第 28 卷 第 6 期
2002 年 11 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 28, No. 6
Nov., 2002

竞争式 Takagi-Sugeno 模糊再励学习[1)]

晏雄伟　　邓志东　　孙增圻

(清华大学计算机科学与技术系智能技术与系统国家重点实验室　北京　100084)

(E-mail: yxw@s1000e.cs.tsinghua.edu.cn)

摘　要　针对连续空间的复杂学习任务，提出了一种竞争式 Takagi-Sugeno 模糊再励学习网络(CTSFRLN)，该网络结构集成了 Takagi-Sugeno 模糊推理系统和基于动作的评价值函数的再励学习方法. 文中相应提出了两种学习算法，即竞争式 Takagi-Sugeno 模糊 Q-学习算法和竞争式 Takagi-Sugeno 模糊优胜学习算法，其把 CTSFRLN 训练成为一种所谓的 Takagi-Sugeno 模糊变结构控制器. 以二级倒立摆控制系统为例，仿真研究表明所提出的学习算法在性能上优于其它的再励学习算法.

关键词　再励学习，函数逼近，T-S 模糊推理系统

中图分类号　TP18

COMPETITIVE TAKAGI-SUGENO FUZZY REINFORCEMENT LEARNING

YAN Xiong-Wei　DENG Zhi-Dong　SUN Zeng-Qi

(*Department of Computer Science & Technology, State Key Laboratory of Intelligent Technology & System, Tsinghua University, Beijing*　100084)

(E-mail: yxw@s1000e.cs.tsinghua.edu.cn)

Abstract　This paper proposes a competitive Takagi-Sugeno fuzzy reinforcement learning network (CTSFRLN) for solving complicated learning tasks of continuous domains. The proposed CTSFRLN is constructed by combining Takagi-Sugeno type fuzzy inference systems with action-value-based reinforcement learning methods. Two competitive learning algorithms are derived, including the competitive Takagi-Sugeno fuzzy Q-learning and the competitive Takagi-Sugeno fuzzy advantage learning. These learning methods lead to so called Takagi-Sugeno fuzzy variable structure controllers. Simulation experiments on the double inverted pendulum system demonstrate the superiority of these learning methods.

Key words　Reinforcement learning, function approximation, Takagi-Sugeno fuzzy inference systems

---

1) 高等学校优秀青年教师教学科研奖励计划资助

收稿日期　2000-11-27　　收修改稿日期　2001-10-24

1 引言

基于动作的评价值函数的再励学习方法主要有三种:Q-学习、R-学习和优胜学习,一般地,这些学习方法只能接受离散化的状态输入,产生离散值的控制动作.但是,智能体所处的环境通常是空间连续的,对连续的状态空间和动作空间进行离散化,会导致维数灾问题.另一方面,模糊推理系统能够将连续状态映射为实值动作.Glorennec 提出了一种特殊的模糊 Q-学习结构[1],Jouffe 介绍了两种模糊再励学习方法,模糊动作器-评价器学习和模糊 Q-学习(FQL)[2],类似的规则结构在 Kim 提出的在线模糊 Q-学习方法中称为扩展的规则[3].以上方法使用了经典的模糊推理系统,能够完成简单的学习任务;但是,当学习任务涉及复杂的多变量环境时,这些方法就失去了效力.

为完成具有离散时间和连续空间的复杂学习任务,本文将基于动作的评价值函数的再励学习方法与 T-S(Takagi-Sugeno)模糊逻辑系统进行了有机结合,提出了一种新的再励学习网络结构,即竞争式 Takagi-Sugeno 模糊再励学习网络(Competitive Takagi-Sugeno Fuzzy Reinforcement Learning Network,简称 CTSFRLN).CTSFRLN 采用 T-S 模糊神经网络表达智能体的结构,由于 T-S 模糊推理系统只需少量的模糊规则,就能充分逼近多变量非线性函数[4],因此 CTSFRLN 能够利用这一优势处理多变量的复杂学习问题.

针对 CTSFRLN 结构,文中提出了两种高效的学习算法,即竞争式 Takagi-Sugeno 模糊 Q-学习算法以及竞争式 Takagi-Sugeno 模糊优胜学习算法.待学习结束之后,CTSFRLN 从功能上将变为一种 Takagi-Sugeno 模糊变结构控制器,能够较大地改善系统的控制性能.本文对上述学习方法在二级倒立摆控制系统中的应用进行了仿真研究,并与其它相关的再励学习算法进行了比较,结果令人满意.

2 竞争式 Takagi-Sugeno 模糊再励学习网络

CTSFRLN 结构利用竞争式 T-S 模糊神经网络来表达学习器,因而能够感知连续空间的输入状态,并在连续空间产生控制动作.另一方面,CTSFRLN 也是一种利用再励信号调节 T-S 模糊逻辑系统的自适应方法,将在线调节 T-S 模糊规则后件的线性增益.而规则前件的模糊划分以及规则后件的候选线性增益,则根据任务的先验知识或者借助于线性系统理论来进行预先设定,且在学习过程中保持不变.

