中国科学: 信息科学 2016 年 第 46 卷 第 3 期: 325–337
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
论 文

基于直接强化学习的面向目标的仿生导航模型

于乃功[①②], 李倜[①②*], 方略[①②]

① 北京工业大学电子信息与控制工程学院, 北京 100124
② 计算智能与智能系统北京市重点实验室, 北京 100124
* 通信作者. E-mail: bgdliti@163.com

收稿日期: 2015–10–09; 接受日期: 2015–12–01; 网络出版日期: 2016–02–22
国家自然科学基金 (批准号: 61573029) 资助项目

摘要 针对连续动作和状态空间中面向目标的导航问题, 依据海马结构中位置细胞相关特性和相关信息传递通路, 构建海马位置细胞到前额叶皮层假设的动作细胞的脉冲神经网络模型. 连续的状态空间和动作空间分别由位置细胞和动作细胞进行表征, 模型采用直接强化学习与脉冲响应模型相结合的算法进行面向目标的自主导航. 在 Morris 水迷宫环境中的仿真实验结果表明, 该模型能够解决连续状态空间中面向目标导航问题, 所采用算法在性能上优于传统的时间差分学习算法. 调整网络中动作细胞的数量, 模型的收敛性能不变, 在改变状态空间和目标位置时, 也可以实现面向目标的导航.

关键词 直接强化学习 位置细胞 动作细胞 脉冲神经网络 面向目标导航

1 引言

自主定位和面向目标导航的能力是动物赖以生存的重要能力. 动物和人类能够使用来自于感知器官不完整的环境传感信息在没有任何先验知识的环境中进行快速自主定位, 即使在目标位置不可见时, 动物也能使用内在的环境地图表达, 即所谓的认知地图 (cognitive map), 进行面向目标位置的导航 (如归巢和觅食), 这种导航策略被称为局部导航 (locale navigation) [1]. 传统的移动机器人进行局部导航实验时通常使用特殊传感器 (如声呐和罗盘) 或人为直接输入先验知识的方法 [2], 智能化程度和可扩展性差.

理解并模仿生物大脑导航机理, 开发能更加智能化的快速定位和导航的自主移动机器人是当前人工智能和机器人研究领域的研究热点. 生物学研究表明大脑颞叶海马体 (hippocampus) 中 CA1 和 CA3 区存在着与空间定位相关的细胞, 被称为位置细胞 (place cells), 位置细胞放电活动对应的动物在环境中的活动范围被称为位置野 (place field) [1,3]. 动物处于空间环境中特定的位置时, 相对应的位置细胞群的放电 (firing) 频率会显著提高, 位置细胞的群放电活动在异我为中心的参考系下编码了动物的当前位置 [4], 建立了脑区和外界物理世界稳定的一一映射关系 [5]. 位置细胞被认为是构成认知地图的基本要素 [6]. 损毁海马体之后, 动物无法进行有效的局部导航 [7].

引用格式: 于乃功, 李倜, 方略. 基于直接强化学习的面向目标的仿生导航模型. 中国科学: 信息科学, 2016, 46: 325–337, doi: 10.1360/N112015-00217

© 2016《中国科学》杂志社 www.scichina.com info.scichina.com

局部导航的空间学习任务在于寻找空间位置和隐藏目标位置之间的相关联系, 这种联系的确定通过奖励评价信号进行调节, 使用强化学习模型 (reinforcement learning models) [8] 可以解决这类学习任务. 传统的强化学习模型如时间差分 (temporal difference, TD) 模型, 大多采用离散的状态和有限的动作去解决局部导航问题 [9~11]. 而在实际的物理世界中, 动物所处的状态空间是连续的, 能够选择任意的运动方向在物理环境中移动. 在连续的状态和动作空间中采用 TD 模型, 智能体 (agent) 会陷入 "维数灾", 算法收敛缓慢, 且无法解决延迟奖励信号带来的时间信度分配问题. 传统强化学习模型通常引入资格迹 (eligibility trace) [12,13] 或采用函数估计 [8] 来避免维数灾, 加快学习速度, 利用资格迹的短时存储特性解决时间信度分配问题.

脉冲神经网络 (spiking neural network, SNN) 是将信息编码为脉冲的时间结构, 而不是采用传统的平均脉冲发放率, 高度模拟了生物神经元的动态发放特性, 其动态性反映在它对外界输入信号的积累发放过程, 能够解决动态绑定问题 [14]. 本文根据海马体到前额叶皮层中奖励调节的动作选择现象, 构建海马位置细胞到前额叶皮层假设的动作细胞的前向连接脉冲神经网络模型, 动作细胞呈环状分布, 利用其脉冲放电频率代表的群向量 (population vector) 指示下一步的活动方向. 模型的学习算法采用基于脉冲响应模型的直接强化学习规则 [15]. 仿真实验结果显示, 经过 20 次左右的训练, 智能体能够学习到面向隐藏目标的导航策略, 与老鼠在 Morris 水迷宫中的表现类似 [16].

文章第 2 节说明了模型的生理学依据和相关理论基础, 第 3 节详细描述了模型构建以及学习算法流程, 第 4 节在 Morris 水迷宫环境下进行了模型的仿真实验, 说明了模型的有效性, 第 5 节对模型结果进行了简单的讨论.

2 构建模型的生理学和行为学依据

1971 年, O'Keefe 和 Dostrovsky 发现在海马体中存在着空间特定位置选择性放电的位置细胞 [3]. 位置细胞的位置野在动物进入新的环境中会迅速生成并随着动物对环境的遍历而覆盖整个环境 [17]; 与视觉皮层上的神经细胞不同, 位置细胞在脑中的相对位置与对应的位置野并无直接联系, 换言之, 两个相邻的位置细胞可能所对应的实际地理位置的位置野并不相邻 [4]; 外源性信息 (如视觉、嗅觉等) [18] 和内源性信息 (如前庭信息、躯体感觉) [19] 都能使位置细胞发放, 并形成稳定的位置野, 在没有外源性信息输入情况下 (如黑暗环境) 位置细胞也能发放并形成稳定的位置野 [20]. 动物在探索环境过程中, 海马体中位置细胞位置野逐渐发展形成代表空间环境的认知地图.

