的性能曲线,点划线表示一般 Q 强化学习导航方法的性能曲线,虚线表示普通模糊推理导航方法的性能曲线.

从仿真结果可知,所提方法能够成功的实现机器人导航. 同一般强化学习方法相比,该算法由于引入了先验信息,在算法的早期就可以得到比较高的每步平均奖励,同时由性能曲线也可以看出,该算法能够比较快的达到较优的学习状态,说明先验信息的引入有效的提高机器人的学习效率与收敛速度. 同普通模糊导航相比,具有在线自学习能力

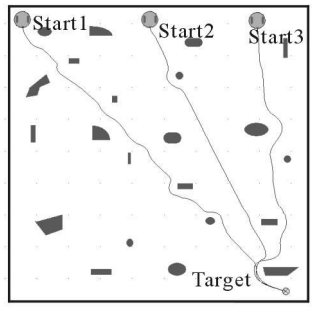


图 4 仿真环境

Fig. 4 Circumstance of simulation

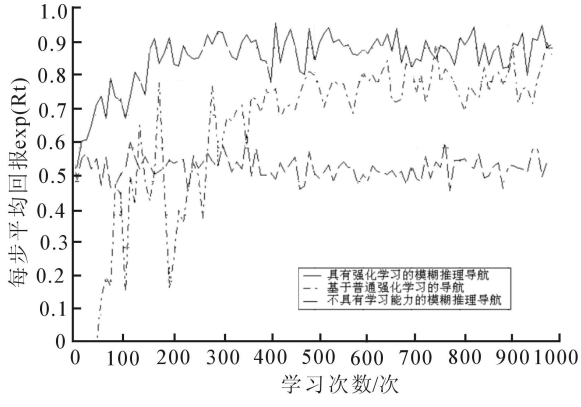


图 5 三种导航方法的性能对比

Fig. 5 Performance comparison of three navigation methods

的机器人的每步平均奖励随着时间的推移不断提高并且最终高于普通模糊导航的每步平均奖励,说明该算法可以在线提高机器人的导航性能,实现决策的逐步求精.

5 结 论

提出了一种将模糊规则与 Q 强化学习相结合实现未知不确定环境下机器人导航的方法. 通过 Q 强化学习,可以在线改善模糊导航的性能,使模糊导航系统对不确定环境具有一定的适应能力. 同时,由于模糊规则的使用,使 Q 强化学习具备了一定的先验知识,从而加快 Q 强化学习的收敛速度,克服传统 Q 强化学习学习速率低下的缺点. 仿真实验表明,该方法结合了模糊导航与 Q 强化学习的各自优点,实现了未知不确定环境下移动机器人的可靠导航,具有较高的实用价值.

参 考 文 献:

[1] 徐昕. 增强学习及其在移动机器人导航与控制中的应用研究[D]. 长沙:国防科学技术大学研究生院,2002.

XU Xin. Reinforcement Learning and Its Applications in Navigation and Control of Mobile Robots[D]. Changsha: College of Post-graduates National University of Defense Technology, 2002. (in Chinese)

[2] Xu W L, Tso S K. Real-time Self-reaction of a Mobile Robot in Unstructured Environments Using Fuzzy Reasoning[J]. Engineering Application of Artificial Intelligence, 1996, 9(5): 475.

[3] Adedeji B, John Y. Fuzzy Engineering Expert Systems with Neural Network Applications[M]. New York: John Wiley & Sons Inc., 2002.

[4] 陈卫东, 席裕庚, 顾冬雷. 自主机器人的强化学习研究进展[J]. 机器人, 2001, 23(4): 379.
CHEN Wei-dong, XI Yu-geng, GU Dong-lei. A Survey of Reinforcement Learning in Autonomous Mobile Robots[J]. ROBOT, 2001, 23(4): 379. (in Chinese)

[5] Lionel J. Fuzzy Inference System Learning by Reinforcement Method[J]. IEEE Transaction on System: man and Cybernetics, 1998, 28(3): 338.

[6] 杜春侠, 高云, 张文. 多智能体系统中具有先验知识的 Q 学习算法[J]. 清华大学学报: 自然科学版, 2005, 45(7): 981.
DU Chun-xia, GAO Yun, ZHANG Wen. Q-learning with Prior Knowledge in Multi-agent Systems[J]. Journal of Tsinghua University: Science and Technology, 2005, 45(7): 981. (in Chinese)

[7] 祖丽楠, 田彦涛, 梅昊. 基于分层强化学习的多移动机器人避障算法[J]. 吉林大学学报: 工学版, 2006, 36(增 2): 108.
ZU Li-nan, TIAN Yan-tao, MEI Hao. Obstacle Avoidance of Multi-mobile Robots Based on Hierarchical Reinforcement Learning[J]. Journal of Jilin University: Engineering and Technology Edition, 2006, 36(S2): 108. (in Chinese)

[8] Sutton R, Barto A. Reinforcement Learning: an Introduction[M]. Cambridge: MIT Press, 1998.

Fuzzy Logic Navigation of Mobile Robot with On-line Self-learning

HE Dong-feng[1,2], *SUN Shu-dong*[2]

(1. School of Mechatronic Engineering, Xi'an Technological University, Xi'an 710032, China;
2. School of Mechatronic, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: To overcome defect of existing fuzzy logic navigation of mobile robot, which lacks adaptability in unknown or uncertainty circumstance, a new approach of mobile robot fuzzy logic navigation with on-line self-learning was established. Firstly, a initial fuzzy logic navigation system has been constructed by designing fuzzy rules and transcendent value of action. Then, in process of system running, on-line increment learning of action value function in fuzzy rules has become true by Q-reinforcement learning to make fuzzy decision-making more precise. Simulation experiments have proved that the robot using this method can adjust his navigation strategy ceaselessly in running process to accommodate unknown or uncertainty circumstance. At the same time, shortcoming of reinforcement learning with low-learning-rate and low-convergence-rate due to blindness search in start moment also is handled and reliable navigation of robot has been achieved.

Key words: robot navigation; fuzzy logic; on-line self-learning; Q reinforcement learning

(责任编辑、校对　魏明明)