2.1 网络结构

从网络结构而言,CTSFRLN 建立在竞争式 T-S 模糊神经网络的基础上,而 T-S 模糊神经网络则是 T-S 模糊推理系统的神经网络实现.在基本的 T-S 模糊规则中,前件的模糊划分使用三角形隶属函数,规则的真值为各输入隶属度的乘积.

T-S 模糊推理系统由 N 个如下形式的模糊规则组成

$$R_i:\text{如果 } s_1 \text{ 是 } L_1^i \text{ 且 } s_2 \text{ 是 } L_2^i \text{ 且 } \cdots \text{ 且 } s_n \text{ 是 } L_n^i,\text{则}$$

$$y^i = p_{i0} + p_{i1}s_1 + \cdots + p_{in}s_n$$

其中 $s_1,\cdots,s_n$ 为输入变量,y^i 为第 i 条规则的后件输出变量,$L_1^i,\cdots,L_n^i$ 为三角形隶属函数,$p_{i0},\cdots,p_{in}$ 为线性增益.T-S 模糊系统的输出量为各激活规则后件的加权和,即

$$y=\sum_{R_i\in A}\left(\alpha_{R_i}\sum_{j=0}^{n}p_{ij}s_j\right) \tag{1}$$

其中 α_{R_i}表示规则真值,A 表示所有激活规则的集合,$s_0\equiv 1$.

这里的 CTSFRLN 将接受 n 个状态输入,产生控制动作及其评价值等两个输出. 在 CTSFRLN 中,规则后件的线性增益 p_{ij}的值是从一个给定的离散集合中动态选取的,这个离散集合本文称之为候选动作增益集合,它附带一个相伴评价增益集合. 集合中的候选动作增益根据各自的相伴评价增益进行随机竞争,只有一个候选动作增益被选中,成为相应输入变量的实际增益系数,参与合成全局的控制动作. 与此同时,选中的动作增益所对应的相伴评价增益以同样的推理方式计算出全局动作的评价值. 这种结构方式把多维动作增益参数空间上的组合搜索简化为多个独立的、一维动作增益空间上的并行搜索,使学习算法的搜索空间大幅度减小,学习速度大大加快.

图 1 给出了 CTSFRLN 的网络结构,它输出一个二元组$[u,v]$,即连续动作 u 及其评价值 v. 输入节点 s_j 到后件节点 y^i 的连接权$[p_{ij},w_{ij}]$,代表动作增益-评价增益二元组,该增益单元是从候选动作增益集合 $P(ij)$及其相伴评价增益集合 $W(ij)$的 M 个元素中随机选取的.

