1 引 0

多智能体系统(Multi-agent System,MAS)是多个智能 1 组成的集合,其目标是将大而复杂的系统建设成小而彼此互相通信协调的易于管理的系统。多智能体系统自 20 世纪 70 年代被提出以来,就在智能机器人、交通控制、分布式决策、商业管理、软件开发、虚拟现实等各个领域迅速地得到了应用,目前已经成为一种对复杂系统进行分析与模拟的工具。多智能体系统由分布式人工智能演化而来,其研究目的是解决大规模的、复杂的现实问题。在现实问题中,单智能体的决策能力远远不够。使用一个中心化的智能体解决问题时,会遇到各种资源和条件的限制,导致单个智能体无法应对错综复杂的现实环境;而使用多个智能体相互协作可以解决很多问题[1]。强化学习(Reinforcement Learning,RL)是机器学习的一种重要方法,它是一种以环境反馈作为输入目标,用试错方法发现最优行为策略的学习方法。在强化学习的数学基础研究取得了突破性的进展后,对强化学习的研究和应用日益增多[2]。目前,强化学习已被广泛应用于手工业制造、机器人控制、优化与调度、仿真模拟、游戏博弈等领域[3]。目前,结合多智能体系统和强化学习方法形成的多智能体强化学习正逐渐成为强化学习领域的研究热点之一,并在各个领域得到广泛应用[4-6]。

多智能体强化学习(Multi-Agent Reinforcement Learning,MARL)是将强化学习的思想和算法应用到多智能体系统中。20 世纪 90 年代,Littman[7] 提出了以马尔可夫决策过程(Markov Decision Process,MDP)为环境框架的 MARL,为解决大部分强化学习问题提供了一个简单明确的数学框架,后来研究者们大多在这个模型的基础上进行了更进一步的研究。最近,随着深度学习的成功,人们将深度学习的方法与传统的强化学习算法相结合,形成了许多深度强化学习算法,使单智能体强化学习的研究和应用得到迅速发展。比如,DeepMind 公司研制出的围棋博弈系统 AlphaGO 已经在围棋领域战胜了人类顶级选手,并以较大优势取得了胜利,这极大地震撼了社会各界[8],也促使研究人员在多智能体强化学习领域投入更多的精力。以 DeepMind,Open AI 公司为代表的企业和众多高校纷纷开发 MARL 的新算法,并将其应用到实际生活中,目前主要应用于机器人系统[9-10]、人机对弈[11-13]、自动驾驶[14]、互联网广告[15]和资源利用[16-17]等领域。

本文第 2 节简单介绍了多智能体强化学习的基础理论;第 3 节结合深度强化学习的最新算法,从可扩展性、智能体意图等不同角度对多智能体强化学习的最新研究进展进行了综述;第 4 节对多智能体强化学习在现实领域中的应用和前景进行了探讨;最后总结全文。

2 MARL 的基础理论

2.1 单智能体强化学习

根据反馈的不同,机器学习方法可以分为监督学习、非监督学习和强化学习。强化学习的智能体通过不断地与动态环境交互和不断地试错来进行学习。如图 1 所示,在每一个过程中,智能体感知环境的完整状态并采取行动,然后使环境进入一个新的状态。之后智能体会收到一个反馈,用于评估此次状态转移。这种反馈相比监督学习中的样本标记,其信息量要少且具有延时性。这是因为在监督学习中,智能体总是被告知采用什么动作是正确的;同时,这种反馈的信息量又比没有标记的无监督学习要多,因为在无监督学习中,智能体需要自己去发现正确的行动,并且得不到关于这次行动的任何明确的反馈[18-19]。

单智能体强化学习的目标是智能体通过与环境的不断互来学习一个最优的策略,使累计回馈最大。图 1 中,一个完整的强化学习任务有几个重要的组成部分:动作、状态、反馈、环境。强化学习的环境是马尔可夫过程,执行策略及价值函数是决策过程中比较重要的概念,最终强化学习的目标可以转化为求解最优贝尔曼方程。

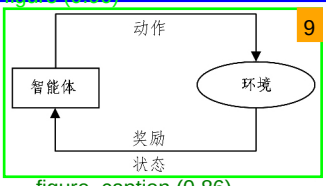

图 1 强化学习的基本框

Fig. 1 Basic framework of reinforcement learn

2.1.1 马尔可夫决策过

强化学习的环境是用马尔可夫决策过程描述的。马尔可夫决策过程是一种无记忆的随机过程,它对完全可观测环境进行描述。

马尔可夫决策过程是这样一个元组:$\langle S,A,R,P\rangle$,其中 S 表示状态空间,A 表示动作空间,R 表示奖励值,P 表示转移概率,R 与 P 都与具体行为 A 对应,数学表达式如下:

$$P^{a}_{SS'}=P[S_{t+1}=S'\mid S_t=s,A_t=a] \tag{1}$$

$$R^{a}_{S}=E[R_{t+1}\mid S_t=s,A_t=a] \tag{2}$$

其中,$P^{a}_{SS'}$ 是状态转移概率函数,R^{a}_{S} 是期望奖励函数。智能体转移到下一个状态的概率以及得到的奖励与当前状态和在此状态下采取的行为相关。

2.1.2 策

智能体采取的策略是概率的集合或分布,其元素 $\pi(a\mid s)=P[A_t=a\mid S_t=s]$ 为对过程中某一状态 s 采取可能行为 a 的概率。π 仅与当前的状态有关,与历史信息无关;同时,某一确定的策略是静态的,与时间无关。给定一个 $MDP[S,A,R,P]$ 和一个策略 π,智能体在 MDP 环境下执行策略 π 时的概率函数和奖励函数为

$$P^{\pi}_{SS'}=\sum\pi(a\mid s) \tag{3}$$

$$R^{\pi}_{S}=\sum\pi(a\mid s) \tag{4}$$

其中,$P^{a}_{SS'}$ 是指执行策略 π 时,执行某一行为的概率与采取一行为的状态转移概率的乘积和,代表了执行策略 π 时智能体从 s 转移到 s' 的概率;R^{π}_{S} 指执行策略 π 时得到的奖励是所有可能执行行为的概率与采取这一行为得到的奖励的乘积和。

2.1.3 最优价值

MDP 下的基于策略 π 的状态价值函数 $v_\pi(s)$ 表示执行策略 π 时个体在状态 s 的价值大小,其数学表达式为: