DOI:10.15938/j.emc.2010.06.016

第14卷 第6期 电 机 与 控 制 学 报 Vol. 14 No. 6
2010年6月 ELECTRIC MACHINES AND CONTROL June 2010

移动机器人模糊 Q - 学习沿墙导航

徐明亮[1,2]， 柴志雷[2]， 须文波[2]

（1. 无锡城市职业技术学校 电子信息系，江苏 无锡 214063；2. 江南大学 信息工程学院，江苏 无锡 214122）

摘 要：针对在基于行为的移动机器人沿墙导航控制器的设计中缺乏足够的先验知识的问题，采用 Q - 学习方法让机器人通过学习来自动构建导航控制器。将模糊神经网络和 Q - 学习相结合，用模糊神经网络直接逼近连续状态和动作空间中的 Q 值函数。利用对 Q 值函数的优化获得控制输出。模糊神经网络中的节点根据状态动作对的各个分量和时间差分的新颖性进行自适应地添加和构造，这样不仅能克服节点选择的困难还能使网络保持适度的规模。网络中的参数采用扩展卡尔曼滤波方法进行自适应调整。基于 Khepera 2 机器人的沿墙导航实验验证了该方法的有效性和优越性。

关键词：Q - 学习；模糊神经网络；沿墙导航；移动机器人

中图分类号：TP 391.41 **文献标志码**：A **文章编号**：1007-449X(2010)06-0083-06

Wall-following control of a mobile robot with fuzzy Q-learning

XU Ming-liang[1,2]， CHAI Zhi-lei[2]， XU Wen-bo[2]

(1. Department of Electronic Information Engineering, Wuxi City College of Vocational Techonology, Wuxi 214063, China;
2. Institute of Information Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: The Q-learning was introduced into navigation control of the wall-following task of mobile robots where there was no enough priori knowledge available. The Q-value function was approached directly using Fuzzy Neural Network (FNN). The optimization method was used to search the greedy action with maximum Q-value. The nodes of FNN were created incrementally and adaptively according to every element of the current pair of state-action and Temporal Difference (TD), which overcame the difficulties of the choice of nodes and ensured an economic size of the network. Moreover the parameters of the FNN were updated using Extended Kalman Filter (EKF). The results of the wall-following task of Khepera 2 mobile robot demonstrate the superiority and validity of the proposed method.

Key words: Q-learning; fuzzy neural network; wall-following navigation; mobile robots

0 引 言

导航是移动机器人的一项重要的功能，是移动机器人完成其他智能行为的基础。沿墙导航控制是指机器人在和墙保持一定距离的情况下沿墙运动，从更一般意义上来说实际上是机器人与物体保持一定距离并沿物体轮廓运动[1]。因此沿墙导航实际上既可以使机器人实现障碍物的避碰[2]，也可以实现在未知环境的导航[3]。

移动机器人的反应式导航是一种直接在机器人

收稿日期：2009-11-24
基金项目：国家自然科学基金(60703106)
作者简介：徐明亮(1973—)，男，博士，讲师，研究方向为机器学习、智能控制；
柴志雷(1975—)，男，博士，副教授，研究方向为嵌入式系统、智能控制；
须文波(1946—)，男，教授，博士生导师，研究方向为嵌入式系统、计算机控制技术。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的感知和行为之间建立映射关系的导航方法。它具有灵活和执行快速的特点而成为移动机器人在未知和快速变化环境中导航的重要方法。已有许多学者提出了不同的反应式导航方法,比如文献[4]采用引力势场法进行导航。文献[5]采用基于模糊规则的反应式导航控制器。这些方法通常基于具体的环境模型,需要较多的先验知识,同时对变化的环境不具有自适应能力。

强化学习能够在没有先验知识的情况下通过与环境的交互获得由状态到动作的策略,因此基于强化学习的机器人导航受到众多研究者的广泛关注。文献[6]中采用 Q-学习来对模糊规则进行调整,但模糊规则则是根据机器人的系统特性手工建立。文献[7]也采用类似技术,其特点是用 RBF 网络逼近选定的若干个离散动作的 Q-值,网络权值利用 Q-学习来调整。而 RBF 网络隐层节点的中心和宽度却要由样本来确定。文献[8]采用 CMAC 神经网络实现了 Q-值函数的逼近,该方法涉及到输入参数的离散化,离散化的粒度也将影响系统的性能。文献[9]利用 FNN 来逼近 Q - 值函数和策略函数,而这些函数都是建立在若干个选定的离散动作的基础之上,使得系统过于复杂。文献[10]是用模糊推理系统来逼近 Q-值函数,每一条规则对应一个由若干个选定的离散动作所构成的向量,每一个规则的输出动作由规则内部的离散动作通过竞争的方法产生,控制器的输出动作由各个规则的输出动作根据当前状态在各个规则所导出的状态值进行加权。在这些方法中,导航控制器输出取决于预先选定的离散动作。这些离散动作的选择影响导航控制器的性能,而如何选择这些种子动作也没有任何先验知识可用。另外这些方法中的 Q-学习从本质上来说是基于 actor-critic 方法的[11]。