根据海马位置细胞的电活动并不能完全正确预测未来行为的方向, 而海马与大脑的命令和控制中心内侧前额皮层 (medial prefrontal cortex, mPFC) 的动态联系才是正确预测未来行为的关键因素 [21]. 大脑腹侧被盖区 (ventral tegmental area, VTA) 存在的多巴胺能神经元 (dopaminergic neurons) 与奖励预测的误差信号有关 [22,23], 这些神经元进一步投射到伏隔核 (nucleus accumbens, NA), 而伏隔核的主要输入信号来源于海马体, 伏隔核与前额叶皮层之间存在着双向投射 [24]. 换言之, 伏隔核从海马体接收空间环境信息, 从腹侧被盖区接收奖励预测误差信息, 并与前额叶皮层相互作用确定动物的运动行为. 这些生物学研究支持这样一种假设: 面向目标的导航学习的神经基础可能是海马位置细胞神经元与 NA 神经元之间的与奖励信号相关的突触调节, NA 进一步投射至前额叶皮层与初级运动皮层互联, 从而提供了一种可能的面向目标的信息用来控制行动. 本文中所描述的导航模型是根据这些生理学研究提出的.

在行为学研究中, 通常采用 Morris 水迷宫模型测试动物的导航能力, 利用啮齿类动物 (如老鼠) 天生会游泳但怕水的天性, 将老鼠置于有着不透明液体的水池中, 水池中有一个略低于水面的隐藏平

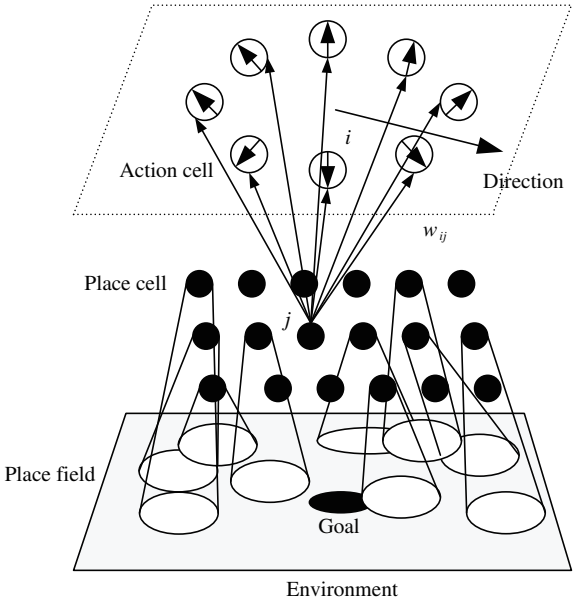

图 1 位置细胞到动作细胞的导航模型示意图

Figure 1 The sketch map of navigation model from place cells to action cells

台, 老鼠为了逃离水池, 需要寻找到隐藏平台 [25], 这个隐藏平台可以被看作是奖励信息. 损毁实验表明, 损毁海马 [26] 的老鼠不能顺利完成水迷宫导航实验. 实验研究表明, 在无路标的水迷宫环境中, 老鼠在黑暗环境中进行探索, 随机寻找到隐藏平台, 而在多次探索实验后, 老鼠能够记住平台位置, 在不同的初始位置也能顺利到达平台, 这表明老鼠对空间环境产生了某种内部表达, 并能利用这种表达进行快速导航.

本文根据上述的生理学和行为学研究基础, 构建基于脉冲神经元的局部导航模型, 学习算法采用基于脉冲响应模型的直接强化学习, 并利用 Morris 水迷宫环境进行仿真实验验证.

3 模型的构建与学习算法

3.1 模型的构建

根据第 2 节所述的生理学研究, 针对连续空间的局部导航问题, 本文构建如图 1 所示的模型, 位置细胞 j 前向投射到动作细胞 i, 给定老鼠的位置, 下一步移动方向由动作细胞群向量表示. 假设动物位于位置细胞 j 形成的位置野的中心, 那么动作细胞的群活动由位置细胞 j 投射到不同的动作细胞突触连接强度控制, 位置细胞 j 到动作细胞 i 的连接强度越大, 动作细胞 i 所代表的动作被选择的可能性越大.

连接权值由奖励调节的基于脉冲响应模型的直接强化学习进行更新; 到达目标位置的奖励信号是延迟的, 即, 动物完成一系列动作后, 得到一个正或者负的奖励信号.

连续位置空间被认为是状态空间, 由海马位置细胞重叠的位置野进行表征, 位置野对整个二维环境进行密集编码, 位置细胞 j 的优先位置定义为 $P_j = (m_j, n_j)$, 位置细胞 j 的放电率为 [27]

$$r_{pj} = \gamma \mathrm{e}^{\left[-\frac{(m-m_j)^2+(n-n_j)^2}{2\sigma^2}\right]}, \tag{1}$$

