高技术通讯 2013 年 第23 卷 第4 期:407 ~ 412

doi:10.3772/j.issn.1002-0470.2013.04.012

基于模糊小波网络的强化学习及其在多机器人决策策略中的应用①

段 勇②* 李 程* 徐心和**

(*沈阳工业大学信息科学与工程学院 沈阳 110870)

(**东北大学信息科学与工程学院 沈阳 110004)

摘 要 给出了一种基于模糊小波神经网络(FWNN)的强化学习方法,并研究了应用该方法解决多机器人足球比赛中的决策策略问题。首先,使用 FWNN 来实现强化学习状态空间到动作空间的映射,从而解决大规格或连续状态空间所导致的学习速度过慢甚至难以收敛等问题。然后,研究了提出的方法在机器人足球比赛的复杂决策策略学习中的应用,证明机器人球员能够通过学习掌握根据比赛状态信息选择合理动作的能力。最后,通过实验验证了该学习方法的有效性,它能够满足机器人足球比赛的需要。

关键词 强化学习(RL),模糊小波神经网络(FWNN),机器人足球比赛,动作选择,决策

0 引 言

如何组织和控制多个机器人协作完成一个复杂任务,比如完成一场机器人足球比赛,已成为机器人学研究的新课题。机器人足球比赛也像人类足球比赛一样,要求机器人球员动作准确、敏捷,战术灵活多变,所以对整个机器人足球队的控制决策至关重要,它直接影响比赛的成败。通常决策系统是根据专家知识和经验来设计的,然而由于机器人足球比赛环境的复杂性和不确定性,因而很难考虑到比赛所有可能出现的情况,这往往使得主观设计的决策系统缺乏完备性和灵活性。而强化学习(reinforcement learning, RL)方法(即智能系统从环境到行为的映射的学习方法)不需要精确的环境模型和完备的专家知识,能够使机器人在同环境的交互过程中学习决策能力和行为能力,因此它为机器人足球的研究提供了一个新的途径,目前也得到了较为广泛的关注和研究[1-3]。

在应用强化学习算法时,往往面临着状态空间或动作空间过大,导致算法执行速度过慢甚至难以收敛等问题(称为"维数灾难"问题)。这是因为为了保证学习精度不得不把连续的状态变量转换为大量的离散值,并由此构成表格型强化学习系统。而强化学习收敛的基本条件是无限多次地访问每个状态-动作对来进行学习,所以过多的离散值将导致学习表的规模过于庞大,导致学习时间过长,甚至不能收敛。解决这类问题的一种有效方法是利用函数逼近算法来实现强化学习状态空间到动作空间的映射[4]。考虑到神经网络的输入为连续变量,而且它将学习表格转化为神经元结构,将表格型的强化学习过程转化为神经网络的训练,而模糊小波神经网络(fuzzy wavelet neural network, FWNN)[5-8]是模糊推理系统和小波神经网络[9-11]的一种结合形式,具有两者的优点,而且它在模糊神经网络的基础上引入小波尺度函数作为模糊隶属度函数,因而具有强大的空间映射能力,也提高了学习的收敛速度[8],本文提出了一种基于 FWNN 的强化学习算法,该算法利用 FWNN 的函数逼近特性实现强化学习状态空间到动作空间的映射,从而解决了强化学习的维数灾难问题。此外将该方法应用于足球机器人系统决策策略的学习,可以实现机器人足球比赛中球员动作选择问题。

1 基于 FWNN 的强化学习

1.1 基本强化学习算法

强化学习是指 agent 从环境状态到动作映射的

① 国家青年科学基金(60905054)和辽宁省高等学校杰出青年学者成长计划(LJQ2011006)资助项目。

② 男,1978 年生,博士,副教授;研究方向:智能机器人,机器学习等;联系人,E-mail: duanyong0607@126.com

(收稿日期:2012-05-30)

学习,目的是从环境中获得的累积强化信号(回报)最大[4]。Agent 状态集合用 $S = \{s_i \mid s_i \in S\}$ 表示,它所执行的动作集合可以描述为 $A = \{a_i \mid a_i \in A\}$。当 agent 根据当前状态 s_t 选择并执行动作 a_t 时,agent状态转移到 s_{t+1},并从外部环境获得强化信号 r_t。

