基于 LSPI 和滚动窗口的移动机器人反应式导航方法

刘春明，李兆斌，黄振华，左磊，吴军，徐昕

(国防科技大学 机电工程与自动化学院，湖南 长沙，410073)

摘要：结合最小二乘策略迭代(Least-squares policy iteration，LSPI)的算法特性和基于滚动窗口的实时重规划，提出一种新的基于 LSPI 和滚动窗口的反应式导航学习控制方法。仿真和实验结果表明：该方法对移动机器人在未知环境中的运动控制有效，并且对未知环境具有自适应性。

关键词：移动机器人；反应式导航；增强学习；LSPI；滚动窗口
中图分类号：TP181 **文献标志码**：A **文章编号**：1672−7207(2013)03−0970−08

A reactive navigation method of mobile robots based on LSPI and rolling windows

LIU Chunming, LI Zhaobin, HUANG Zhenhua, ZUO Lei, WU Jun, XU Xin

(College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha 410073, China)

Abstract: Combining the advantages of least-squares policy iteration (LSPI) and path planning based on rolling windows, a novel reactive navigation method based on LSPI, and rolling windows was presented. The results show that the proposed method is effective for reactive navigation of mobile robots in unknown environment and the adaptation for unknown environments of the proposed method is verified.

Key words: reactive navigation; reinforcement learning; LSPI; rolling windows

自主式移动机器人是一种具有高度自规划、自组织、自适应能力，并且适合于在复杂的非结构化环境中工作的移动机器人。它的目标是在没有人的干预、无需对环境做任何规定和改变的条件下，有目的地移动并完成相应的任务。反应式导航是指基于传感器和执行器直接映射的一类运动控制方法，提高移动机器人在未知环境的实时性和灵活性的重要手段。目前已提出了多种移动机器人反应式导航控制方法，如模糊逻辑方法[1]、神经网络方法[2]，模糊神经网络方法[3]等。但已有方法往往要求较多的先验知识，如何构造和优化移动机器人的反应式导航控制器以及提高它对未知环境的适应性，仍然是有待解决的关键技术问题。近年来，国内外学者已开始重视利用增强学习(Reinforcement learning)与近似动态规划方法来研究解决移动机器人自主导航控制问题[4−7]。利用增强学习方法解决移动机器人自主导航控制问题主要有以下优点[8−9]：一是增强学习与监督学习不同，不需要专门的教师信号；二是增强学习在马尔可夫决策过程(Markov decision process，MDP)的模型框架下能有效地利用动态规划的算法和理论对问题进行求解；三是增强学习能够直接地利用观测数据建立从状态到动作的优化控制策略。目前，基于增强学习的移动机器人自主导航控制研究已取得了一些研究成果，例如，采用改进 Q 学习算法与模糊逻辑的导航控制[10]，基于神经网络增

收稿日期：2012−02−21；**修回日期**：2012−04−25
基金项目：国家自然科学基金资助项目(61075072，90820302)；教育部新世纪优秀人才支持计划(NCET-10-0901)
通信作者：刘春明(1981−)，男，黑龙江讷沙人，博士研究生，从事模式识别与智能系统的研究；电话：1397491442；E-mail: lummm@126.com

强学习算法的移动机器人自主导航[11−13]等。但这些方法都存在梯度学习算法的局部极值问题，并且算法的收敛性和泛化性能等问题仍然有待进一步研究解决。而最小二乘策略迭代(Least-squares policy iteration, LSPI)算法[14]能够通过观测数据直接对策略进行评价和改进，已成为增强学习中的一类重要的学习方法。基于滚动窗口的路径规划方法能使得移动机器人在行进的过程中在线地根据传感器信息进行重规划，获得新的次目标点，因此，针对未知环境中移动机器人的导航与控制问题，本文作者提出了一种新的基于 LSPI 和滚动窗口的移动机器人反应式导航方法。

1　增强学习中的 LSPI 算法概述

在增强学习中，整个系统环境被看作为一个可以用四元组 $\{S,A,R,P\}$ 来表示的 MDP 模型，其中 S 是有限状态集，A 是有限动作集，P 是状态转移概率，R 是回报函数。对一个 MDP 模型，策略 π 定义为 $\pi: S\to Pr(A)$，$Pr(A)$ 是动作集的概率分布所组成的集合。增强学习方法的目标就是对于一个 MDP 模型，获得最优策略 π^* 来满足下式：

$$J^* = \max_{\pi} J_\pi = \max_{\pi} E_\pi[\sum_{t=0}^{\infty}\gamma^t r_t] \tag{1}$$

其中，$\gamma\in[0,1)$ 为折扣因子；r_t 为单步回报；$E_\pi[\cdot]$ 为关于策略 π 的期望；J_π 为关于策略 π 的期望折扣总回报。

关于策略 π 的状态−动作值函数 $Q^\pi(s,a)$ 定义为在状态 s 下采取动作 a，并且后续动作都按照策略 π 执行，所获得的期望折扣总回报。那么 $Q^\pi(s,a)$ 应满足 Bellman 方程[14]：

$$Q^\pi(s,a) = \sum_{s'}[p(s,a,s')r(s,a,s')] + \gamma\sum_{s',a'}\left[p(s,a,s')\pi(a'|s')Q^\pi(s',a')\right] \tag{2}$$

