文章编号: 1001-0920(2008)09-0961-08

从知识的表达和运用综述强化学习研究

陈宗海, 杨志华, 王海波, 盛　捷
(中国科学技术大学 自动化系, 合肥 230027)

摘　要: 为推进强化学习研究的进一步深入和扩大其实际应用范围,从强化学习研究的理论基础——知识表示和运用的角度对强化学习进行分类,并就经典随机强化学习、模糊强化学习、定性强化学习以及灰色强化学习作了较详细的探讨与比较. 最后从知识表达和运用的角度对强化学习的发展进行了展望.

关键词: 强化学习; 知识表示; 模糊理论; 定性推理; 灰色系统理论
中图分类号: TP24　　**文献标识码**: A

Overview of reinforcement learning from knowledge expression and handling

CHEN Zong-hai, YANG Zhi-hua, WANG Hai-bo, SHENG Jie
(Department of Automation, University of Science and Technology of China, Hefei 230027, China. Correspondent: CHEN Zong-hai, E-mail: chenzh@ustc.edu.cn)

Abstract: In order to advance reinforement learning (RL) research and expand its practical application scope, it's necessary to classify RL from RL research theory base, knowledge expressed and the handling angle. Based on classical stochastic RL, fuzzy RL, qualitative RL and grey RL, the detailed discussion and comparions are given. Finally, RL development is forecasted from the knowledge expression and handling angle.

Key words: Reinforcement learning; Knowledge representation; Fuzzy theory; Qualitative reasoning; Grey system theory

1 引　　言

任何 agent 要实现自主智能,学习能力无疑是其关键技术之一. 对于学习的一般论述,Simon 表述为[1]:如果 agent 通过执行某个过程能够改进自身的智能,这就是学习. 对于不同系统的学习,可利用不同的学习方法. 在机器学习领域,根据反馈的不同,学习技术可分为监督学习、非监督学习和强化学习(RL). 其中强化学习的适应性较强,在各种应用尤其在机器人领域得到了广泛应用,成为机器学习的研究热点之一.

2 强化学习的基本概念

上世纪五六十年代,对动物刺激反应的心理学研究促生了强化学习. 其奠基性成果[2]有:只用强化信号反馈的联想奖惩 (ARP) 算法[3],由 ASE 和 ACE 构成的 AHC 算法[4,5],根据时间序列进行预测的瞬时差分(TD)算法[6],利用状态-动作对映射的 Q-学习算法[7]. 近 10 年来,强化学习引起了广泛的研究,并在一些实际应用中取得了较好的效果.

强化学习[8]是 agent 在与动态环境的交互过程中,通过反复试错来学习适当的行为. 它介于监督式学习和无监督式学习之间,是一种在策略学习,通过与环境的即时交互来获得环境的状态信息,并通过反馈强化信号对所采取的行动进行评价,通过不断的试错和选择,从而学习到最优的策略. 强化学习的模型如图 1 所示[9],其基本要素是:智能主体(agent)、环境、策略、报酬函数、值函数以及环境模型 W(可未知).

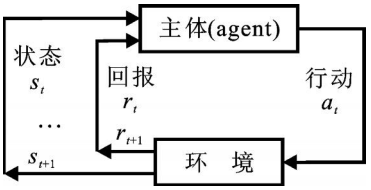


图 1　强化学习的结构

收稿日期: 2007-05-29; **修回日期**: 2007-07-27.
基金项目: 国家自然科学基金项目(60575033); 国家 863 计划项目(2007AA04Z227).
作者简介: 陈宗海(1963—),男,安徽桐城人,教授,博士生导师,从事复杂系统建模、模式识别等研究; 杨志华(1983—),男,安徽砀山人,硕士生,从事模式识别、人工智能的研究.

在图 1 中,策略 $\pi: S \rightarrow A$ 表示从状态到动作的映射,是强化学习的核心;报酬 r 是对所采取动作的即时评价,通常用一标量值来表示;值函数则是对策略的长远评估,即目标函数. agent 与环境通过状态(state)、动作(action)以及奖罚(reward)进行交互,其过程如图 2 所示. 即 agent 由当前策略选择一个动作,转移到下一状态获得即时报酬,进而改进策略. 其目标是最大化长期回报,获得对应的最优策略.

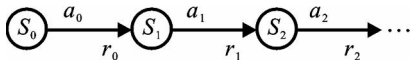


图 2 agent 的交互过程

强化学习具有自适应、在线学习、试错以及自我选择的特点,在控制、决策及规划中得到广泛的应用[10]. 但一般的强化学习方法很难满足实际问题对其泛化能力的要求,且应用中环境经常是未知不确定的. 为此,许多学者从新的角度对强化学习进行研究. 如 Glorennec[11,12] 和 Berenji[13-15] 等将强化学习与模糊理论结合起来,构成模糊强化学习方法; Arkady[16] 则将定性推理应用于强化学习;Chen 等[9] 对灰色强化学习进行初步研究,并应用于移动机器人的导航. 这些方法对于推动强化学习在实际中的应用起到了积极作用,它们是从知识表达和运用的角度对强化学习进行研究,是强化学习发展的重要方向.

3 强化学习方法分类

agent 智能的实现过程其实是对知识的处理过程. 知识描述对象和状况的关系主要包括 3 个方面[17]:知识的获取,知识的表示,知识的运用. 对于智能机器而言,首先通过传感器获取周围及其自身的信息,并以一定的方式表示出来;然后 agent 运用所表示的知识实现机器的智能化. 因此,知识表示是 agent 处理信息的一个重要任务,其表示方法的不同对于知识的利用程度和效率都有很大的不同.

对可完全认知的简单而确定性知识的表示较为容易,一般利用纯粹的定量知识即可,而且人们对定量知识运用的理论研究也较成熟. 然而,现实中的知识一般是复杂、不确定和不完备的,仅用定量的描述方式难以获得良好的效果. 为此,人们根据知识的特点,分别利用相对应的不确定性的知识表示方法,如灰色、定性、模糊等理论. 其特点的分析和比较如表 1 所示[18].

