Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2013, 2, 24-28
http://dx.doi.org/10.12677/airr.2013.21004 Published Online February 2013 (http://www.hanspub.org/journal/airr.html)

Multi-Robot Reinforcement Learning Based on LCS and LS-SVM*

Jie Shao[1,2], Lijuan Du[3], Haixia Lin[1]

[1]Department of Information Engineering, Zhengzhou Chenggong University of Finance and Economics, Zhengzhou
[2]School of Computer Science, Nanjing University of Science and Technology, Nanjing
[3]School of Information and Electonic, Shangqiu Institute of Technology, Shangqiu
Email: sj012328@163.com

Received: Jan. 5th, 2013; revised: Jan. 25th, 2013; accepted: Feb. 4th, 2013

Abstract: This paper presents a multi-robot reinforcement learning method combination LCS and LS-SVM, the optimal learning strategy LS-SVM obtained as an initial rule set of LCS. LCS interact with the environment, which can quickly find the guiding rules for multi-robot reinforcement learning, provide real-time, dynamic feedback, so that multi-robot autonomously learn the optimal strategy of mutual cooperation. Algorithm analysis and simulation show that a large space for multi-robot learning, the learning speed converges slowly, uncertainties and other learning problems can get a great improvement.

Keywords: Learning Classifier System; LS-SVM; Reinforcement Learning; Multi-Robot

基于 LCS 和 LS-SVM 的多机器人强化学习*

邵　杰[1,2]，杜丽娟[3]，林海霞[1]

[1]郑州成功财经学院信息工程系，郑州
[2]南京理工大学计算机科学与技术学院，南京
[3]商丘工学院信息与电子学院，商丘
Email: sj012328@163.com

收稿日期：2013 年 1 月 5 日；修回日期：2013 年 1 月 25 日；录用日期：2013 年 2 月 4 日

摘　要：本文提出了一种 LCS 和 LS-SVM 相结合的多机器人强化学习方法，LS-SVM 获得的最优学习策略作为 LCS 的初始规则集。LCS 通过与环境的交互，能更快发现指导多机器人强化学习的规则，为强化学习系统的动作选择提供实时、动态的反馈，使多机器人自主地学习到相互协作的最优策略。算法的分析和仿真表明多机器人学习空间大、学习速度收敛慢、学习效果不确定等问题得到很大的改善。

关键词：学习分类器；协同最小二乘支持向量机；强化学习；多机器人

1. 引言

机器人强化学习问题自提出至今已有众多学者做了多年的深入研究并产生了大量研究成果[1-4]。*Q* 学习方法作为一典型的强化学习方法，且不需要建立环境和任务的精确数学模型，已被广泛地应用于机器人领域[5]。但在多机器人的学习过程中，经常出现由于学习空间大、造成学习速度慢、学习效果不确定等问题[6]。

基于上述分析，本文提出了将 LCS 和 LS-SVM 结合用于解决多机器人的强化学习问题。LS-SVM 获得的最优学习策略作为 LCS 的初始规则集。LCS 通过与环境的交互，可以发现一组用于指导机器人学习的规则，为 LCS 系统的动作选择提供实时、动态的反馈，使机器人自主地学习到最优路径规划策略。

*基金项目：河南省教育厅重点资助项目(12B520047)。

2. 相关技术原理

2.1. 基于学习分类器的多机器人强化学习模型

多机器人强化学习的任务是利用环境回报来学习针对某个环境的最优策略。其原理是通过对感知的环境状态采取各种试探动作，获得此种试探动作对此种环境状态的回报，并能不断调整学习策略以获得较大的回报，同时反馈给其它机器人，最终获得最好协作的强化学习策略。

权值调整：Q 学习是著名的强化学习之一，该算法无需任何数学模型，Q 值和最优策略的算法更新如下：

Step1：LCS 系统观察当前状态 s_t，按一定的学习策略选取学习规则 a_t；

Step2：LCS 从环境中获得即时回报 r；

Step3：LCS 系统状态 s_t 由转换为 s_{t+1}，更新 Q 值。

$$\begin{aligned} &Q_{t+1} = Q_t\left(s_t, a_t\right) + Q_{t+1} \\ &= Q_t\left(s_t, a_t\right) + \beta^*\left(r + \gamma^* \max Q_{t+1}\left(s_{t+1}, a_{t+1}\right) - Q_t\left(s_t, a_t\right)\right) \end{aligned} \tag{1}$$

$$\pi\left(s_{t+1}, a_{t+1}\right) = \arg\max_{a_t \in A} Q\left(s_t, a_t\right) \tag{2}$$

