「智能博弈对抗方法」最新2022综述-博弈论与强化学习综合视角对比分析

专知
www.zhuanzhi.ai 专业可信的AI知识分发服务

8 人赞同了该文章


智能博弈对抗方法:博弈论与强化学习综合视角对比分析

袁唯淋　罗俊仁　陆丽娜　陈佳星　张万鹏　陈　璟
国防科技大学智能科学学院　长沙 410073
(yuanweilin12@nudt.edu.cn)

智能博弈对抗是人工智能认知决策领域亟待解决的前沿热点问题.以反事实后悔最小化算法为代表的博弈论方法和 以虚拟自博弈算法为代表的强化学习方法,依托大规模算力支撑,在求解智能博弈策略中脱颖而出,但对两种范式之间的关联 缺乏深入发掘. **文中针对智能博弈对抗问题,定义智能博弈对抗的内涵与外延,梳理智能博弈对抗的发展历程,总结其中的关 键挑战.**从博弈论和强化学习两种视角出发,介绍智能博弈对抗模型、算法.多角度对比分析博弈理论和强化学习的优势与局 限,归纳总结博弈理论与强化学习统一视角下的智能博弈对抗方法和策略求解框架,旨在为两种范式的结合提供方向,推动智 能博弈技术前向发展,为迈向通用人工智能蓄力.

jsjkx.com/CN/10.11896/j...

1. 导论

对抗是人类社会发展与演进的主旋律,广泛存在于人与 自然、人与人、人与机器之间,是人类思维活动特别是人类智 能的重要体现.人工智能浪潮中,对抗的形式不断发生变化, 贯穿计算智能、感知智能和认知智能３个阶段[１].以对抗关 系为主的博弈研究,为探索认知智能的关键技术原理提供了 有效工具.在认知智能层面,信息环境复杂、对抗对手复杂、 策略求解复 杂 等 愈 发 逼 近 真 实 世 界 的 复 杂 场 景 应 用 需 求, 推动了博弈对抗能力的不断提升.高 度 保 留 真 实 世 界 特 性 (巨复杂、高动态、强对抗)的智能博弈对抗技术逐渐成为了金 融、经济、交通等民用领域的技术引擎和军事智能化实现的重 要助推力.在民用领域,尤其是在保护各种关键公共基础设 施和目标的挑战性任务[３]中,智能博弈对抗技术不可或缺, 例如博物馆、港口、机场等安全机构部署有限的防护资源,在 入口处或者外围路网设置安检口进行警力的巡逻防控[４].在 军事领域,智能博弈技术积极推动了指挥与控制的智能化发 展[５],美国先后启动了“深绿”[６]、指挥官虚拟参谋[７]、“终身学习机器”“指南针”(COMPASS)等项目,旨在缩短“观察G判 断G决策G行动”(OODA)的循环时间.

近年来,在 人 机 对 抗 场 景 中,AlphaGo[８]、AlphaStar[９]、 Pluribus[１０]、Suphx[１１]、绝 悟[１２]等 一 大 批 高 水 平 AI在 游 戏 验证平台中战胜了人类玩家,智能博弈发展取得了显著突破. 智能博弈技术的巨大成功主要依赖于博弈论和强化学习两种 范式的结合[１３]:博弈论提供了有效的解概念来描述多智能体 系统的学习结果,但主要是在理论上发展,应用于实际问题的 范围较窄;深度强化学习算法为智能体的训练提供了可收敛 性学习算法,可 以 在 序 列 决 策 过 程 中 达 到 稳 定 和 理 性 的 均 衡[１４G１５].一方面,反事实后悔值最小化算法(CounterFactual RegretMinimization,CFR)[１６]是一种迭代搜索算法,其依托 大规模算力支撑,在求解大规模不完美信息博弈策略中脱颖 而出,逐渐成为了智能博弈中博弈论范式下的先进代表性算 法之一.另 一 方 面,虚 拟 自 博 弈 算 法 (FictitiousSelfGPlay, FSP)[１７]依托大规模分布式计算框架,在求解多智能体系统问 题中成为了一种通用的强化学习框架,先后被成功应用到雷 神之锤III[１８]、星际争霸[１９]、王者荣耀[１２]、德州扑克[２０]等复杂 大规模对抗场景.CFR 与 FSP是博弈范式和强化学习范式 中的典型方法,也是连接两种范式的算法基础.本文将深挖 博弈范式与强化学习范式的深层联系,为两种范式的结合提 供方向,推动智能博弈技术前向发展,为迈向通用人工智能 蓄力.