针对强化学习的随机特点,定量知识用概率统计来表示,即经典的概率统计本质上是一种知识的定量表示方式,且具有坚实的理论基础和广泛的应用,是其他各种知识表示的基础,其不确定性即为经典的随机不确定性;模糊数学是描述信息的外延模糊的知识,是处理模糊知识的理论基础;灰色系统理论描述的信息具有部分已知、部分未知所导致的不确定性;定性推理是利用系统之间的关系及其因果性进行知识的表示与推理.

强化学习有多种分类方式:根据其环境模型的有无,可分为基于模型和模型无关的强化学习;根据其状态的转移,可分为马尔可夫环境、非马尔可夫环境和半马尔可夫环境下的强化学习;根据其映射方式的不同,可分为 TD- 学习、Q- 学习等. 这些分类方式对强化学习的发展发挥了很大作用,但均未重视知识表示这一强化学习问题的基础.

强化学习作为 agent 自学习、实现其智能化的一种重要方式,知识表示是进行强化学习设计的基础. 本文从知识表示和运用(推理过程)的角度,结合环境知识的特点,分别从知识表示的概率、模糊、定性及灰色入手,将强化学习分为:经典随机强化学习(SRL),模糊强化学习(FRL),定性强化学习(QRL)和灰色强化学习(GRL),如表 2 所示. 其中 MDPs 是马尔可夫决策过程的简称. 当然,各种表示方式和推理方式之间都可相互结合,扩大强化学习

表 1 4 种知识表示方法的比较

项 目	概率统计(定量)	模糊数学	灰色系统	定性推理
研究对象	随机不确定	认知不确定	贫信息不确定	关系可认知
基础集合	康托尔集	模糊集	灰色朦胧集	关系集
方法依据	概率分布	隶属函数	信息覆盖	因果关系
途径手段	频率统计	截 集	灰序列生成	定性推理
数据要求	典型分布	隶属度可知	任意分布	系统结构可知
侧 重	内 涵	外 延	内 涵	结构和功能
目 标	历史统计规律	认知表达	现实规律	人类常识推理
特 色	大样本	凭经验	小样本	连续状态
理论基础	完 善	较完善	不完善	一 般
表示特点	纯定量	半定量	半定量	纯定性

表 2　基于知识的强化学习的构成

基于知识的强化学习	知识表示	推理方法
经典随机强化学习	精确数字量	各种经典 MDPs
模糊强化学习	模糊数、模糊概率	模糊理论、模糊 MDPs
定性强化学习	定性量、定性关系	定性推理、定性 MDPs
灰色强化学习	灰色数、灰色表示	灰色理论、灰色 MDPs

的适应性,使强化学习更易于在实际中应用.

以下分别对各种知识表示方式以及 SRL, FRL, QRL 和 GRL 进行较详细的阐述. 除了对经典随机强化学习和模糊强化学习的研究较多外,人们对定性、灰色强化学习以及各种方式结合起来的研究还很不够. 希望本文的讨论能促进这些方面的研究进展.

4　经典随机强化学习(SRL)

SRL 是最基本的强化学习方法,是各种强化学习方法的基础. 它有多种学习方式,如 TD-学习、Q-学习、Sarsa 算法、Dyna-Q 学习等. 其中以 TD-学习和 Q-学习最为基本,且实际应用也最为广泛.

4.1　TD-学习

1988 年, Sutton 发表了"Learning to predict by the methods of temporal difference"的经典论文[6], 提出了 TD 方法,解决了强化学习中根据时间序列进行预测的问题,并证明了在系统满足马尔可夫属性、α 绝对递减条件下, TD 方法收敛于最优. Dayan[19] 则对含有延迟回报的 TD(λ) 方法的收敛性给予了证明.

TD-学习结合了蒙特卡罗和动态规划两种方法的优点,是一种模型无关的学习. 它一方面可以不需要系统模型,直接与环境交互并从 agent 经验中学习;另一方面又利用动态规划的思想,根据估计的值函数进行迭代,直至得到最优策略[9]. 在动态规划中,状态转移和报酬函数是已知的,对于从任意选择的初始策略开始,均可利用策略迭代的方法逼近最优的值函数 V^* 和最优策略 π^*,即

$$\pi_k(s) = \arg\max_a \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V^{\pi_{k-1}}(s')], \quad (1)$$

$$V^{\pi_k}(s) \leftarrow \sum_a \pi_{k-1}(s,a) \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V^{\pi_{k-1}}(s')]. \quad (2)$$

而蒙特卡罗方法在当前状态转移概率函数和报酬函数未知时,采用逼近方法进行值函数的估计,即

$$V(s_t) \leftarrow V(s_t) + \alpha[R_t - V(s_t)]. \quad (3)$$

在 TD-学习中,最简单的为一步 TD 算法(即 TD(0) 算法),其迭代公式为

$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)). \quad (4)$$

其中: γ 为折扣因子, $\alpha \in (0,1]$ 为学习率, $V(s_t)$ 是 agent 在 t 时刻状态 s_t 时估计的状态值函数, r_{t+1} 是 agent 从 s_t 向 s_{t+1} 转移时获得的瞬时报酬值. 为提高 TD(0) 的收敛速度,人们提出了将瞬时奖赏值后退多步的 TD(λ) 算法,即

$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)) e(s), \quad (5)$$

其中 $e(s)$ 为状态 s 的资格迹. 实际应用时可用下式计算:

$$e(s) = \begin{cases} \gamma\lambda e(s) + 1, & s \text{ is current state;} \\ \gamma\lambda e(s), & \text{otherwise.} \end{cases} \quad (6)$$

其中 $\lambda \in [0,1]$, 当 $\lambda = 0$ 时,式(5) 即为 TD(0) 算法. 由式(6) 可以看出,被访问次数越多的状态,其选举度 $e(s)$ 越大,从而对奖赏值的贡献也越大,因此最优策略可更快地获得.

4.2　Q-学习

1992 年, Watkins 等提出一种著名的强化学习方法: Q-learning[7]. 不同于 TD-学习的迭代仅考虑状态的值函数, Q-学习利用状态-动作对的值函数 $Q(s,a)$ 进行迭代,利用其奖赏并作为估计函数来选择下一动作,即直接优化 Q-函数. Q-学习也是通过多步迭代学习逼近最优值函数,其迭代公式为[9]

$$Q(s_t,a_t) = (1-\alpha)Q(s_t,a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1},a)) = Q(s_t,a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)).$$

其中: α 为学习率, γ 为折扣因子. Q-学习首先初始化 $Q(s,a)$ 的值; 然后 agent 根据某一探索策略(如 ε-greedy 策略),在状态 s 选择某一动作 a,执行后得到下一状态 s' 及其报酬值 r; 最后据此报酬并通过迭代公式改进 Q 值,直到实现目标状态或达到限制的迭代次数而结束一次循环. 接着从初始状态开始下一循环,最终可得到最优策略. 即在状态 s 下,最优动作使 $Q(s,a)$ 取得最大值的动作

$$\pi^*(s) = \arg\max_a Q(s,a). \quad (7)$$

为提高学习速度, Q-学习可扩展为 $Q(\lambda)$ 算法[20]. Watkins 等[7] 证明了当 α 满足绝对收敛条件时 Q-学习算法的收敛性.

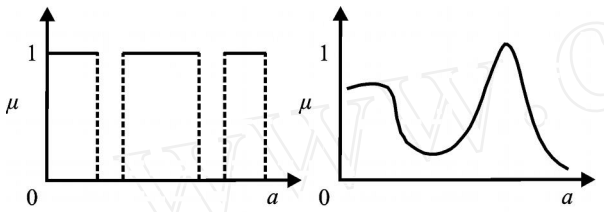
Q-学习具有较强的适应性,已成为应用最普遍的强化学习之一. 此外, Rummery 等[21] 提出的利用 Q 值迭代的基于模型的改进的 Q-学习 ——Sarsa 算法,也取得了良好的应用效果. 同时, TD 方法和 Q-学习作为较基础的强化学习方法,也为其他强化学习方法的设计奠定了基础.

经典的随机强化学习的训练需要大量的精确数据,且运算量的指数增加会出现维数灾问题,较难处

理连续状态和动作空间问题,泛化能力较弱,它仅适用于一些理论情况及其非常简单的离散问题. 同时,对在实际中获得的环境的不完备信息或不精确信息,经典强化学习的处理能力往往很弱,并且不易有效地导入与融合一些先验知识,不利于加速和指导学习过程.

5 模糊强化学习

模糊理论[22] 起源于 Zadeh 教授建立的模糊集理论,是研究不确定性(特别是模糊性) 推理与知识表示的理论基础及其应用的学科,它以模糊集理论为基础,通过模糊推理求得结果. 模糊集合的基本思想是把经典集合中的隶属关系加以扩充,使元素对集合的隶属程度由只可取 0 和 1 这两个值(见图 3(a)),推广到可取单位区间[0,1] 中的任意一个数值(见图 3(b)),从而实现定量地刻画模糊性对象.



(a) 经典集合的隶属关系 (b) 模糊集合的隶属关系

图 3 经典集合与模糊集合的隶属关系

模糊推理不需要系统精确的数学模型,它以模糊判断为前提, 运用模糊语言规则推出模糊结论. 模糊推理系统利用有限的模糊集合来描述连续的状态空间,因此可视为一种函数逼近器,将连续状态映射为连续动作;但它不具备自学习、自适应的能力,因而可与强化学习结合起来,以解决一些具有不确定性或连续的状态、动作空间的实际问题. 模糊推理系统可表达模糊和不确定知识,比较符合人类的思维方式,因此可通过将先验知识加入模糊规则库,以提高学习速度和正确性.

模糊推理方法中具有代表性的是 Takagi-Sugeno 模糊推理规则,其推理形式为[23]

R_i 如果输入 x_1 为 X_{i1}, …, x_n 为 X_{in},则输出 y_1 为 Y_{i1}, …, y_m 为 Y_{im}. 其中:R_i 为规则集的第 i 个规则;$x = (x_1, x_2, \cdots, x_n)$ 为 n 个输入变量;X_{ij} 为规则 R_i 中对应于输入变量 x_j 的模糊值,其隶属函数为 $\mu(x_j)$;$y_j (j = 1,2, \cdots, m)$ 为 m 个输出变量;Y_{ij} 为由规则 R_i 推出的输出变量 y_i 的模糊值.

自经典强化学习理论提出以来,模糊强化学习便引起了一些学者的重视[13,24]. 由于模糊系统理论的突出特点及其较好的适用性,其理论研究与实际应用也逐渐成熟起来[25,26]. 学者们从模糊推理的角度将强化学习用于在线模糊推理[25,27],并将模糊逻辑、强化学习与神经网络或遗传算法结合起来,构建了模糊神经强化学习[28],或模糊遗传强化学习[29]. Berenji 更是证明了在满足一定条件下模糊强化学习的收敛性[30].

模糊理论与强化学习相结合有不同的方式,如利用模糊性提高强化学习的学习速率及其适用范围;利用强化学习的自学习能力调整模糊输出的隶属函数,使其适用于更复杂的情况. 然而,本质上都是利用强化学习在线调整模糊输出的隶属函数,并发挥模糊理论的优点加速学习过程. 如 Yung 等[31] 利用简单的环境来学习构造模糊规则集,进而将 FRL 应用于复杂环境.

模糊强化学习的推理规则与一般的模糊推理规则类似,如假定任一规则对应于唯一的行为输出,其推理方式可描述为[11,32]

$$R_i: \text{if state } s \text{ is } S_i, \text{ then action } = a_i, \text{ with } Q\text{-value } = fq(S_i, a_i),\ a_i \in A_i. \quad (8)$$

其中:对于模糊标签 $(S_{i,j})_{j=1}^{n}$,状态 S_i 为 x_1 is $S_{i,1}$ and …and x_n is $S_{i,n}$;A_i 是在状态 S_i 时所有可能行动的集合;值函数为 $fq(S_i, a_i)$.

假定有 N 个规则,对于输入向量 $S = (S_1, S_2, \cdots, S_n)$,其行为输出结果为

$$a(S) = \Big(\sum_{i=1}^{N}\mu_i(S) \times a_i\Big) \Big/ \sum_{i=1}^{N}\mu_i(S), \quad (9)$$

对应的模糊 Q 值为

$$Q(S,a) = \Big(\sum_{i=1}^{N}\mu_i(S) \times fq'(S,a_i)\Big) \Big/ \sum_{i=1}^{N}\mu_i(S). \quad (10)$$

其中:$\mu_i(S)$ 为输入对应模糊值 S_i 的隶属函数,a_i 为行为 a_i 的量化值,$fq'(S,a_i)$ 为模糊 Q 值 $fq(S,a_i)$ 的量化值. 当然,这些模糊值及量化值既可由专家事先确定,也可在学习过程中调整.

模糊值 Q 的更新过程与经典强化学习类似,可表示为[9]

$$\Delta q(i,j) = \gamma \times \Delta Q \times e(i,j). \quad (11)$$

其中

$$\Delta Q = r + \gamma Q(S',a) - Q(S,a); \quad (12)$$

资格迹

$$e(i,j) = \begin{cases} \lambda\gamma e(i,j) + \mu_i(S) \Big/ \sum_{i=1}^{N}\mu_i(S), & j = i; \\ \lambda\gamma e(i,j), & \text{otherwise.} \end{cases} \quad (13)$$

式中:r 为在状态 S 执行行为 a 所获得的即时报酬,$Q(S',a)$ 为下一状态的 Q 值.

agent 在学习过程中,其行动选择的探索 / 利用

策略(EEP) 可采用 Boltzman 方程

$$P(x,a) = e^{FQ(x,a_i)/T} \Big/ \sum_{i}^{n} e^{FQ(x,a_i)/T}. \quad (14)$$

其中: $P(x,a)$ 为在状态 x 选择行为 a 的概率, T 为可随时间变化的温度参数, n 为在状态 x 的可能行为数. 也可采用 ε 贪心搜索策略,使 agent 达到探索与利用之间的均衡.

模糊理论具有对复杂不精确知识较强的表达能力,以及对实际问题较好的处理能力,且易于先验知识的加入,使之一直受到研究者的重视,并在实际中取得了广泛应用,特别是在移动机器人的避障[33]、导航[34,35]方面取得了较好的实验效果.

模糊强化学习适用于对象认知模糊的情况,在一些简单离散的实例中,它不如经典随机强化学习方便. 对于规模较大的系统,模糊强化学习也会出现维数灾问题. 人们经常无法较好地获得许多对象的模糊状态,仅能得知一些相关的定性信息或不精确的边缘信息,这时模糊强化学习的应用也会受到限制. 模糊强化学习存在的难题体现在:1) 模糊规则的确定问题,即模糊规则的个数以及如何划分的问题;2) 如何使模糊规则能通过在线自学习获得和更新.

6　定性强化学习

定性推理源于对物理现象的常识推理,是一种从物理系统的结构描述出发,导出行为描述和功能描述,预测物理系统的行为,并给出因果关系解释的一种非定量推理方法,是解决知识不完备的复杂系统仿真和控制难题的有效方法. Reiger 最先发表了定性推理的论文; Kleer 等关于定性推理的奠基性文章,标志着定性推理走向成熟. 随后,定性推理得到人工智能界的普遍关注[36].

定性方法可表达物理系统的因果关系,通过局部传播能在较高层次上给出系统的宏观描述. 结构描述和行为描述可理解为定量方程及其解的抽象,功能描述是对实际物理系统行为表现的一种理解. 定性推理的论域是离散变化的符号集,最简单的定性论域是 $\{+,0,-\}$, 相当于把实数轴离散化为 $\{(-\infty,0),[0,0],(0,+\infty)\}$. 开域内的值表现了定性一致的行为性质,而边界值反映的则是转折点.

如今,人们已探索出多种定性推理方法,如 Kleer 的 Envision 方法,Forbus 的定性进程理论(QPT),Kuipers 的 QSIM 方法,Iwasaki 等的因果分析法等.

定性推理可充分利用定性及不完全、不精确的信息来推理系统的定性行为,给出易于理解的行为描述和因果解释,为信息不完全的复杂系统产生行为预测,便于先验知识的加入,从而加快推理过程. 如何将定性推理应用于强化学习是一个诱人的研究方向. 1992 年,经典的强化学习方法提出不久,便有学者开始了对定性强化学习的初步研究[37],也有研究者对含有定性知识的强化学习进行探讨[38,39],但是直到 2006 年,Arkady[16] 才对定性强化学习的最主要、最突出的一种算法理论作了较好的研究与阐述.

Arkady 提出一种定性框架,它允许专家根据随机优势约束指定转移概率的不完全知识,该算法能有效解决定性问题或减少定量学习所需的探索时间. 它将定性马尔可夫决策过程(QMDPs) 与定量的概率表示有效地结合起来,使定性状态有了明确的概率解释. 该方法对于具有先验知识的模型仅需知道环境是如何工作的,而无需知道如何去探索它,这对于那些具有动态环境的一些相关信息而不知如何去解决的问题具有实际意义.