其中 $\beta\left(0 \le \beta \le 1\right)$ 是学习率，$\gamma\left(0 \le \gamma \le 1\right)$ 是折扣系数。

2.2. 基于最小二乘支持向量机的 Q 学习

典型强化学习系统的映射关系包括：状态空间到动作空间的映射 $\left(S \to A\right)$、状态空间到回报函数的映射 $\left(S \to r\right)$、(状态，动作)对到值函数的映射 $\left(S \times A \to Q\right)$ 以及(状态，动作)对到转移函数的映射 $\left(S \times A \to P\right)$。王雪松等[7]提出了协同最小二乘支持向量机的 Q 学习算法。该学习系统由一个最小二乘支持向量回归机(Least squares support vector regression machine, LS-SVRM)和一个最小二乘支持向量分类机(Least squares support vector classification machine, LS-SVCM)，分别采用 LS-SVRM 和 LS-SVCM 实现对 $S \times A \to Q$ 和 $S \to A$ 映射的学习估计。

协同最小二乘支持向量机的 Q 学习过程如下[7]：

Step1：在环境的初始状态 s_t 下，LCS 系统直接选择执行由 LS-SVCM 给出的初始动作 a_t，并依据环境反馈信号 r_t，更新此时状态下的 Q_t。

Step2：Q_t 与 LCS 系统的 $\left(s_t, a_t\right)$ 组合后形成 LS-SVRM 训练样本集合 S1，对 LS-SVRM 模型进行在线训练；$\left(s_t, a_t\right)$ 则送入 LS-SVCM 的训练样本集 S2，对其进行在线训练。

Step3：在 $\left(\left(s_t, a_t\right), Q_t\right)$ 和 $\left(s_t, a_t\right)$ 加入样本集 S1 和 S2 时，需要对样本进行有效性判断，当实际动作执行前后系统的变化满足下列条件(3)时，$\left(\left(s_t, a_t\right), Q_t\right)$ 加入 S1，满足条件(4)时，$\left(s_t, a_t\right)$ 加入集合 S2。

$$\left\|s_{t+1} - s_t\right\| > \theta \tag{3}$$

$$\left\|s_{t+1} - s_d\right\| < \left\|s_t - s_d\right\| \tag{2}$$

其中，θ 为状态阈值，s_d 表示目标状态。

Step4：当 LS-SVRM 给出的动作选择与 LS-SVCM 给出的动作相同时，此时 LS-SVCM 失去作用，LS-SVRM 需对 Q 值进行在线估计，学习系统根据得到的数据 $\left(\left(s_t, a_t\right), Q_t\right)$，依据动作 a_t 被选择的概率 p，及时对 LS-SVRM 模型作出调整，重复上述 Step1、Step2、Step3 过程，直至学习到准确的最优学习策略。

$$p\left(\tilde{a} = a_t \middle| s_t\right) = \frac{\exp\left(\dfrac{Q\left(s_t, a_t\right)}{T_t}\right)}{\sum_a \exp\left(\dfrac{Q\left(s_t, a\right)}{T_t}\right)} \tag{5}$$

其中：$T_t > 0$ 为温度系数，控制动作选择的随机程度，一般在学习初期选择较高的温度，在学习的过程中逐渐降低温度，以保证较好的学习效果。

$$\begin{cases} T_0 = T_{\max} \\ T_{t+1} = T_{\min} + \beta\left(T_t - T_{\min}\right) \end{cases} \tag{6}$$

其中：$0 \le \beta \le 1$ 为退火因子，$T_{\max}$ 和 $T_{\min}$ 为设定的最大和最小温度。

3. 基于学习分类器和最小二乘支持向量机的 Q 学习

LCS[8,9]是一个集成强化学习和遗传算法的新型进化算法，其主要功能是如何通过信用分配和基于遗传算法的规则发现机制对环境不断学习，不断产生出新的学习规则。LCS 系统主要由规则与消息系统、基于信用分配系统的强化学习和基于遗传算法(GA)的规则发现系统三部分组成。

规则及消息系统可用 **IF**<Condition>**THEN** <ACTION>，strength 表示的特殊产生式，直接与外部

环境发生作用。信用分配系统用于调整现有 LCS 的学习规则强度，并对分配算法进行更新，同时产生三种 LCS 学习规则：匹配规则集中未中标的 LCS 学习规则、匹配规则集中竞标获胜的 LCS 学习规则、未匹配的 LCS 学习规则。

三种 LCS 学习规则强度的信用分配算法如下：

$$B_i(t)=\rho S_i(t)f_i \tag{7}$$

其中 $0<\rho\le 1$，投标风险系数。$S_i(t)$ 是当前学习规则的权值强度。$B_i(t)$ 为规则参与投标的投标值。f_i 表示规则参与投标的力度，$0\le f_i\le 1$。

$$f_i=l/L \tag{8}$$

其中 l 为规则条件部分中非“#”的位数，L_1 为规则条件部分的长度，$L=16$。

$$S_i(t+1)=(1-Tax_{\text{life}})S_i(t)+R_i(t)-Tax_{\text{bid}}*B_i(t) \tag{9}$$

Tax_{life} 为规则的生存税，防止其权值人为地增长。Tax_{bid} 为规则需支付的投标税，用于调节规则的权值强度。

3.1. 学习规则编码

机器人的学习策略规由三元组 $\langle 0,1,\#\rangle$ 的 22 位字符位串组成的 (s_t,a_t) 数据，前 16 位是学习规则的条件部分，后 6 位是学习规则的动作部分。机器人传感器的探测区域为一定偏转度的扇形区域，机器人共有 16 个声纳传感器，可感知每一个扇形区域中障碍物和机器人分布情况。学习规则编码系统用 6 位二进制代码表示机器人动作，其中 4 位代码表示转向角，1 位代码控制速度快慢，1 位代码表示前进或后退。

学习策略规则编码描述如下：

$$\langle \text{Condition}\rangle ::= \{0,1,\#\}^l$$

$$\langle \text{Action}\rangle ::= \{0,1\}^m$$

环境信息的编码格式为：

$$\langle \text{Condition}\rangle ::= \{0,1\}^k$$

其中 l,m 分别为规则的条件和动作部分的长度，这里 $l=16,m=6,k=16$。

3.2. 学习优化模型

本文的强化学习采取 ε-Greedy 探索策略，无限范围衰减模型，基于一定的衰减因子 $\gamma(0\le\gamma<1)$ 进行几何衰减，多机器人系统最优化的期望奖励值为：

$$\delta=\sum_{t=0}^{\infty}\gamma^t r_t \tag{10}$$

其中 r_t 代表将来的第 t 步接收到的奖励。

3.3. 适应度函数

遗传算法是 LCS 内嵌的集成算法，在 LCS 系统的每一次工作周期中，如果 LCS 受到了奖励或惩罚，权值分配系统都会相应调整匹配规则集中学习规则的强度，而权值强度则作为内嵌遗传算法的适应度函数来产生新的学习规则。

在不考虑环境其它因素的情况下，基于 LCS 和 LS-SVM 的多机器人强化学习的适应度函数为：

$$f=\delta*S_i(t+1) \tag{11}$$

3.4. 遗传算子策略的改进

LCS 系统在进化过程中，将权值最大的学习策略规则直接传递，能有效防止学习“早熟收敛”和“搜索迟钝”，同时引入冲突消解功能的竞争择优交叉操作，加快 LCS 系统学习的收敛速度，为滤除相似的个体，减小学习空间，采取学习策略规则合并的策略，并对遗传算法的交叉算子进行了如下改进：

Step1：在 LCS 的匹配规则集中选定两个父类分类器 $p_1=(a_1^{p_1},a_2^{p_1},\cdots,a_l^{p_1})$ 和 $p_2=(a_1^{p_2},a_2^{p_2},\cdots,a_l^{p_2})$，其中 $a_i^{p_i}=\left[l_{a_i}^{p_i},u_{a_i}^{p_i}\right]$。

Step2：设定一随机值 $\alpha\in[0,0.5]$。

Step3：计算父代分类器的最低下限 $c_{\min}^i=\min\left(l_{a_i}^{p_1},l_{a_i}^{p_2}\right)$ 和最高上限 $c_{\max}^i=\max\left(u_{a_i}^{p_1},u_{a_i}^{p_2}\right)$ 值，以及 $c_{\max}^i$ 和 $c_{\min}^i$ 的距离：$I=c_{\max}^i-c_{\min}^i$。

Step4：产生新分类器 o_1 和 o_2 的上限和下限值

$$l_{a_i}^{o_1}=c_{\min}+\alpha*I*\text{rand}(\{-1,1\})$$

$$u_{a_i}^{o_1}=c_{\min}+\alpha*I*\text{rand}(\{-1,1\})$$

$$l_{a_i}^{o_2}=c_{\min}+(1-\alpha)*I*\text{rand}(\{-1,1\})$$

$$u_{a_i}^{o_2}=c_{\min}+(1-\alpha)*I*\text{rand}(\{-1,1\})$$

其中函数 $\text{rand}(\{-1,1\})$ 的值为 1 或−1，当 α 的值较小时，产生的子代分类器与父代分类器很相似，适用于机器人局部搜索；当 α 的值较大时，产生的子代分类

