第22卷 第1期 自 动 化 学 报 Vol. 22, No.1
1996 年 1 月 ACTA AUTOMATICA SINICA Jan., 1996

短 文

联想记忆的自强式学习控制[1)]

许 力 蒋静坪 诸 静 朱炎新
(浙江大学电机系 杭州 310027)

摘 要

在 ASE/ACE 模型中引入监督式学习的局部性网络作联想记忆机制，从而构造一种新的自强式学习控制器模型 RELCAM．仿真结果表明，使用该模型能显著改善学习和控制性能．

关键词：自强式学习，联想记忆，ASE/ACE 模型，RELCAM 模型．

1 引言

自强式学习（Reinforcement Learning）是在没有教师的示范而只有评论家的评判的情况下进行的学习控制．学习中获取的反馈信息只是对或错，而不是具体的误差信息．ASE/ACE 模型[3] 是在文 [1, 2] 基础上发展起来的一种自强式学习控制器模型，其学习是一种完全独立的表格查询与修改过程，缺乏记忆的联想性．而“实现了联想记忆就等于实现了半个大脑[6]”，因此，本文提出一种带联想记忆机制的自强式学习控制器模型 RELCAM（Reinforcement Learning with Associative Memory），并对文献 [3] 介绍的倒立摆进行了在线学习控制的仿真研究．

2 ASE/ACE 模型的分析

ASE/ACE 模型由关联式搜索单元 ASE 和自适应评判单元 ACE 构成．ASE 的作用是确定控制信号 y，而 ACE 则对自强信号 r 进行改善．ASE 和 ACE 各有 n 路输入通道，由系统状态 s 解码而成，且每一时刻只选一个，即 $x_i(t)$，$i=1, 2, \cdots, n$．控制信号的确定和各通道权值修正的公式如下[3]：

$$y=f\left[\sum_{i=1}^{n} w_i(t)x_i(t)+\text{noise}(t)\right], \tag{1}$$

$$p(t)=\sum_{i=1}^{n} v_i(t)x_i(t), \tag{2}$$

$$\hat{r}(t)=r(t)+\gamma p(t)-p(t-1), \tag{3}$$

$$w_i(t+1)=w_i(t)+\alpha\hat{r}(t)e_i(t), \tag{4}$$

---

1) 获国家自然科学基金资助课题．
本文于 1993 年 6 月 19 日收到

$$e_i(t+1)=\delta e_i(t)+(1-\delta)y(t)x_i(t), \tag{5}$$

$$v_i(t+1)=v_i(t)+\beta\hat{r}(t)\overline{x}_i(t), \tag{6}$$

$$\overline{x}_i(t+1)=\lambda\overline{x}_i(t)+(1-\lambda)\overline{x}_i(t). \tag{7}$$

其中 w_i 和 v_i 分别为 ASE 和 ACE 各通道的权值; $\hat{r}$ 是经改善的自强信号; $\alpha, \beta, \gamma, \delta$ 和 λ 是有关系数; noise 是随机噪声.

显然, 各单元的输出几乎完全取决于被选通道的权值 (ASE 略受噪声的影响). 但是, 各权值的学习却是相互独立的, 只有那些曾被选中的通道的权值才会得到修正, 其它的则不变. 这样, 一旦碰到新情况, 即选中一个从未选中的通道, 就可能输出一个完全错误的控制信号, 导致系统运行的失败.

3 RELCAM 模型

以上分析表明, 应对权值进行联想记忆, 即任一通道权值的学习会同时影响另几路的权值, 其中包括未曾选中的通道. 这样, 系统在失败几次而稍具经验教训后, 即使碰到新情况, 也能联想出一个比较好的 (尽管可能不是最好的) 控制信号, 以保证系统能继续成功地运行和学习.

RELCAM 模型就是基于以上考虑提出的, 它由 ASE/ACE 模型和联想记忆机制 AM 构成, 如图 1 所示. 每个通道的权值通过一个子网络来记忆, 各子网络既独立又关联地组成一个监督式学习 (Supervised Learning) 的局部性网络. 全局性网络 (如 BP 网) 因其完全联结机制而较适合于对固定样本集的学习[5]. 研究还表明, 直接将 BP 网用作 AM 是不合适的; 而局部性网络因其部分联结机制而较适合于实时控制[5]. 在 RELCAM 模