不同于通常的 MDPs,文献[16] 利用一种可泛化的近视 MDPs 框架,该 MDPs 对更早接收到的报酬更感兴趣,其结构与普通的 MDPs 既类似又有不同,为一个七元组结构: $(S,A,P,R,\otimes,\oplus,\mathrm{Next})$. 其中: S 为有限状态集, A 为有限行为集, R 为报酬函数; $P: S \times A \times S \to [0,1]$ 为转移概率函数; $\mathrm{Next}: S \times A \to S$ 为在状态 $s \in S$ 执行行为 $a \in A$ 后,能以非零概率到达的状态集;算子 $\oplus$ 定义了基于后续状态值的转移值;算子 $\otimes$ 则定义了基于所有状态 - 行为对的状态值.

Bellman 方程的泛化形式描述如下: 对于任意一个状态 s,其最优值函数可表示为

$$V^*(s) = \otimes_a^{(s)}([KV])(s,a)). \quad (15)$$

其中: $[KV](s,a) = R(s,a) + \oplus_{s'}^{(s,a)} V(s')$,而

$$\oplus_{s'}^{(s,a)} g(s') = a' \sum_{s'} P(s' \mid s,a)\, g(s'), \quad (16)$$

$$\otimes_a^{(s)} f(s,a) = \max_a f(s,a). \quad (17)$$

式(16) 和(17) 替代了传统的 MDPs 公式,其中折扣因子 $a' \in (0,1)$. 在近视 MDPs 框架中,对于固定策略 $\pi: S \to A$,其值函数 $V_\pi(s)$ 为 N 维向量; $V_{\pi,i}(i \in 1,2,\cdots,N)$ 表示 agent 从状态 s 出发,采用策略 π 后在第 i 步的期望报酬.

强化学习与马尔可夫过程密切相关,定性强化学习算法的一个重要特点是采用定性近视 MDPs 框架. 近视 MDPs 的单调属性为: 如果 $V^t(s_1) \prec V^t(s_2)$,对于所有策略评价以后序列的迭代 $q > t$,则有 $V^q(s_1) \prec V^q(s_2)$. 其中: s_1 和 s_2 为任意两种状态, $V^0 = 0$.

根据以上性质可知:1) 定性策略迭代算法仅保

持各状态值的次序轨迹,而不保留状态的具体值;2)该策略仅需将报酬的次序作为输入,不必求其状态的具体值. 定性策略迭代算法与传统的迭代算法最关键的区别在于:它是对成对状态而不是对单个状态进行操作处理.

定性强化学习利用定性策略迭代,迭代过程的输入不同于传统的 RL,输入是成对状态,通过 QMDPs 决定顺序关系 Order $\in\{$‘ < ’,‘ > ’,‘ = ’,‘ ? ’$\}$,从而选择策略和行为. QRL 利用定性 MDPs 放弃了许多不好的状态,仅保留那些较优的状态,因而学习过程更快. 由于随机优势约束对状态间通过评价概率转移的知识具有较准确的可能性解释,使得定性与定量知识的结合成为可能. Arkady 将该学习方案用于山地车和平衡摆问题[16],取得了较好的实验效果;Franklin[37] 将滚动球的轨迹分为 3 个定性状态,并用路标值表示,在有少量先验知识的情况下,取得了较好的实验效果;文献[38,39] 则将定性与定量结合起来,以扩展强化学习的适用范围. 由于定性强化学习对大系统复杂问题具有较强的处理能力,它将成为强化学习的一个重要研究方向.

定性强化学习对复杂大空间问题无疑是一种较好的处理方式,但它不适用于一些简单离散的问题,对认知不确定以及贫信息不确定的系统也不适用,并且往往难以获得对象的精确解. 因此,实际中经常需要将定量与定性信息结合起来,共同应用于强化学习.

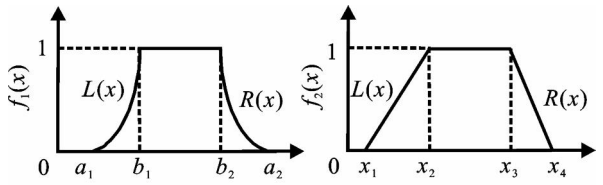
7 灰色强化学习

邓聚龙教授于 1982 年创立的灰色系统理论,是以部分信息已知、部分信息未知的小样本、贫信息不确定系统为研究对象,主要通过对部分已知信息的生成和开发,提取有价值的信息,实现对系统运行行为、演化规律的正确描述和有效监控[40]. 灰色系统模型对实验观测数据没有特殊的要求和限制,因此应用领域宽广.

灰数一般指只知道大概范围而不知其确切值的数. 在应用中,灰数实际上是指在某一区间或某个一般的数集内取值的不确定数,通常以记号 $\otimes$ 表示. 灰数有以下几种类型:仅有下界的灰数,如取数域(灰域)为 $[a,\infty)$ 的灰数 $\otimes\in[a,\infty)$,其中确定数 a 为灰数 $\otimes$ 的下确界;仅有上界的灰数,$\otimes\in(-\infty,\bar{a})$;区间灰数,$\otimes\in(a,\bar{a})$. 黑数 $\otimes\in(-\infty,+\infty)$,白数 $\otimes\in[a,\bar{a}](a=\bar{a})$ 可看作特殊的灰数.

某些灰数可找到一个白数作为其代表,该白数称为相应灰数的白化值,记为 $\hat{\otimes}$,并用 $\otimes(a)$ 表示以 a 为白化值的灰数. 对于一般的区间灰数 $\otimes\in[a,b]$,其等权白化的白化值 $\hat{\otimes}$ 取为 $\hat{\otimes}=\alpha a+(1-\alpha)b,\alpha\in[0,1]$. 当 $\alpha=0.5$ 时,所得到的白化值称为等权均值白化.

为更好地描述一个灰数对其取值范围内不同数值的偏爱程度,可根据已知信息设计其白化权函数. 图 4(a) 是一种典型的白化权函数,图 4(b) 给出了实际常用的白化权函数.