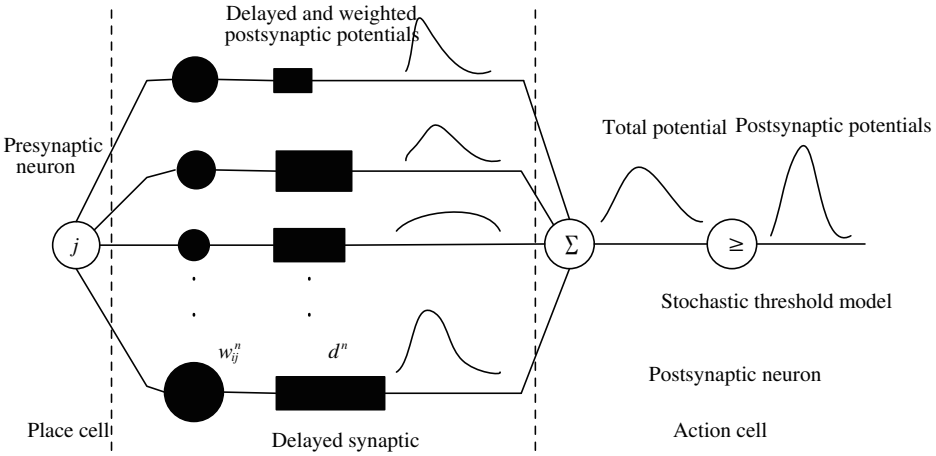

图 2　位置细胞到动作细胞突触连接示意图

Figure 2　The sketch map of synaptic from place cells to action cells

其中, γ 为最大放电率, 设定为 100 Hz, (m,n) 是当前位置, σ 表示位置野的宽度. 由 (1) 式可知, 当前位置可由位置细胞群联合编码, 通过这种密集编码方式, 位置细胞对整个环境进行了表征. 位置细胞被建模为 Poisson 神经元, 所以, 瞬时放电率为 r_p 的位置细胞在无穷小的持续时间 (Δt) 产生脉冲的可能性为 $P\left(\text{spike}\right)=\frac{\mathrm{e}^{-r_p\Delta t}(r_p\Delta t)^1}{1!}\approx r_p\Delta t$, 当 $P\left(\text{spike}\right)$ 大于一个在 0 到 1 之间均匀分布的随机采样值时, 位置细胞产生一个脉冲.

模型假设前额叶皮层中存在着代表动物运动方向的动作细胞 (action cells), 将动作细胞构建成环状模型, 不同于文献 [12, 13, 27], 动作细胞之间不存在着横向突触连接. 动作细胞被建模成脉冲响应模型 (spike response model, SRM) [28], 位置细胞到动作细胞的突触信息传递示意图如图 2 所示, 动作细胞 i 的膜电位为

$$u_i\left(t\right)=u_{\text{rest}}+\sum_{j=1}^{N}w_{ij}\sum_{t_j^f\in x_j}\varepsilon\left(t-t_j^f\right)+\sum_{t_i^f\in y_{i,t}}\eta\left(t-t_i^f\right),\tag{2}$$

其中, u_{rest} 是静息电位 (resting potential), 等于 -70 mv, N 是位置细胞的个数, w_{ij} 是位置细胞到动作细胞的突触连接权值, x_j 是突触前脉冲序列, $y_{i,t}$ 是时间 t 之前的突触后脉冲集合, t_j^f 和 t_i^f 分别是动作细胞和位置细胞的第 f 次放电时间, $\varepsilon\left(t\right)=\varepsilon_0\mathrm{e}^{\left(-\frac{t}{\tau_m}\right)}$ 和 $\eta\left(t\right)=\eta_0\mathrm{e}^{\left(-\frac{t}{\tau_m}\right)}$ 分别表示兴奋性突触后电位和脉冲后电位的响应核, ε_0 和 η_0 相应的时间系数, 设定为 1 ms, τ_m 为膜时间常数. 注意到模型中并没有考虑实际的突触延时问题, 建模过程中将突触延时作为内在噪声进行处理.

考虑到突触内在噪声问题, 用随机阈值取代固有阈值强度, 使用概率密度 (probability density) 的好处是可以扩大编码范围, 脉冲的产生由一个指数函数确定 [14],

$$D_i=D_0\mathrm{e}^{\left(\frac{u_i-u_\theta}{\Delta u}\right)},\tag{3}$$

其中, $D_0=1/\text{ms}$ 是放电率的比例因子, u_θ 是正常放电阈值, 设定为 -50 mv, Δu 控制脉冲响应的噪声水平, 设定为 5 mv. 那么位置细胞 j 输入信号 (此处以集合 X 表示) 导致动作细胞 i 在 t 时刻前产生一个脉冲序列 $y_t=\{t_i^1,t_i^2,\ldots,t_i^f\}$ 的概率密度为 [29]

$$P\left(y_t|X\right)=P\left(t_i^1,t_i^2,\ldots,t_i^f\right)S\left(t|y_t\right),\tag{4}$$

其中, $P(t_i^1,t_i^2,\ldots,t_i^f)$ 是在 $t_i^1,t_i^2,\ldots,t_i^f$ 时间有 f 个脉冲的概率分布, $S\left(t|y_t\right)=\mathrm{e}^{\left[-\int_{t_i^f}^{t}D(t'|y_{t'})\mathrm{d}t'\right]}$ 是 t_i^f 到 t 的时间段内不产生脉冲的概率分布, 而可以由条件分布表示,

$$P\left(t_i^1,t_i^2,\ldots,t_i^f\right)=P\left(t_i^1\right)\cdot\prod_{i=2}^{f}P\left(t_i^f|t_{i-1}^f,\ldots,t_1^f\right). \tag{5}$$

将 (5) 式带入 (4) 式, 可以得到

$$P\left(y_t|X\right)=\left[\prod_{t_i^f\in y_t}D\left(t_i^f|y_{t_i}^f\right)\right]\cdot\mathrm{e}^{\left[-\int_0^t D\left(t'|y_{t'}\right)\mathrm{d}t'\right]}. \tag{6}$$

将 (6) 式转换可得

$$P(y_t|X)=\mathrm{e}^{\left(\int_0^t\log D\left(t_i^f|y_{t_i}^f\right)Y_i(t)-D\left(t_i^f|y_{t_i}^f\right)\mathrm{d}t\right)}, \tag{7}$$

其中, $Y_i(t)=\sum_f\delta(t-t_i^f)$ 是动作细胞 i 在 t_i^f 的时间内产生的全部突触后脉冲序列. 通过 (7) 式, 整个脉冲响应序列的概率密度得到了表述.

3.2 学习算法的提出

在传统强化学习理论中, 状态空间通过状态动作值函数与动作空间产生关联, 而状态动作值由期望的未来奖励表达, 这种基于值函数预测的方式能够为相应的状态空间提供最优的动作, 但在连续状态和动作空间迭代中传统强化学习方法不可避免地会陷入维数灾难. 直接强化学习方法, 预测能够提高所使用策略表现的梯度方向, 直接对一个长时程奖励的梯度预测值进行调节, 而不是计算值函数, 避免了基于策略迭代方法的一些缺点 [15]. 本文采用直接强化学习算法调节 SRM 神经元突触, 即动作神经元突触, 产生的输出脉冲确定了智能体下一步的移动方向.