其中，$p(s,a,s')$ 表示在状态 s 下采取动作 a 转移到状态 s' 的概率；$r(s,a,s')$ 表示在状态 s 下采取动作 a 转移到状态 s' 的单步回报；$\pi(a'|s')$ 表示策略 π 在状态 s' 下采取动作 a' 的概率。如果策略 π 在任意状态 s 下都存在一个动作 a，使得采取这个动作的概率为 1，采取其他动作的概率为 0，那么称策略 π 为确定性的策略，并记 $a=\pi(s)$。本文中讨论的策略均为确定性的策略。

对任意一个 MDP 模型，都存在一个最优策略 π^*，使得在状态 s 下的期望折扣总回报最大：

$$\pi^*(s) = \arg\max_a Q^{\pi^*}(s,a) \tag{3}$$

策略迭代就是通过不断地迭代改善已有策略来得到最优策略的过程，每一次迭代包括 2 个步骤：第 1 步是计算当前 t 时刻的状态−动作值函数 $Q^{\pi[t]}(s,a)$，第 2 步是通过贪婪的方法改善策略 $\pi[t]$：

$$\pi[t+1](s) = \arg\max_a Q^{\pi[t]}(s,a) \tag{4}$$

这 2 个步骤反复进行直到前后两次策略 $\pi[t]$ 和 $\pi[t+1]$ 没有差别为止。策略迭代收敛以后，就得到了最优策略。

然而对于大规模或者连续状态空间，需要对关于策略 π 的状态−动作值函数 $Q^\pi(s,a)$ 进行逼近。为了解决这样的问题，通常采用一组线性结构的逼近器进行逼近，即用 $m\cdot n_a$ 个状态−动作基函数对状态−动作值函数 $Q^\pi(s,a)$ 进行线性加权表示：

$$\begin{cases}\boldsymbol{\phi}(s,a) = \left(\underbrace{0,\cdots,0}_{m\cdot(l-1)},\phi_1(s),\cdots,\phi_m(s),\underbrace{0,\cdots,0}_{m\cdot(n_a-l)}\right)^{\mathrm{T}}\\ \hat{Q}^\pi(s,a,\boldsymbol{w}) = \boldsymbol{\phi}(s,a)^{\mathrm{T}}\boldsymbol{w}\end{cases} \tag{5}$$

其中，n_a 是动作的个数，动作 a 被标记为第 l 个动作；$\boldsymbol{w}=\left[w_1,\cdots,w_{m\cdot n_a}\right]^{\mathrm{T}}$ 为权值向量；$\phi_i(s)(i=1,\cdots,m)$ 为状态基函数；$\boldsymbol{\phi}(s,a)$ 为状态−动作基函数向量。

若给定一个样本集：

$$D=\{(s_i,a_i,s_i',r_i)|i=1,\cdots,L\} \tag{6}$$

令：

$$\boldsymbol{\Phi} = [\phi(s_1,a_1)\cdots\phi(s_L,a_L)]^{\mathrm{T}} \tag{7}$$

$$\boldsymbol{R}_{\mathrm{e}} = (r_1\cdots r_L)^{\mathrm{T}} \tag{8}$$

$$\boldsymbol{\Phi}' = [\phi(s_1',\pi[t](s_1'))\cdots\phi(s_L',\pi[t](s_L'))]^{\mathrm{T}} \tag{9}$$

则 LSPI 算法采用如下的迭代求解过程[14]：

$$\begin{cases}\boldsymbol{w}^{\pi[t]} = (\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi}-\gamma\boldsymbol{\Phi}')^{-1})\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{R}_{\mathrm{e}}\\ \pi[t+1](s) = \arg\max_a(\boldsymbol{\phi}(s,a)^{\mathrm{T}}\boldsymbol{w}^{\pi[t]})\end{cases} \tag{10}$$

由式(10)可知：LSPI 算法不依赖于系统的环境模型，它通过观测数据直接对策略进行评价和改进，有利于移动机器人在未知环境中完成导航控制任务。

2 基于 LSPI 和滚动窗口的反应式导航方法

由于 LSPI 算法是针对 MDP 模型逼近最优策略的一类有效方法，因此，首先把移动机器人的行为决策建模为 MDP 模型，然后再结合基于滚动窗口的路径规划方法，就得到了基于 LSPI 和滚动窗口的移动机器人反应式导航方法。

2.1 移动机器人行为决策过程的 MDP 建模

关于将移动机器人的行为决策过程建模为一个 MDP 模型的方法，一些学者进行了专门的研究[15−16]。本文结合移动机器人的 MDP 特性，即移动机器人在移动的过程中，下一时刻点所处的位姿只与当前的位姿和在当前位姿状态下采取的动作有关，而与其他的信息无关，所以可以直接建模为 MDP 模型。