(a) 典型白化权函数 (b) 简化白化权函数

图 4 典型白化权函数和简化白化权函数

对于白化权函数为 $f[a_1,b_1,b_2,a_2]$ 的灰数 $\otimes\in[a_1,a_2],a_1<a_2$,其灰度定义为

$$g(\otimes)=\frac{2|b_1-b_2|}{b_1+b_2}+\max\left\{\frac{a_1-b_1}{b_1},\frac{a_2-b_2}{b_2}\right\}. \tag{18}$$

可见,g 受两部分的影响:峰区大小和非峰区的面积. 峰区越大、非峰区面积越大,则 g 越大,它所表示的灰度也越大. 当 $g=0$ 时,灰数 $\otimes$ 退化为一个白数.

由于灰色系统对非完备信息具有较好的处理能力,对学习结果具有较好的泛化能力,并易于引入先验知识,使得灰色系统理论在强化学习中具有较大的发展空间. 灰色系统理论可在很多方面对强化学习进行改进,如输入输出的灰色表示,构建灰色强化函数,利用灰色预测等[9].

强化学习的输入可用概率灰数来表示,如图 5(a) 所示. 概率灰数是一种以概率密度函数为白化权函数的特殊区间灰数,它由 3 个要素确定:区间 $[a,b]$,测度 $1-\alpha$ 和分布参数 σ. 对于定义在区间 $[a,b](a<b)$ 上的区间概率灰数 $X_{[a,b]}$,其白化函数为

$$f(x)=N\left[\frac{1}{2}(a+b),\sigma^2\right]=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{[x-0.5(a+b)]^2}{2\sigma^2}\right\}. \tag{19}$$

其中:σ^2 满足

$$\int_{-\infty}^{a}\frac{1}{\sqrt{2\pi}\sigma}e^{(x-0.5(a+b))^2/2\sigma^2}dx=\frac{\alpha}{2};$$

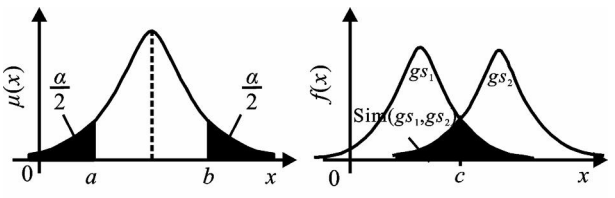
$1-\alpha$ 称为概率灰数 $X_{[a,b]}$ 的测度,记作

$$\mu(X_{[a,b]})=1-\alpha,\ \alpha\in[0,1].$$

该概率灰数可通过一定的规则由 agent 获得的数据转换而成,如图5(b)所示. 对于给定的两个概

表 3　SRL, FRL, QRL 和 GRL 的特性比较

基于知识的强化学习	处理对象	不完备信息处理能力	处理连续状态空间问题	自适应性	先验知识导入能力	泛化能力	学习速率	理论基础	维数灾问题
经典 SRL	简单离散事件	弱	弱	无	弱	无	慢	成熟	有且不易消除
模糊 FRL	认知不确定	中等	较强	一般	较强	较强	较快	较成熟	不易产生
定性 QRL	复杂连续	中等	强	强	强	强	快	不太成熟	有但易消除
灰色 GRL	贫信息不确定	强	较强	较强	较强	较强	较快	不太成熟	不易产生



(a) 概率灰数　　(b) 概率灰数的相似性

图 5　概率灰数及其相似性

率灰数,可定义其相似性

$$\mathrm{Sim}(gs_1, gs_2) = \mu(gs_{1[c,+\infty]}) + \mu(gs_{2[-\infty,c]}), \tag{20}$$

其中 c 为概率灰数 gs_1 和 gs_2 白化函数的交点. 当 $\mathrm{Sim}(gs_1, gs_2) = 1$ 时, $gs_1 = gs_2$.

利用概率灰数的相似性, 可将灰色强化函数 $GR(gs, ga)$ 定义为 gs' 与灰色目标 gg 的相似性的函数

$$GR(gs, ga) = f(\mathrm{Sim}(gs', ga)), \tag{21}$$

其中 gs' 为在状态 gs 执行动作 ga 后到达的下一状态. 从而可得强化学习(以 Q-学习为例)的更新规则

$$\mathrm{Grey}\ Q(gs, ga) \leftarrow (1-\alpha_k)\, GR(gs, ga) + \alpha_k (GR(gs, ga) + \gamma \max_{ga'} \mathrm{Grey}\ Q(gs', ga')). \tag{22}$$

其中: $\alpha_k (0 \le \alpha_k < 1)$ 为学习速率, $\gamma (0 \le \gamma \le 1)$ 为折扣因子, 灰色状态集合 $GS = (gs_1, gs_2, \cdots, gs_m)$ 和灰色动作集合 $GA = (ga_1, ga_2, \cdots, ga_n)$ 分别完全覆盖状态和动作空间.

灰色系统理论还有另外一个优点: 可用现在已知的(或不确定的)信息,并用灰色预测器来建立一个灰色模型,从而可由过去的信息预测系统未来的输出信息. 这种预测机制可为强化学习的收敛提供较好的方向(尤其是在初始阶段),因此可加快强化学习的收敛速度, 使系统具有更好的适应性能. Chen[41] 利用一个解码器、一对协作单元自适应行为选择器、灰色预测器等组成系统,实现倒立摆的平衡实验,在仅有很少先验知识的情况下取得了较好的控制性能.

由于灰色系统理论的特点,灰色强化学习仅对特定的存在贫信息的实际问题处理效果较好. 在实际中,将一些信息用灰色形式恰当地表示出来是一件困难的事情,这使得灰色强化学习的实际应用受到制约. 再加上灰色系统理论本身发展尚不成熟,它与强化学习的结合方式、理论的完备性等,都还有待进一步深入研究.

8　总结与展望

从知识表达和运用的角度,强化学习可分为经典随机强化学习(SRL),模糊强化学习(FRL),定性强化学习(QRL),灰色强化学习(GRL)等. 其特点与适用范围可简单地概括为表 3.