本文第２节简要介绍了智能博弈对抗,包括智能博弈对 抗的内涵与外延、智能博弈对抗发展历史以及智能博弈对抗 中的关键挑战;第３节介绍了智能博弈对抗模型,包括博弈论 的基础模型——扩展式博弈模型和强化学习的基础模型—— 部分可观随机博弈模型,以及结合扩展式博弈模型与马尔可夫模型的通用模型——观察信息可分解的随机博弈模型,从 模型上梳理了博弈理论和强化学习的内

在联系;第 4 节进行 了博弈论与强化学习的对比分析,首先详细梳理了博弈论和 强化学习视角下的典型对抗方法,分别以 CFR 和 FSP 为代 表介绍其具体原理,分析变体改进思路,然后多角度对比分析 博弈理论与强化学习的优缺点,探讨后悔值与值函数等基础 概念的联系,归纳总结博弈理论与强化学习的结合方法和结 合框架;第 5 节介绍了智能博弈对抗研究前沿,归纳了当前热 点前沿智能博弈问题,分别从面向复杂博弈场景的智能博弈 模型、多智能体场景下博弈论与强化学习融合的智能博弈方法、结合对手建模 的 智 能 博 弈 方 法,以 及 结 合 元 学 习 的 多 任务场景泛化 4 个角度讨论了智能博 弈 前 沿 研 究;最 后 总 结全文.

2. 智能博弈对抗简介

智能博弈对抗的内涵与外延

广义上的智能概念涵盖了人工智能、机 器 智 能、混 合 智 能和群体智能.本 文 的 智 能 概 念 特 指 认 知 智 能 中 机 器 的 自主决策能力,即机器智能,表现为机器模 拟 人 类 的 行 为、思考方式,通过摄像头、话筒等传感器接收 外 界 数 据,与 存 储器中的数据进行对比、识别,从而进行判断、分 析、推理、 决策.机器智能智能水平的高低可 分 为 若 干 层 次,如 从 最 简单的应激反射算法到较为基础的控 制 模 式 生 成 算 法,再 到复杂神经网 络 和 深 度 学 习 算 法.博 弈 对 抗 指 代 以 对 抗 关系为主的博弈,在 冲 突 为 主 的 背 景 下 博 弈 方 (拥 有 理 性 思维的个体或群体)选 择 行 为 或 策 略 加 以 实 施,并 从 中 取 得各自相应的结果或收 益[２１].博 弈 与 对 抗 是 人 类 演 化 进 程中的重要交互活动,是人类智能和人 类 思 维 方 式 的 重 要 体现.这种 交 互 活 动 广 泛 存 在 于 个 体 与 个 体、个 体 与 群 体、群体与群体之间.