MDP 模型的状态集定义为环境信息(距离)所组成的集合，动作集定义为速度和朝向角构成的向量所组成的集合。为使移动机器人始终处于安全位置，算法应该最大化机器人与障碍物之间的距离。表 1 所示为 5 种移动机器人与障碍物之间的位置关系。

表 1 移动机器人与障碍物之间的位置关系

Table 1 Positional relationship between mobile robot and obstacles

条件	表 达	说 明
1	$d\le D_{\mathrm{us}}$	机器人与障碍物之间的距离小于等于紧急停车距离
2	$d_{\mathrm{l}}\le D$ $d_{\mathrm{r}}\le D$ $d>D_{\mathrm{us}}$	机器人与左、右侧障碍物之间的距离均小于等于安全避障距离
3	$d_{\mathrm{l}}>D$ $d_{\mathrm{r}}\le D$ $d>D_{\mathrm{us}}$	机器人与左侧障碍物之间的距离小于等于安全避障距离， 机器人与右侧障碍物之间的距离大于安全避障距离
4	$d_{\mathrm{l}}>D$ $d_{\mathrm{r}}>D$ $d>D_{\mathrm{us}}$	机器人与左侧障碍物之间的距离大于安全避障距离， 机器人与右侧障碍物之间的距离小于等于安全避障距离
5	$d_{\mathrm{l}}\le D$ $d_{\mathrm{r}}\le D$ $d>D_{\mathrm{us}}$	机器人与左、右侧障碍物之间的距离均大于安全避障距离

注：D_{us} 为紧急停车距离；D 为安全避障距离，且 $D>D_{\mathrm{us}}$；d 为机器人与障碍物之间的距离；d_{l} 为机器人与左侧障碍物之间的距离；d_{r} 为机器人与右侧障碍物之间的距离。

针对上述 5 种不同情况，回报函数的设计如下：

$$r_{\mathrm{t}}=\begin{cases}-1\,000 & \text{条件1}\\ -k(D-d_{\mathrm{l}})-k(D-d_{\mathrm{r}}) & \text{条件2}\\ -k(D-d_{\mathrm{l}}) & \text{条件3}\\ -k(D-d_{\mathrm{r}}) & \text{条件4}\\ 0 & \text{条件5}\end{cases}\tag{11}$$

其中，k 为比例常数。

2.2 基于滚动窗口的路径规划方法

为了使基于 LSPI 的导航算法能够引导移动机器人顺利到达目标，提高导航效率，引入基于滚动窗口的路径规划算法。基于滚动窗口的路径规划算法是一类针对未知环境的移动机器人路径规划方法[17]。该方法充分利用移动机器人探测到的局部环境信息，以滚动窗口的方式进行在线实时重规划，实现了反馈与优化的合理结合。

假设机器人只能探测到以当前位置 p_t 点为中心，r_w 为半径的局部环境信息，并且机器人按照其动力学约束行进的步长为 $\varepsilon(0<\varepsilon\le r_w)$，那么 $\mathrm{Win}(p_t)=\{p\,|\,p\in W,d(p,p_t)\le r_w\}$ 为机器人在 t 时刻的视野区域，即滚动窗口，其中 W 为机器人的工作区域，$d(p,p_t)$ 为点 p 和点 p_t 之间的距离。

基于滚动窗口的路径规划算法主要有以下 2 个步骤：(1) 刷新当前窗口信息，生成局部优化次目标点；(2) 根据局部次目标点以及探测到的环境信息，生成到达次目标点的可通行路径。

在算法中，若目标点位于当前滚动窗口内，则设置目标点为次目标点，否则，依据启发式函数 $f(p)=g(p)+h(p)$ 来选择使得 $f(p)$ 最小的点 p 为次目标点，即：

$$\min_{p} f(p)=g(p)+h(p)$$
$$\text{s.t.}\quad p\in W\tag{12}$$

其中，$h(p)$ 为机器人从点 p 到目标点的代价，由于探测区域外的信息对机器人未知，所以可以用点 p 到目标点的实际距离来估计；$g(p)$ 为从当前位置 p_t 到达点 p 的代价。

文献[17]证明滚动窗口算法的可行性以及收敛性。

2.3 基于 LSPI 和滚动窗口的反应式导航算法

当传感器的探测距离 $d\le D$ 时，由于 LSPI 算法的学习目标是最大化期望折扣总回报，也就是移动机器人最大化自己与障碍之间的期望折扣总距离，所以这样的避障行为可能导致偏离目标的结果。若将 LSPI 学习算法与基于滚动窗口的路径规划方法相结合，就

可以较好地实现移动机器人在确保安全的状态下到达目标点，提高导航效率，从而完成任务。基于 LSPI 和滚动窗口的移动机器人反应式导航算法描述如下。

(1) 初始化样本数为 0；

(2) 初始化环境和移动机器人的位姿；

(3) 循环：

1) 刷新移动机器人的位姿信息；

2) 运用滚动窗口法生成次目标点；

3) 向着次目标点产生动作并执行：如果移动机器人与障碍之间的距离大于安全避障距离，则不需要避障，直接进行趋向于次目标点的控制；否则进行避障控制，随机选择避障动作并收集样本；