为避免种子动作的选择,我们用模糊神经网络直接对强化学习中的 Q-值函数进行逼近,即网络的输入为状态动作对,而非相关文献中的状态,利用函数优化技术产生控制器输出动作。同时在学习过程中引入网络节点自适应构建和参数自适应调整方法,减少人工干预。

1 基于模糊神经网络的 Q-学习

Q-学习的主要目标是通过与环境的交互获得表征策略的状态动作对的 Q-值函数。Q-学习中状态动作对的 Q 值按照下式进行更新:

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\,\delta_{\mathrm{TD}}。\tag{1}$$

其中:s_t 为当前状态;a_t 为当前状态下选择执行的动作;α 为学习率;δ_{TD} 为时间差分(temporal difference, TD)。一步时间差分 δ_{TD} 计算式为

$$\delta_{\mathrm{TD}}=r_{t+1}+\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t)。\tag{2}$$

其中:γ 为折扣因子,r_{t+1} 为学习 agent 在状态 s_t 执行动作 a_t 后转移到状态 s_{t+1} 时所获得的立即奖赏。经典的 Q-学习是以查找表来描述离散空间状态动作对的值函数。对于连续空间下的 Q 学习,直接的方法是将连续空间进行离散化处理。而对于离散的粒度选择,往往没有任何先验知识可用。离散粒度过大将会导致系统性能下降,甚至学习不成功;过小也会使学习速度下降。为克服离散化所产生的弊端,研究者普遍采用具有泛化功能的神经网络或模糊推理系统来逼近 Q 值函数。

模糊神经网络是模糊推理系统和神经网络相结合的产物,它既拥有模糊推理系统便于知识的表达和便于在系统中嵌入已有知识的优点,也拥有神经网络的自学习自组织的特点,因此在函数逼近中得到广泛应用。因此我们采用模糊神经网络来逼近 Q 值函数。

1.1 网络结构

用于对 Q 值函数进行直接逼近的模糊神经网络结构如图 1 所示。第一层为输入层,它将由状态 s 和动作 a 构成的向量 $x=(s_1,\cdots,s_m,a)^{\mathrm{T}}$ 直接传送到下一层。状态空间 s 为 m 维,记向量 x 维数为 n,则 $n=m+1$。第二层为模糊化层,其中每个节点代表一个语言变量。该层的作用是计算各个分量在不同语言变量中的隶属度。各个语言变量的隶属度函数采用高斯函数。输入向量第 i 个分量的第 j 个语言变量的隶属度函数为

$$MF_{ij}(x_i)=\exp\left(\frac{(x_i-\mu_{ij})^2}{\sigma_{ij}^2}\right),i=1,\cdots,n\ \text{and}\ j=1,\cdots,J。\tag{3}$$

其中:μ_{ij} 和 σ_{ij} 分别为该隶属度函数的中心和宽度;J 为该分量的语言变量的个数。

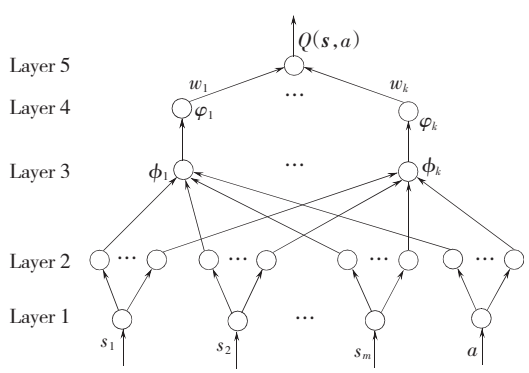

图 1 网络结构

Fig. 1 The architecture of network

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第三层为 T-norm 运算层，该层计算每个规则的发射强度。第 k 条规则的发射强度为

$$\phi_k(x)=\exp\left(\sum_{i=1}^{n}\frac{(x_i-\mu_{ki})^2}{\sigma_{ki}^2}\right),\quad i=1,2,\cdots,p \text{ and } k=1,2,\cdots,K。\tag{4}$$