器与父代分类器差异较大，适用于机器人对未知领域环境的探索。

3.5. 基于 LCS 和 LS-SVM 强化学习策略描述

Step1：最小二乘支持向量机的 Q 学习产生的最优学习策略作为 LCS 的初始规则集，并随机执行其中的学习策略。

Step2：调用拍卖模块判断当前学习策略的条件部分与消息列表中消息是否匹配，匹配的学习策略规则送入匹配规则集。

Step3：从匹配规则集中选择较优的学习策略规则送往效应器并产生相应动作，LCS 系统依据环境反馈值来判断学习效果。

Step4：重复 Step2~Step3 过程，直到匹配规则集为空。

Step5：启动进化(规则发现)系统构造新的学习策略规则。

经规则合并后，发送到 LCS 公共规则集中，供其它机器人 LCS 产生新学习规则时共享。

Step6：如果机器人未达到学习效果，则返回 Step2。

4. 实验及仿真结果

我们以多机器人路径规划效果来验证本文提出的混合学习策略。图 1 内 O1-O3 机器人均是用直径为 0.3 m 的圆柱体表示，各机器人根据自己的路况，可进行 0˚、左曲线 5˚、左曲线 10˚、左曲线 15˚、右曲线 5˚、右曲线 10˚、右曲线 15˚或停止当前动作的八种学习策略。活动区域内圆形和方形图形均为静态障碍物。图 2 为基于 LCS 学习的多机器人收敛学习曲线；图 3 是基于 LCS 和 LS-SVM 混合方法的多机器人的学习收敛曲线。与图 2 相比图 3 明显地提高了学习效率和收敛速度。

一般情况下，一个强化学习算法的性能需要两方面的判定，一个是算法的收敛性，一个是算法的收敛速度。从图 4、图 5 中可以看出，系统经过若干次数的学习后，算法出现收敛的趋势，在学习初始过程，$R(t)$ 的震荡幅度较大。经过若干次数的学习之后，$R(t)$ 振荡幅度就很小了，我们可以认为已经近似地收敛了。图 4 为基于传统 LCS 强化学习的收敛情况，体现了传统强化学习时间较长、收敛速度慢的缺陷；图 5 为基于梯度的 LCS 和 LS-SVM 的强化学习情况，不管是机器人个体强化学习或是团队学习，均能在较短的时间内收到理想的学习效果。

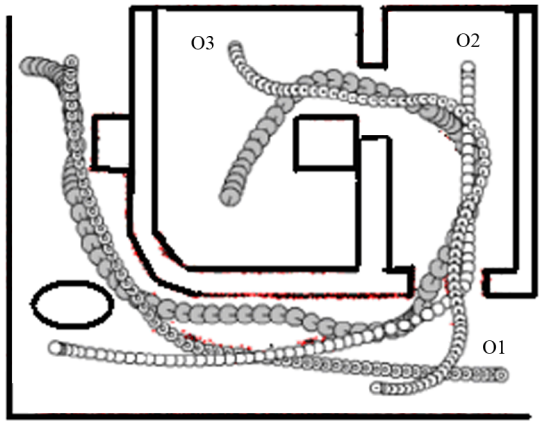

Figure 1. The evolution algebra and fitness curve based on LCS
图 1. 进化代数与适应度曲线图(LCS 学习)

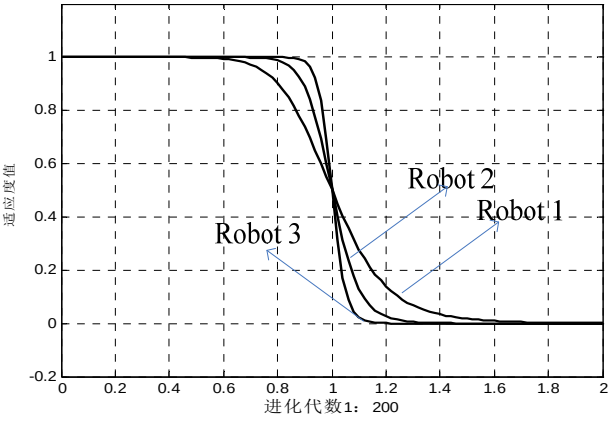

Figure 2. Reinforcement learning convergence based on the traditional LCS
图 2. 基于传统的 LCS 强化学习收敛情况

