引用格式:刘赟,刘星.智能博弈对抗算法及其在情报领域中的应用[J].指挥控制与仿真,2024,46(6):49-54.LIU G,LIU X.Study of intelligent game adversarial algorithms and their applications in the intelligence field[J].Command Control & Simulation,2024,46(6):49-54.

智能博弈对抗算法及其在情报领域中的应用*

刘 赟[1], 刘 星[2]

(1. 国防科技大学外国语学院, 江苏 南京 210000; 2. 海军航空大学岸防兵学院, 山东 烟台 264000)

摘 要:智能博弈对抗算法不仅充分利用了博弈模型的刻画精度,还通过神经网络的强大计算能力和强化学习的试错机制求解均衡解,使得智能博弈对抗算法在诸多领域都取得了不错的效果。通过多智能体博弈学习、多智能体博弈强化学习和多智能体博弈深度强化学习三个层面对智能博弈对抗算法进行了系统梳理,并与情报领域的工作进行了相应映射,论证了智能博弈对抗算法运用在情报领域的可行性和必要性,最后给出了智能博弈对抗算法在情报领域的具体应用以及后续提升质效的有效措施。

关键词:智能对抗; 博弈论; 强化学习; 情报处理

中图分类号:E917 **文献标志码**:A **DOI**:10.3969/j.issn.1673-3819.2024.06.009

Study of intelligent game adversarial algorithms and their applications in the intelligence field

LIU Geng[1], LIU Xing[2]

(1. National University of Defense Technology, Nanjing 210000, China; 2. Naval Aviation University, Yantai 264000, China)

Abstract: Intelligent game adversarial algorithms not only make full use of the portrayal accuracy of the game model, but also solve the equilibrium solution through the powerful computational ability of neural network and the trial-and-error mechanism of reinforcement learning, which makes the intelligent game adversarial algorithms achieve good results in many fields. Through the multi-intelligence body game learning, multi-intelligence body game reinforcement learning and multi-intelligence body game deep reinforcement learning three levels of intelligent game confrontation algorithm is systematically sorted out, and the corresponding mapping with the intelligence field of work, demonstrates the feasibility and necessity of intelligent game confrontation algorithm in the field of intelligence, and finally gives the specific application of the intelligent game confrontation algorithm in the field of intelligence and the effective measures of the follow-up to improve the quality and efficiency. Finally, it gives the specific application of intelligent game confrontation algorithm in the field of intelligence, as well as the effective measures to improve the quality and efficiency.

Key words: intelligent game confrontation; game theory; reinforcement learning; intelligence processing

博弈强化学习结合博弈论和强化学习各自的优势,通过各类博弈强化模型在多个领域内取得了不错的成果[1-2],特别是在有限注德州扑克、无限注德州扑克和网格世界等场景中,相继打败人类顶级选手,涌现出大量的优质博弈强化学习算法。随着强化学习技术不断进步,基于值函数方法和策略梯度方法的不断优化迭代,通过结合深度学习技术,博弈强化学习能够处理更复杂的博弈场景和策略优化问题。多智能体博弈强化学习作为研究智能博弈对抗的前沿课题,面临对抗性环境、非平稳对手、不完全信息和不确定行动等挑战。目前,多智能体博弈强化学习框架已经涵盖了基础模型、元博弈模型、均衡解和博弈动力学等多个方面,在团队博弈、有限零和博弈、不完全信息扩展式博弈中均取得部分突破。但与深度强化学习相比,博弈强化学习要想进一步扩展其应用范围,提高算法的适配性,仍需在以下几个方面深入研究。

1)如何保证博弈模型能够存在稳定的纯策略纳什均衡?对于智能体而言,混合策略纳什均衡的执行往往需要进行多次决策,并以频率替代概率,这种做法会花费更多的时间和算力。因此,混合策略一直以来饱受诟病,更高效、便捷和泛化的博弈强化学习更需要纯策略纳什均衡。

2)如何求解大规模博弈对抗问题,以及如何处理博弈对抗过程中的不确定性?军事博弈对抗中“战争迷雾”是普遍存在的,并且存在大量的非线性、非逻辑成分,动态表征与强弱推理相互交织,由此产生的不确定性问题自然无法避免。如何在仅有局部信息的情况下做出较为合理的选择,如何在智能体的数量无法确定时,保证计算的科学性和高效性,都是博弈强化学习