4) 如果样本数目足够多，采用 LSPI 学习算法生成避障策略，跳至(4)；否则，样本数加 1；

5) 如果已经到达目标、紧急停车，或者达到最大允许的步数时，跳至(2)；

(4) 将 LSPI 学习算法得到的避障策略与基于滚动窗口的目标趋向性控制策略相结合，得到了基于 LSPI 和滚动窗口的反应式导航策略。

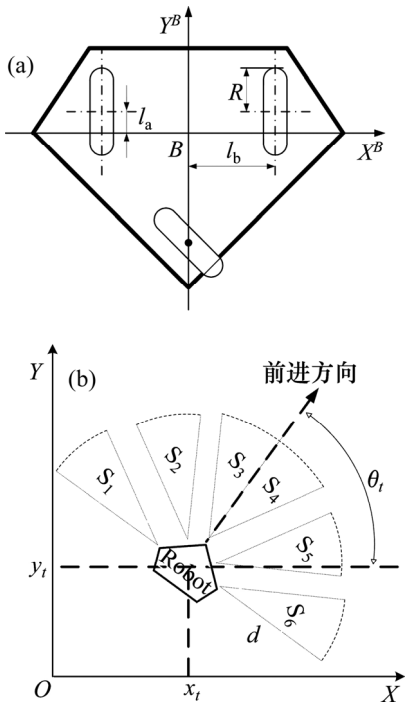
从算法中可以看出：基于 LSPI 和滚动窗口的反应式导航算法在样本采集的过程中，如果移动机器人与障碍物之间的距离大于安全避障距离时，直接采取基于滚动窗口的目标趋向性控制，否则随机选择避障动作并生成样本；当算法获得了足够多的样本后，将采用 LSPI 学习算法生成避障策略，并把该避障策略与基于滚动窗口的目标趋向性控制相结合，从而得到了基于 LSPI 和滚动窗口的反应式导航策略。该策略既能够使得移动机器人在移动的过程中与障碍物保持一定的安全距离，又能够使得移动机器人对目标进行趋向性控制。由于基于 LSPI 和滚动窗口的反应式导航算法既发挥了 LSPI 学习算法具有良好的收敛性、泛化性及不依赖于环境模型的优点，又具有滚动窗口的实时重规划特性，所以它能够使得移动机器人在未知的环境中提高导航效率，并且顺利地完成导航任务。

3　仿真研究

针对未知环境的移动机器人导航控制问题，本节利用 SimRobot 仿真环境[18]对基于 LSPI 和滚动窗口的反应式导航方法进行性能评估。

3.1　移动机器人描述

如图 1(b)所示，移动机器人 t 时刻在全局坐标系 OXY 下的位姿定义为$[x_t, y_t, \theta_t]$，其中 x_t 和 y_t 分别为移动机器人在全局坐标系下的横坐标和纵坐标，θ_t 为机器人的前进方向与 OX 轴正向的夹角。该移动机器人具有 1 个万向随动轮和 2 个驱动轮，2 个驱动轮的角速度均可控，如图 1(a)所示。它的正前方装有 6 个超声传感器，每个传感器的探测距离为 d，探测角度为 30°，如图 1(b)所示。



(a) 机器人的运动机构；(b) 距离传感器的配置

图 1　移动机器人系统俯视图

Fig.1　Top view of mobile robot system

移动机器人的车体参数如下：车轮半径 R=1 m，车轮左右对称，重心到车轮的距离 l_a 为 3 m，重心到车轴中心的距离 l_b 为 0 m，w_l 和 w_r 分别为左右车轮的滚动角速度，每次动作的作用时间为 0.1 s。机器人行进时的运动学约束在车体坐标系 BX^BY^B 下表现为车体速度：

$$\begin{cases} v_x = (R/(2\cdot l_a))\cdot(-l_b\cdot w_l + l_b\cdot w_r) \\ v_y = (R/(2\cdot l_a))\cdot(-l_a\cdot w_l - l_a\cdot w_r) \end{cases} \tag{13}$$

转换到全局坐标系 OXY 下为：

$$\begin{cases} \dot{x}_t = -v_x\cdot\sin\theta_t - v_y\cdot\cos\theta_t \\ \dot{y}_t = v_x\cdot\cos\theta_t - v_y\cdot\sin\theta_t \end{cases} \tag{14}$$

转向角速度为：

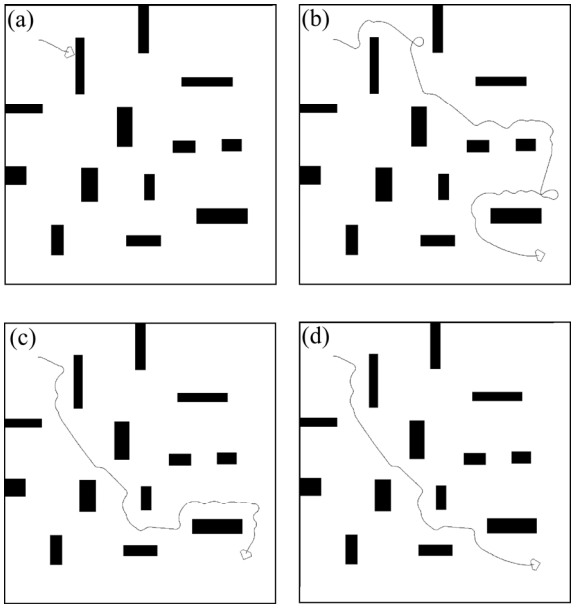
$$\dot{\theta}_t = (R/(2\cdot l_a))\cdot(-w_l + w_r) \tag{15}$$