智能体进行强化学习的目标是最大化其预期的未来奖励 [30], 那么, 随机参数化的脉冲响应策略为 (7) 式所示的产生一系列脉冲的概率密度, 在此策略下, 预期的奖励信息能够定义为

$$\hat{R}_{Xy}=\sum_{X,y}R\left(X,y\right)P\left(y|X\right)P\left(X\right), \tag{8}$$

其中, $R\left(X,y\right)$ 为环境反馈的奖励信号, $P\left(X\right)$ 为位置细胞的放电概率. 通过调节期望奖励对于突触权值的梯度, 可以获取最大奖励. 根据直接强化学习算法, 求解最优的策略梯度, 为了方便计算, 对似然函数 $P\left(y|X\right)$ 取对数似然函数 $\log P\left(y|X\right)$, 并对权值求取偏导数:

$$\frac{\partial}{\partial w_{ij}}\log P\left(y|X\right)=\frac{1}{\Delta u}\int_0^T\left[Y_i\left(t_i^f\right)-D\left(t_i^f|y_{t_i}^f\right)\right]\sum_{t_j^f\in X_j}\varepsilon\left(t-t_j^f\right)\mathrm{d}t, \tag{9}$$

其中, $\frac{\partial u_t}{\partial w_{ij}}$ 使用 SRM 模型求得 $\frac{\partial u_t}{\partial w_{ij}}=\sum_{t_j^f\in X_j}\varepsilon(t-t_j^f)$, 所以位置细胞 j 到动作细胞 i 的突触连接的策略梯度为

$$\Delta w_{xy}=\alpha R\left(X,y\right)\cdot\frac{1}{\Delta u}\int_0^T\left[Y_i\left(t_i^f\right)-D\left(t_i^f|y_{t_i}^f\right)\right]\sum_{t_j^f\in X_j}\varepsilon\left(t-t_j^f\right)\mathrm{d}t. \tag{10}$$

因此, 基于脉冲响应模型的直接强化学习 (RL-SRM) 的脉冲神经元权值更新规则为

$$\begin{cases}w_{ij,t+1}=w_{ij,t}+\alpha R_{t+1}e_{ij,t+1}, & \text{(11a)}\\ e_{ij,t+1}=\beta e_{ij,t}+\dfrac{1}{\Delta u}\displaystyle\int_0^T\left[Y_i\left(t_i^f\right)-D\left(t_i^f|y_{t_i}^f\right)\right]\sum_{t_j^f\in X_j}\varepsilon\left(t-t_j^f\right)\mathrm{d}t, & \text{(11b)}\end{cases}$$

其中, α 是学习率, R_{t+1} 是 $t+1$ 时刻从环境获得的强化信号, 即奖励 (惩罚) 信息; $e_{ij,t+1}$ 是 $t+1$ 时刻的资格迹, β 是资格迹的衰减因子.

由于强化信号只有在智能体达到目标状态时才能由环境给定, 强化信号在转移过程中缓慢的回传到先前到达过的状态, 在由初始状态到目标状态的转移过程中必然存在着延时奖励 (latency reward), 这就造成了强化信号的时间信度分配问题. 状态空间由海马位置细胞重叠的位置野表达, 而动作空间由动作细胞代表的动作信息表示, 资格迹是对过去的位置细胞和动作细胞活动的短时记录, 记录的信息随时间逐渐衰减, 只有被资格迹记录的状态动作信息才有资格进行时间信度的分配, 使得强化信号不必遍历整个状态和动作空间, 从而加快了学习速度.

4 仿真实验

Morris 水迷宫实验作为导航实验的标准范式, 用于测试本文模型的有效性. 仿照大鼠的行为学实验, 每次探索的最大时长为 100 s, 每次仿真实验, 智能体被随机置于环境靠近边界的任意位置, 在逐步探索环境的过程中, 获取环境的反馈信号 (奖励或惩罚信号), 当智能体到达隐藏平台时, 先前经历的状态空间通过反馈信号进行强化, 多次训练后, 智能体在环境的任意位置, 均能以最优路径到达隐藏平台.

智能体的运动由动作细胞族确定, 每个动作细胞 i 代表了一个特定的方向 ϕ_i, 这些方向是 0 到 2π 之间的均匀分布. 动作细胞群向量的角度 ϕ^{AC} 确定下一次的移动方向. $\bar{r}_i^{\mathrm{AC}}$ 是动作细胞 i 一定时间内的平均放电率, 角度 ϕ^{AC} 可表示为 [31]

$$\phi^{\mathrm{AC}} = \tan^{-1}\left[\frac{\sum_i \bar{r}_i^{\mathrm{AC}} \cdot \sin\left(\frac{2\pi i}{N^{\mathrm{AC}}}\right)}{\sum_i \bar{r}_i^{\mathrm{AC}} \cdot \cos\left(\frac{2\pi i}{N^{\mathrm{AC}}}\right)}\right], \tag{12}$$

其中, N^{AC} 是动作细胞的个数, 仿真实验设定为 360, $\bar{r}_i^{\mathrm{AC}}$ 可由 (13) 式表示:

$$\bar{r}_i^{\mathrm{AC}} = \frac{1}{T}\int_0^T Y_i(t)\mathrm{d}t, \tag{13}$$

其中, T 为海马 θ 节律震荡的周期 [32], 此时取 θ 频率为 5 Hz, 那么 T=200 ms. 行为学研究表明, 大鼠在水迷宫中的速度是 48 cm/s [33], 仿真实验的每个 θ 周期, 智能体移动 9.6 cm.

仿真实验算法流程如图 3 所示, 仿真实验参数设定如表 1 所示. 为了使智能体充分地探索环境, 在决策过程中加入了 ξ 贪婪算法, 贪婪搜索参数 ξ 设为 0.01. 在大鼠的水迷宫行为学实验中, 一次探索的最大时长一般设置为 60 s 到 120 s, 在本文的仿真实验中, 一次探索实验的最大时长设置为 100 s, 超过一次探索时长未找到目标位置, 则重置智能体于一个新的初始点, 开始下一次探索实验.