智能博弈对抗发展历史

图 1　智能博弈对抗发展历程与典型应用

博弈对抗不断推动着智能水平的发展,对抗场景从早期 的“图灵测试”到目前的“通用场景”探索,不断向真实世界场 景靠拢. ２０１６年,DeepMind基于深度强化学习和蒙特卡洛树搜 索开发的智能围棋博弈程序 AlphaGo [８],以４ ∶ １的分数战胜 了人类顶级围棋选手李世石,这标志着人工智能的发展重点 逐渐由感知智能向认知智能过渡.同年,辛辛那提大学基于 遗传模糊树构建的 AlphaAI空战系统[２２]在空战对抗中击败 人类飞行员,这成为了无人系统博弈对抗能力生成的推动性 进展.２０１７年,DeepMind提出的基于自博弈强化学习的棋 类 AIAlphaZero [２３]可以从零开始自学围棋、国际象棋和将 棋,并击败了 AlphaGo.以围棋为代表的完全信息博弈已基 本得到解决,智能博弈的研究开始转向德州扑克和星际争霸 等不完全信息博弈.同年,阿尔伯塔大学和卡内基梅隆大学 先后开发了智能 德 州 扑 克 博 弈 程 序 DeepStack [２４]和 LibraG tus [２５],在人 机 对 抗 中 击 败 了 职业玩家.２０１８ 年,DeepG Mind在雷神之 锤III夺 旗 游 戏 中 提 出 了 一 种 基 于 种 群 训 练的多智能 体 强 化 学 习 框 架[１８],训 练 构 建 的 AIFTW 的 性能超越了人类玩家水平.随后,智 能 博 弈朝着多智能体参与、通用场景扩 展 迁 移 等 方 向 不 断 发 展,高 效 海 量 数 据 的实时采样(数据)、大 规 模 算 力 加 速 采 样 和 优 化 (算 力)、 大规模集群架构算 法(算 法)成 为 了 多 智 能 体 强化学习成功的关键.博弈均衡的方法在多智能体博弈中仍存在理论上的局限性,但基于两

人框架的多人博弈扩展依旧在实验中具有较好表现,如2019年卡内基梅隆大学的六人德州扑克智能博弈程序Pluribus[10]击败了多名职业玩家.随后,智能博弈的研究趋势开始形成“高质量对抗数据引导”+“分布式强化学习训练”的模式(如麻将AISuphx,星际争霸AIAlphaStar[19],谷歌足球AI觉悟GWeKick),并逐渐摆脱先验知识,直接完成“端到端”的学习(如捉迷藏AI[26]、斗地主AI DouZero[27]、两人德州扑克AI AlGphaHoldem[20]).2021年,DARPA举办的AlphaDogFight挑战赛[28]推动了无人系统博弈对抗能力的提升.另一方面,DARPA开始布局通用AI的探索性项目,推动智能博弈向强人工智能迈进.智能博弈对抗发展历程与典型应用总结如图1所示.

智能博弈对抗中的关键挑战

复杂博弈环境难评估

(1)不完全信息与不确定性

环境中的不完全信息与不确定性因素提高了博弈决策的难度.战争迷雾造成的不完全信息问题中,关于其他智能体的任何关键信息(如偏好、类型、数量等)的缺失都将直接影响智能体对世界状态的感知,并间接增加态势节点评估的复杂性.不仅如此,考虑不完全信息带来的“欺骗”(如隐真、示假等[29G31])行为,将进一步扩展问题的维度.此外,不确定性引入了系统风险,任何前期积累的“优势”都可能因环境中随机因素的负面影响而“落空”.如何综合评估当前态势进行“风险投资”,以获得最大期望回报,成为了研究的另一个难点.另一方面,在策略评估与演化过程中,如何去除不确定因素带来的干扰[32]成为了“准确评价策略的好坏、寻找优化的方向”的难点.

(2)对抗空间大规模

在一些复杂博弈环境中,状态空间和动作空间的规模都非常庞大(见表1),搜索遍历整个对抗空间,无论是在时间约束上还是在存储空间约束上[33]都难以满足要求.

表1 典型大规模对抗空间博弈场景

Table 1 Typical game scenarios with large adversarial space

博弈场景	状态空间	备注
桥牌	10^{67}	—
斗地主	10^{83}	—
德州扑克	10^{160}	两人无限注
围棋	10^{170}	—
王者荣耀	10^{600}	1对1场景
兵棋推演	10^{793}	城镇居民地想定[38]
星际争霸	10^{1685}	128×128地图,仅考虑400个单元的位置

模型抽象[34G35]的方法在一定程度上可以降低问题的规模,但缺乏理论保证,往往以牺牲解的质量为代价[36].即使以求解次优策略为目标,部分优化算法(如EGT[37]、一阶(FirstOrder)算法)仍旧难以直接应用到抽象后的模型.蒙特卡洛采样可以有效地加快算法的速率,但在复杂环境下,如何与其他方法结合并减小搜索中的方差依旧是研究的难点.