第四层为归一化层，对每一条规则的发射强度进行归一化处理。第 k 条规则的归一化后的发射强度为

$$\varphi_i=\frac{\phi_i}{\sum_{k=1}^{K}\varphi_k}。\tag{5}$$

第五层为解模糊层即输出层，采用重心法进行解模糊，输出为输入状态动作对的 Q 值，即

$$Q(s,a)=\sum_{k=1}^{K}w_k\phi_k。\tag{6}$$

w_k 是第 k 条规则的后件。

1.2　结构与参数学习

模糊规则可以根据样本的 ε 完整性[2]来建立，但 ε 完整性不能充分体现每个分量对系统性能有不同影响。在文献[3]中，RBF 网络隐层节点根据当前样本与隐层节点中心最小距离和误差进行自适应添加。在这些方法中没有考虑系统性能对输入向量中的不同分量有不同的敏感程度。如果对这些分量不加区别，为保证系统性能，就要照顾到对系统性能敏感的分量，因此必须采用较细的分辨率，这样将导致规则数或隐层节点数目过多。对系统性能敏感的分量应采用较细的分辨率，而对系统性能不敏感的分量应采用较粗的分辨率。这样可以减少节点数目，减小系统计算量，加快学习速度。因此考虑不同分量具有不同分辨率的模糊规则构造的条件为

$$\begin{cases}|\delta_{\mathrm{TD}}|>e_{\mathrm{T}},\\ \exists i,d_i>\rho_i,i\in(1,\cdots,n)。\end{cases}\tag{7}$$

式中：ρ_i 为第 i 个分量的分辨率；e_{T} 为 TD 误差阈值；d_i 为输入状态动作对的第 i 个分量和该分量全部语言变量的隶属度函数中心的最小距离，即

$$d_i=\min|x_{ti}-\mu_{ij}|,i\in\{1,\cdots,n\},j\in\{1,\cdots,J\}。\tag{8}$$

其中：x_{ti} 为在时间 t 的状态动作对的第 i 个分量。

如果当前状态动作对满足式(7)的条件，系统就构造一个新的模糊规则。为描述方便，记

$$g=\min_j|x_{ti}-\mu_{ij}|,j=1,\cdots,J。\tag{9}$$

新增规则的第 i 个分量语言变量的隶属度函数的中心 $\mu_{i\mathrm{new}}$ 为

$$\left.\begin{aligned}&\mu_{i\mathrm{new}}=\mu_{ig},\text{if } d_i<\rho_i\\ &\mu_{i\mathrm{new}}=sa_{ti},\text{if } d_i\geqslant\rho_i\end{aligned}\right\},j=1,\cdots,J。\tag{10}$$

该规则第 i 个分量的隶属度函数的宽度为

$$\left.\begin{aligned}&\sigma_{\mathrm{new}i}=\sigma_{ig},\text{if } \min|x_{ti}-\mu_{ki}|<\eta_i,\\ &\sigma_{\mathrm{new}i}=\tau|x_{ti}-\mu_{ig}|,\text{ if } \min|x_{ti}-\mu_{ki}|\geqslant\eta_i。\end{aligned}\right\}\tag{11}$$

式中参数 τ 为重叠系数。η_i 为第 i 个分量的分辨率。新规则的后件 w_{new} 为

$$w_{\mathrm{new}}=\delta_{\mathrm{TD}}=r_{t+1}+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1})-Q(s_t,a_t)。\tag{12}$$

在学习的初始阶段，系统没有任何规则，可以以第一个状态动作对作为隶属度函数中心建立第一个规则，规则的宽度设为 $[\tau\rho_1,\cdots,\tau\rho_n]$。对于后续的状态动作对则根据式(7)来判断是否创建新的规则。

当无需增加新规则时，即状态动作对不满足规则构建的条件时，可以采用梯度下降法对规则的前件和后件中的相关参数进行调整。但由于梯度下降法容易陷入局部最优，因此采用卡尔曼滤波法对相关参数进行调整[4]。

令规则中全部可调参数构成的向量为 $\boldsymbol{v}=[w_1,\mu_1^{\mathrm{T}},\sigma_1^{\mathrm{T}},\cdots,w_K,\mu_K^{\mathrm{T}},\sigma_K^{\mathrm{T}}]^{\mathrm{T}}$，根据卡尔曼滤波方法其更新式为

$$\boldsymbol{v}_h=\boldsymbol{v}_{h-1}+\delta_{\mathrm{TD}_h}\boldsymbol{K}_h。\tag{13}$$

其中 $\boldsymbol{K}_h$ 为卡尔曼增益，其计算式为

$$\boldsymbol{K}_h=[\boldsymbol{R}_h+\boldsymbol{\Gamma}_h^{\mathrm{T}}\boldsymbol{P}_{h-1}\boldsymbol{\Gamma}_h]^{-1}\boldsymbol{P}_{h-1}\boldsymbol{\Gamma}_h。\tag{14}$$

其中 $\boldsymbol{\Gamma}_h$ 为各个参数的梯度向量，由式(3)，(4)，(5)，(6)可以推导出

$$\begin{aligned}\boldsymbol{\Gamma}_h=\Big[&\varphi_1,w_1\varphi_1(1-\varphi_1)\frac{x_1-\mu_{11}}{\sigma_{11}^2},\cdots,w_1\varphi_1(1-\varphi_1)\frac{x_n-\mu_{n1}}{\sigma_{n1}^2},\\ &w_1\varphi_1(1-\varphi_1)\frac{(x_1-\mu_{11})^2}{\sigma_{11}^3},\cdots,w_1\varphi_1(1-\varphi_1)\frac{(x_n-\mu_{n1})^2}{\sigma_{n1}^3},\\ &\cdots,\\ &\varphi_K,w_K\varphi_K(1-\varphi_K)\frac{x_1-\mu_{11}}{\sigma_{1K}^2},\cdots,w_K\varphi_K(1-\varphi_K)\frac{x_n-\mu_{nK}}{\sigma_{nK}^2},\\ &w_K\varphi_K(1-\varphi_K)\frac{(x_1-\mu_{11})^2}{\sigma_{1K}^3},\cdots,w_K\varphi_K(1-\varphi_K)\frac{(x_n-\mu_{nK})^2}{\sigma_{nK}^3}\Big]^{\mathrm{T}}\end{aligned}\tag{15}$$

其中 $\boldsymbol{R}_h$ 为测量噪声方差，$\boldsymbol{P}_h$ 为协方差矩阵，其更新式为

$$\boldsymbol{P}_h=[\boldsymbol{I}-K_h\boldsymbol{\Gamma}_h^{\mathrm{T}}]\boldsymbol{P}_{h-1}+U\boldsymbol{I}。\tag{16}$$

其中 U 是一个标量，表示在梯度方向上的随机步长，$\boldsymbol{I}$ 为单位矩阵。如果参数 $\boldsymbol{v}$ 的长度为 N，$\boldsymbol{P}_h$ 是一个 $N\times N$ 正定对称矩阵，当新增一个规则时，可调参数也将增加，矩阵 $\boldsymbol{P}_h$ 的维数按下式增加，即

$$\boldsymbol{P}_h=\begin{pmatrix}\boldsymbol{P}_{h-1}&0\\0&\boldsymbol{P}_0\boldsymbol{I}\end{pmatrix}。\tag{17}$$

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

其中 P_0 为新增规则的初始化参数。

1.3 动作输出与搜索和利用平衡

为获得最大的累积报酬，学习主体在任何状态下总是选择具有最大 Q 值的动作，即贪婪动作：

$$a_{\text{greedy}} = \arg\max_b Q(s,b), s_i = s_{ti}, (i = 1,2,\cdots,n-1)。 \tag{18}$$

其中 b 为可选动作。该贪婪动作的求解实际上是一个优化问题，即

$$\max(Q(s,a)), a \in A, s = s_t。 \tag{19}$$

式中 A 为动作定义域。将当前状态 s_t 代入上式有

$$\max(Q(s,a)) = \max\left(\frac{\sum_{k=1}^{K} w_k \phi_k}{\sum_{p=1}^{K} \phi_k}\right) =$$

$$\max\left(\frac{\sum_{k=1}^{K} w_k \exp\left(\sum_{i=1}^{n} \frac{(x_i - \mu_{ki})^2}{\sigma_{ki}^2}\right)}{\sum_{k=1}^{K} \exp\left(\sum_{i=1}^{n} \frac{(x_i - \mu_{ki})^2}{\sigma_{ki}^2}\right)}\right) =$$

$$\max\left(\frac{\sum_{k=1}^{K} w_k \exp\left(\sum_{i=1}^{m} \frac{(s_i - \mu_{ki})^2}{\sigma_{ki}^2}\right)\exp\left(\frac{(a - \mu_{ka})^2}{\sigma_{kn}^2}\right)}{\sum_{k=1}^{K} \exp\left(\sum_{i=1}^{m} \frac{(s_i - \mu_{ki})^2}{\sigma_{ki}^2}\right)\exp\left(\frac{(a - \mu_{kn})^2}{\sigma_{kn}^2}\right)}\right)。$$

令 $c_k = \exp\left(\sum_{i=1}^{m-1} \frac{(s_i - \mu_{ki})^2}{\sigma_{ki}^2}\right)$，则有

$$\max\left(\frac{\sum_{k=1}^{K} w_k c_k \exp\left(\frac{(a - \mu_{km})^2}{\sigma_{km}^2}\right)}{\sum_{k=1}^{K} c_k \exp\left(\frac{(a - \mu_{km})^2}{\sigma_{km}^2}\right)}\right), a \in A。 \tag{20}$$