3.2　仿真实验

在仿真实验中，MDP 模型的状态定义为由 6 个超

声雷达探测到的环境信息(距离)所组成的数值向量，动作定义为左右车轮角速度的固定组合：1) [0.5, 0.5], 2) [0，0.5]，3) [0.5，0]。滚动窗口方法中的安全避障距离 D 为 8 m，滚动窗口半径 r_w 为 8 m，紧急停车距离 D_{us} 为 0.2 m。当任意超声传感器满足条件 1 时，采取紧急制动措施，且获得的回报值为−1 000。LSPI 算法中的状态基函数设置为状态的四次多项式，折扣因子 γ=0.9，回报函数中的比例常数 k=0.9。

采取随机控制策略的移动机器人几乎很难越过第一个障碍，如图 2(a)所示。当基于 LSPI 和滚动窗口的反应式导航算法的学习样本数增加到 2 500 时，移动机器人在图 2 所示的环境中能够始终保持与障碍之间的安全距离并且顺利到达目标，较好地完成自主导航任务，如图 2(b)所示。随着学习样本数的进一步增加，当学习样本数达到 5 000 时，移动机器人的导航策略进一步优化，可以得到较为理想的运行轨迹，如图 2(c)所示。当学习样本数达到 10 000 时，移动机器人可以“游刃有余”的在障碍之间穿行，以近似最优路径到达预定目标，如图 2(d)所示。



(a) 未学习；(b) 学习了 2 500 个样本；(c) 学习了 5 000 个样本；(d) 学习了 10 000 个样本

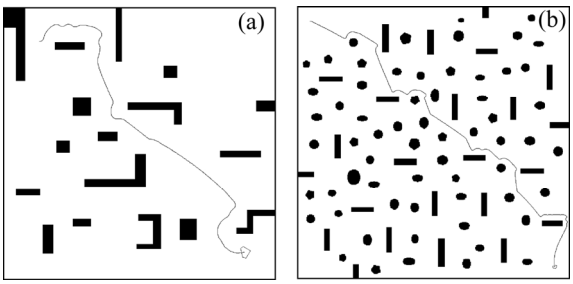
图 2 基于 LSPI 和滚动窗口的反应式导航算法性能

Fig.2 Performance of reactive navigation algorithm based on LSPI and rolling windows

为了检验基于 LSPI 和滚动窗口的反应式导航方法的泛化能力，这里把学习样本数为 10 000 时，所得到的导航控制策略分别在图 3(a)所示的全新的环境中和图 3(b)所示的 500 m×500 m 大范围地图环境中进行测试。从图 3 可以看出：基于 LSPI 和滚动窗口的反应式导航算法所获得的导航控制策略具有较强的泛化能力，在环境变化后依然可以得到较为理想的运动轨迹，在运动过程中也始终与障碍物保持一定的安全距离，并顺利地到达目标,而且还适应于大范围、更复杂的工作环境。

将基于 LSPI 和滚动窗口的反应式导航方法与传统的势场法进行了实验对比。众所周知，人工势场法在导航过程中难以克服一些死区，致使移动机器人无法到达目标，并且当移动机器人在障碍物或狭窄通道附近会产生抖动。如图 4(b)和图 5(b)所示，采用势场法的移动机器人难以克服环境中的局部极小点，不能完成导航控制任务。为了进行实验结果的对比，在同样的地图环境中对基于 LSPI 和滚动窗口的反应式导航算法进行了测试，结果如图 4(a)和图 5(a)所示。从实验结果可以看出：基于 LSPI 和滚动窗口的反应式导航方法可以克服局部死锁，顺利地到达目标。

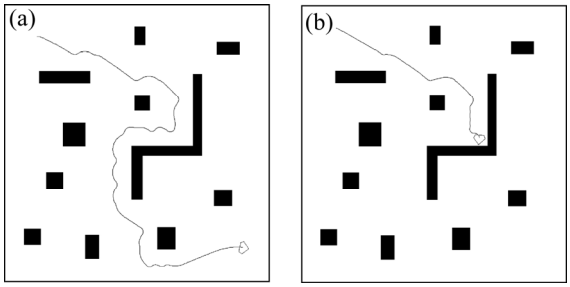
为进一步测试基于 LSPI 和滚动窗口的反应式导航方法与学习样本数之间的关系，本节还对学习结果的成功率和碰撞率进行了统计测试。所谓成功就是移