Q 学习[4,11]是一种常用的强化学习算法,其基本思想是不去估计环境模型,而是直接优化一个可以递推计算的 Q 函数,使用该函数 $Q(s,a)$ 来表达与状态相对应的各个动作的评估。

Q 学习算法的迭代计算公式为[4,12]

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \eta_t \cdot [r_t + \gamma \cdot \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t,a_t)] \quad (1)$$

其中,$< s_t, a_t >$ 为 t 时刻的状态 - 动作对,η_t 表示学习率,γ 为折扣因子。

1.2 基于 FWNN 的强化学习系统结构

当强化学习状态空间规模较大或连续时,基本的 Q 学习方法面临维数灾难问题。因此可以通过函数逼近的方法来实现 Q 学习,利用函数逼近方法的非线性映射能力来实现强化学习的状态空间到动作空间的映射。强化学习的状态和动作分别作为函数逼近器的输入变量和输出变量,因此 Q 学习算法可以转化为函数逼近器的内部结构来实现。

FWNN 除了具有神经网络的特点外,还利于引入专家知识,适合于表达模糊或定性的知识,同时网络具有良好的尺度变换和伸缩平移特性,可以实现任意的非线性映射,并具有更高的函数逼近精度和适应性。

基于模糊小波神经网络的强化学习(FWNN - RL)系统使用 FWNN 的非线性逼近能力实现强化学习状态空间到动作空间的映射,同时可以应用强化学习进行 FWNN 的模糊规则结构辨识和模糊隶属度参数整定。强化学习的实现方式是将它的状态矢量作为 FWNN 的输入变量,系统的输出部分为强化学习的动作空间。由系统执行动作从环境中获得的强化信号构成 FWNN 的误差代价函数,然后通过系统的学习来确定模糊规则的结论部分和调整模糊隶属度函数参数。

FWNN-RL 系统在基本模糊神经网络五层结构的基础上进行改进,得到六层结构。用 $I_i^{(j)}$ 和 $O_i^{(j)}$ 分别表示第 j 层网络的第 i 个神经元的输入和输出。网络的第 1 层为输入层,它将输入的状态变量传送到下一层,即

$$O_i^{(1)} = I_i^{(1)} = s_i \quad (2)$$

第 2 层为语言变量层。其中每个节点表示一个语言变量,其作用是计算输入状态分量 s_i 的隶属度。这里模糊隶属度函数采用 Marr 小波基函数,其表达式为

$$\psi(s) = (1 - s^2)e^{-s^2/2} \quad (3)$$

通过平移和伸缩小波基函数得到每个语言变量的隶属度函数。与输入变量 s_i 相关的第 j 个节点的输出模糊隶属度为

$$O_{ij}^{(2)} = \mu_{ij} = \psi_{ij}(O_{i(1)}) = \psi_{ij}(s_i) = \psi(\alpha_{ij}s_i - b_{ij}) \quad (4)$$

其中,α_{ij} 和 b_{ij} 分别表示相应的伸缩系数和平移系数。

第 3 层表示规则层。层中每个节点为一条模糊规则,用来计算每条模糊规则前提部分的适应度 μ_j,即

$$O_j^{(3)} = \mu_j = \mu_{1i_1}\mu_{2i_2}\cdots\mu_{ni_n} \quad (5)$$

第 4 层为归一化层。它实现本层输出的归一化处理,其节点的输出为

$$O_j^{(4)} = \bar{\mu}_j = O_j^{(3)} / \sum_{k=1}^{m} O_k^{(3)} = \mu_j / \sum_{k=1}^{m} \mu_k \quad (6)$$

