$$v_\pi(s)=E_\pi[G_t|S_t=s] \tag{5}$$

最优状态价值函数定义为:

$$v_*=\max_\pi v_\pi(s) \tag{6}$$

MDP 下的行为价值函数 $q_\pi(s,a)$ 被用于衡量在当前策略 π 下对当前状态 s 执行行为 a 的最优价值。其数学表达式为:

$$q_\pi(s,a)=E_\pi[G_t|S_t=s,A_t=a] \tag{7}$$

最优行为价值函数定义为:

$$q_*(s,a)=\max_\pi q_\pi(s,a) \tag{8}$$

2.1.4　最优贝尔曼方程

智能体通过最大化最优价值函数,将得到以下两个方程,其也被称为贝尔曼方程。

$$v_\pi(s)=E_\pi[R_{t+1}+\gamma V_\pi(S_{t+1})|S_t=s] \tag{9}$$

$$Q_\pi(s,a)=E_\pi[R_{t+1}+\gamma Q_\pi(S_{t+1},A_{t+1})|S_t=s,A_t=a] \tag{10}$$

其中,γ 是折扣因子,用于保证越后面的回报对回报函数的影响越小,刻画了未来回报的不确定性,同时也使得回报函数是有界的。通过式(9)、式(10)可以看出,贝尔曼方程由两部分组成:1)该状态的即时奖励期望;2)下一时刻状态的价值期望乘上衰减系数。贝尔曼方程是对于某一个给定策略,求其状态价值函数和行为价值函数,也即对某一策略进行评估,而强化学习最终的目标是寻找最优策略,于是引入最优贝尔曼方程:

$$V_\pi(s)=\max_a R_S^a+\gamma\sum_{s'\in S}P_{SS'}^a v_*(s') \tag{11}$$

$$Q_*(s,a)=R_S^a+\gamma\sum_{s'\in S}P_{SS'}^a\max_{a'} q_*(s',a') \tag{12}$$

强化学习的问题最终可以转化为求解最优贝尔曼方程。由于方程是非线性的,因此需要通过一些方法来求解。由此产生了两个重要分类,所有强化学习问题的解决方法基本都可以归结为基于价值的和基于策略的,其中基于价值函数的代表方法是 Q-learning[20]。Q-learning 最早由 Watkins 和 Dayan 于 1992 年提出,它的价值函数的迭代方式为:

$$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha[r+\gamma\max_{a'} Q(s',a')] \tag{13}$$

其中,α 是学习速率。

2.2　多智能体强化学习

多智能体系统由分布式人工智能演变而来,具有自主性、分布性、协调性等特点,并具备学习能力、推理能力和自组织能力。尽管智能体的概念在 20 世纪 40 年代就已经出现,但在 20 世纪 70 年代之前将多个智能体作为一个整体系统的研究却很少。直到 20 世纪 80 年代后期,分布式人工智能开始显著发展,建立在博弈论概念之上的分布式人工智能逐渐演变并最终形成了多智能体系统。之后,在多智能体系统的分布式问题求解模型中,分布式约束推理(DCR)模型(如分布式约束满足问题(DCSP)和分布式约束优化问题(DCOP))的研究和使用较为广泛。DCR 在各种分布式问题上都有应用,比如分布式传感器任务分配和分布式会议安排策略等。

最近,大量的研究关注于寻找解决多智能体系统不确定性问题的方法。在各种模型和求解方法中,分布式马尔可夫决策过程(Dec-MDP)和分布式部分可观测马尔可夫决策过程(Dec-POMDP)是不确定性情形下最常用的两种模型。不幸的是,求解 Dec-POMDP 通常是很难的。强化学习的发展给多智能体系统解决不确定性等问题提供了一种全新的思路,多智能体强化学习正逐渐成为 MAS 众多子领域中最受关注的领域。下面对多智能体强化学习的基本概念和经典方法进行了简单的介绍。

首先,MARL 的环境是以马尔可夫决策过程为基础的随机博弈框架,它是这样一个元组 $\langle S,A_1,\cdots,A_n,R_1,\cdots,R_n,P\rangle$。其中,$n$ 指多个智能体的数量;A 是所有智能体的联合动作空间集,$A=A_1\times\cdots\times A_n$;$R_i$ 是每个智能体的奖励函数,$R_i:S\times A\times S\to R$;$P$ 是状态转移函数,$P:S\times A\times S\to[0,1]$。我们假设奖励函数是有界的。

在多智能体情况下,状态转换是所有智能体共同行动的结果,因此智能体的奖励也取决于联合策略。定义策略 H 是智能体的联合策略 $H_i:S\times A\to H$,相应地,每个智能体的奖励为:

$$R_i^H=E[R_{t+1}|S_t=s,A_{t,i}=a,H] \tag{14}$$

其贝尔曼方程为:

$$v_i^H(s)=E_i^H[R_{t+1}+\gamma V_i^H(S_{t+1})|S_t=s] \tag{15}$$

$$Q_i^H(s,a)=E_i^H[R_{t+1}+\gamma Q_i^H(S_{t+1},A_{t+1})|S_t=s,A_t=a] \tag{16}$$

根据任务类型,多智能体强化学习可以分为完全合作、完全竞争和混合型。在完全合作的随机博弈中,奖励函数对于所有智能体都是相同的,即 $R_1=R_2=\cdots=R_n$,因此回报也相同,多智能体的目标就是最大化共同回报。如果 $n=2$,$R_1=-R_2$,那么两个智能体有相反的目标,随机博弈就是完全竞争的。此外,存在既不完全竞争也不完全合作的策略,称其为混合策略。

在完全合作的随机博弈中,回报可以共同最大化。在其他情况下,智能体的回报通常是不同且相关的,它们不可能独立最大化。因此,指定良好的、通用的 MARL 目标是一个难题。回顾已有文献中对学习目标的定义,其主要可以概括为两个方面:稳定性和适应性。

稳定性指智能体的学习动力的稳定性以及策略会收敛于固定。适应性确保智能体的表现不会因为其他智能体改变策略而下降。收敛至均衡态是稳定性的基本要求,即所有智能体的策略收敛至协调平衡状态,最常用的是纳什均衡。适应性体现在理性或无悔两个准则上。理性是指当其他智能体稳定时,智能体会收敛于最优反馈;无悔是指最终收敛的策略的回报不能差于任何其他策略的回报。在确定学习目标以后,我们根据不同的任务类型对经典的强化学习算法做了分类和回顾。

2.2.1　完全合作

在完全合作的随机博弈中,智能体有相同的奖励函数,此时学习目标可以表述为:

$$Q_{t+1}(s_t,a_t)=Q_t(s_t,a_t)+\alpha[r_{t+1}+\gamma\max Q_{t+1}(s_{t+1},a')-Q_t(s_t,a_t)] \tag{17}$$

与单智能体一样,智能体会采用贪心策略来最大化回报:

$$h_i(x)=\arg\max_{a_i}\max_{a_i\cdots a_n} Q^*(s,a) \tag{18}$$

然而,各智能体在做决策时是非独立的,即使它们平行学习一个共同的目标,因此考虑智能体之间的协作问题变得很有必要。Team-Q 算法[21]通过假设最优的联合行动是唯一的来避免协作问题。Distributed-Q 算法[22]在不假设协调的