收稿日期:2024-05-11 **修回日期**:2024-06-03

作者简介:刘 赟(1985—),男,硕士研究生,研究方向为情报处理、任务规划。

刘 星(1982—),男,博士,讲师。

算法取得实效的关键所在。

3)如何表征博弈对抗过程的指挥艺术?军事博弈对抗不仅是作战双方兵力的较量,也是博弈意志和指挥艺术的对抗。智能化战争中的制胜因素逐渐从信息优势转向智能优势,这使得我方决策目标也应从“阻断敌方信息通道”变为“毁伤敌方作战体系”,实现此目标的前提是我方作战行动的真实意图不易被察觉,故考量指挥艺术的重要性不言而喻。

情报领域内对抗涉及的主体为敌我双方,充满了多种真真假假的策略[3-5],博弈强化学习能够应用在情报领域的主要原因有以下几点:一是情报领域内的多数工作可以建模成为博弈问题,适合利用博弈强化学习算法求解计算。二是情报工作是一个多阶段、动态的活动过程,需要在不同阶段根据环境和对手的变化进行决策调整。博弈强化学习结合了强化学习的动态决策能力和博弈论的策略分析能力,能够支持情报工作者在不同阶段做出最优决策。三是在情报工作中的策略选择直接影响最终的收益和竞争优势,博弈强化学习通过建立准确的博弈模型,考虑各方面影响因素,构建更准确的博弈支付函数,从而提高策略选择的准确性,为情报工作者提供更有价值的决策支持。

本文通过对博弈强化学习的深度解析,提炼出当前博弈强化学习的重难点问题和发展方向,将核心问题与情报工作的具体要求相互映射,从理论上论证了博弈强化学习算法可应用在情报领域中,提升了情报工作的质效,加速了战斗力形成。

1 基本概念

博弈强化学习涉及的主要概念有博弈论中的纳什均衡和用于表述强化学习问题的马尔科夫决策过程,具体的定义如下。

1.1 纳什均衡

博弈论中,最核心的环节就是求解纳什均衡[6]。纳什均衡本质上是所有博弈玩家策略形成的策略集,在该策略集的加持下,每个玩家在其他玩家策略不变的情况下,该玩家的收益会因为自身策略更换而减少,即 $\forall s \in S, i = 1,2,\cdots,n$,都有如下不等式:

$$R^i(s,\sigma_*^1,\cdots,\sigma_*^n) \geqslant R^i(s,\sigma_*^1,\cdots,\sigma_*^{i-1},\sigma^i,\sigma_*^{i+1},\cdots\sigma_*^n)$$

其中,$\sigma^i \in \Pi^i$,Π^i 是玩家 i 所有可能的策略集合。

1.2 马尔科夫决策过程[7-8]

MDP 由五元组 (S,A,P,R,r) 构成。其中:S 是包含所有状态的有限集合;A 是包含智能体所有可选动作的有限集合;P 定义为 $S \times A \times S \to [0,1]$ 的状态转换函数,表示智能体从某一状态采取某动作后变为下一状态的概率,如果概率为1,则表示采取该动作后一定会到达该状态;R 定义为 $S \times A \times S \to R$ 的回报函数,回报函数指的是智能体从一个状态变换成另一个状态后,环境给他的奖励值,可能是正向奖励也可能是负向奖励;$r \in [0,1]$ 是奖励折扣系数,该系数充分考虑动作与奖励的时效性,使得智能体能够兼顾长期回报和瞬时回报,以获得最大的长期累积回报的期望。

MDP 的最终求解目标是最优策略 σ^*,而最优策略的量化指标就是期望回报值最大,该值的量化一般用最优状态动作值函数形式化表示:

$$Q^*(S,a) = \max_{\pi} E[R_t \mid S_t = s, a_t = a, \sigma]$$

如果智能体的数量大于等于1,且每个智能体所采取的动作都会对其他智能体的回报和环境产生影响,此时一般称之为多智能体马尔可夫决策过程。而单智能体马尔可夫决策过程则是多智能体的退化版,下一状态的变化仅与上一状态有关,多智能体的情况要比单智能体复杂得多。