在半径为 0.5 m 的圆形水迷宫环境中, 通过迭代仿真实验训练, 前向连接的突触强度的群向量的方向在任意位置均指向隐藏目标位置, 对应了智能体的导航地图 (navigation map). 在没有障碍物的情况下, 从当前位置直接指向目标位置就是最佳运动方向. 如图 4 所示, 分别为 10 次、20 次和 50 次训练后的导航地图. 大圈表示水迷宫环境边界, 小圈表示目标位置, 10 次训练后, 智能体已经具有了向目标位置运动的趋势, 20 次训练之后, 水迷宫环境中, 靠近目标位置的前向连接权值的方向均有着面向目标位置的趋势, 且越靠近目标位置, 相对权值越大, 而在环境的边界位置, 智能体的权值方向并没有全部面向目标位置, 这表明, 智能体还并未遍历整个环境, 通过进一步地探索训练, 智能体逐渐学会了在水迷宫环境中任意初始位置, 面向目标的导航运动. 如图 4(c) 为 50 次训练后的导航地图, 可以

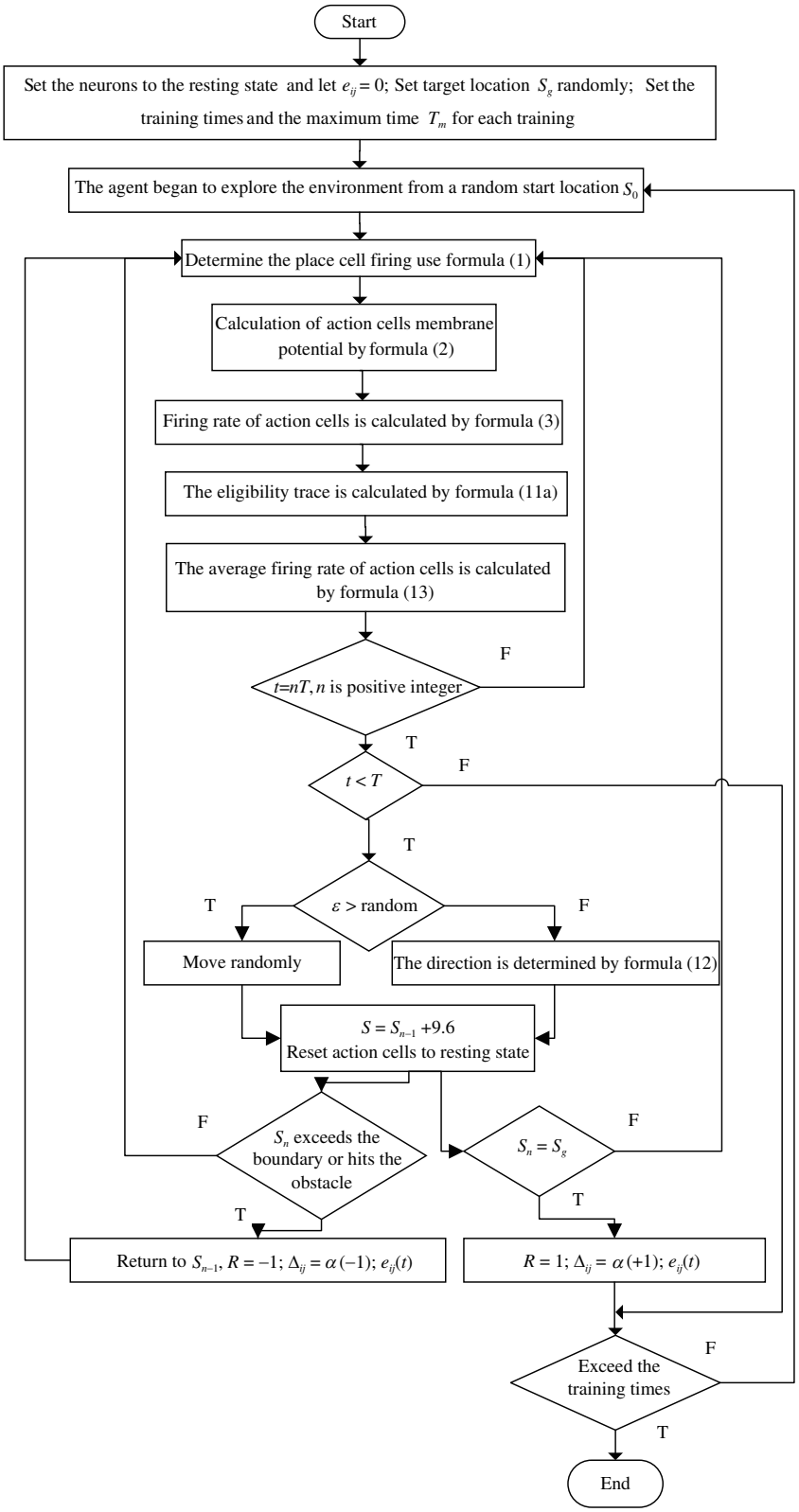

图 3 SRM-RL 算法流程图

Figure 3 The flowsheet of SRM-RL algorithm

表 1　仿真实验参数设定

Table 1　The parameters of simulation experiment

Parameter	Δt	σ	τ_m	α	β	N
Value	1 ms	12 cm	10 ms	0.02	0.95	100

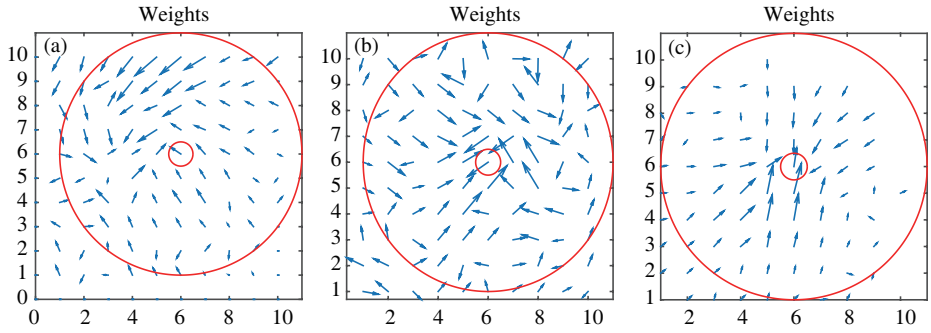

图 4　(网络版彩图) 智能体的导航地图

Figure 4　(Color online) The navigation map of the agent. Navigation map after (a) 10 times training, (b) 20 times training, and (c) 50 times training