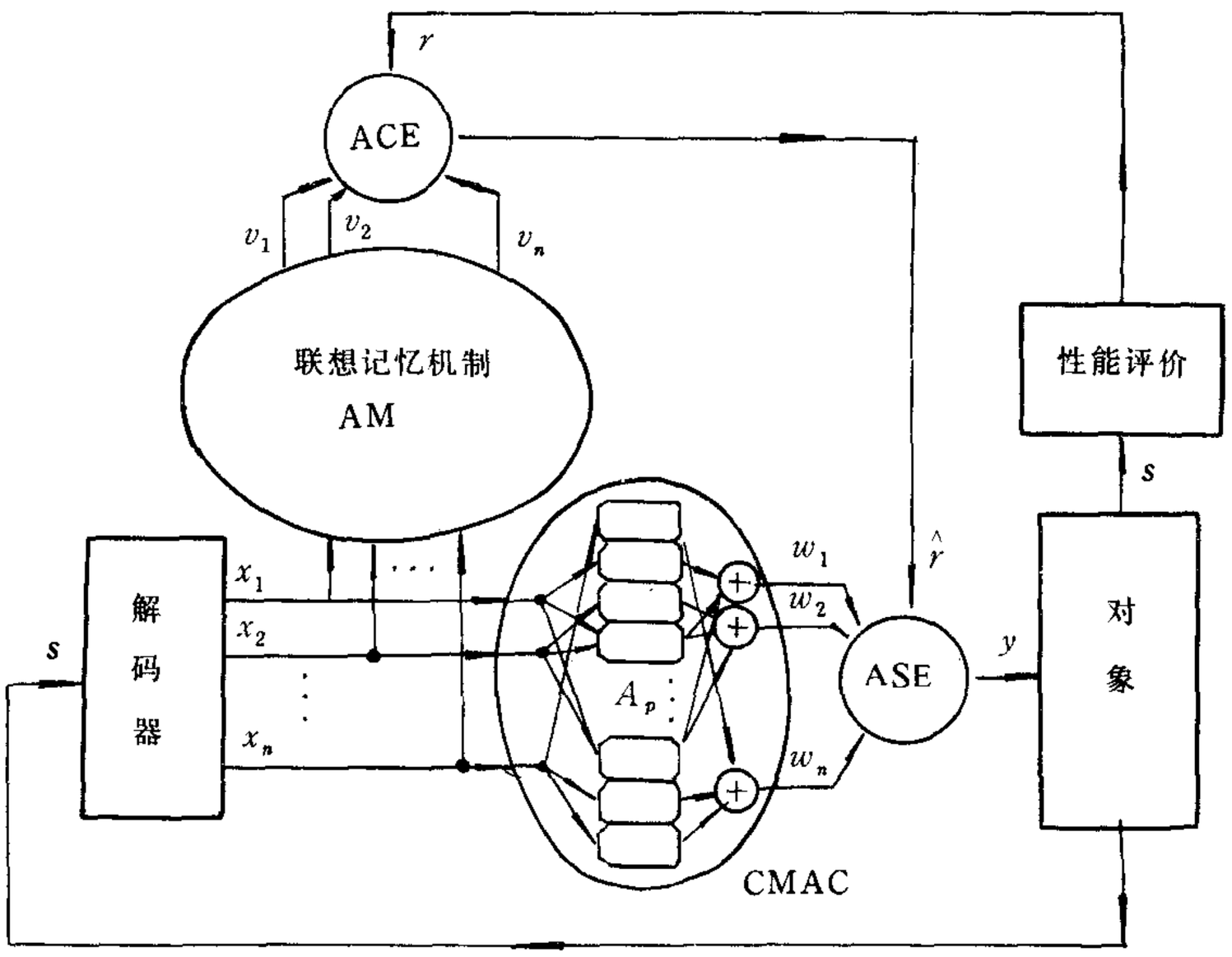

图 1 RELCAM 模型

型中; AM 可采用 CMAC[4] 或局部化的 BP 网 (LBP)[8] 等局部性网络.

下面以 CMAC 为例说明 RELCAM 的工作原理. 系统状态 $s(t)$ 经解码选中通道 $x_i(t)$, 经杂凑映射 (Hashmapping)[4] 选中 CMAC 记忆空间 A_p 中的 A^* 个记忆单元, 该 A^* 个单元中的信号之和即为本时刻本通道的权值, 相应子网络学习的教师信号分别由式 (4) 和 (6) 给出. 由于子网络间的关联性, 任何一个权值的一部分同时也是另几个通道权值的一部分. 因而, 没有一个通道的权值是完全独立的, 学习是联想式的.

控制信号的确定和权值的修正依旧采用式 (1)—(7).