1.3 博弈强化学习

博弈强化学习是博弈论和强化学习的结合体,其基本组成部分包含博弈智能体 N、激励函数 f、状态集 S、动作集 A、状态转移概率 T、折扣因子 γ、信息 I、行动顺序和环境。博弈强化学习是将博弈模型、均衡策略与强化学习的试错机制相结合,但学习的方式还是试错机制,属于强化学习的范畴,而此学习过程的目标是使博弈中所有智能体的长期累积回报最大,这是与经典强化学习不同的地方。

此过程应当注意的是博弈强化学习最终学到的均衡策略也是在该策略下,所有智能体没有从单方面改变自身策略的动机,故该均衡策略不一定是最优策略,此条件下智能体期望的回报值定义为

$$E[f_i(\pi_*,t) \mid S_t = s, \pi_*] \geqslant E[f_i(\pi',t) \mid S_t = s, \pi']$$

其中,π_* 为最优策略,π' 与最优策略的差异是第 i 个智能体的策略不同。博弈强化学习与强化学习、博弈论的内在关系如图1所示。

2 智能博弈对抗算法研究现状

2.1 研究现状

智能博弈对抗的场景复杂,涉及多个、多种角色,所以解决智能博弈对抗的方法和角度也存在多样性。基于智能博弈对抗的场景和方法,可从多智能体博弈学习、多智能体博弈强化学习和多智能体博弈深度强化学习三个层次梳理当前智能博弈对抗算法的研究现状。

2.2 智能博弈对抗算法

2.2.1 多智能体博弈学习

多智能体博弈学习涉及多个智能体在博弈环境中

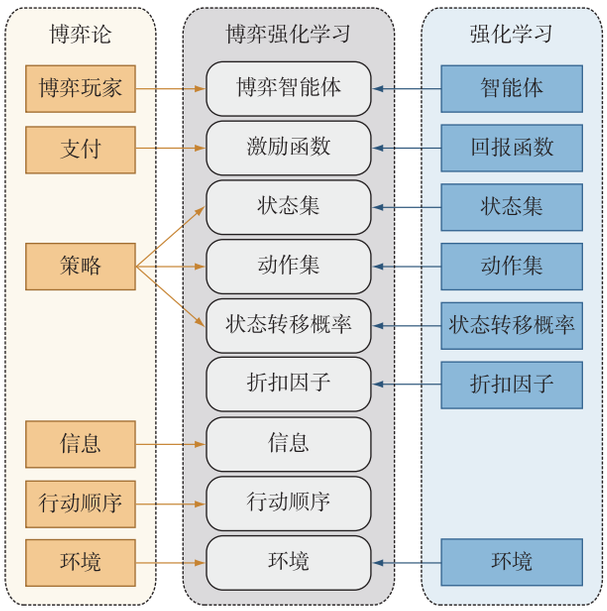

图 1　博弈强化学习与强化学习、博弈论的内在关系

Fig. 1　Game reinforcement learning with intrinsic relationship between reinforcement learning and game theory

的学习，这些智能体通过与环境的交互来适应并优化自身策略。博弈论是其理论基础，定义了动作、收益等基本概念，并侧重分析理性智能体的博弈结果，即均衡。多智能体博弈学习主要关注智能体之间的交互和协作，以及如何通过策略学习达到纳什均衡或其他稳定状态。在多智能体博弈学习方面，较为经典的算法有 Minmax-Q[9]、CFR[10-11] 和 GDA[12]。

Minmax-Q 算法的基本思想是在每个状态下，智能体都尝试找到一个动作，该动作能够最大化其未来可能的最小回报。实现方式为博弈树，其中，每个节点代表一个状态，每个边代表一个可能动作，而叶子节点则包含在该状态下采取特定动作的预期回报。CFR 算法的核心思想是通过模拟博弈过程，计算每个决策点上每个动作的“遗憾值”，并通过迭代更新每个动作的遗憾值，基于这些遗憾值重新计算每个动作的选择概率，以逐渐优化策略。GDA 算法通过计算每个类别数据的均值和协方差矩阵来估计高斯分布的参数。然后，对于新的数据点，GDA 算法根据新数据点属于各个类别的概率来进行分类。上述三类算法的特点如表 1 所示。