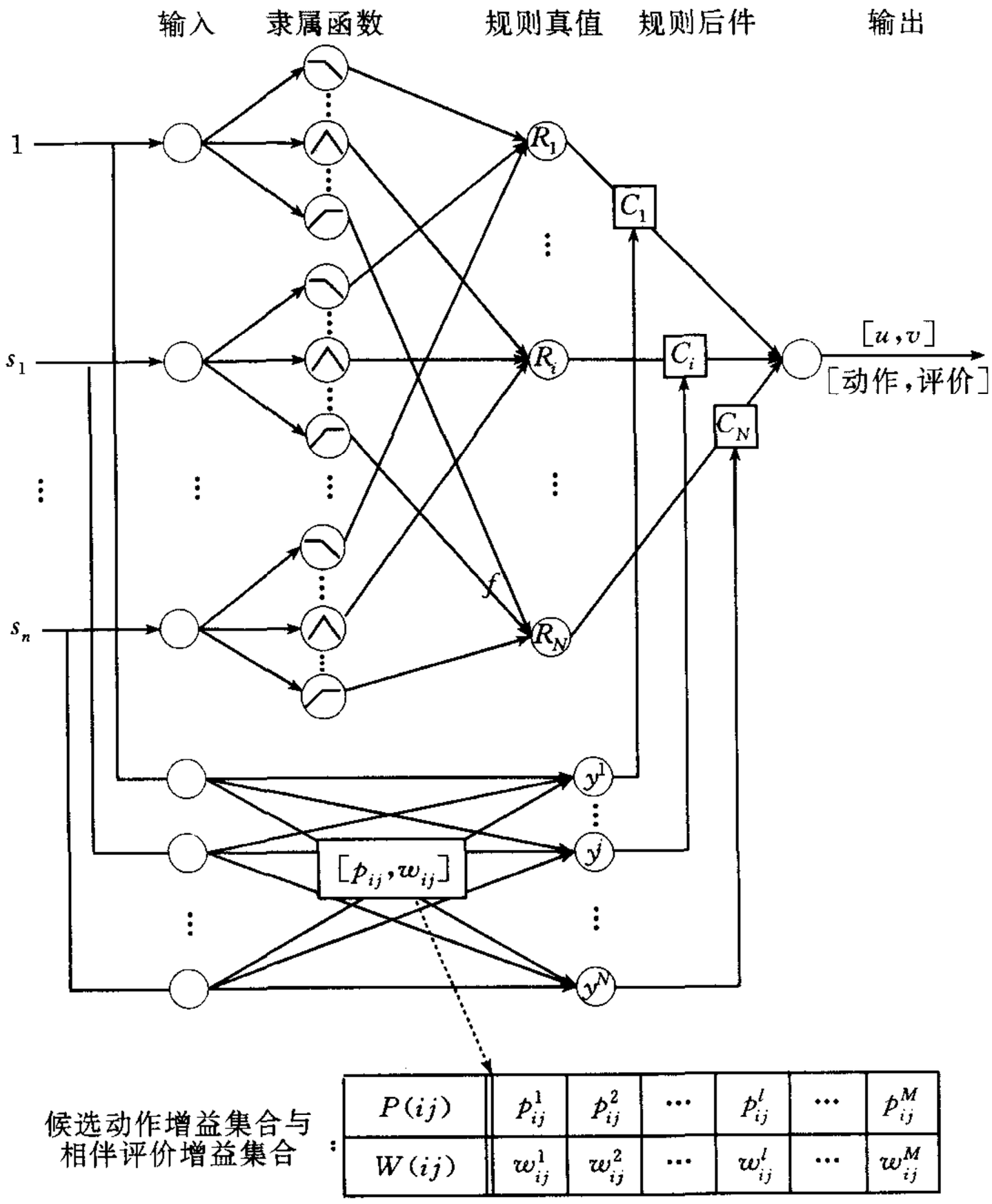


图 1　CTSFRLN 结构

2.2　探索策略

CTSFRLN 中,线性动作增益 p_{ij}从集合 $P(ij)$的候选动作增益中随机取值,p_{ij}的价值是

根据乘积项 $w_{ij}s_j$ 的值进行衡量的. 基于以上认识,本文提出了一种新的探索策略,即最大-最小 Boltzmann 探索策略,以实现候选动作增益之间的随机竞争.

根据输入变量 s_j 的符号,首先将集合 $W(ij)$中的元素 w_{ij}^l转化为变量 $\overline{w}_{ij}^l$,

$$\overline{w}_{ij}^l = \begin{cases} w_{ij}^l, & s_j > 0 \\ 0, & s_j = 0,\ \forall\ w_{ij}^l \in W(ij),\ l = 1,2,\cdots,M;\ j = 0,1,\cdots,n;\ i = 1,2,\cdots,N \\ -w_{ij}^l, & s_j < 0 \end{cases} \tag{2}$$

然后利用 Sigmoid 函数,将派生变量 $\overline{w}_{ij}^l$映射至区间[0,1],即

$$\hat{w}_{ij}^l = \frac{1}{1 + \mathrm{e}^{-\overline{w}_{ij}^l}},\quad l = 1,2,\cdots,M \tag{3}$$

从而集合 $P(ij)$中各候选动作增益的概率分布可由变量 $\hat{w}_{ij}^l$决定,此时有

$$Prob\ (p_{ij}^l) = \frac{\mathrm{e}^{\frac{\hat{w}_{ij}^l}{T}}}{\sum_{k=1}^{M} \mathrm{e}^{\frac{\hat{w}_{ij}^k}{T}}},\quad \forall\ p_{ij}^l \in P(ij),\ l = 1,2,\cdots,M \tag{4}$$

其中温度 $T>0$ 用于调节动作增益的选择随机性. 式(2),(3)及(4)一起组成了最大-最小 Boltzmann 探索策略.

智能体在学习阶段使用上述探索策略,而在测试阶段则使用贪婪策略[5],运行经学习得到的最优动作器. 在当前输入状态下,最优动作器输出的动作具有最大的评价值. 在最优动作器函数中,线性动作增益 p_{ij}^* 为集合 $P(ij)$中的最优候选动作增益,即

$$p_{ij}^* = \begin{cases} \arg \max\limits_{p_{ij}^l \in P(ij)} (w(p_{ij}^l)), & s_j \geqslant 0 \\ \arg \min\limits_{p_{ij}^l \in P(ij)} (w(p_{ij}^l)), & s_j < 0 \end{cases} \tag{5}$$

其中,$w(p_{ij}^l)$(即 w_{ij}^l)表示与候选动作增益 p_{ij}^l相对应的相伴评价增益. 在 t 时刻,最优动作器函数为

$$u_t^*(\boldsymbol{s}_t) = \sum_{R_i \in A_t} \left(\alpha_{R_i} \sum_{j=0}^{n} p_{ij}^*(t) s_j(t) \right) \tag{6}$$

式中,$\boldsymbol{s}_t = [s_1(t) \quad s_2(t) \quad \cdots \quad s_n(t)]^{\mathrm{T}}$,表示 t 时刻的输入状态矢量. 相应的评价器函数表示为

$$v_t^*(\boldsymbol{s}_t) = \sum_{R_i \in A_t} \left(\alpha_{R_i} \sum_{j=0}^{n} w(p_{ij}^*(t)) s_j(t) \right) \tag{7}$$

其中 $v_t^*(\boldsymbol{s}_t)$表示状态 $\boldsymbol{s}_t$ 的最优评价值函数.