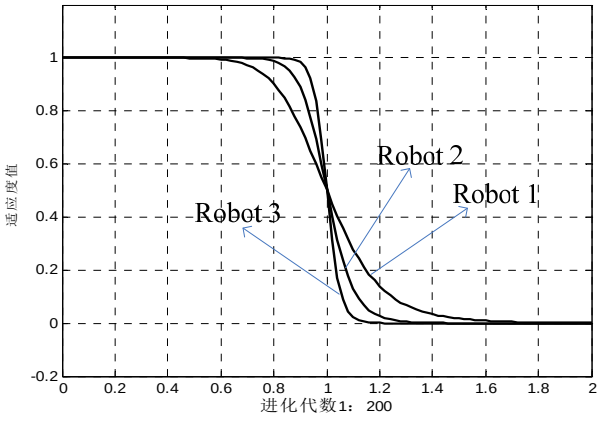

Figure 3. Algebra and fitness curves based on LCS and LS-SVM
图 3. 化代数与适应度曲线图(LCS 和 LS-SVM)

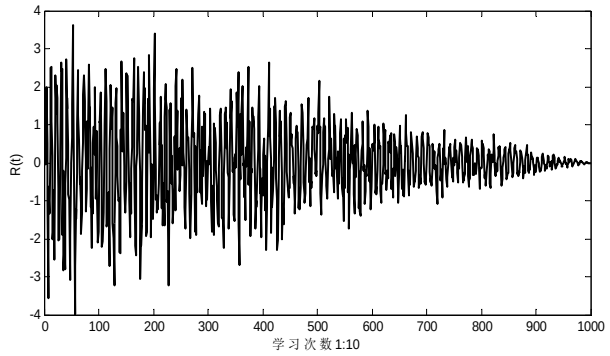

Figure 4. Reinforcement learning convergence based on the traditional LCS
图 4. 基于传统的 LCS 强化学习收敛情况

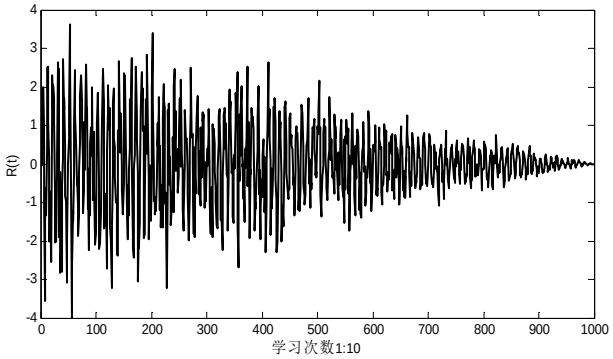

Figure 5. Reinforcement learning convergence based on LCS and LS-SVM
图 5. 基于 LCS 和 LS-SVM 的强化学习收敛情况

5. 结束语

本文提出了 LCS 系统和 LS-SVM 算法结合的混合算法，解决多机器人在复杂环境下的强化学习问题。LCS 内嵌的遗传算法采用了特有的遗传算子策略，加强了 LCS 系统的强化学习效果，由学习空间大造成的学习速度慢、学习效果不确定等问题得到很大的改善。仿真实验结果也表明将 LCS 用于多机器人的强化学习是有效的。

参考文献 (References)

[1] J. shao, J. Y. Yang. Multi-robot reinforcement learning based on learning classifier system with gradient descent methods. Journal of Computational Information Systems, 2010, 6(8): 2449-2455.

[2] 高阳, 陈世福, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004, 30(1): 86-100.

[3] 沈晶, 程晓北, 刘海波等. 动态环境中的强化学习[J]. 控制理论与应用, 2008, 25(1): 71-74.

[4] 邵杰, 杨静宇, 万鸣华, 黄传波. 基于学习分类器的多机器人路径规划收敛性研究[J]. 计算机研究与发展, 2010, 47(5): 948-955.

[5] 焦殿科, 石川. 共享经验的多主体强化学习研究[J]. 计算机工程, 2008, 34(11): 219-221.

[6] 陈卫东, 席玉庚, 顾东雷. 自主机器人的强化学习进展[J]. 机器人, 2001, 23(4): 379-384.

[7] 王雪松, 田西兰, 程玉虎, 易建强. 基于协同最小二乘支持向量机的 Q 学习[J]. 自动化学报, 2009, 35(2): 215-219.

[8] X.-L. Wang, Z.-J. Yin, Y.-B. Lv and S.-F. Li. Operating rules classification system of water supply reservoir based on learning classifier system. Expert Systems with Applications, 2008, 36(3): 5654-5659.

[9] P. Musilek. Enhanced learning classifier system for robot navigation. International Conference on Intelligent Robots and Systems, Edmonton, 2-6 August 2005: 3390-3395.