第 5 层为动作选择层。与每条模糊规则前提部分相对应的可能动作作为该规则的可选结论部分,则模糊规则用如下形式表示[13,14]:

$$\begin{aligned} R_j: \text{If } s \text{ is } F^j \text{ Then } a \text{ is } a_{j1} \text{ with } q_{j1} \\ \text{Or } a \text{ is } a_{jl} \text{ with } q_{jl} \\ \cdots\cdots \\ \text{Or } a \text{ is } a_{jL} \text{ with } q_{jL} \end{aligned} \quad (7)$$

a_{jl} 和 q_{jl} 分别表示状态 s 的可能动作及其评估值。第 4 层的每个规则节点可以对应第 5 层的多个动作节点,节点之间的连接权表示相应动作的激活程度。在强化学习过程中,采用搜索策略激活第 l^* 个结论部分的动作 a_{jl*} 作为第 j 条规则的结论部分。

第 6 层是输出层。该层的每个节点与第 5 层所有动作节点相连,但在每次学习时,只有被激活的动作节点被选择,其输出值为强化学习执行的动作 a 及相应的评价值 $Q(s,a)$。动作由所有模糊规则的局部激活动作通过解模糊操作所得到,对应输出动作的评价值 $Q(s,a)$ 由激活动作对应的评估值 q_{jl*} 解模糊得到,按如下方法计算:

$$a(s) = \sum_{j=1}^{m} \bar{\mu}_j(s) \times a_{jl*} \quad (8)$$

$$Q(s,a) = \sum_{j=1}^{m} \bar{\mu}_j(s) \times q_{jl*} \quad (9)$$

1.3 网络学习算法

FWNN - RL 系统采用两阶段混合学习算法:

首先更新网络第5层每条模糊规则候选动作对应的评估值 q_{jl},以确定模糊规则的结论部分;然后调整小波模糊隶属度函数的伸缩和平移参数。

定义瞬时差分的 Bellman 均方残差[15]作为网络学习误差代价函数,即

$$E_t = \frac{1}{2}\sum_s [r_t + \gamma \cdot \max_{a\in A} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)]^2 \tag{10}$$

残差算法的网络连接权增量迭代公式为

$$\begin{aligned}\Delta w &= -\beta \cdot \frac{\partial E}{\partial w} \\ &= -\beta \cdot \vartheta_t \cdot \left[\phi\gamma \frac{\partial Q(s_{t+1}, a_{t+1})}{\partial w} - \frac{\partial Q(s_t, a_t)}{\partial w}\right]\end{aligned} \tag{11}$$

其中,$\vartheta_t = -\frac{\partial E_t}{\partial Q(s_t)} = r_t + \gamma \cdot \max_{a\in A} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)$,$w$ 表示待学习的参数,β 为学习率。ϕ 是 0~1 之间的实数,它用来表示残差算法逼近残差梯度和直接梯度的程度[15]。

利用 Q 学习更新评估值 q,定义对应的资格迹更新规则如下:

$$e_{jl}(t+1) = \begin{cases}\gamma\lambda e_{jl}(t) + \nabla_q Q_t, & l = l^* \\ \gamma\lambda e_{jl}(t), & \text{其他}\end{cases} \tag{12}$$

其中 $\nabla_q Q(s,a) = \frac{\partial Q(s,a)}{\partial q} = \bar{\mu}_j$,$\lambda$ 表示资格迹学习率。候选动作对应 q 值更新方法如下:

$$q_{jl}(t+1) = \begin{cases}q_{jl}(t) + \eta_t \cdot \nabla_q Q_t \cdot e_{jl}(t+1), & l = l^* \\ q_{jl}(t), & \text{其他}\end{cases} \tag{13}$$

其中 η_t 表示学习率。

下面讨论如何计算 $\frac{\partial E}{\partial \alpha_{ij}}$ 和 $\frac{\partial E}{\partial b_{ij}}$。利用神经网络的误差反向传播更新隶属度函数伸缩和平移参数的计算过程如下:

$$\begin{aligned}\frac{\partial Q(\mathbf{s}_t,a_t)}{\partial a_{ij}(t)} &= \frac{\partial Q}{\partial O_j^{(4)}} \cdot \frac{\partial O_j^{(4)}}{\partial O_j^{(3)}} \cdot \frac{\partial O_j^{(3)}}{\partial O_{ij}^{(2)}} \cdot \frac{\partial O_{ij}^{(2)}}{\partial \alpha_{ij}(t)} \\ &= \sum_{i=1}^{M} q(j,i) \cdot [1/(\sum_{k=1}^{N}\mu_k)^2] \cdot (\sum_{l=1,l\neq j}^{N}\mu_i \\ &\quad - \sum_{m=1,m\neq j}^{N}\mu_i) \cdot \psi'(\alpha_{ij}x_i - b_{ij}) \cdot x_i\end{aligned} \tag{14}$$