在式(5)中,最优动作增益 p_{ij}^* 的取值根据输入变量 s_j 的符号而切换. 式(6)所描述的最优动作器称为 Takagi-Sugeno 模糊变结构控制器(TSFVSC). 此外,值取为零的 p_{ij}^* 意味着从规则后件的线性组合中删除了相应的输入变量 s_j. 在 TSFVSC 中,只有重要的输入变量才会进入后件的线性组合,这些输入变量的选取是由具体的学习过程决定的.

2.3 动作器函数和评价器函数

在学习阶段,CTSFRLN 利用 T-S 模糊推理,产生实值的控制动作. 在 t 时刻,动作器函数输出的控制动作为

$$u_t(\boldsymbol{s}_t) = \sum_{R_i \in A_t} \left(\alpha_{R_i} \sum_{j=0}^{n} p_{ij}(t) s_j(t) \right) \tag{8}$$

其中,$p_{ij}(t)$为最大-最小 Boltzmann 探索策略从候选动作增益集 $P(ij)$中选出的实际动作增益. 在 CTSFRLN 中,评价器函数根据学习任务,对动作器函数在当前状态下输出的实值动作进行评价. 评价器函数由构成全局动作的实际动作增益所对应的相伴评价增益经过模糊推理输出,即

$$v_t(\boldsymbol{s}_t,u_t)=\sum_{R_i\in A_t}\left(\alpha_{R_i}\sum_{j=0}^{n}w(p_{ij}(t))s_j(t)\right) \tag{9}$$

动作的评价值函数 $v_t(\boldsymbol{s}_t,u_t)$相对于相伴评价增益的梯度为

$$\frac{\partial v_t(\boldsymbol{s}_t,u_t)}{\partial w_t(p_{ij}^l)}=\begin{cases}\alpha_{R_i}s_j(t), & \text{如果 } p_{ij}^l=p_{ij}(t),\ R_i\in A_t\\ 0, & \text{否则}\end{cases} \tag{10}$$

其中 $p_{ij}(t)$为在 t 时刻参与合成全局动作 U_t 的实际动作增益,A_t 表示在 t 时刻所有激活规则的集合.

3　学习算法

TD(λ)学习算法采用传导性迹加快学习过程,不仅能调节当前时刻所对应的学习参数,而且允许调节以前时刻所涉及的学习参数. CTSFRLN 采用累加传导性迹记录当前和过去的梯度值,令 $\Phi_w(t)$表示与相伴评价增益 $w_t(p_{ij}^l)$相关的累加传导性迹,则

$$\Phi_w(t)=\gamma\lambda\Phi_w(t-1)+\frac{\partial v_t(\boldsymbol{s}_t,u_t)}{\partial w_t(p_{ij}^l)} \tag{11}$$

其中,γ 为折扣因子,λ 为传导率(或称新近因子),且二者都用于时间步的加权.

CTSFRLN 采用 T-S 模糊神经网络表达再励学习智能体的结构,因而其学习机制融合了基于动作的评价值函数的时差(Temporal Difference, TD)学习方法与基于神经网络的梯度反传算法. 根据 TD(λ)学习算法,CTSFRLN 适用的一种通用学习规则为

$$w_{t+1}(p_{ij}^l)=w_t(p_{ij}^l)+\eta\tilde{\varepsilon}_{t+1}\Phi_w(t),\quad \forall\ p_{ij}^l\in P(ij) \tag{12}$$

式中,η 为学习率,$\tilde{\varepsilon}_{t+1}$表示在 $t+1$ 时刻得到的 TD 误差. 两种基于动作的评价值函数的再励学习方法,即 Q-学习和优胜学习,均可用来计算 TD 误差.

1) 竞争式 Takagi-Sugeno 模糊 Q-学习算法(CTSFQL). 利用 Q-学习方法计算 TD 误差. 在这种情况下,评价器函数输出当前状态-动作对应的 Q-值. 计算 TD 误差时,动作的评价值函数 $v_t(\boldsymbol{s}_t,u_t)$也可记为 $Q_t(\boldsymbol{s}_t,u_t)$. 此时 CTSFQL 算法为

$$\tilde{\varepsilon}_{t+1}^{Q}=r_{t+1}+\gamma Q_t^*(\boldsymbol{s}_{t+1})-Q_t(\boldsymbol{s}_t,u_t) \tag{13}$$

$$w_{t+1}(p_{ij}^l)=w_t(p_{ij}^l)+\eta\tilde{\varepsilon}_{t+1}^{Q}\Phi_w(t),\quad \forall\ p_{ij}^l\in P(ij) \tag{14}$$

其中,r_{t+1}为智能体在时刻 $t+1$ 收到的再励信号,γ 为折扣因子,$Q_t^*(\boldsymbol{s}_{t+1})$代表最优控制策略对应的 Q-值函数. 式(7)在输入状态矢量为 $\boldsymbol{s}_{t+1}$时,根据当前相伴评价增益 $w_t(p_{ij}^l)$可计算出 $Q_t^*(\boldsymbol{s}_{t+1})$.