2.2.2　多智能体博弈强化学习

多智能体博弈强化学习是多智能体博弈学习和强化学习的结合。强化学习的核心原理是通过智能体与环境的交互，学习如何在给定情境下做出最优决策。在多智能体环境中，每个智能体都会根据环境的反馈来调整自身行为策略，以最大化累积奖励，除了考虑单个智能体的学习和决策过程外，还需要特别关注智能体间的交互和协作，以共同优化系统性能。常见的多智能体博弈强化学习算法有 NFSP[13-14]、FOF[15] 和 Nash-Q-learning 算法[16]。

表 1　典型的多智能体博弈学习算法特点

Tab. 1　Characteristics of typical multi-intelligence body game learning algorithms

算法	特点
Minmax-Q	Minmax-Q 算法适用于可以完全观察到状态信息和行动结果的场景，如井字棋、国际象棋等。然而，对于更复杂的场景，如围棋或视频游戏，由于状态空间过大，直接应用 Minmax-Q 算法的效果一般。
CFR	CFR 算法通过反事实概率来计算玩家的策略，适用于大规模不完全信息博弈。
GDA	GDA 算法在处理符合高斯分布的数据时表现良好，且计算效率较高。然而，当数据不符合高斯分布时，GDA 算法的性能会下降。此外，GDA 算法还需要计算协方差矩阵，这在处理高维数据时可能会导致计算成本过高。

NFSP 算法是一种利用强化学习技术从自我博弈中学习近似纳什均衡的方法。它结合了虚拟博弈和神经网络近似函数，适用于不完美信息二人零和博弈。FOF 算法主要是利用简化思想将所有智能体划分为朋友和敌人，作为朋友的智能体会帮助自身，反之作为敌人则会阻碍目标的实现。该算法的优势在于可以简化其他智能体对自身的影响，可以处理智能体数量较多的博弈问题，并且能够获得稳定且相对有效的解。Nash-Q-Learning 算法在 Q-Learning 的基础上将 Q 值替换为 Nash-Q 值，并要求智能体的策略选择基于纳什均衡，即每次迭代的策略是当前阶段的均衡策略，任一智能体不会独自改变自身策略，由于对均衡点的强制要求，导致该算法要求博弈的每个阶段都具有纳什均衡点，多数复杂博弈问题不一定总是满足该要求。上述三类算法的特点如表 2 所示。

表 2　典型的多智能体博弈强化学习算法特点

Tab. 2 Characteristics of typical reinforcement learning algorithms for multi-intelligent body games

算法	特点
NFSP	算法可以解决无先验知识、不依赖局部搜索的近似纳什均衡问题，适用于连续行动博弈，但计算复杂度较高。
FOF	用二人零和博弈对复杂问题进行建模，模型考虑因素较少，简化程度较高。
Nash-Q-Learning	利用均衡解引导策略选择，使最终策略具有合理性，要求博弈各阶段均有鞍点或全局最优点，收敛要求苛刻。

2.2.3　多智能体博弈深度强化学习

多智能体博弈深度强化学习在多智能体博弈强化学习的基础上引入了深度学习，深度学习允许智能体处理更复杂、更高维度的状态空间和动作空间，从而提

高决策准确性。深度强化学习使用深度神经网络来近似值函数或策略函数,使得智能体能够学习更复杂的策略和行为。在多智能体环境中,每个智能体都需要考虑其他智能体的行为对自身的影响,并制定相应的博弈策略。因此,多智能体博弈深度强化学习需要解决更复杂的决策问题,包括如何平衡竞争与合作的关系、如何对其他智能体进行建模以及如何适应策略的动态变化等。常见的多智能体博弈强化学习算法有Mean-Field[17-18]、Minimax-DQN[19]和LOLA算法[20]。