其中,$\psi'(\alpha_{ij}x_i - b_{ij}) = (\alpha_{ij}x_i - b_{ij}) \times [(\alpha_{ij}x_i - b_{ij})^2 - 3]e^{-(\alpha_{ij}x_i - b_{ij})^2/2}$,则

$$\begin{aligned}\frac{\partial Q(\mathbf{s}_t,a_t)}{\partial b_{ij}(t)} &= \frac{\partial Q}{\partial O_j^{(4)}} \cdot \frac{\partial O_j^{(4)}}{\partial O_j^{(3)}} \cdot \frac{\partial O_j^{(3)}}{\partial O_{ij}^{(2)}} \cdot \frac{\partial O_{ij}^{(2)}}{\partial b_{ij}(t)} \\ &= \sum_{i=1}^{M} q(j,i) \cdot [1/(\sum_{k=1}^{N}\mu_k)^2] \cdot \\ &\quad (\sum_{l=1,l\neq j}^{N}\mu_i - \sum_{m=1,m\neq j}^{N}\mu_i) \cdot [-\psi'(\alpha_{ij}x_i - b_{ij})]\end{aligned} \tag{15}$$

执行网络的输出动作 $a(s_t)$ 使状态转移到 s_{t+1},同理可以计算得到 $\frac{\partial Q(\mathbf{s}_{t+1},a_{t+1})}{\partial \alpha_{ij}(t)}$ 和 $\frac{\partial Q(\mathbf{s}_{t+1},a_{t+1})}{\partial b_{ij}(t)}$。

从而由式(11)得到模糊隶属度函数参数调整算法如下:

$$\begin{aligned}\alpha_{ij}(t+1) &= \alpha_{ij}(t) + \Delta\alpha_{ij}(t) \\ &= \alpha_{ij}(t) - \beta \cdot \frac{\partial E_t}{\partial \alpha_{ij}(t)}\end{aligned} \tag{16}$$

$$\begin{aligned}b_{ij}(t+1) &= b_{ij}(t) + \Delta b_{ij}(t) \\ &= b_{ij}(t) - \beta \cdot \frac{\partial E_t}{\partial b_{ij}(t)}\end{aligned} \tag{17}$$

通过以上学习过程,选择第5层具有最大 q 值对应的动作节点作为规则的结论部分节点,并删去其他相连接的动作节点,从而能够确定每条模糊规则的结论部分。此外,小波模糊隶属度函数的伸缩和平移参数也得到整定。

2　基于 RL 的足球机器人动作选择策略

本研究将以上的 WFNN-RL 方法应用于机器人足球比赛决策策略学习,以实现机器人足球比赛中球员动作的选择。本文将 FIRA 系列比赛中的微型机器人足球比赛(MiroSot)5 对 5 比赛作为研究平台。机器人足球的决策系统可以看作从比赛状态空间到机器人球员动作空间的映射。也就是说,每个机器人球员能够根据与自身相关的状态信息决定相应的动作(动作选择机制)[16]。

使用强化学习方法实现足球机器人动作选择机制的学习,首先需要根据应用系统对强化学习的状态变量、动作变量和强化信号函数进行定义。决定机器人球员选择动作的状态因素可以用图 1 来描述,其中浅色机器人为我方球队,深色为对方球队。图中 d_{B2HG} 和 d_{B2OG} 分别表示球与我方球门和对方球门的距离;θ_{BV} 为球运动方向与对方球门连线的角度;d_{R2HG} 和 d_{R2OG} 为学习主体机器人与我方球门和对方球门的距离;d_{R2B} 和 α_{RB} 为主机器人与球的距离和角度;d_{HR2B} 和 α_{HRB} 为我队除主机器人外另一个与球位姿最好的机器人与球的距离和角度;d_{OR2B} 和 α_{ORB} 为对方与球位姿最好的机器人球员与球的距离