上述优化问题可以用 PSO 或 GA 算法或最简单的格点法进行求解。

为平衡强化学习中特有的搜索和利用两难问题，贪婪动作并不直接作用于环境，而是在其上叠加一个干扰动作 a_d [5]，即执行动作 a_{exe} 为

$$a_{\text{exe}} = a_{\text{greedy}} + a_d。 \tag{21}$$

而干扰动作 a_d 服从如下的正态分布：

$$a_d \sim n(0, \sigma_Q(t))。 \tag{22}$$

式中

$$\sigma_Q(t) = \frac{\lambda}{(1 + 2\exp(\max(Q(s_t,a))))}。 \tag{23}$$

λ 为比列因子。

2 沿墙导航

和文献［0］一样，用 Khepera 2 [6] 机器人仿真器进行实验，该机器人的结构如图 2 所示，该机器人有 6 对红外传感发射、接收器，用于测量相应方向上的障碍物与机器人之间的距离。

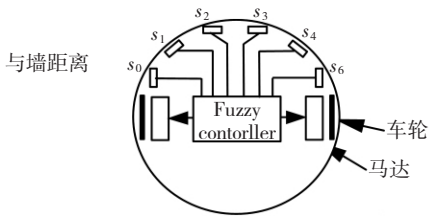

图 2 Khepera Ⅱ 机器人传感器及其布置 [0]

Fig. 2 The displace of the sensor of the robot Khepera Ⅱ

实验目的是希望机器人在没有任何先验知识的情况下通过强化学习获得一个能控制机器人按顺时针方向沿墙绕行的控制器，因此只需考虑 s_0，s_1，s_2，s_3 这四个传感器的输出作为机器人状态。机器人的动作为转角。将距离进行归一化处理后，机器人离墙的最小距离为 0.15 m，最大距离为 0.85 m。以传感器 s_0 的测量值为机器人与墙的距离。

仿真实验环境如图 3 所示。在这个环境中，有三个 90°拐角，一个 270°拐角，一个 116.57°钝角拐角，一个 63.43°锐角拐角，可见该环境比文献［0］要复杂。实验中 4 个传感器的最大分辨率为 1 m，最小分辨率为 0.5 m。输出转角在 −30°和 30°之间，其最大分辨率为 60°，最小分辨率为 30°。所有分辨率衰减因子均为 0.82。TD 误差阈值 e_T 为 0.01，TD 误差计算式(2)中的折扣因子 $\gamma = 0.48$，Q 值更新式(1)中学习因子 $\alpha = 1$。重叠因子 $\tau = 0.82$。卡尔曼滤波器参数 $P_0 = R_0 = 1.0$，$Q_0 = 0.000\,5$。式(23)中 $\lambda = 0.4$。机器人速度为 0.2 m/s。机器人决策时间间隔为 200 ms。为简单起见，贪婪动作及最大 Q 值的计算采用格点法，即按 $a_i = -30° + i + f$ $(i = 0°, 1°\cdots, 60°, f$ 是一个在区间 $[-0.5, 0.5]$ 上服从均匀分布的随机数。如果 a_i 大于 30°，取 a_i 为 30°，如果 a_i 小于 −30°，取 a_i 为 −30°)。比较 60 个输出动作的 Q 值，取 Q 值最大的动作为贪婪动作。

强化信号定义为

$$r = \begin{cases} -1, & s_0 < d_{\min} \text{ or } s_0 > d_{\max}, \\ 0, & \text{其他}。 \end{cases} \tag{24}$$

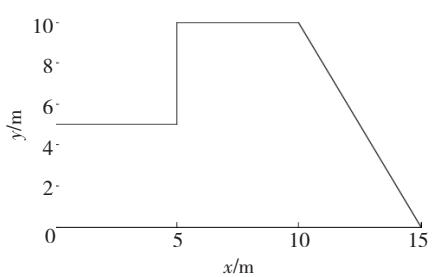

图 3 仿真环境

Fig. 3 The enviroment of simulation

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

为加快学习速度同时减小计算量，采用经验重放的方法[7]进行学习。

一次运行中最大尝试次数为 500，当机器人能绕墙成功行走 1 200 步（1 200 步能确保机器人绕墙至少一周）就视为学习成功。每次尝试时机器人的初始位姿为（0.5，0.5，90°）。

在 25 次运行中，每次机器人都能经过不到 500 次尝试就能完成绕墙至少一周的任务。平均尝试次数为 67.6，最大尝试次数为 480，最小尝试次数为 5。控制器平均规则数为 14，最大为 40，最小为 8。性能优于文献[10]的 DFQL 方法。如果直接采用文献[3]中的方法进行规则增加，即取总的最大分辨率为 1，最小分辨率为 0.5，在 500 次尝试中并不能次次成功，同时由于规则数过多，计算量大。

图 4 给出了在学习成功后绕墙行走的轨迹，该轨迹起点为（0.5，0.5）出发，绕墙一周后终点为（2.66，4.42）。在这次运行中，控制器经过 9 次尝试后就能成功地控制仿真机器人实现沿墙行走。