多智能体博弈难求解

(1)均衡特性缺失

纳什均衡作为非合作博弈中应用最广泛的解概念,在两人零和场景中具有成熟的理论支撑,但扩展到多智能体博弈时具有较大局限性.两人零和博弈具有纳什均衡存在性和可交换性等一系列优良特性[39].然而,多人博弈的纳什均衡解存在性缺乏理论保证,且计算复杂,两人一般和博弈的纳什均衡是PPAD难问题[40],多人一般和的计算复杂度高于PPAD.即使可以在多人博弈中有效地计算纳什均衡,但采取这样的纳什均衡策略并不一定是“明智”的.如果博弈中的每个玩家都独立地计

算和采取纳什均衡策略,那么他们的策 略组合可能并不是纳什均衡,并且玩家可能具有偏离到不同 策略的动机[４１Ｇ４２].

(２) 多维学习目标

对于单智能体强化学习而言,学习目标是最大化期望奖 励,但是在多智能体强化学习中,所有智能体的目标不一定是 一致的,学习目标呈现出了多维度[１３].学习目标可以分为两 类[４３]:理性和收敛性.当对手使用固定策略时,理性确保了 智能体尽可能采取最佳响应,收敛性保证了学习过程动态收 敛到一个针对特定对手的稳定策略,当理性和收敛性同时满 足时,会达到新的纳什均衡.

(３) 环境非平稳

当多个智能体同时根据自己的奖励来改进自身策略时, 从每个智能体角度来看,环境变得非平稳,学习过程难以解 释[４４].智能体本身无法判断状态转移或奖励变化是自身行 为产生的结果,还是对手探索产生的.完全忽略其他智能体独立学习,这种方法有时能产生很好的性能,但是本质上违背 了单智能体强化学习理论收敛性的平稳性假设[４５].这种做 法会失去环境的马尔可夫性,并且静态策略下的性能测度也 随之改变.例如,多智能体中单智能体强化学习的策略梯度 法的收敛 结 果 在 简 单 线 性 二 次 型 博 弈[４６](LinearＧQuadratic Games)中是不收敛的.

３ 智能博弈对抗模型

扩展式博弈模型

扩展式博弈适用于序贯决策中建模智能体与环境的重复 交互过程,尤其是存在“智能体对其他智能体之前的决策节点 不可分辨(含有隐藏信息)”或者“智能体遗忘之前的决策(不 完美回忆)”的情景.

部分可观随机博弈

与扩展 式 博 弈 的 树 结 构 不 同,马 尔 可 夫 博 弈 (Markov Game)也称随机博弈(StochasticGame),具有马尔可夫链式 结构.

通用模型

博弈理论和强化学习理论并不互斥,在模型上,博弈论的 扩展式博弈模型和强化学习的部分可观随机博弈两种模型之 间具有一定的联系,例如都可以通过放宽某些条件限制转化 为观察 信 息 可 分 解 的随 机 博 弈 (FactoredＧObservationStoＧ chasticGames,FOSG)[５８].FOSG 是 POSG 的一种扩展性变 体,模型聚焦于公共信息(PublicInformation)的表示和分解, 如图３所示.

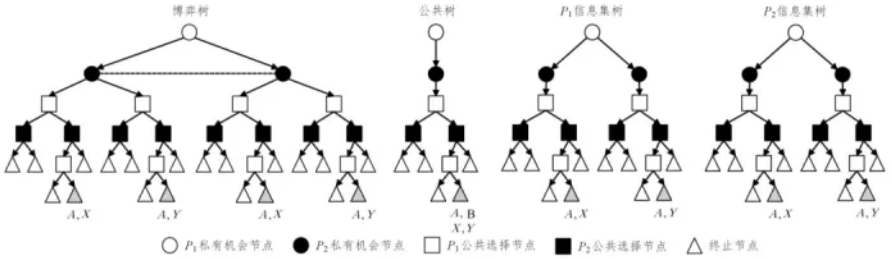


图３ 观察信息分解博弈树[80]