和角度；d_{R2OR} 和 α_{ROR} 为主机器人与对方机器人的距离和角度。

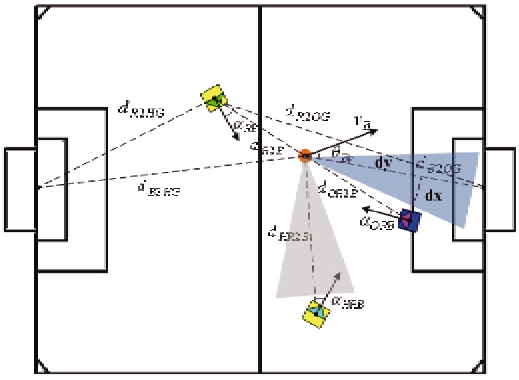

图 1 机器人足球比赛状态信息

完善的动作选择策略需要考虑比赛环境中的诸多因素，从中我们提取对策略学习有重要意义的特征信息作为强化学习的状态空间，RL 的状态变量定义为

$$s_1 = e^{-\tau_1(d_{B2HG}/d_{B2OG})} \tag{18}$$

$$s_2 = \tau_2 \mid \theta_{Bv} \mid /\pi \tag{19}$$

$$s_3 = \frac{1}{2}\left[e^{-\tau_{31}(d_{R2B}/d_{R2OR})} + e^{-\tau_{32}(\alpha_{RB}/\alpha_{ROR})}\right] \tag{20}$$

$$s_4 = \frac{1}{2}\left[e^{-\tau_{41}(d_{R2B}/d_{OR2B})} + e^{-\tau_{42}(\alpha_{RB}/\alpha_{ORB})}\right] \tag{21}$$

$$s_5 = \frac{1}{2}\left[e^{-\tau_{51}(d_{R2H}/d_{HR2H})} + e^{-\tau_{52}(\alpha_{RH}/\alpha_{HRH})}\right] \tag{22}$$

$$s_6 = e^{-\tau_6(d_{R2HG}/d_{R2OG})} \tag{23}$$

$$s_7 = e^{-\tau_7(OFI_R/OFI_G)} \tag{24}$$

在式(18)－(24)中，变量 s_1，s_2，s_3，s_4，s_5，s_6，$s_7 \in [0,1]$，其中 τ_1，τ_2，τ_{31}，τ_{32}，τ_{41}，τ_{42}，τ_{51}，τ_{52}，τ_6 及 τ_7 均为正常数。状态变量 s_1 可以表示球在球场的位置，当 $s_1 \to 0$ 时，球接近对方球门；当 $s_1 \to 1$ 时，球接近我方球门。状态变量 s_2 用来表示球的运动方向。状态变量 s_3 可以体现主机器人与球和对方机器人的位姿关系，当 $s_3 \to 0$ 时，主机器人更容易接近球；当 $s_3 \to 1$ 时，主机器人更容易接近对方机器人。这里考虑的对方机器人与球有最好的位姿关系。状态变量 s_4 表示我方球员和对方球员的竞争关系，该式综合考虑了双方竞争球员与球的距离因素和角度因素。当 $s_4 \to 1$ 时，我方球员更有机会接触球；相反，当 $s_4 \to 0$ 时，对方球员与球有更好的位姿关系。状态变量 s_5 显示了我方球员与队友的协作关系，当 $s_5 \to 1$ 时，表示主机器人球员与球有更好的位姿关系；相反，当 $s_5 \to 0$ 时，我方另一球员与球有更好的位姿关系。状态变量 s_6 可以表示学习机器人在球场的位置，当 $s_6 \to 0$ 时，主机器人接近对方球门；当 $s_6 \to 1$ 时，主机器人接近我方球门。状态变量 s_7 可以评估踢球的目标，式中 OFI_R 和 OFI_G 分别表示对方球员阻挡我方进攻球员传球和射门的可能程度。当对方机器人位于球和目标连线附近时，有较大可能拦截到球，见图 1 中的阴影部分。使用 OFI(obstruction-free-index)值[7]来表示阻挡的可能程度，其计算方法可参阅文献[17]，OFI 的值越小表示对方球员阻挡可能性越大。当 $s_7 \to 0$ 时，更利于射门；当 $s_7 \to 1$ 时，更适合传球。将状态变量 $s_1 - s_7$ 作为系统的输入，并将它们模糊化为三个语言变量 PS、PM 和 PB，其隶属度函数采用小波函数，参数通过学习过程来整定。