2) 竞争式 Takagi-Sugeno 模糊优胜学习算法(CTSFAL). 利用优胜学习方法计算 TD 误差. 在这种情况下,评价器函数 $v_t(\boldsymbol{s}_t,u_t)$输出当前状态下的动作优胜值,优胜值函数标记为 $A_t(\boldsymbol{s}_t,u_t)$. CTSFAL 的学习规则为

$$\tilde{\varepsilon}_{t+1}^{A}=\Gamma[r_{t+1}+\gamma A_t^*(\boldsymbol{s}_{t+1})]-(\Gamma-1)A_t^*(\boldsymbol{s}_t)-A_t(\boldsymbol{s}_t,u_t),\quad \Gamma=\frac{1}{\Delta tK} \tag{15}$$

$$w_{t+1}(p_{ij}^l)=w_t(p_{ij}^l)+\eta\tilde{\varepsilon}_{t+1}^{A}\Phi_w(t),\quad \forall\ p_{ij}^l\in P(ij) \tag{16}$$

其中，Δt 为控制动作之间的时间间隔，K 为时间单位放大因子，Γ 称为相对放大因子. $A_t^*(\boldsymbol{s}_{t+1})$和 $A_t^*(\boldsymbol{s}_t)$分别表示输入状态 $\boldsymbol{s}_{t+1}$和 $\boldsymbol{s}_t$ 对应的状态优胜值，由式(7)根据当前相伴评价增益 $w_t(p_{ij}^l)$求得.

4 二级倒立摆系统

二级倒立摆系统将用于验证上述学习算法的有效性，并与其它再励学习方法进行比较. 控制系统的学习目标是，对于上摆存在一定范围内的任意初始偏角的倒立摆系统，智能体能够维持双摆平衡并保持小车的运动不超出导轨的长度. 二级倒立摆系统有 6 个状态变量：上摆偏离垂直方向的角度 θ_2 和角速度 $\dot{\theta}_2$，下摆偏离垂直方向的角度 θ_1 和角速度 $\dot{\theta}_1$，以及小车的位置 x 和线速度 $\dot{x}$. 二级倒立摆系统的非线性动力学方程为

$$M\begin{bmatrix}\ddot{x}\\ \ddot{\theta}_1\\ \ddot{\theta}_2\end{bmatrix}=A\begin{bmatrix}\dot{x}\\ \dot{\theta}_1\\ \dot{\theta}_2\end{bmatrix}+\begin{bmatrix}f\\ (m_1l_1+m_2L_2)g\sin\theta_1\\ m_2gl_2\sin\theta_2\end{bmatrix} \tag{17}$$

$$M=\begin{bmatrix}m_c+m_1+m_2 & (m_1l_1+m_2L_1)\cos\theta_1 & m_2l_2\cos\theta_2\\ (m_1l_1+m_2L_1)\cos\theta_1 & J_1+m_1l_1^2+m_2L_1^2 & m_2L_1L_2\cos(\theta_2-\theta_1)\\ m_2l_2\cos\theta_2 & m_2L_1L_2\cos(\theta_2-\theta_1) & J_2+m_2l_2^2\end{bmatrix} \tag{18}$$

$$A=\begin{bmatrix}-f_0 & (m_1l_1+m_2L_1)\sin\theta_1\cdot\dot{\theta}_1 & m_2l_2\sin\theta_2\cdot\dot{\theta}_2\\ 0 & -(f_1+f_2) & m_2L_1l_2\sin(\theta_2-\theta_1)\cdot\dot{\theta}_2+f_2\\ 0 & -m_2L_1l_2\sin(\theta_2-\theta_1)\cdot\dot{\theta}_1+f_2 & -f_2\end{bmatrix} \tag{19}$$

其中，重力加速度 $g=9.8\text{m/s}^2$，小车质量 $m_c=1.0\text{kg}$，上摆及下摆的质量 $m_2=m_1=0.1\text{kg}$，两摆的半杆长 $l_2=l_1=0.5\text{m}$，小车与导轨的摩擦系数 $f_0=0.01\text{N}\cdot\text{s/m}$，上摆与下摆各自转轴处的摩擦阻力矩系数 $f_2=f_1=0.01\text{N}\cdot\text{s}\cdot\text{m}$，两摆长度 $L_2=2l_2$ 和 $L_1=2l_1$，且 J_2 和 J_1 分别为上摆和下摆相对于各自质心的转动惯量. 作用在小车上的控制力 $f\in[-60\text{N}, 60\text{N}]$. 仿真方法为四阶龙格-库塔法，采样周期取为 20ms.