(a) 环境 1；(b) 环境 2

图 3 基于 LSPI 和滚动窗口的反应式导航算法的泛化性能

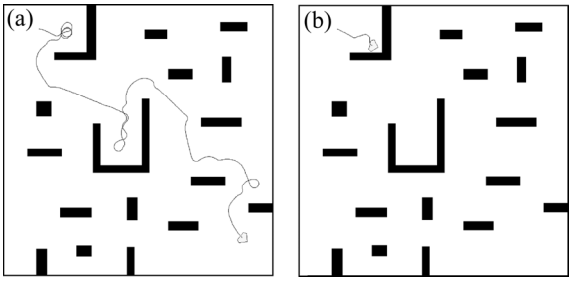
Fig.3 Generalization performance of reactive navigation algorithm based on LSPI and rolling windows



(a) 本文方法；(b) 传统势场法

图 4 本文方法与传统势场法的导航轨迹对比 1

Fig.4 Case 1 for navigation trajectory comparison between proposed method and traditional potential field method



(a) 本文方法；(b) 传统势场法

图 5　本文方法与传统势场法的导航轨迹对比 2

Fig.5　Case 2 for navigation trajectory comparison between proposed method and traditional potential field method

动机器人能在 2 000 步以内无碰撞的顺利到达目标，仿真中针对 10 种不同的地图环境进行了测试，每种环境测试 10 次，得到的成功率和碰撞率分别如图 6 和 7 所示。从图 6 和图 7 中可以看出，基于 LSPI 和滚动窗口的反应式导航方法经过 2 300 个样本的学习就可以得到较好的自主导航策略，能够在与障碍之间保持一定的安全距离(碰撞率为 0)的条件下顺利地到达目标。

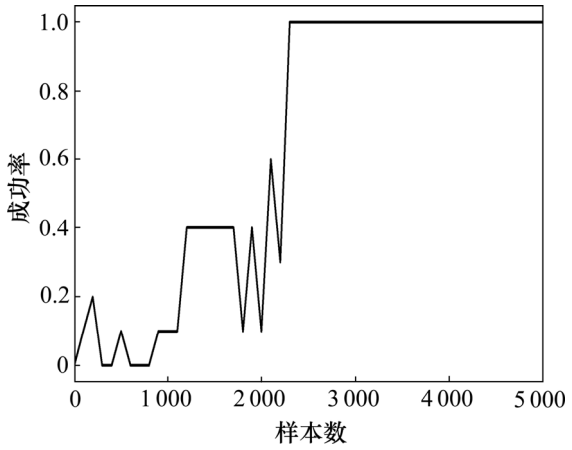


图 6　成功率与学习样本数之间的关系

Fig.6　Relationship between success rate and number of learning samples

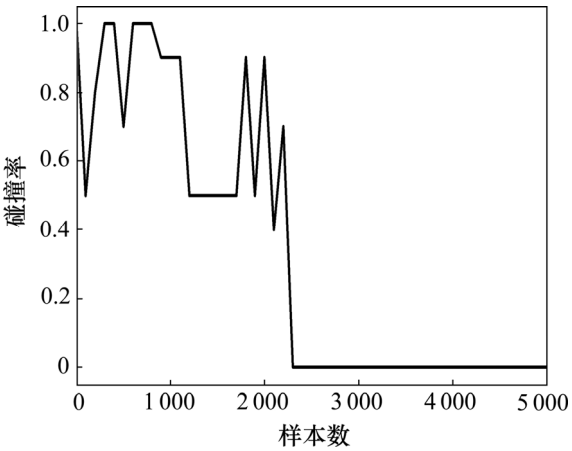


图 7　碰撞率与学习样本数之间的关系

Fig.7　Relationship between collision rate and number of learning samples

由以上仿真结果可以看出，基于 LSPI 和滚动窗口的反应式导航方法在未知的仿真环境中表现出较好的泛化性能。

4　实车实验

为了更好地对基于 LSPI 和滚动窗口的反应式导航方法进行性能测试，选用 MobileRobots 公司研制的 Pioneer3-AT(P3-AT)型移动机器人(如图 8 所示)进行实车实验。