机器人所执行的动作是进行比赛的基本单元，也是决策策略的输出结果。这里讨论的候选动作是指足球机器人的技术动作，而战术动作可以通过这些技术动作的组合来实现。根据比赛经验，定义最基本、最重要的技术动作包括：射门、带球、传球、解围和盯人。定义 $Act = \{1,2,3,4,5\}$ 表示射门、带球、传球、解围和盯人的动作编号，强化学习的动作变量为球员的候选动作编号 Act_l，$l = 1,\cdots,5$。RL 的状态变量 $s_1,\cdots,s_7$ 作为学习系统的输入变量，动作变量作为第 5 层的动作选择节点。用 Act 值来表示模糊规则的结论部分。FWNN-RL 学习系统的输出动作为相应的动作编号的解模糊值 $\overline{Act}$。然后根据 $\overline{Act}$ 与动作编号的接近程度来确定机器人球员所要执行的动作。接近程度 e_l 用 $\overline{Act}$ 和动作编号的差值来表示，即 $e_l = \mid \overline{Act} - Act_l \mid$，$1 \leqslant l \leqslant 5$。其中最小的差值为 $e_{\min} = \min\{e_1,\cdots,e_5\}$，则最终的动作选择准则为

$$A = \begin{cases} \text{射门}, & e_1 = e_{\min} \\ \text{带球}, & e_2 = e_{\min} \\ \text{传球}, & e_3 = e_{\min} \\ \text{解围}, & e_4 = e_{\min} \\ \text{盯人}, & e_5 = e_{\min} \end{cases} \tag{25}$$

机器人足球比赛的目标是取得比赛的胜利，因此我方球队进球应该获得奖励，而失球应该给予惩罚。但在比赛中通常双方进球数量是较少的，并且进球是一系列决策和动作执行的结果。为了提高强化学习的速度，考虑增加其他因素来指导强化学习。所以，强化信号函数按照如下方式定义：

$$r_t^1 = \begin{cases} 1.0, & \text{我方进球} \\ -1.0, & \text{对方进球} \end{cases} \tag{26}$$

$$r_t^2 = \tau \cdot \Delta \bar{d}_{B2G} \tag{27}$$

$$r_t^3 = \begin{cases} +0.3, & PB_{t0} < PB_{t1} \\ -0.3, & \text{其他} \end{cases} \tag{28}$$

$$r_t^4 = \begin{cases} 0.2, & \text{我方控球} \\ -0.2, & \text{对方控球} \end{cases} \tag{29}$$

$$r_t = \sum_{i=1}^{4} r_t^i \tag{30}$$

式(27)中,$\Delta \bar{d}_{B2G}$ 为球与双方球门平均距离的变化,即 $\Delta \bar{d}_{B2G} = \bar{d}_{B2HG} - \bar{d}_{B2OG}$,其中 τ 为系数。用 $\Delta \bar{d}_{B2G}$ 来表示球在球场中的位置,当球在对方半场时应获得正的强化信号,球在我方半场获得负的强化信号,并且当球越接近对方球门所获得的奖励越大。式(28)表示由控球时间决定的回报函数,PB_{t0} 和 PB_{t1} 分别表示相邻学习周期我方球队的控球时间,当我方控球时间增加时应获得奖励。r_t^4 表示学习主体执行动作后,我方获得控球权,则获得奖励;相反,应该得到惩罚。