Mean-Field算法主要实行的是少数服从多数的原则,策略的选择依托于大多数智能体的策略选择,该算法的优势在于可以解决智能体的数量问题,即使智能体的数量较多时,使用该原则可以获得其他智能体及环境对于某一智能体的影响,但需要注意过度假设带来的误差问题。Minimax-DQN算法结合了DQN(Deep Q-Networks)和Minimax原理。DQN算法使用神经网络来近似Q函数,从而能够处理高维状态空间。而Minimax原理则是一种在零和博弈中寻找最优策略的方法,通过考虑对手的最坏情况来制定自己的策略。LOLA算法是让智能体在更新自己策略的同时考虑其他智能体的学习过程。通过建模其他智能体的参数更新过程来调整自己的策略,LOLA算法能够实现智能体之间的合作与双赢。上述三类算法的特点如表3所示。

表3 典型的多智能体博弈深度强化学习算法特点

Tab. 3 Characteristics of typical deep reinforcement learning algorithms for multi-intelligent body games

算法	特点
Mean-Field	Mean-Field算法通常用于处理具有大量相互作用的系统,如统计物理、神经网络等。算法通过引入一个平均场来近似系统中各个元素之间的相互作用,简化计算并提高效率。
Minimax-DQN	Minimax-DQN算法需要大量的计算资源来训练神经网络,并且对于非零和博弈问题不太适用。
LOLA	需要考虑其他智能体的学习过程,算法计算量相较于其他算法比较大。

3 智能博弈对抗技术在情报领域中的应用

3.1 智能博弈对抗技术运用在情报领域的必要性

情报领域内的诸多工作,在结构上与博弈问题极度相似,从工作质效方面考虑,智能博弈对抗技术的契合主要体现在三个方面。

一是海量数据处理。随着信息化程度的提高,情报领域产生的数据量呈爆炸性增长。传统的数据处理和分析方法已无法满足快速、准确获取有价值情报的需求。智能博弈对抗算法,如深度学习、强化学习等,能够高效处理和分析这些海量数据,为情报分析提供有力支持。

二是提高决策效率。情报分析往往需要面对复杂的决策环境,如多目标冲突、不确定性因素等。智能博弈对抗算法能够模拟真实世界的复杂情况,为决策者提供多种可能的决策方案,并评估其优劣,从而提高决策效率和准确性。

三是适应动态变化。情报环境具有高度的动态性和不确定性,需要情报分析系统能够实时响应并作出调整。智能博弈对抗算法具有自适应性和学习能力,能够根据环境变化自动调整策略,保持对情报环境的持续监控和分析。

3.2 智能博弈对抗技术在情报领域中的具体应用

人工智能技术的不断发展逐渐解决了多个难题,问题呼吁技术,技术又可以运用于实践。军事对抗或战争是技术实践的重要领域,现代战争的复杂度日益提升,决定战争走向的因素种类繁多,战场数据量也较以往大幅提升,美国防部披露每年使用超过万次无人机收集数以千万的实时数据。为在海量数据中挖掘有效情报,算法战跨职能小组(Algorithmic warfare cross functional team, AWCFT)于2017年迅速成立,主要任务就是使用机器学习算法Maven将从战场采集的各类数据加以处理和分析,并与海军陆战队的Minotaur系统相互结合实时跟踪目标。2018年,美空军研究实验室计划利用三年时间和1亿美元重点攻关网络情报智能处理和分类等人工智能技术,类似的项目还有美情报高级研究项目组的水银项目(Mercury Program)。

故智能技术在情报领域的作用正在日渐凸显,情报领域涉及大量的信息收集、整理、分析和传播,这些信息往往具有不确定性、动态性和复杂性。同时,情报领域还涉及多个利益主体之间的博弈和竞争,因此,需要使用智能博弈对抗技术对情报进行高效处理和分析。

3.2.1 情报收集与分析

智能对抗算法在情报收集与分析中的应用,主要体现在对海量数据的处理、模式识别以及深度分析上。例如,在网络安全领域,智能对抗算法可以帮助安全团队快速识别和过滤出潜在的威胁信息,如恶意软件、网络钓鱼攻击等。通过对这些威胁信息的深度分析,可以了解攻击者的行为模式、攻击路径以及目标,从而采取相应防护措施。

同时,通过对卫星图像、雷达数据以及电子侦察数据的处理和分析,算法能够自动识别出敌方的重要设施、装备部署以及兵力分布等信息。这些信息对于指挥员制定作战计划、评估战场态势以及做出决策具有