４ 博弈论与强化学习的对比分析

博弈论和强化学习是求解序贯决策问题的有效工具,然 而它们在算法特性(泛化性、可解释性、收敛性保证)、应用场 景(多人博弈、序贯博弈、即时策略博弈)以及硬件资源(算力 需求)等 方 面 各 有 所 长,本 文 总 结 了 近 ５ 年 AAAI,IJCAI, NeuralPS,AMMAS,ICRL等人工智能顶刊顶会中与智能博弈技术相关的博弈论与强化学习文章,按专家打分的方法,绘 制对比分析雷达图,如图７所示.博弈理论在两人零和博弈 问题上已经具有较为成熟的理论,包括纳什均衡(以及其他解概念)的等价性、存在性、可交换性(Interchangeability)[３９]等, 但在多人博弈问题中还需要新的解概念以及相关理论的支 撑.CFR算法通过后悔值迭代更新生成策略,模型具有可解 释性.但是,完美回放和终端可达的

强烈假设限制了 CFR 的 使用场景[９２].强化学习结合深度学习,直接实现端到端的学 习,具有很强的泛化性,在多智能体博弈中已取得较多成功应 用.但网络的训练往往需要超大规模的算力支撑,且模型的 可解释性不强.本节将对两种方法的具体局限性进行深入剖 析,为两种方法的结合互补提供方向.

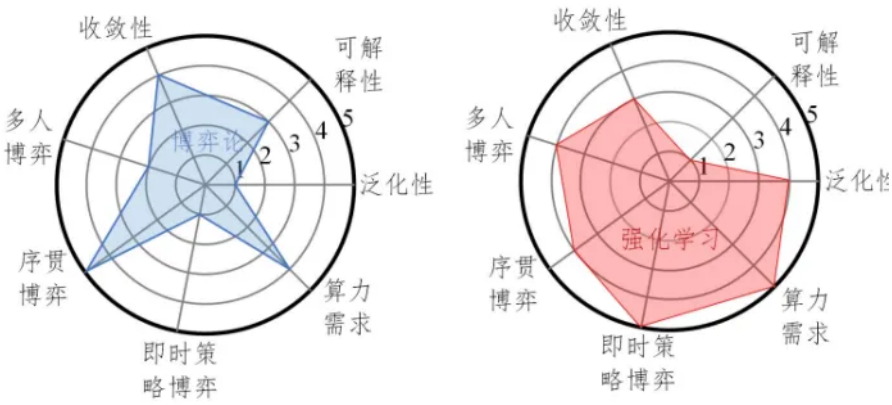


图 7　博弈论与强化学习方法的特性对比

5. 智能博弈对抗研究前沿

面向复杂博弈场景的智能博弈模型

博弈论中的“信息集”和强化学习中的“观察函数”都是针 对智能博弈场景中不完全信息的形式化描述.现实世界中,博 弈场景更加复杂,不完全信息引发了博弈信息“不可信”等问 题——智能体通常是不完全理性的,并且不同认知层次的智能 体参与的博弈具有 “欺诈[１１４Ｇ１１５]”“隐藏”“合谋”“认知嵌套(建 模对手的同时,对手也在建模利用己方)”[１１６Ｇ１１７]等新挑战.如 何针对认知博弈对抗中的新挑战,形式化描述“欺骗”等复杂博 弈要素,建立复杂博弈信息的量化评估体系,成为了智能博弈 向真实世界迁移应用的模型基础.

多智能体场景下博弈论与强化学习融合的智能博弈方法

虽然博弈论提供了易于处理的解决方案概念来描述多智 能体系统的学习结果,但是纳什均衡是一个仅基于不动点的 静态解概念,在描述多主体系统的动态特性方面(如循环集 (RecurrentSet)[１１８]、周 期 轨 道 (Periodic Orbits)和 极 限 环 (LimitCycles)[５５])具有局限性.一方面,寻求具有更多优良 特性的多人博弈新解概念,探索多人局部纳什均衡点求解方 法,将是博弈视角下求解多智能体博弈问题的新突破口;另一 方面,发挥深度学习和强化学习在信息表征、复杂函数拟合 方面的优势,基于自博弈求解框架、值函数评估方法、强化学 习结合 CFR等方法,探索博弈论方法与强化学习的有效融合 机制,将是突破多智能体博弈学习瓶颈的前沿方向.