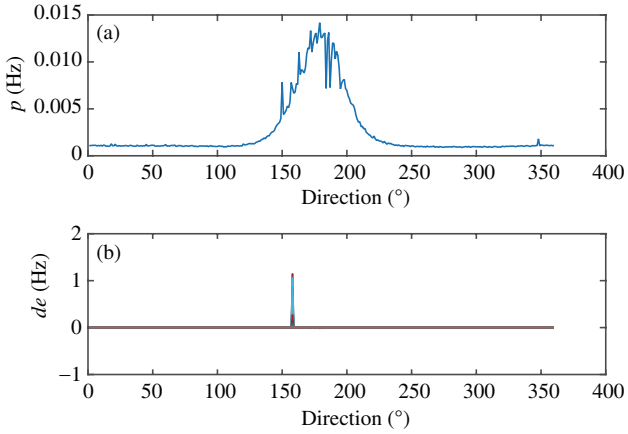

图 5　(网络版彩图) 动作细胞瞬时放电率和对应的资格迹

Figure 5　(Color online) (a) Firing rate of action cell and (b) the corresponding momentary value of the eligibility

看到从水迷宫的边界任意位置出发, 智能体都能以最优路径面向目标位置移动. 大鼠行为学实验中, 在没有外在路标参考的情况下, 大鼠能在 36 次左右的探索学习后, 从环境任意初始位置以最短路径的方向快速地到达隐藏平台 [16], 这与仿真实验学习结果基本相同.

在第 50 次实验中, 对动作细胞瞬时放电和资格迹的记录如图 5 所示, 横坐标表示动作细胞代表的方向, 纵坐标分别表示动作细胞放电率和资格迹的瞬时变化量. 在当前记录的时刻, 动作细胞的放电频率在 170° 左右的方向上达到峰值, 而放电频率越高, 动作细胞产生脉冲的可能性越大, 动作细胞的瞬时群编码活动指示的方向与资格迹单位时间内的瞬时变化量所指示方向一致, 这说明资格迹能够有效地记忆动作细胞的群编码活动.

在 Morris 水迷宫环境中, 老鼠到达隐藏平台的时间被称为逃避潜伏期 (escape latency), 是检测算

1. Calculate state action values: $Q(r_p(t),a_i) = r^i_{ac} = \Sigma_j w_{ij} \cdot r_{pj}$

2. Environmental detection by greedy exploration strategy to balance exploration and exploitation. Determine the neurons firing during the decision time T in a theta cycle.

3. Action space generalization: $r^2_{ac}(t) = \exp(-\Delta\phi_i^2/2\sigma^2_{ac})$

4. Update eligibility trace: $e_{ij,t+1} = \beta e_{ij,t} + r^2_{ac}(t) r_{pj}$

5. Calculate reward prediction error:

$\delta(t) = R(t) + \gamma Q(r_p(t),a^0(t)) - Q(r_p(t-1),a^x(t-1))$

6. Update synaptic strengths: $\Delta w_{ij} = \eta \cdot \delta(t) \cdot e_{ij,t-1}$

图 6 Q(λ) 算法流程图

Figure 6 Q(λ) algorithm flow chart

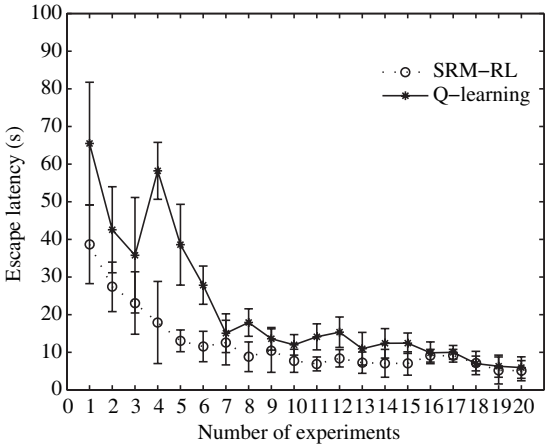

图 7 逃避潜伏期随实验次数的增加的变化过程

Figure 7 The change of escape latency with the increase of the number of experiments

法有效性和收敛性能的指标. 利用传统 Q-leaning 结合资格迹的算法 (简称为 Q(λ)) 进行相同的学习任务 [12,13], 算法流程图如图 6 所示.

统计基于 SRM 的直接强化学习和 Q(λ) 算法相对于试验次数的逃避潜伏期, 计算 10 次实验结果求取曲线误差棒图, 仿真实验结果如图 7 所示. 可见, 两种算法均能够有效地对面向目标的导航任务进行学习. SRM-RL 算法直接计算奖励信号的策略梯度, Q(λ) 算法是对奖励预测误差值的反复迭代, 而不是奖励值本身, 所以, SRM-RL 算法在局部导航中有着更好的收敛性和稳定性. 但 SRM-RL 算法由于脉冲神经网络的存在, 有着更大的计算量 [15]. SRM-RL 算法和 Q(λ) 算法的局限在于, 只有在资格迹确定的影响半径内, 学习算法才能影响学习速率. 当超过资格迹记录的时间半径后, 两种算法都不能进行有效的权值更新.