CTSFRLN 结构包含 64 条模糊规则，使用 6 个状态变量作为输入，每个状态变量的论域上有 2 个三角形隶属函数，规则后件为所有状态变量的线性组合，每个状态变量的候选动作增益集合包括 5 个候选动作增益，它们来自线性化系统的线性二次型最优调节器的设计结果. 学习参数设置为：$\gamma=0.95$，$\lambda=0.9$，$\Gamma=20$ 和 $\eta=0.001$.

仿真结果为 30 次运行的平均结果. 一次运行包括学习阶段和测试阶段，两个阶段都由一系列试探组成. 一个试探是指一个动态的控制过程，它起始于下摆保持中立而上摆存在有限的随机偏角的初始状态，结束于失败的控制状态或者平衡时间达到 100 秒的情况. 失败的控制状态是指上摆角度 $|\theta_2|>12°$，或者下摆角度 $|\theta_1|>18°$，或者小车位置 $|x|>2.4\text{m}$. 平衡时间达到 100 秒的试探称为成功的试探. 当连续出现 50 个成功的试探或者试探总数超过 50 000 时，学习阶段结束. 测试阶段由 100 个试探组成. 两个阶段各对应一套随机的初始状态：θ_2 的初始值在区间$[-12°, 12°]$上均匀分布，其它状态变量的初始值取为零. 学习阶段的性能指标为学习速度，即该阶段所需的试探总数；而测试阶段则用于统计试探成功率，即成功的试探数占 100 个试探总数的百分比. 试验中，对于失败的控制状态，再励信号的值取为 -1，否则为零.

仿真试验研究了 CTSFQL 和 CTSFAL 的性能. 对于 Jouffe 的 FQL 和作者以前提出的 FAL[6],如果每个状态变量的论域上有 2 个三角形模糊子集,每一规则的后件配有 10 个或 20 个候选动作,研究发现,智能体经过 50 000 个试探的学习阶段后,试探成功率仍不足 10%. 为了改善学习性能,在 FAL 和 FQL 中,每一状态变量的论域必须划分为 3 个模糊子集,而模糊规则也剧增到 729 条.

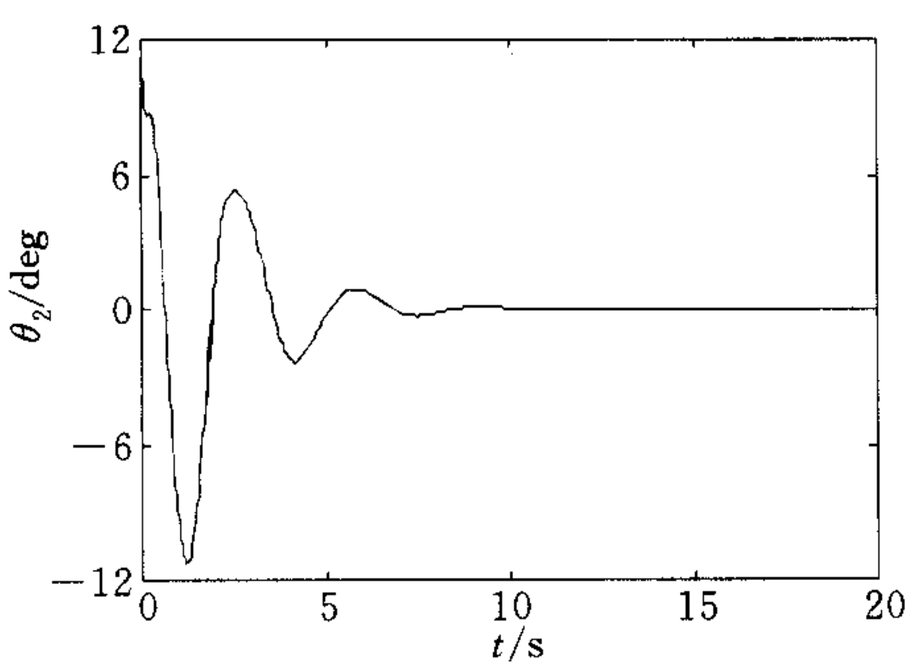
表 1 列出了 4 种学习算法的试验结果. 本文提出的两种学习算法取得了满意的性能,且 CTSFAL 的学习速度明显快于 CTSFQL. 其原因在于,对于控制动作的时间间隔非常小的