4　仿真结果

为了验证 RELCAM 模型, 采用文献 [3] 介绍的倒立摆的数学模型和参数值在 SUN 工作站上进行在线学习控制的仿真研究. 系统状态 $\boldsymbol{s}$ 取为 $[x, \dot{x}, \theta, \dot{\theta}]^T$, 其中 x 和 θ 分别为位移和偏角. ASE 和 ACE 各有 162 个输入通道. 当 θ 超出 $[-12°, 12°]$, 或者 x 超出 $[-2.4\text{m}, 2.4\text{m}]$ 的范围时, 尝试 (Trial) 失败. 以不失败地运动 8 万步 (相当于实际时间 26′40″, 步长为 0.02 秒) 为一次成功的尝试, 每次运行 (Run) 从零初值开始, 即 $\boldsymbol{w}(0)$、$\boldsymbol{v}(0)$ 和 $\boldsymbol{s}(0)$ 都为零, 到成功的尝试结束. 每次尝试的初值 $\boldsymbol{s}(0)$ 为零, $\boldsymbol{w}(0)$ 和 $\boldsymbol{v}(0)$ 为本次运行中的上次尝试失败时的数值. 不同运行的差别在于随机数“种子”的不同.

联想记忆机制有两个主要参数, A_p 和 A^*. 在 CMAC 中, A_p 是整个网络记忆空间的大小, 而 A^* 则为各子网络中记忆单元的个数; 在 LBP 中, A_p 表示整个网络的隐单元总数, 而 A^* 则为各子网络的隐单元数. A_p 与 A^* 没有严格限制, 其典型取值为 $A_p \geqslant 100A^*$[4]. 而研究表明应满足 $A_p \approx 10A^*$ 的关系[7], 当 (A_p, A^*) 取为 (50, 5), (100, 10) 和 (300, 30) 等多种组合, 均能很好地工作. 本文统一选取 $A_p = 100$ 和 $A^* = 10$.

对 ASE/ACE, ASE/ACE + CMAC 和 ASE/ACE + LBP 三种情况各作 6 次运行, 并对成功尝试中的下列指标进行统计: (1) 尝试成功的最快时刻 T_s; (2) 尝试成功的平均时刻 T_{sm}; (3) 偏角均方根的平均值 θ_{sm}; (4) 最大偏角 $\theta_{\max}$; (5) 最大偏角的平均值 θ_{mm}.

表 1　6 次运行的统计

模　　型	T_s (次)	T_{sm} (次)	θ_{sm} (度)	$\theta_{\max}$ (度)	θ_{mm} (度)
ASE/ACE	21	49.83	2.5895	11.2223	7.1734
ASE/ACE + CMAC	6	7	2.2878	6.4668	5.0178
ASE/ACE + LBP	5	7	2.1984	5.4431	4.9262

从表 1 统计结果可以看到, 联想记忆机制的引入, 使尝试成功前的失败次数和偏差明显减少; 同时, 记忆空间也可减小, 如采用 CMAC, 记忆单元总数从原来的 324 (即 162×2) 减少到 200 (即 100×2). 这一特点, 在状态空间较大而输入通道较多时显得尤其重要.

5 结论

在自强式学习控制中引入联想记忆机制，不仅可缩小记忆空间，更能显著地改善系统的学习控制性能．此外，自强式与监督式学习并非完全独立的，两机的有机结合，对改善复杂非线性系统的学习和控制性能是大有益处的．

参 考 文 献

[1] Barto A G, Sutton R S, Brouwer P S. Associative search network: A reinforcement learning associative memory. *Biol. Cybern.*, 1981, **40**: 201—211.

[2] Barto A G, Sutton R S. Landmark learning: An illustration of associative search. *Biol. Cybern.*, 1981, **42**: 1—8.

[3] Barto A G, Sutton R S, Anderson C W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst., Man., Cybern.*, 1983, SMC-**13**(5): 834—846.

[4] Albus J S. A new approach to the manipulator control: The cerebellar model articulation controller (CMAC). *Trans. ASME, J. Dynamic Syst. Meas. Contr.*, 1975, **97**: 220—227.

[5] Werbos P J. Neurocontrol and elastic fuzzy logic: Capabilities, concepts and Applications. *IEEE Trans. Industr. Electronics*, 1993, **40**(2): 170—180.

[6] 中野馨〔日〕，卫作人．联想记忆工程．国防工业出版社，1992.

[7] 许力，蒋静坪．CSTR系统的基于CMAC神经元网络的学习控制研究．控制与决策，1992，**7**(2)：131—136.

[8] 许力．一种局部化的反向传播网络．控制与决策，1995，**10**(2)：**148—152**.

REINFORCEMENT LEARNING CONTROL WITH ASSOCIATIVE MEMORY

XU LI JIANG JINGPING ZHU JING ZHU YANXIN

(*Department of Electrical Engineering, Zhejing University, Hangzhou* 310027)

ABSTRACT

The supervised learning local nets are employed in the ASE/ACE model as the associative memory to construct a novel reinforcement learning controller model called RELCAM, simulation results show that the learning ability and control performance can be improved by using this model.

Key words: Reinforcement learning, associative memory, ASE/ACE model, RELCAM model.