改变模型动作细胞的数量, 统计 5 次实验、20 次实验和 50 次实验的逃避潜伏期, 实验结果如图 8 所示, 在动作细胞数量由 60 个变化到 480 个的过程中, 逃避潜伏期与动作细胞在 360 个时没有明显的变化, 原因在于, 动作细胞对动作空间的表达是连续的, 数量的变化仅仅影响每个动作细胞在均匀分布中表达的动作的范围, 而其放电活动取决于位置细胞到动作细胞的放电权值, 动作的选择取决于

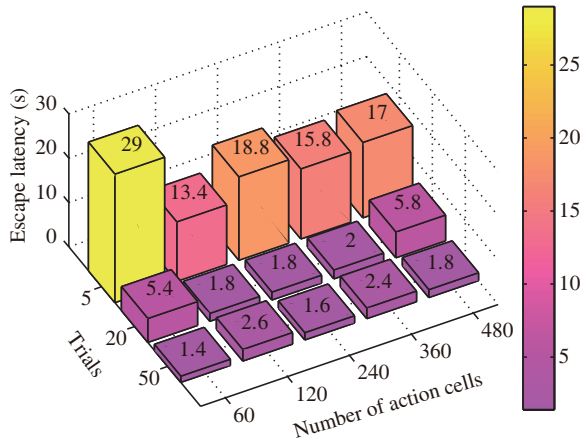

图 8　逃避潜伏期随实验次数和动作细胞数量变化的过程

Figure 8　The change of escape latency with the alteration of the numbers of trials and the action cells

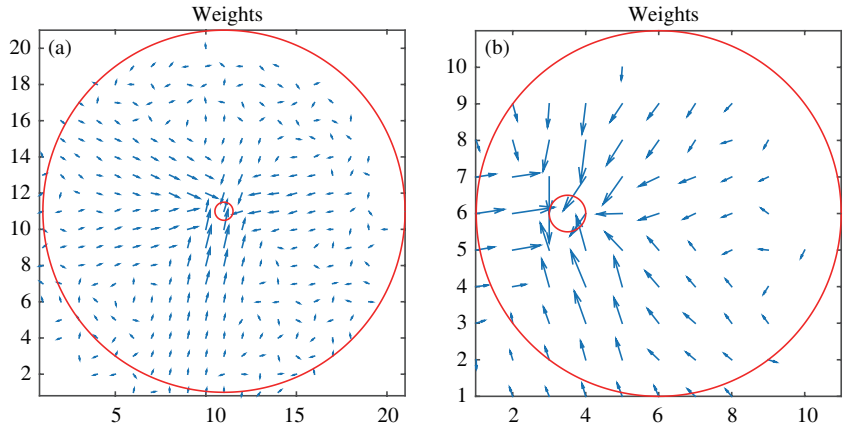

图 9　网络扩展和目标位置改变后的导航地图

Figure 9　The navigation map after change (a) the scaling properties and (b) the goal location of the network

动作细胞的群编码, 所以数量变化不影响动作的选择. 这与离散状态和动作空间的强化学习的表现相反, 离散状态下, 当动作和状态空间的数量增加时, 学习速率下降 [8].

改变位置细胞的数量, 在不改变位置野大小的情况下, 对状态空间同样进行密集型编码, 水迷宫环境大小发生了变化, 图 9(a) 为改变环境大小后, 30 次学习后的导航地图. 这说明在改变位置细胞数量后, 模型同样能够进行面向目标的导航学习活动. 图 9(b) 为改变目标位置, 进行 30 次训练后的导航地图, 实验结果表明, 模型能够在任意改变目标位置的情况下, 通过训练, 学习到面向目标位置的导航地图.

为了验证模型在复杂环境中的导航能力, 构建了一个有障碍物的水迷宫仿真环境进行仿真实验, 隐藏平台被障碍物包围, 只能由一个方向接近目标位置. 如图 10 所示, 为智能体在该环境中训练 30 次的导航地图, 导航地图显示, 在越接近目标点的位置, 权值越大, 且权值方向避开了包围目标位置的障碍物. 实验结果表明, 导航模型能够使智能体在有障碍物的环境中实现面向目标位置的无碰撞运动, 以最快的路径到达目标位置, 说明了模型在有障碍物的环境中的有效性.

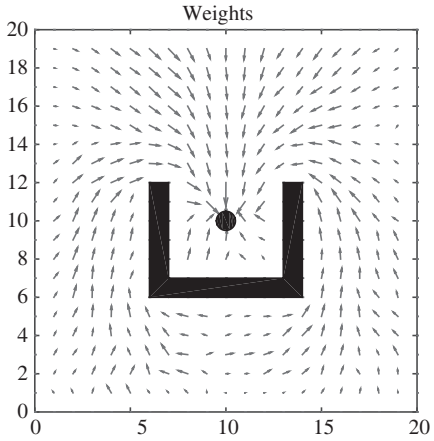

图 10 有障碍物环境中的导航地图

Figure 10 Navigation map with obstacle environment

5 结论

本文根据海马体到前额叶皮层的生理学研究, 构建位置细胞到假设动作细胞的脉冲神经网络模型, 在无先验知识的条件下, 在连续的状态和动作空间中进行面向目标位置的导航, 其中, 状态空间由位置野构成, 动作空间由动作细胞所代表的运动方向表示. 本文使用基于脉冲响应模型的直接强化学习, 调节位置细胞到动作细胞的突触连接权值, 构成整个状态空间的导航地图. 仿真实验结果表明, 该模型能够有效的学习到连续状态和动作空间面向目标位置的导航策略, 所采用的方法在收敛性上优于传统的强化学习方法. 在改变模型中位置细胞和动作细胞的数量、目标位置以及在环境中加入障碍物后, 模型也能够有稳定的表现, 能够实现在连续状态和动作空间中的有效的学习和导航活动.

模型采用脉冲神经网络, 更加切合生物学事实, 但也使得整个模型的计算量有所增加. 仿真实验中, 初始位置和目标位置是随机给定的二维坐标, 使得模型输入信息不是智能体完全自主探测所得. 因此, 对模型的后续改进将集中在基于海马认知机理对环境的表达上, 使模型能够更加自主的探索和生成对环境的认知地图, 从而实现更加智能化的局部导航.