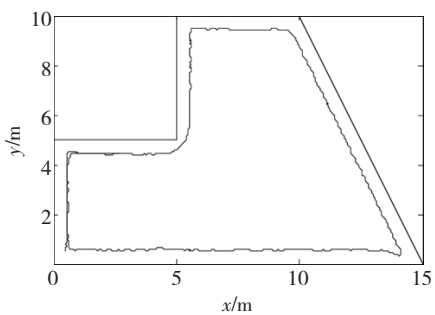

图 4　机器人绕墙行轨迹

Fig. 4　The tragectory of the robot

图 5 给出了机器人方向角的变化情况，图 6 给出了控制器在 1 200 步的沿墙运动中输出的控制量变化。图 7 所示为每个传感器在沿墙行走过程中的测量值。图 8 给出了在学习后自动建立的 9 条规则中每个输入变量隶属度函数。

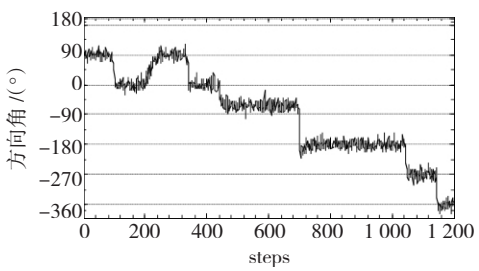

图 5　机器人方向角变化曲线

Fig. 5　The azimuth of the robot

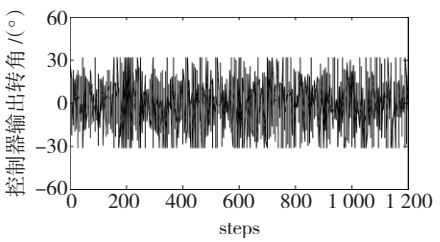

图 6　控制器输出控制量

Fig. 6　The output of the controller

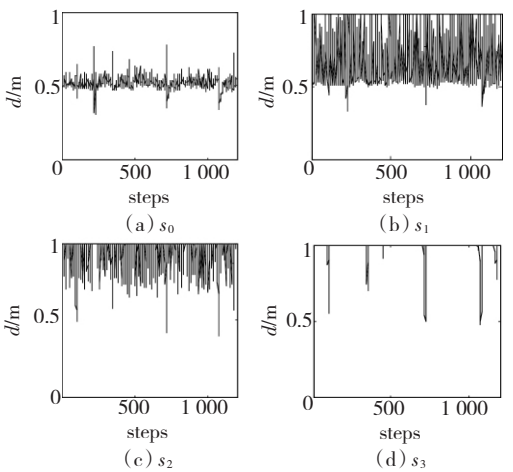

图 7　各传感器测量值变化

Fig. 7　The sensors value during moving

注：传感器测量值范围为（0，1），测量值为 1 表示已超去其测量范围。

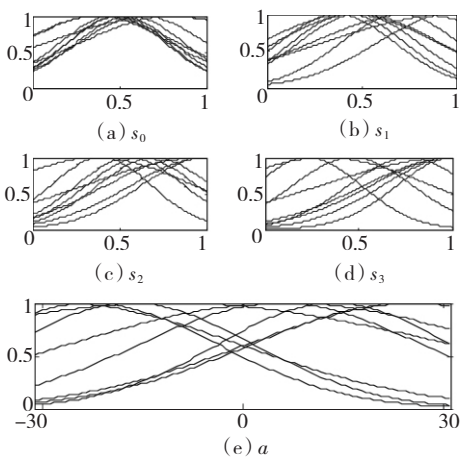

图 8　各输入变量的隶属度函数曲线

Fig. 8　Member function of ench variable

在 25 次运行过程中，一旦控制器通过学习能控制机器人过 90°拐角和 270°拐角后，就可以直接过 116.57°钝角拐角，这说明该学习系统具有较强的泛化性能。对于 63.43°锐角拐角有时还需要进行学习。同时控制器在学习通过拐角的过程中，对沿直线墙行走控制一般没有影响，这源于 FRBF 网络具有局部逼近的特性。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

从图5的机器人方位角变化情况来看，机器人在直线段行走过程中，方向角总的来说是以墙的方向角为基准波动，这种波动一是由于强化学习中已有知识的应用和新知识的探索之间的平衡所致。为能搜索到更好的策略，学习主体每次执行的动作并非是最优动作(已有知识的利用)，通过选择贪婪动作之外的动作来发现更好的动作(搜索)，但这将会导致系统性能下降。二是由于传感器 s_0 的测量值并非是机器人和墙之间的距离，传感器的测量值对机器人的姿态非常敏感，姿态的细微变化会导致测量值的巨大波动，为克服这种测量波动，机器人基本上是以一种之字形的路线沿墙行走，以保持其不出"轨道"。这一点从图6控制器输出控制量变化情况也可以看出。尽管传感器 s_0 的测量值不是机器人与墙的真正距离，用本文所提的方法仍然能获得一个可以控制该机器人实现沿墙行走的控制器。