结合对手建模的智能博弈方法

对抗胜负的本质是超越对手的相对优势,决策的制定必 须以对手的行动或策略为前提.纳什均衡是应对未知通用对 手时最小化最坏可能性,用最“保险”的策略应对,而并不是寻 求最优应对策略.放宽纳什均衡中“未知通用对手”的设定, 考虑不完全理性对手的最佳应对,一些新的解 概 念[６１]被提 出,结合显式对手建模(ExplicitOpponentModeling)[１１９]和均 衡近似,平衡利用性与剥削性,实现多目标优化,为融合对手 建模的博弈学习提供参考.此外,在一些更加复杂的对抗场 景中,如 对 手 具 有 学 习 意 识 (OpponentGLearningAwareGness)[１２０]等,最大熵多智能体强化学习[１２１]成为研究如何进行 博弈均衡对抗策略选择的新趋势.不仅如此,反对手建模方 法也在同步发展.基于意图识别设计[１２２]的对抗意图识别、 包含意图隐藏与欺骗的对抗意图识别方法等[１２３]反对手建模 方法在欺骗路径规划[２９]等问题中得到了一定的研究.在复 杂博弈对抗场景中,如何基于对手模型安全利用对手,以及如 何保全自我反对手建模成为了新的探索性研究.

结合元学习的多任务场景泛化

学习模型如何更好地泛化到差异很大的新领域中,是一 种更加高效和智能的学习方法.元学习(MetaLearning)逐 渐发展成为让机器学会学习的重要方法.元学习是通用人工 智能(GeneralAI)的分支,通过发现并推广不同任务之间的 普适规律来解决未知难题.元学习的输入是一个任务集合, 目的是对每个任务的特性和任务集合的共性建模,发现任务 之间的共性和内在规律,以追求在差异较大任务之间的迁移, 且不会产生 较 大 的 精 度 损 失,目 前 其 已 经 扩 展 到 元 强 化 学 习[１２４Ｇ１２５]、元模仿学习[１２６]、元迁移学习、在线元学习[１２７]、无监 督元学习[１２８Ｇ１２９]等.如 何 结 合 博 弈 理 论 和 元 强 化 学 习 的 优 势,构建高效、可解释性强、具有收敛性保障和泛化性的近似 纳什均衡求解体系,将是未来智能博弈技术发展的巨大推动 力之一.

结束语

本文针对智能博弈对抗问题,介绍了智能博弈 对抗的内涵与外延,梳理了智能博弈对抗发展历程,总结了其 中的关键挑战.从博弈论和强化学习两种视角出发,介绍了 智能博弈对抗模型和算法,多角度对比分析了博弈理论和强 化学习的优势与局限,归纳总结了博弈理论与强化学习统一 视角下的智能博弈对抗方法和策略求解框架,旨在为两种范 式的结合提供方向,推动智能博弈技术前向发展,为迈向通用 人工智能蓄力.

智能博弈对抗方法:博弈论与强化学习综合视角对比分析 - 专知VIP
www.zhuanzhi.ai/vip/1612538fbdd96a366a4527ffce339···

「智能博弈对抗方法」最新2022综述-博弈论与强化学习综合视角对比分析
mp.weixin.qq.com/s/us0AYQg8KRDvawPF···

编辑于 2022-08-29 15:06

博弈论

写下你的评论...

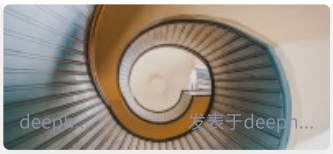
评论内容由作者筛选后展示

还没有评论，发表第一个评论吧

推荐阅读

博弈论(5) 自动机、单步偏离原理、重复博弈

在上一节我们介绍了动态博弈的框架、子博弈精炼纳什均衡与逆向递



deepl... 发表于deepl...

归法的关系. 在本节我们讨论一类更具体的动态博弈: 重复博弈. 重复博弈在实证研究中起到了很好的指引

TOVAR... 发表于高级微观经...

机器之心 发表于机器之心

钟定胜 发表于高级博