强化学习研究取得了可喜的成果,并使其在实际中得到了较广泛的应用. 在理论和应用方面,作者认为仍有很多问题需要研究,如下几个方面将成为近期的研究热点:

1) 定性强化学习和灰色强化学习的理论分析;

2) 结合模糊理论和定性推理探索模糊定性强化学习;

3) 结合灰色系统理论和定性推理构建灰色定性强化学习策略;

4) 从知识运用的角度论证强化学习理论的完备性;

5) 扩大强化学习在复杂不确定大系统中的应用实践.

强化学习还有很多理论及实际问题需要深入研究. 由于知识表示对强化学习的重要性,人们从这方面进行研究将对强化学习的发展起到重要作用,并促进强化学习的实际应用.

参考文献(References)

[1] Tom M M. 机器学习[M]. 北京: 机械工业出版社, 2003.
(Tom M M. Machine learning [M]. Beijing: China Machine Press, 2003.)

[2] 张汝波, 顾国昌, 刘照德, 等. 强化学习理论、算法与应用[J]. 控制理论与应用, 2000, 17(5): 637-642.
(Zhang R B, Gu G C, Liu Z D, et al. Reinforcement learning theory, algorithms and its application [J]. Control Theory and Applications, 2000, 17(5): 637-642.)

[3] Barto A G, Sutton R S, Brouwer P S. Associative search network: A reinforcement learning associative

memory[J]. Biological Cybernetics, 1981, 40(2): 201-211.

[4] Barto A G, Sutton R S, Anderson C W. Neuronlike adaptive elements that can solve difficult learning control problems [J]. IEEE Trans on Systems, Man and Cybernetics, 1983, 13(5): 834-846.

[5] Sutton R S. Temporal credit assignment in reinforcement learning [D]. Amherst: University of Massachusetts, 1984.

[6] Sutton R S. Learning to predict by the methods of temporal difference[J]. Machine Learning, 1988, 3(1): 9-44.

[7] Watkins J C H, Dayan P. *Q*-learning [J]. Machine Learning, 1992, 8(2): 279-292.

[8] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. Cambridge: MIT Press, 1998.

[9] Chen C L, Dong D Y, Chen Z H. Grey reinforcement learning for incomplete information processing [J]. Lecture Notes in Computer Science, 2006, 3959: 399-407.

[10] 陈宗海, 文锋. 基于复杂过程简化模型的DHP学习控制[J]. 控制与决策, 2006, 21(10): 1087-1091.
(Chen Z H, Wen F. Learning control of DHP method based on complex process simplified model[J]. Control and Decision, 2006, 21(10): 1087-1091.)

[11] Glorennec P Y, Jouffe L. Fuzzy *Q*-learning[C]. 6th IEEE Int Conf on Fuzzy Systems. Piscataway, 1997, 2: 659-662.

[12] Glorennec P Y. Fuzzy *Q*-learning and dynamical fuzzy *Q*-learning[C]. 3rd IEEE Int Conf on Fuzzy Systems. Piscataway, 1994, 1: 474-479.

[13] Berenji H R, Khedkar P. Learning and tuning fuzzy logic controllers through reinforcement [J]. IEEE Trans on Neural Networks, 1992, 3(5): 724-740.

[14] Berenji H R. Fuzzy *Q*-learning: A new approach for fuzzy dynamic programming[C]. 3rd IEEE Int Conf on Fuzzy Systems. Piscataway, 1994, 1: 486-491.

[15] Berenji H R. Fuzzy *Q*-learning for generalization of reinforcement learning [C]. 5th IEEE Int Conf on Fuzzy Systems. Piscataway, 1996, 3: 2208-2214.

[16] Arkady E, Gerald D J. Qualitative reinforcement learning[C]. 23rd Int Conf on Machine Learning. New York, 2006: 305-312.

[17] 赵瑞清. 知识表示与推理[M]. 北京: 气象出版社, 1991.
(Zhao R Q. Knowledge representation and reasoning [M]. Beijing: China Meteorological Press, 1991.)

[18] 黄元亮. 灰色定性仿真基础的研究[D]. 合肥: 中国科学技术大学, 2004.
(Huang Y L. Research on grey qualitative simulation base[D]. Hefei: University of Science and Technology of China, 2005.)

[19] Dayan P. The convergence of TD(λ) for general λ[J]. Machine Learning, 1992, 8(2): 341-362.

[20] Jing P, Ronald J W. Increment multi-step *Q*-learning [J]. Machine Learning, 1996, 22(4): 283-291.

[21] Rummery G, Niranjan M. On-line *Q*-learning using connectionist system [R]. Cambridge: Cambridge University, 1994.

[22] 陈水利, 李敬功, 王向公. 模糊集理论及其应用[M]. 北京: 科学出版社, 2005.
(Chen S L, Li J G, Wang X G. Fuzzy set theory and its application[M]. Beijing: Science Press, 2005.)

[23] Jouffe L. Fuzzy inference system learning by reinforcement methods[J]. IEEE Trans on System, Man and Cybernetics: Part C, 1998, 28(3): 338-355.

[24] Lin C T, Lee C S G. Reinforcement structure/parameter learning for neural-network-based fuzzy logic control systems[J]. IEEE Trans on Fuzzy Systems, 1994, 2(1): 46-63.

[25] Er M J, Deng C. Online tuning of fuzzy inference systems using dynamic fuzzy *Q*-learning [J]. IEEE Trans on Systems, Man and Cybernetics: Part B, 2004, 34(3): 1478-1489.

[26] Yan X W, Deng Z D, Sun Z Q. Fuzzy advantage learning[C]. 9th IEEE Int Conf on Fuzzy Systems. Piscataway, 2000, 2: 865-870.

[27] Jouffe L. Fuzzy inference system learning by reinforcement methods[J]. IEEE Trans on System, Man and Cybemnetic: Part C, 1998, 28(3): 338-355.

[28] Luciana F K, Marley V, Marco P. Reinforcement learning-hierarchical neuro-fuzzy politree model for autonomous agents-evaluation in a multi-obstacle environment[C]. 5th Int Conf on Hybrid Intelligent Systems. Piscataway, 2005: 551-554.

[29] Lin C T, Jou C P. GA-based fuzzy reinforcement learning for control of a magnetic bearing system[J]. IEEE Trans on Systems, Man and Cybernetics: Part B, 2000, 30(2): 276-289.

[30] Berenji H R, Vengerov D. On convergence of fuzzy reinforcement learning [C]. 10th IEEE Int Conf on Fuzzy Systems. Melbourne, 2001, 2: 618-621.

[31] Yung N H C, Ye C. An intelligent mobile vehicle navigator based on fuzzy logic and reinforcement learning [J]. IEEE Trans on Systems, Man and Cybernetics: Part B, 1999, 29(2): 314-321.

[32] Deng C, Er M J. Real-time dynamic fuzzy *Q*-learning and control of mobile robots[C]. 5th Asian Control Conf. New York, 2004, 3: 1568-1576.

(下转第 975 页)

Practice, 2003, 11(10): 1099-1111.

[3] Yue D, Han Q L, Lam J. Network-based robust H_∞ control of systems with uncertainty[J]. Automatica, 2005, 41(6): 999-1007.

[4] Zhao H, Wu M, Liu G P, et al. *H*-infinity control for networked control system (NCS) with time-varying delays[J]. J of Control Theory and Applications, 2005, 3(2): 157-162.

[5] 姜培刚, 姜偕富, 李春文, 等. 基于LMI方法的网络化控制系统的 H_∞ 鲁棒控制[J]. 控制与决策, 2004, 19(1): 17-26.
(Jiang P G, Jing X F, Li C W, et al. Robust H_∞ control for the networked control systems based on LMI [J]. Control and Decision, 2004, 19(1):17-26.)

[6] 樊卫华, 蔡骅, 胡维礼, 等. 时延网络控制系统的稳定性[J]. 控制理论与应用, 2004, 21(6): 880-884.
(Fan W H, Cai H, Hu W L, et al. Stability of networked control systems with time-delay[J]. Control Theory & Applications, 2004, 21(6): 880-884.)

[7] 邱占芝, 张庆灵. 一类不确定时延网络控制系统最优 H_∞ 控制[J]. 信息与控制, 2006, 35(1): 64-72.
(Qiu Z Z, Zhang Q L. Optimal H_∞ control for a class of networked control system with uncertain time-delay[J]. Information and Control, 2006, 35(1): 64-72.)

[8] 张先明, 吴敏. 线性多时滞不确定离散时间线性系统的时滞相关 H_∞ 控制[J]. 控制理论与应用, 2006, 23(6): 918-922.
(Zhang X M, Wu M. Delay-dependent *H*-infinity control for linear discrete-time uncertain systems with multiple unknown delays [J]. Control Theory & Applications, 2006, 23(6): 918-922.)

[9] Witrant E, Canudas C, Georges D. Remote output stabilization under two channels time-varying delays[C]. Proc of 4th IFAC Workshop on Time Delay Systems. Rocquencourt, 2003: 1-6.

[10] Wu M, He Y, She J H, et al. Delay-dependent criteria for robust stability of time-varying delay systems[J]. Automatica, 2004, 40(8): 1435-1439.

[11] Laurent E I, Francois O, Mustapha A, et al. A cone complementarity linearization algorithm for static output-feedback and related problems[J]. IEEE Trans on Automatic Control, 1997, 42(8): 1171-1176.

[12] Gao H J, Wang C H. Comments and further results on "A descriptor system approach" to *H*-infinity control of linear time-delay systems [J]. IEEE Trans on Automatic Control, 2003, 48(3): 520-525.

(上接第 968 页)

[33] Ye C, Yung N H C, Wang D. A fuzzy controller with supervised learning assisted reinforcement learning algorithm for obstacle avoidance[J]. IEEE Trans on System, Man and Cybernetics: Part B, 2003, 33(1): 17-27.

[34] Zhang R B, Shi Y. Research on intelligence robot formation based on fuzzy *Q*-learning[C]. Int Conf on Machine Learning and Cybernetics. New York, 2004, 3: 1936-1941.

[35] Beom H R, Cho H S. Sensor-based navigation for a mobile robot using fuzzy logic and reinforcement learning [J]. IEEE Trans on System, Man and Cybernetics, 1995, 25(3): 464-477.

[36] 石纯一, 廖士中. 定性推理方法[M]. 北京: 清华大学出版社, 2002.
(Shi C Y, Liao S Z. Qualitative reasoning methods [M]. Beijing: Tsinghua University Press, 2002.)

[37] Franklin J A. Qualitative reinforcement learning control[C]. 31st IEEE Conf on Decision and Control. Piscataway, 1992, 1: 870-877.

[38] Sabbadin R. Towards possibilistic reinforcement learning algorithms[C]. 10th IEEE Int Conf on Fuzzy Systems. New York, 2001, 1: 404-407.

[39] Colombini E L, Ribeiro C H C. An analysis of feature-based and state-based representation for module-based learning in mobile robots[C]. 5th Int Conf on Hybrid Intelligent System. Piscataway, 2005: 163-168.

[40] 刘思峰, 党耀国, 方志耕, 等. 灰色系统理论及其应用[M]. 北京: 科学出版社, 2004.
(Liu S F, Dang Y G, Fang Z G, et al. Grey system theory and its application[M]. Beijing: Science Press, 2004.)

[41] Chen Y J, Lin J H, Hwang K S, et al. A grey evaluation function for reinforcement learning [C]. IEEE Int Conf on Neural Networks and Signal Processing. Piscataway, 2003, 1: 58-61.