P3-AT 型移动机器人的车身前部配置有 8 个超声波雷达传感器，其分布如图 9 所示。

MDP 模型的状态定义为车身前部的编号为 1~6 的雷达探测距离所组成的向量，动作定义为：1) 前行

图 8　Pioneer3-AT(P3-AT)型移动机器人

Fig.8　Pioneer3-AP(P3-AT) mobile robot

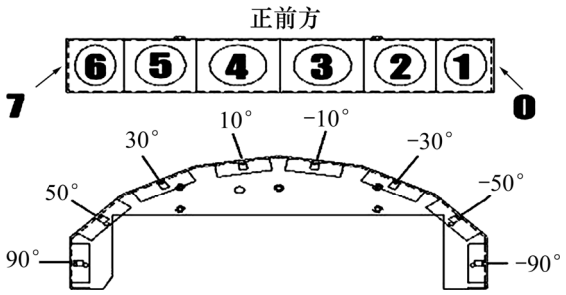


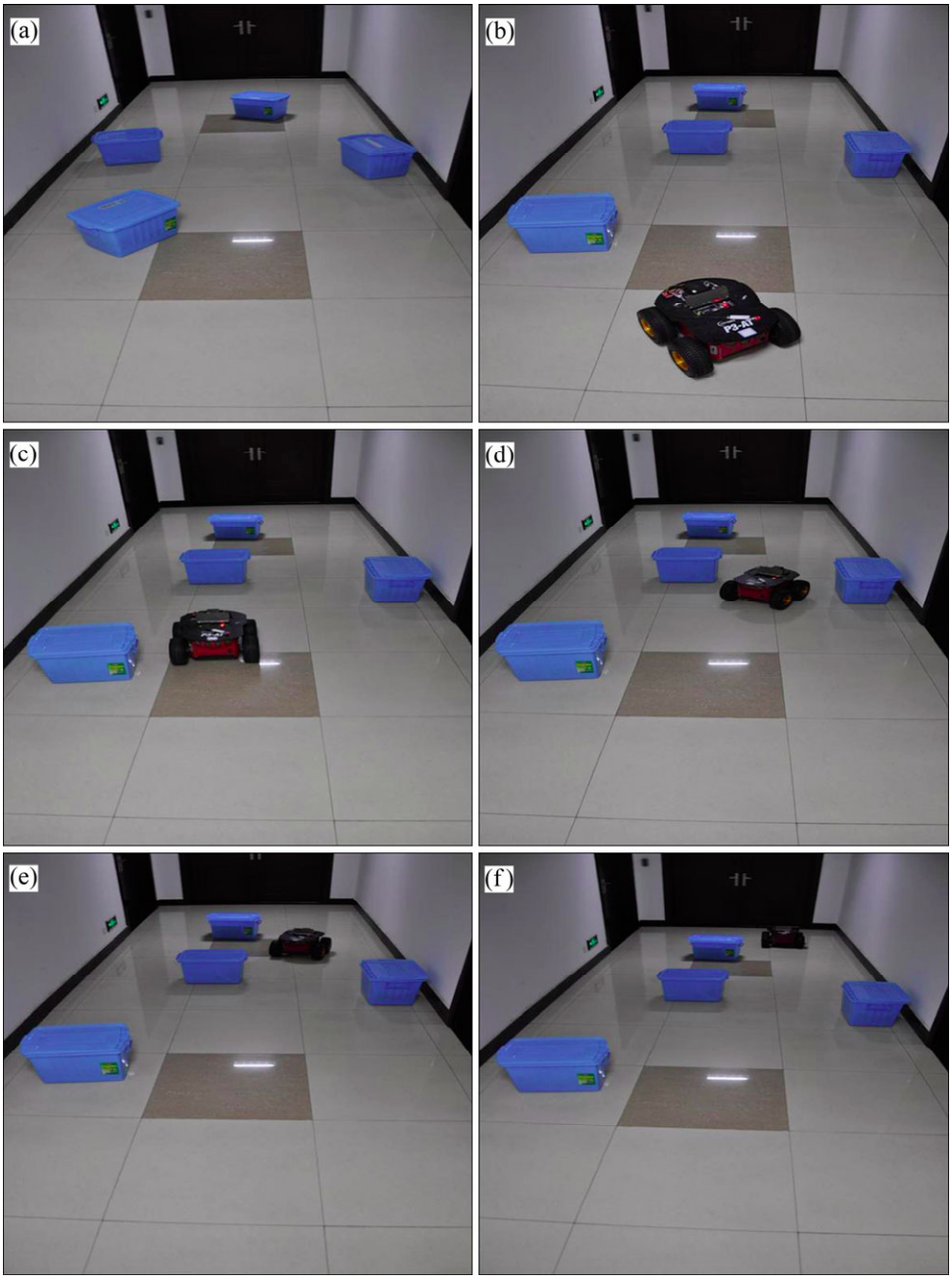
图 9　P3-AT 型移动机器人的雷达分布俯视图

Fig.9　Overlooking figure of radar distribution on P3-AT mobile robot

0.1 m；2) 左转 10°；3) 右转 10°。回报函数中的比例常数 k=0.9，安全避障距离 D 为 0.65 m，紧急停车距离 D_{us} 为 0.2 m；折扣因子 γ 为 0.9。LSPI 算法中的状态基函数设置为状态的四次多项式。基于 P3-AT 的样本采集环境与性能测试结果如图 10 所示。

P3-AT 型移动机器人在图 10(a)所示的样本采集环境中采集样本。当基于 LSPI 和滚动窗口的反应式导航算法的学习样本数增加到 2 000 时，对所获得的导航策略进行测试，如图 10(b)~(f)所示。移动机器人的初始位置如图 10(b)所示，目标点位置如图 10(f)所示，图 10(c)~(e)展示了移动机器人的导航过程。

从图 10 可以看出：在全新的测试环境中，移动机器人能够始终与障碍物保持一定的安全距离并且顺利地到达目标，即基于 LSPI 和滚动窗口的反应式导航算法在未知的实物环境中表现出了较好的泛化性能。