图7和图8分别给出机器人在1 200步的运行过程中四个传感器输出的测量数值和状态动作对的五个分量的八个语言变量的隶属度函数。

在前面所述文献中，如文献[0]，控制器的输出动作总的来说是基于预先选定的若干个离散的加权构成，假设离散种子动作集合 $B=\{a_1,a_2,\cdots,a_m\}$，规则数为 n，则控制器输出动作为

$$a=\sum_{i=1}^{n}\beta_i b_i,\text{其中}\sum_{i=1}^{n}\beta_i=1,b_i\in B。\quad(25)$$

其中 b_i 为第 i 条规则输出动作，β_i 为第 i 条规则后件。将上式进行如下处理并按 a_i 进行同类项合并，即有：

$$\sum_{i=1}^{n}\beta_i b_i=\beta_1 b_1+,\cdots,+\beta_n b_n+0\times a_1+,\cdots,0\times a_m=\sum_{j=1}^{m}\xi_j a_j=\langle\xi,A\rangle。\text{其中}A=[a_1,a_2,\cdots,a_n],$$

既 A 是由种子动作所构成的向量，而向量 $\xi=[\xi_1,\xi_2,\cdots,\xi_n]$，且 $\sum_{j=1}^{m}\xi_j=1$。(26)

由上式可见，由离散种子动作加权获得输出动作本质上是由种子动作构成的向量和由规则后件构成的向量的内积，后者需要通过学习来确定，其参数空间的维数为 $m-1$。为获得连续的动作，离散动作数 m 一般都取得较大，如在文献[0]中，m 为13，这样参数空间维数比原来动作空间维数要大，因此学习速度慢，且同一规则内的离散动作之间没有泛化能力。

3 结 论

本文提出了一种基于模糊 Q-学习导航控制方法，与已有的连续空间中的 Q 学习方法相比，其特点之一是以模糊神经网络对 Q 值函数进行直接逼近，通过函数优化的方法获得输出动作，这就克服了相关文献中种子动作的选择问题，实现了强化学习和函数优化技术的结合。特点之二是提出了一种新的模糊规则自适应构造方法，实现了在学习过程中模糊规则的自动构建和相关参数的自适应调整，避免了模糊规则手工建立的困难。此外学习经验不但能在邻近的状态中可以进行泛化，还可以在相邻的动作之间进行泛化，因此学习速度也得到提高。实验结果表明该方法可以成功实现机器人沿墙行走的导航控制。

今后还将该方法应用到其他控制问题，此外本文所提的模糊规则自适应方法还可以应用到监督学习等方面，这都是今后的研究内容。

参 考 文 献：

[1] TURENNOUT P, HONDERD G, SCHELVEN L J. Wall-following control of a mobile robot [C]//*Proceedings of the* 1992 *IEEE International Conference on Robotics and Automation*, May 12 - 14, 1992, Nice, France. 1992: 280 - 285.

[2] BORENSTEIN J, KERON Y. The vector filed histogram fast obstacle avoidance for mobile robots [J]. *IEEE Transactions on Robotics and Automation*, 1991, 7 (3):278 - 288.

[3] 彭一准，原魁，刘俊承，等. 室内移动机器人的三层规划导航策略 [J]. 电机与控制学报，2006，10(4)：380 - 384.
PENT Yizhun, YUAN Kui, LIU Juncheng, *et al*. A three-layer planning navigation method for indoor mobile robot [J]. *Electric Machines and Control*, 2006, 10(4): 380 - 384.

[4] KHATIB O. Real-time obstacle avoidance for manipulator and mobile robots [J]. *International Journal of Robotic Research*, 1986, 5 (1):90 - 98.

[5] Lee P S, Wang L L. Collision avoidance by fuzzy logic control for automated guided vehicle navigation [J]. *Journal of Robotic Systems*, 1994, 11(8):743 - 760.

[6] ZHANG Wenzhi, LU Tiansheng. Reactive fuzzy controller design by *Q*-learning for mobile robot navigation [J]. *Journal of Harbin Institute of Technology*, 2005, 12(3):319 - 324.

[7] 吴洪岩，刘淑华，张嵩. 基于RBFNN的强化学习在机器人导航中的应用 [J]. 吉林大学学报:信息科学版，2009，27(2)：185 - 190.
WU Hongyan, LIU Shuhua, ZHANGYu. Application of reinforcement learning based on radial basis function neural networks in robot navigation [J]. *Journal of Jilin University*: *Information Science Edition*, 2009, 7(2): 185 - 190.

[8] 陆军，徐莉，周小平. 强化学习方法在移动机器人导航中的应用 [J]. 哈尔滨工程大学学报，2004，25(2)：176 - 179.
LU Jun, XU Li, ZHOU Xiaoping. Research on reinforcement learning and its application to mobile robot [J]. *Journal of Harbin Engineering University*, 2004, 25(2): 176 - 179.