参考文献

1 Packard M, McGaugh J. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. Neurobiol Learn Mem, 1996, 65: 65–72
2 Franz M O, Mallot H A. Biomimetic robot navigation. Robot Auton Syst, 2000, 30: 133–153
3 O'Keefe J, Dostrovsky J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res, 1971, 34: 171–175
4 Wilson M A, McNaughton B L. Dynamics of the hippocampal ensemble code for space. Science, 1993, 26: 1055–1058
5 Wagatsuma H, Yamaguchi Y. Neural dynamics of the cognitive map in the hippocampus. Cogn Neurodynamics, 2007, 1: 119–141
6 O'Keefe J, Nadel L. The Hippocampus as a Cognitive Map. Oxford: Clarendon Press, 1978
7 Morris R, Garrard P, Rawlins J, et al. Place navigation impaired in rats with hippocampal lesions. Nature, 1982, 297: 681–683
8 Sutton R S, Barto A G. Reinforcement Learning: an Introduction. Cambridge: MIT Press, 1998
9 Barrera A, Weitzenfeld A. Biologically-inspired robot spatial cognition based on rat neurophysiological studies. J Auton Robots, 2008, 25: 147–169

10 Hasselmo M E, Eichenbaum H B. Hippocampal mechanisms for the context-dependent retrieval of episodes. Neural Netw, 2005, 18: 1172–1190
11 Barrera A, Tejera G, Llofriu M, et al. Learning spatial localization: from rat studies to computational models of the hippocampus. Spat Cogn Comput, 2015, 15: 27–59
12 Sheynikhovich D, Chavarriaga R, Strosslin T, et al. Spatial representation and navigation in a bio-inspired robot. Lecture Notes Comput Sci, 2005, 3575: 245–264
13 Strosslin T, Sheynikhovich D, Chavarriaga R, et al. Robust self-localisation and navigation based on hippocampal place cells. Neural Netw, 2005, 18: 1125–1140
14 Andrew A M. Spiking neuron models: single neurons, populations, plasticity. Encyclopedia Neurosci, 2002, 4: 277–280
15 Queiroz M S D, Berrêdo R C D, Braga A D P. Reinforcement learning of a simple control task using the spike response model. Neurocomputing, 2006, 70: 14–20
16 Foster D, Morris R, Dayan P. Models of hippocampally dependent navigation using the temporal difference learning rule. Hippocampus, 2000, 10: 1–16
17 Jeffery K, Hayman R. Plasticity of the hippocampal place cell representation. Rev Neurosci, 2004, 15: 309–331
18 Muller R U, Kubie J L. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. J Neurosci, 1987, 7: 1951–1968
19 Etienne A, Jeffery K. Path integration in mammals. Hippocampus, 2004, 14, 180 –192
20 Wallace D, Gorny B, Whishaw I. Rats can track odors, other rats, and themselves: implications for the study of spatial behavior. Behav Brain Res, 2002, 131: 185–192
21 Hyman J M, Zilli E A, Paley A M, et al. Working memory performance correlates with prefrontal-hippocampal theta interactions but not with prefrontal neuron firing rates. Front Integr Neurosci, 2010, 4: 2
22 Schultz W, Dayan P, Montague P R. A neural substrate of prediction and reward. Science, 1997, 275: 1593–1599
23 Schultz W. Predictive reward signal of dopamine neurons. J Neurophysiol, 1998, 80: 1–27
24 Sesack S R, Pickel V M. In the rat medial nucleus accumbens, hippocampal and catecholaminergic terminals converge on spiny neurons and are in apposition to each other. Brain Res, 1990, 527: 266–279
25 Morris R, Garrard P, Rawlins J, et al. Place navigation impaired in rats with hippocampal lesions. Nature, 1982, 297: 681–683
26 Pearce J M, Roberts A D L, Good M. Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors. Nature, 1998, 396: 75–77
27 Vasilaki E, Frémaux N, Urbanczik R, et al. Correction: spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. Plos Comput Biol, 2009, 5: 1–17
28 Gerstner W. The spike response model. mNN4, 1999, 9: 1–51
29 Pfister J P, Toyoizumi T, Barber D, et al. Optimal spike-timing dependent plasticity for precise action potential firing in supervised learning. Neural Comput, 2006, 18: 1309–1339
30 Frémaux N. Reinforcement learning using a continuous time actor-critic framework with spiking neurons. Plos Comput Biol, 2013, 9: 570
31 Strosslin T, Gerstner W. Reinforcement learning in continuous state and action space. In: Reinforcement Learning. Berlin: Springer, 2003. 207–251
32 O'Keefe J, Recce M L. Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus, 1993, 3: 317–330
33 Arleo A, Gerstner W. Modeling rodent head-direction cells and place cells for spatial learning in bio-mimetic robotics. In: Meyer J A, Berthoz A, Floreano D, et al, eds. From Animals to Animats VI. Cambridge: MIT Press, 2000. 236–245

Biological plausible goal-directed navigation model based on direct reinforcement learning algorithm

Naigong YU[1,2], Ti LI[1,2*] & Lue FANG[1,2]

1 *College of Electronic and Control Engineering, Beijing University of Technology, Beijing* 100124*, China*;
2 *Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing* 100124*, China*
*E-mail: bgdliti@163.com

Abstract For the problem of the goal-directed navigation in continuous state and action space, a spiking neural network model from hippocampal place cells to the putative action cells in prefrontal cortex is proposed based on the firing characteristics of place cells and the information cycles in the hippocampus. The continuous state and action space are respectively characterized by a population of place cells and action cells, and the direct reinforcement learning algorithm combined with the spike response model has been used to goal-directed autonomous navigation. The simulation results in Morris watermaze task show that the algorithm used in the model can solve the problem of the goal-directed navigation in continuous state and action space, and obtain a better performance when compared to the classic methods based on temporal difference at the same given problem. When the number of the actions cell had been changed, the convergence of model remains the same. The model can still achieve the goal location, when the scale of the watermaze and the goal location had been changed.

Keywords direct reinforcement learning algorithm, place cells, action cells, spiking neural network, goal-directed navigation

Naigong YU was born in 1966. He received his Ph.D. degree from Beijing University of Technology, Beijing, China, in 2004. Currently, he is a professor at Beijing University of Technology. His research interest is in pattern recognition theory and application.

Ti LI was born in 1991. He received his B.E. degree from Jilin Agricultural University, Changchun, in 2013, and now he is pursuing his mater's degree in Beijing University of Technology, Beijing, China. His main research interests are pattern recognition and intelligent system.

Lue FANG was born in 1990. He received his B.E. degree in automation from Luoyang Institute of Science and Technology, Luoyang, in 2014. Currently, he is studying for a master's degree at Beijing University of Technology. His research interest is Bionic robot navigation.