综上所述,本研究完成了强化学习状态空间、动作空间及强化信号函数的定义,然后使用 FWNN-RL 算法实现了足球机器人动作选择策略的学习。

3　实验结果与分析

本实验基于 FIRA 仿真平台 Robot Soccer Simulator 和实际微型组足球机器人进行了强化学习和方法验证,利用研究的基于模糊小波神经网络的强化学习方法来使机器人球员学习动作选择策略。强化学习每进行 1000 次为一个阶段。通过学习可以实现对 FWNN-RL 系统中模糊规则的模糊隶属度函数参数的整定,并能够确定模糊规则的后件结论部分。在完成每个阶段后,利用强化学习建立的模糊推理系统作为机器人球员动作选择策略控制器,然后对学习结果进行测试。测试时进行若干场比赛,比赛对手为强化学习对象。图 2 显示了在学习过程中,各学习阶段利用学习结果进行比赛时我方平均每场比赛的竞胜球数。结果显示了我方球队的竞胜球数逐渐增多,能够反映出我方球队的整体能力在学习过程中逐渐提高,这说明足球机器人能够通过学习来选择正确的动作。图 3 表示通过在学习过程中进行的阶段测试得到的我方控球时间的百分比率。

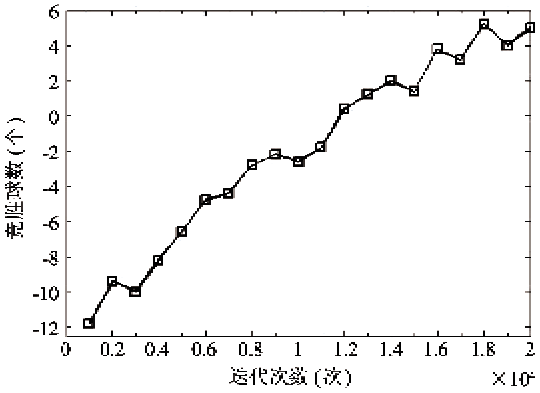

图 2　我方球队竞胜球

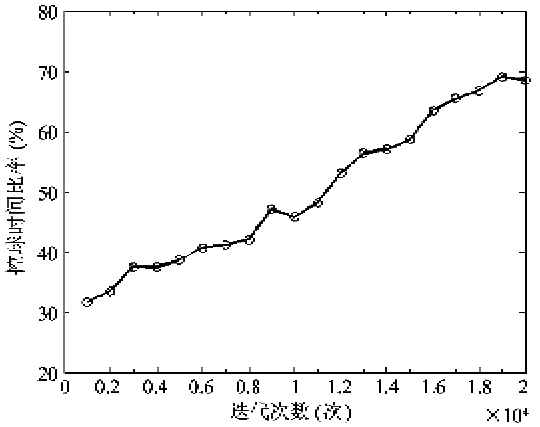

图 3　我方控球时间比率

在学习过程结束后,利用学习结果策略控制球队进行多场比赛测试,表 1 显示了学习得到的动态选择策略和采用固定策略的对比实验数据,数据为经过多场比赛测试计算的平均数据。其中在统计执行动作失误数时,排出了射门动作,这是因为有对手的防守和干扰存在,即使正确执行了该动作,但仍较难成功,所以不能客观反映动作选择机制的有效性。

表 1　对比测试比赛数据

	球在对方半场时间比率	球出现在对方禁区次数	执行动作的失误比率
学习策略	67.5%	11	16.5%
固定策略	53.9%	7	20.8%

从以上实验结果来看,在学习过程中,我方球队的整体实力逐渐提高,最终在同对手的比赛中取得较大的优势。但决策系统的学习速度是比较慢的,这是因为机器人足球系统的复杂性和动态性,因此需要在大量的比赛训练才能建立较为完善的决策系统。此外,强化学习算法的性能和参数的选择也对学习结果有一定的影响,因此提高算法的效率和性

能以及参数的合理性也能够提高系统的决策能力。

4 结 论

本文研究了一种基于模糊小波神经网络的强化学习方法,利用模糊小波神经网络的函数逼近特性实现强化学习状态空间到动作空间的映射。该算法将 Q 学习过程转化成模糊规则表述,并利用了小波网络的多尺度学习能力,从而有效地解决了大规模或连续状态空间的强化学习问题。此外,使用提出的强化学习方法来实现足球机器人动作选择策略的学习,对影响机器人动作选择的决策因素进行了分析,通过与环境的交互学习,使机器人在比赛过程中逐渐掌握动作选择能力,使其能够根据环境状态来选择合理的执行动作。