(下转第97页)

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

more independent uncertain time-delay subsystems

参 考 文 献：

[1] WANG S H, DAVISON E J. On the stabilization of decentralized fixed modes for interconnected systems [J]. *Automatica*, 1983, 19(2): 473 – 478.

[2] YAND G H, WANG J L. SOH Y C. Decentralized control of symmetric systems [J]. *Systems & Control Letters*, 2001, 42(2): 145 – 149.

[3] DUAN Z S, WANG J Z, HUANG L. Special decentralized control problems and effectiveness of parameter-dependent Lyapunov function method [C]//*Proceedings of American Control Conference*. Portland, USA. 2005: 1697 – 1702.

[4] DUAN Z S, WANG J Z, HUANG L. Special decentralized control problems in discrete-time interconnected systems composed of two subsystems [C]//*Proceedings of the 25th Chinese Control Conference*, August 7 – 11, 2006, Harbin, China. 2006: 1080 – 1085.

[5] DUAN Z S, WANG J Z, HUANG L. Special decentralized control problems in discrete-time interconnected systems composed of two subsystems [J]. *Systems and Control Letters*, 2007, 56(3): 206 – 214.

[6] DUAN Z S, WANG J Z, CHEN G R, *et al*. Stability analysis and decentralized control of a class of complex dynamical networks [J]. *Automatica*, 2008, 44(4): 1028 – 1035.

[7] DUAN Z S, HUANG L, WANG J Z, *et al*. Harmonic control between two systems [J]. *Acta Automatica Sinica*, 2003, 29(1): 14 – 22.

[8] YANG Y, DUAN Z S, HUANG L. Design of nonlinear interconnections guaranteeing the absence of periodic solutions [J]. *Systems and Control Letters*, 2006, 55(4): 338 – 346.

[9] NIAN X H, CAO L. BMI approach to the interconnected stability and cooperative control of linear systems [J]. *Acta Automatica Sinica*, 2008, 34(4): 438 – 444.

[10] 王广雄，李连锋，王新生. 鲁棒设计中参数不确定性的描述 [J]. 电机与控制学报，2001, 5(1): 5 – 7.
WANG G X, LI L F, WANG X S. The description of the parameter uncertainty for robust design [J]. *Electric Machines and Control*, 2001, 5(1): 5 – 7.

[11] 王常虹，奚伯齐，李清华，等. 网络化控制系统鲁棒 $L_2 - L_\infty$ 控制器设计 [J]. 电机与控制学报，2010, 14(2): 25 – 30.
WANG C H, XI B Q, LI Q H, *et al*. Robust $L_2 - L_\infty$ controller design for networked control systems [J]. *Electric Machines and Control*, 2010, 14(2): 25 – 30.

（编辑：于智龙）

（上接第 88 页）

[9] 段勇，徐心和. 基于模糊神经网络的强化学习及其在机器人导航中的应用 [J]. 控制与决策，2007, 22(5): 525 – 534.
DUAN Yong, XU Xinhe. Reinforcement learning based on FNN and its application in robot navigation [J]. *Control and Decision*, 2007, 22(5): 525 – 534.

[10] ER M J, DENG C. Online tuning of fuzzy inference systems using dynamic fuzzy *Q*-learning [J]. *IEEE Transactions on Systems, Man and Cybernetics Part B: Cybernetics*, 2004, 34(3): 1478 – 1489.

[11] BARTO A G, SUTTON R S, ANDERSON C W. Neuron like adaptive elements that can solve difficult learning control problems [J]. *IEEE Transactions on Systems, Man, and Cybernetics*, 1983, 13(5): 834 – 846.

[12] LEE C C. Fuzzy logic in control systems: fuzzy logic controller-Part I and II [J]. *IEEE Transactions on Systems, Man and Cybernetics*, 1990, 20(2): 404 – 435.

[13] PLATT J. A resource allocating network for function interpolation [J]. *Neural Computation*, 1991, 3(2): 213 – 225.

[14] SINGHAL S, WU L. Training multilayer perceptrons with the extended Kalman algorithm [C]//*Advances in Neural Processing Systems*, December 12 – 14, 1988, San Mateo, CA. 1988: 133 – 140.

[15] KONDO T, ITO K. A reinforcement learning with evolutionary state recruitment strategy for autonomous mobile robots control [J]. *Robotics and Autonomous Systems*, 2004, 46(2): 111 – 124.

[16] KTEAM S A. Khepera 2 User Manual [R]. Switzerland, 2002.

[17] LIN L J. Self-improving reactive agents based on reinforcement learning, planning and teaching [J]. *Maching Learning*, 1992, 8(3 – 4): 293 – 321.

（编辑：刘素菊）

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net