表 1　学习算法比较

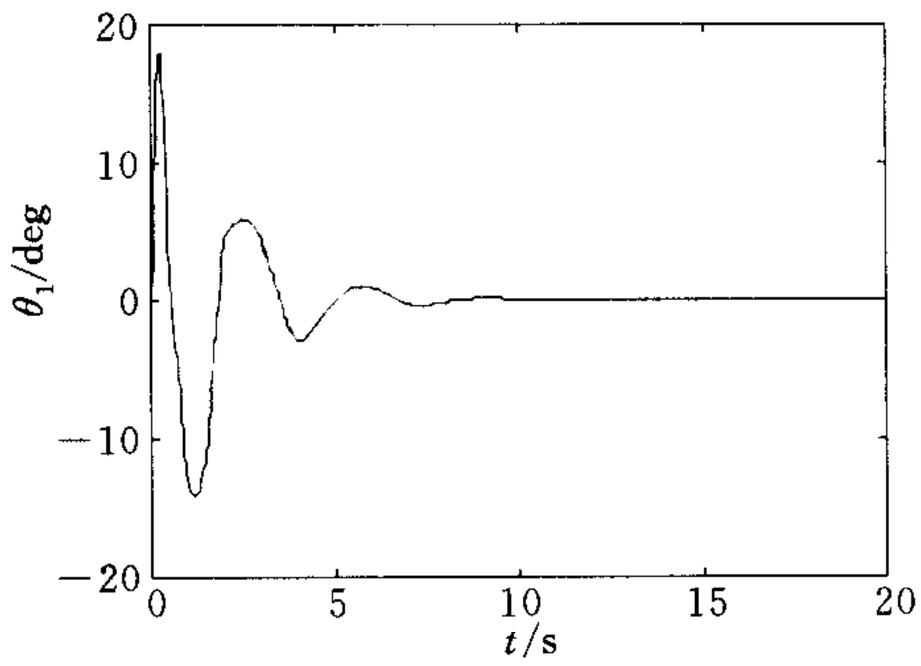
算法	模糊推理系统		学习速度(个试探)	试探成功率(%)	
	模糊分割数	候选动作增益或候选动作的个数		学习之后	学习之前
CTSFAL	2	5	4 325.6	94.7	70.3
CTSFQL	2	5	7 803.2	92.5	70.7
FAL	3	10	23 856.4	80.6	0.0
FQL	3	10	38 274.8	81.3	0.0

控制问题,CTSFQL 算法的探索策略易于受到学习误差的干扰,导致其学习过程需要更多的迭代次数,而 CTSFAL 算法在学习公式中采用了相对放大因子来消除这一影响,因而具有较快的学习速度. CTSFQL 算法的优点在于计算简便. 试验也表明,文中提出的学习算法在性能上要优于其它两种学习方法,且这里的 CTSFRLN 结构较之 FQL 和 FAL 结构需要少得多的模糊规则. 特别地,本文的学习算法在学习之前已经充分利用了线性控制理论的初步设计结果,且学习之后智能体的控制性能又得到了明显提高.

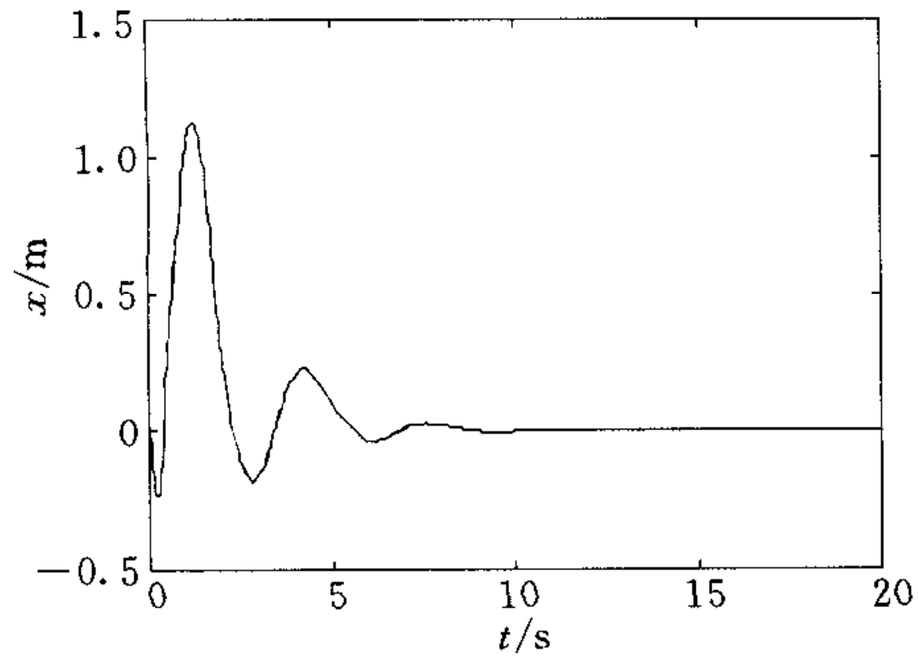
图 2 所示为系统采用 CTSFAL 算法时,获得的 Takagi-Sugeno 模糊变结构控制器在起