参考文献

[1] Stone P, Sutton R S, Kuhlmann G. Reinforcement learning for robocup soccer keepaway. *Adaptive Behavior*, 2005, 13(3): 165-188

[2] Hwang K S, Tan S W, Chen C C. Cooperative strategy based on adaptive Q-learning for robot soccer systems. *IEEE Transaction on Fuzzy Systems*, 2004, 12(4): 569-576

[3] Park K H, Kim Y J, Kim J H. Modular Q-learning based multi-agent cooperation for robot soccer. *Robotics and Autonomous Systems*, 2001, 35: 109-122

[4] Sutton R S, Barto A G. Reinforcement Learning: an Introduction. Cambridge, MA, USA: MIT Press, 1998

[5] Daniel W C H, Zhang P A, Xu J H. Fuzzy wavelet networks for function learning. *IEEE Transactions on Fuzzy Systems*, 2001, 9(1): 200-211

[6] Rahib H A, Okyay K. Fuzzy wavelet neural networks for identification and control of dynamic plants—a novel structure and a comparative study. *IEEE Transactions on Industrial Electronics*, 2008, 55(8): 3133-3140

[7] Leonardo M R. Unification of neural and wavelet networks and fuzzy systems. *IEEE Transaction on Neural Networks*, 1999, 10(4): 801-814

[8] 王耀南. 机器人智能控制工程. 北京: 科学出版社, 2004

[9] Zhang Q, Benveniste A. Wavelet network. *IEEE Transaction on Neural Network*, 1992, 3: 889-898

[10] Ling S H, Iu H H C, Leung F H F, et al. Improved hybrid particle swarm optimized wavelet neural network for modeling the development of fluid dispensing for electronic packaging. *IEEE Transactions on Industrial Electronics*, 2008, 55(9): 3447-3460

[11] 郑晶,王祖林,郭旭静. 基于神经网络的任意延迟 M 带小波设计. 沈阳工业大学学报,2011, 33 (5): 561-565

[12] Watkins C J C H, Dayan P. Q-learning. *Machine Learning*, 1992, 8(3-4): 279-292

[13] Jouffe L. Fuzzy inference system learning by reinforcement methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 1998, 28 (3): 338-355

[14] Juang C F, Lu C M. Ant colony optimization incorporated with fuzzy Q-learning for reinforcement fuzzy control. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 2009, 39(3): 597-608

[15] Baird L C. Residual algorithms: reinforcement learning with function approximation. In: Proceedings of the 12th International Conference on Machine Learning, San Francisco, USA, 1995. 9-12

[16] Chia J W, Lee T L. A fuzzy mechanism for action selection of soccer robots. *Journal of Intelligent and Robotic Systems*, 2004, 39: 57-70

[17] Veloso M, Stone P, Han k. CMUnited-97: RoboCup-97 small-robot world champion team. *AI Magazine*, 1998, 19(3): 61-69

Reinforcement learning based on FWNN and its application in decision-making for multi-robot cooperation

Duan Yong*, Li Cheng*, Xu Xinhe**

(*School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870)

(**School of Information Science and Engineering, Northeastern University, Shenyang 110004)

Abstract

A reinforcement learning (RL) algorithm based on fuzzy wavelet neural networks (FWNN) was proposed. Furthermore, its application in selection of decision-making strategies for robot soccer was studied. Firstly, a FWNN was used to perform the mapping from the state space to the action space of RL, consequently, the problems of slow learning and difficult convergence caused by the large or continuous state space were solved effectively. Then, the application of the presented method in learning of decision-making strategies for robot soccer was studied, achieving the result through learning, the robot players can master the ability of selecting actions based on their states in the game. Finally, the effectiveness of the presented method was verified by experiment. The experimental result shows that it can meet the demands of robot soccer.

Key words: reinforcement learning (RL), fuzzy wavelet neural network (FWNN), robot soccer, action selection, decision-making