重要的参考价值。

3.2.2　情报对抗与反制

在情报对抗与反制方面,智能对抗算法主要用于识别和防御敌方的情报收集、分析和干扰行为。例如,在电子战中,智能对抗算法可以帮助我方快速识别出敌方的电子侦察设备、干扰设备以及通信设备等,并采取相应的反制措施,以应对敌方侦察预警探测。

此外,智能对抗算法还可以用于对抗敌方的网络攻击和信息渗透。通过构建复杂的网络防御体系,算法能够实时监测和分析网络流量,发现异常行为后报警。同时,算法还可以生成虚虚实实的情报,迷惑敌方情报人员,使其做出错误判断。

3.2.3　决策支持与优化

智能对抗算法能够基于当前情报数据和态势为决策者提供科学的情报分析结论,通过对历史数据的学习和分析,算法能够预测未来的战场态势和发展趋势,为指挥员制定作战计划提供科学依据。智能对抗算法还可以根据实时情报数据,对作战计划进行动态调整和优化。例如,在作战过程中,当发现敌方有新的兵力部署或战术调整时,算法可以迅速分析这些变化对作战计划的影响,并提出相应优化建议。

综上所述,智能博弈对抗算法在情报收集与分析、情报对抗与反制以及决策支持与优化这三个方面发挥着重要作用。

3.3　提升智能博弈对抗技术在情报领域质效的措施

智能博弈对抗算法在情报领域中的应用应当尽可能结合对抗实际,切实提升计算的精度和效率,以下三个可能是重点研究的方向。

1)拓展更加广义的策略评估方式。现有的策略评估方式多基于值函数对某单一策略进行评估,但是值函数自身的过估计等问题也会造成评估的不准确[21-22]。在训练过程中,智能体学习到的策略有很多个,抛开训练之初的随机策略,当训练到达特定阶段时,任一策略在特定的状态下动作也许是最优的。因此,可以通过拓展原有的策略评估方式,对多个策略以及策略之间组合的优劣进行评估。策略组合可以通过截断或裁剪的方式提取局部最优策略,形成全局最优策略。

2)引导智能体演绎科学的推理行为。智能体的智能性不应该局限于学习能力,高层次的智能性还应当具有推理能力[23-25],如何基于已有知识和博弈模型推理出自身的最佳策略,并预判可能的风险是未来该领域发展的重大挑战,这也是完成从“机器学习”到“学习机器”转变的关键环节。在此方向的一些尝试还处于初步阶段,例如各类建模理论,目前的建模只能对于底层行为进行推理。除此之外,现有的模型库虽然能够纠正一些错误推理,并对智能体的策略加以引导,但距离自主、有效和科学的推理仍有不小差距。

3)基于大脑分区构建协作型神经网络架构。异步 AC 算法中的多线程模式展示的高效性无疑应该引起更多关注,而在人的大脑中存在多个不同分区,每个分区指导人的不同行为,多个分区的团队协作使得人能够完成多种复杂行为。因此,可以通过引入上述思想,将复杂问题分解为不同类型的子问题,以特定的神经网络解决特定的子问题,该思想可使算法能够解决较复杂的军事博弈对抗问题。

4　结束语

智能博弈对抗算法的优势集中体现在建模和计算上,通过对智能博弈对抗问题的博弈建模,并依托博弈强化学习算法进行求解,有助于发现问题的本质,探索出新的战法。博弈和对抗是普遍的,不仅存在于情报领域,在军事领域的诸多方面均有不同程度的体现,故智能博弈对抗可广泛运用到多个军事领域,支撑辅助决策,提升战斗力。

参考文献:

[1] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7 676): 354-359.

[2] FANG F, LIU S, BASAK A, et al. Introduction to game theory[J]. Game Theory and Machine Learning for Cyber Security, 2021, 12(8): 21-46.

[3] 梁春华. 大数据与人工智能环境下“一主三辅”情报研究工作模式研究[J]. 情报理论与实践, 2021, 44(9): 64-67, 63.
LIANG C H. “One domain and three aid” intelligence analysis model at big-data and artificial intelligence environment[J]. Information Studies(Theory & Application), 2021, 44(9): 64-67, 63.