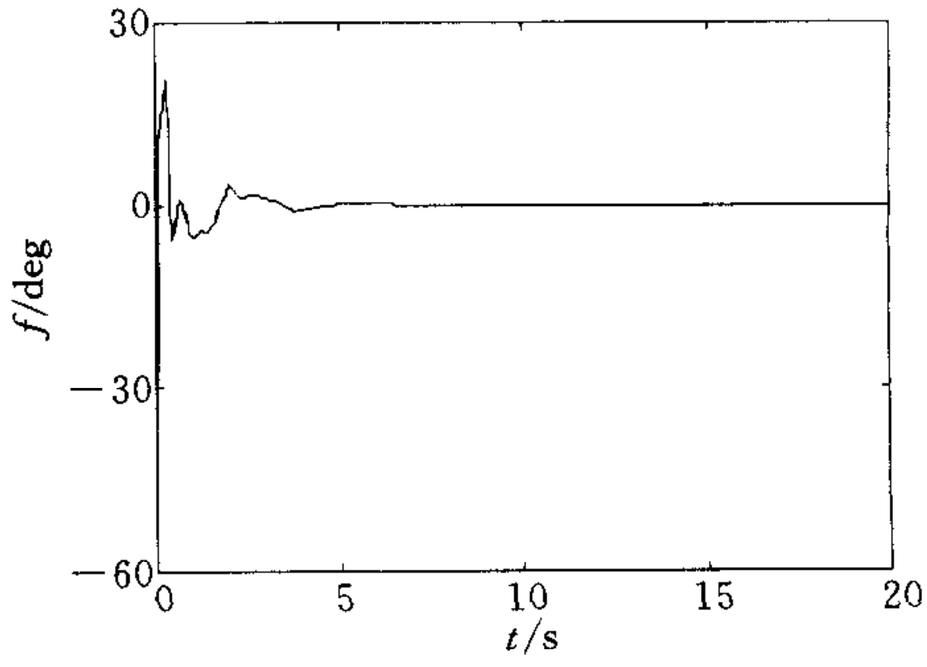
(a) 上摆角度



(b) 下摆角度



(c) 小车位置



(d) 控制力

图 2　TSFVSC 的响应曲线

始 20 秒的仿真结果，其中 θ_2 的初始值为 10.6°，整个系统一直稳定到 100 秒之后. 从图中可以看出，虽然调节过程的波动较大，但所有状态变量的调节时间少于 10 秒.

CTSFRLN 结构具有以下三个特点. 首先，在学习之初充分利用了来自线性控制的分析结果，进而在学习过程中对所有的候选动作增益进行优化重组，从而提高了学习系统的性能. 第二，CTSFRLN 使用了竞争式的 T-S 模糊推理系统来表达学习器. 这种结构显著地减少了所需要的规则数目，增强了系统的泛化能力，并使得每一维动作增益参数在各自的候选动作增益集合内进行独立搜索，这些都加快了学习过程. 第三，学习获得的 TSFVSC 能够输出连续值的控制动作，增强控制系统的鲁棒性，抗干扰性和快速性.

5 结论

为解决连续空间的复杂再励学习问题，本文提出了竞争式 Takagi-Sugeno 模糊再励学习网络. 仿真结果表明，本文提出的学习算法在学习速度与控制性能等方面明显优于其它的再励学习方法. 但是，由于使用了 T-S 模糊神经网络这一函数逼近器来表示评价值函数，学习算法的收敛性目前还没有得到证明. CTSFRLN 将使再励学习技术能够完成连续空间的复杂学习任务，且使模糊逻辑控制器的设计更加方便与实用. 更加重要的是，在有机结合线性控制理论的初步分析结果与智能控制的学习能力方面，CTSFRLN 提供了一种新的思路与方法. CTSFRLN 可望广泛应用于机器学习与模糊逻辑系统等领域.

参 考 文 献

1 Glorennec P Y. Fuzzy Q-learning and dynamic fuzzy Q-learning. In: Proc. IEEE, 3rd Int. Conf. Fuzzy Systems, FL: Orlando, 1994, **1**:474～479

2 Jouffe L. Fuzzy inference system learning by reinforcement methods. *IEEE Trans. SMC—Part C, Applications and Reviews*, 1998, **28**(3):338～355

3 Kim M S, Hong S G, Lee J J. On-line fuzzy Q-learning with extended rule and interpolation technique. In: Proc. 1999 IEEE/RSJ Int. Conf. Intelligent Robotics and Systems, Kyongju: South Korea, 1999, **2**:757～762

4 Takagi T, Sugeno M. Fuzzy identification of systems and its application to modeling and control. *IEEE Trans. Syst., Man and Cybern.*, 1985, **15**(1):116～132

5 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. MA: MIT Press, 1998

6 Yan X W, Deng Z D, Sun Z Q. Fuzzy advantage learning. In: Proc. IEEE, 9th Int. Conf. Fuzzy Systems, TX: San Antonia, 2000, **2**: 865～870

晏雄伟 1994 年和 1997 年于西北工业大学获飞行器控制、制导与仿真专业学士和硕士学位，现为清华大学计算机系博士研究生. 研究领域包括智能控制、机器人、机器学习等.

邓志东 1991 年获哈尔滨工业大学控制工程系博士学位，1992～1994 年在清华大学计算机系从事博士后研究，现为清华大学教授. 主要研究领域为神经网络控制、模糊控制、学习控制等.

孙增圻 1966 年毕业于清华大学自控系并留校任教，1981 年在瑞典获博士学位，现为清华大学计算机系教授，IEEE 高级会员. 长期从事智能控制、机器人、计算智能等方面的研究.