(a) 样本采集环境；(b) 初始点；(c) 算法测试过程 1；(d) 算法测试过程 2；(e) 算法测试过程 3；(f) 目标点

图 10 基于 P3-AT 的样本采集环境与性能测试

Fig.10 Sample collection environment and performance tests based on P3-AT mobile robot

5　结论

(1) 结合最小二乘策略迭代(LSPI)算法具有良好的收敛性和泛化性, 且不依赖于系统环境模型的特点, 以及基于滚动窗口的路径规划方法的实时重规划特能，提出了一种新的基于 LSPI 和滚动窗口的反应式导航方法。

(2) 通过 SimRobot 仿真实验平台以及基于 Pioneer3-AT 型移动机器人的实车实验结果表明该导航方法有效且具有很强的泛化能力。为提高移动机器人在未知环境中的自主行为能力提供了一种有效的技术手段。

(3) 进一步加快 LSPI 算法的学习过程，把已有的先验知识注入到该方法中，从而提高学习效率，获得更好的导航策略，是下一步值得研究的课题。

参考文献：

[1] Faress K N, EL Hagry M T, EL Kosy A A. Trajectory tracking control for a wheeled mobile robot using fuzzy logic controller[J]. WSEAS Transactions on Systems, 2005, 4(7): 1017−1021.

[2] Wang X, Hou Z, Zou A, et al. A behavior controller based on spiking neural networks for mobile robots[J]. Neurocomputing, 2008, 71(4/5/6): 655−666.

[3] Er M J, Tan T P, Loh S Y. Control of a mobile robot using generalized dynamic fuzzy neural networks[J]. Microprocessors and Microsystems, 2004, 28(9): 491−498.

[4] 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: 国防科学技术大学机电工程与自动化学院, 2002: 10−30.
XU Xin. Reinforcement learning and its applications in navigation and control of mobile robots[D]. Changsha: National University of Defense Technology. College of Mechatronics Engineering and Automation, 2002: 10−30.

[5] Busoniu L, Babuska R, de Schutter B. A comprehensive survey of multiagent reinforcement learning[J]. IEEE Transactions on Systems, Man and Cybernetics, 2008, 38 (2): 156−172.

[6] Carrersa M, Yub J K, Batlle J, et al. Application of SONQL for real-time learning of robot behaviors[J]. Robotics and Autonomous System, 2007, 55 (8): 628−642.

[7] Tan A H, Lu N, Xiao D. Integrating temporal difference methods and self-organizing neural networks for reinforcement learning with delayed evaluative feedback[J]. IEEE Transactions on Neural Networks, 2008, 19(2): 230−244.

[8] 徐昕. 增强学习与近似动态规划[M]. 北京: 科学出版社, 2010: 15−35.
XU Xin. Reinforcement learning and approximate dynamic programming[M]. Beijing: Science Press, 2010: 15−35.

[9] Xu X, Hu D W, Lu X C. Kernel based least-squares policy iteration[J]. IEEE Transactions on Neural Networks, 2007, 18 (4): 973−992.

[10] Boubertakh H, Tadjine M, Glorennec P Y. A new mobile robot navigation method using fuzzy logic and a modified Q-learning algorithm[J]. Journal of Intelligent and Fuzzy Systems, 2010, 21(1/2): 113−119.

[11] 乔俊飞, 樊瑞元, 韩红桂, 等. 机器人动态神经网络导航算法的研究和实现[J]. 控制理论与应用, 2010, 27(1): 111−115.
QIAO Junfei, FAN Ruiyuan, HAN Honggui, et al. Research and realization of dynamic neural network navigation algorithm for mobile robot[J]. Control Theory and Applications, 2010, 27(1): 111−115.

[12] Ma X L, Likharev K K. Global reinforcement learning in neural networks[J]. IEEE Transactions on Neural Networks, 2007, 18 (2): 573−577.

[13] Hafner R, Riedmiller M. Neural reinforcement learning controller for real robot application[C]//IEEE International Conference on Robotics and Automation. Roma: IEEE, 2007: 2098−2103.

[14] Michail G L, Parr R. Least-squares policy iteration[J]. Journal of Machine Learning Research, 2003, 4: 1107−1149.

[15] Belker T, Beetz M, Cremers A B. Learning action models for the improved execution of navigation plans[J]. Robotics and Autonomous Systems, 2002, 38(3/4): 137−148.

[16] Beetz M, Belker T. Learning Structured Reactive Navigation Plans from Executing MDP Navigation Policies[C]//Proceedings of the fifth international conference on Autonomous agents. New York, 2001: 19−20.

[17] 张纯刚, 席裕庚. 全局环境位置未知时基于滚动窗口的机器人路径规划方法研究[J]. 中国科学: E 辑, 2001, 31(1): 51−58.
ZHANG Chungang, XI Yugeng. Research on robot path planning methods based on rolling windows when the global environmental position is unknown[J]. Science in China: Series E, 2001, 31(1): 51−58.

[18] SIMROBOT Team. Autonomous mobile robotics toolbox SIMROBOT[EB/OL]. [2012−01−10]. http://www.uamt.feec.vutbr.cz/robotics/simulations/ amrt/simrobot.zip.

(编辑　赵俊)