[4] 储节旺, 李振延, 吴蓉. 面向科技自立自强的情报保障体系研究[J]. 情报理论与实践, 2022, 45(8): 15-22, 53.
CHU J W, LI Z Y, WU R. Study of intelligence assurance system for China's self-reliance in science and technology [J]. Information Studies(Theory & Application), 2022, 45(8): 15-22, 53.

[5] 袁建霞, 冷伏海, 黄龙光, 等. 科技前沿方向的情报监测分析与综合研判方法探讨[J]. 图书情报工作, 2022, 66(19): 92-98.

江苏省地质测绘院印刷厂　\DZ06a\f\指挥控制与仿真\24 指挥 6 期　4 校样　排版:张　芸　时间　2024/11/22

YUAN J X, LENG F H, HUANG L G, et al. Exploration of intelligence monitoring, analysis and comprehensive study and judgment methods for S & T frontier direction [J]. Library and Information Service, 2022, 66(19): 92-98.

[6] KASSAY G, RĂDULESCU V D. Equilibrium problems and applications[M].Pittsburgh:Academic Press,2014.

[7] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[J]. 2nd ed.Massachusetts:MIT Press,1998.

[8] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016.

[9] ZHU Y H, ZHAO D B. Online minimax Q network learning for two-player zero-sum Markov games[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(3): 1228-1241.

[10] SCHMID M, BURCH N, LANCTOT M, et al. Variance reduction in Monte Carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 2 157-2 164.

[11] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-Agent policy gradients [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1):1 585-1 602.

[12] O SEBBOUH, M CUTURI, G PEYRÉ. Randomized stochastic gradient descent ascent[C]//International Conference on Artificial Intelligence and Statistics, Virtual Conference, 2022: 2 941-2 969.

[13] BROWN N, BAKHTIN A, LERER A, et al. Combining deep reinforcement learning and search for imperfect-information games[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2020: 17 057-17 069.

[14] ŠUSTR M, SCHMID M, MORAVČÍK M, et al. Sound algorithms in imperfect information games[EB/OL]. 2020: 2006.08740.http://arxiv.org/abs/2006.08740v2

[15] LITTMAN M. Friend-or-foe Q-learning in general-sum games[J]. ICML, 2001, 1(6): 322-328.

[16] HU J L, WELLMAN M P. Nash Q-learning for general-sum stochastic games[J]. Journal of Machine Learning Research, 2004, 4(6): 1 039-1 069.

[17] MA H, PU Z, PAN Y, et al. Causal mean field multi-agent reinforcement learning[EB/OL]. 2018.1802.05438. http://arxiv.org/abs/1802.05438v5,2018.

[18] TUYLS K, PÉROLAT J, LANCTOT M, et al. Symmetric decomposition of asymmetric games [J]. Scientific Reports, 2018, 8(1): 1 015.

[19] MISHRA B, AGGARWAL A. Opponent hand estimation in gin rummy using deep neural networks and heuristic strategies[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(17): 15 607-15 613.

[20] FOERSTER J N, CHEN R Y, AL-SHEDIVAT M, et al. Learning with opponent-learning awareness [EB/OL]. arXiv preprint arXiv:1709.04326, 2017.

[21] 王军, 曹雷, 陈希亮, 等. 多智能体博弈强化学习研究综述[J]. 计算机工程与应用, 2021, 57(21): 1-13.
WANG J, CAO L, CHEN X L, et al. Overview on reinforcement learning of multi-agent game[J]. Computer Engineering and Applications, 2021, 57(21): 1-13.

[22] CHIU C Y, FRIDOVICH-KEIL D, TOMLIN C J. Encoding defensive driving as a dynamic Nash game [C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an, 2021: 10 749-10 756.

[23] BOGACHEV V I, SMOLYANOV O G. Real and Functional Analysis [M]. Cham: Springer International Publishing, 2020.

[24] N BROWN. Equilibrium finding for large adversarial imperfect-information games[D]. US Army, 2020.

[25] WANG J, CAO L, WANG B, et al. Overview of one-dimensional continuous functions with fractional integral and applications in reinforcement learning[J]. Fractal and Fractional, 2022, 6(2): 69.

(责任编辑:张培培)