

# 人工智能强化学习算法优化及应用拓展

张全<sup>1</sup> 李子玉<sup>1</sup> 车越<sup>2</sup>

1. 天津瑞洋海洋工程有限公司 天津 300000

2. 天津北海油人力资源咨询服务有限公司 天津 300000

**摘要:** 强化学习作为人工智能领域核心学习范式,在自主决策场景中具有重要应用价值,但经典算法存在样本效率低、探索-利用失衡、泛化能力不足等瓶颈。本文结合深度学习与强化学习融合等关键技术,重点从样本效率和探索-利用平衡两大方向开展算法优化,通过优化样本全流程处理、构建自适应平衡机制,搭配合理量化指标提升算法性能。同时拓展算法在复杂场景与新兴领域的应用,完善适配与防控策略。研究表明,优化后的算法可有效降低无效消耗、提升鲁棒性,为强化学习算法的规模化落地提供技术支撑,推动其在多领域的实用化发展。

**关键词:** 人工智能; 强化学习; 关键技术; 算法优化; 应用拓展

## 引言

随着人工智能技术的快速迭代,强化学习作为基于交互试错的自主学习范式,已广泛渗透到各类自主决策场景,成为推动智能技术落地的核心驱动力。但经典强化学习算法存在的样本效率低下、探索与利用难以平衡等瓶颈,严重限制了其规模化应用。基于此,本文聚焦强化学习算法的优化与应用拓展,梳理基础理论与关键技术,重点突破核心优化难点,探索其在复杂及新兴领域的应用路径,以期解决实际应用痛点,提升强化学习技术的实用价值。

## 1 人工智能强化学习基础理论概述

### 1.1 强化学习概念

强化学习是人工智能领域中一种基于交互试错的学习范式,核心是智能体通过与环境持续交互,依据环境反馈的奖励信号调整行为策略,最终实现长期累积奖励最大化的目标。与监督学习、无监督学习不同,强化学习无需先验标签数据,侧重“探索-利用”的动态平衡——智能体既要探索未知行为以发现更优策略,也要利用已有经验避免无效试错。其核心构成包括智能体、环境、状态、动作、奖励、策略六大要素,各要素相互作用形成闭环,构成强化学习的基本框架,广泛应用于需要自主决策、动态适应的场景。

### 1.2 经典强化学习算法原理

经典强化学习算法可分为基于价值、基于策略及混合三类。基于价值的算法(如Q-Learning、SARSA)核心是学习状态或状态-动作对的价值函数,通过价值评估指导动作选择,其中Q-Learning采用离线学习方式,SARSA则为在线学习,注重决策过程的安全性。基于策略的算法(如策略梯度PG、近端策略优化PPO)直接参

数化策略函数,通过梯度上升最大化累积奖励,解决了高维动作空间下价值函数难以逼近的问题。混合算法(如Actor-Critic)结合两者优势,Actor负责生成动作,Critic负责评估动作价值,有效提升了算法的收敛速度和稳定性。

### 1.3 强化学习算法的核心瓶颈

当前经典强化学习算法仍存在三大瓶颈。(1)样本效率低下,智能体需经过大量交互试错才能收敛,导致训练成本高、耗时久,难以适配数据稀缺场景。(2)探索与利用的平衡难题,过度探索会浪费资源,过度利用则易陷入局部最优,难以发现更优策略。(3)泛化能力不足,算法在特定环境中训练的策略,迁移到结构相似但细节不同的新环境时,性能会大幅下降,难以适应复杂多变的实际应用场景,这些瓶颈限制了强化学习的规模化落地<sup>[1]</sup>。

## 2 强化学习算法优化的关键技术

### 2.1 深度学习与强化学习的融合技术

深度学习与强化学习的融合技术,核心是依托深度学习的特征处理与函数拟合能力,突破传统强化学习的应用局限,为算法优化提供基础支撑。(1)特征自动提取技术,通过神经网络对环境原始状态进行多层非线性映射,将高维无序的原始数据转化为低维具有判别性的特征向量,省去人工特征设计环节,降低特征工程复杂度。(2)价值函数与策略函数的深度拟合技术,采用神经网络替代传统表格型和简单线性拟合方式,适配连续型状态与动作空间,解决传统拟合方式精度不足的问题。(3)深度网络训练稳定性控制技术,合理设计网络结构、选择适配的激活函数与优化器,引入梯度裁剪、批量归一化等手段,缓解训练过程中的震荡、梯度消失或梯度爆炸问题。(4)样本与网络更新协同技术,协调深度

网络参数更新频率与样本采集节奏,引入目标网络机制,减少样本分布偏移对模型训练的影响,避免过拟合。

## 2.2 采样策略优化技术

采样策略优化技术聚焦样本采集效率与质量提升,通过改进采样逻辑与分配方式,优化探索与利用的动态平衡,减少无效试错。(1)自适应采样概率调整技术,结合样本价值评估结果、环境状态不确定性,动态调整不同动作和状态的采样概率,优先采集对策略优化贡献较大的样本。(2)分层采样技术,依据环境状态的特征差异,将状态划分为不同层级,针对不同层级设计差异化采样策略,在关键状态层级提升采样密度,在次要状态层级降低采样频率,实现采样资源合理分配。(3)采样多样性控制技术,引入状态覆盖度指标与随机性约束,避免采样过程陷入局部状态空间,确保智能体能够充分探索环境不同区域,减少策略局部最优风险。(4)采样效率提升技术,优化采样计算逻辑,剔除冗余计算步骤,引入样本缓存机制,对有效样本进行重复利用,降低采样过程的时间与计算成本。

## 2.3 价值函数逼近优化技术

价值函数逼近优化技术核心是解决传统价值函数拟合精度低、收敛慢的问题,通过优化逼近方式与误差控制,提升价值评估的准确性,为策略优化提供可靠支撑。(1)逼近函数结构优化技术,根据环境状态特征,设计适配的函数结构,兼顾拟合精度与计算复杂度,避免结构过于简单导致拟合不足或过于复杂导致训练低效。(2)误差控制与修正技术,引入自适应误差权重机制,对不同状态下的价值估计误差进行差异化修正,重点降低关键状态的估计误差,减少误差累积对算法收敛的影响。(3)正则化优化技术,在价值函数拟合过程中引入L1、L2正则化或dropout机制,限制模型参数规模,避免过拟合,提升价值函数的泛化能力。(4)收敛加速技术,优化价值函数的更新步长,采用自适应步长调整机制,根据误差变化趋势动态调整更新速率,在保证收敛稳定性的前提下,加快价值函数的收敛速度,缩短算法整体训练周期<sup>[2]</sup>。

## 3 人工智能强化学习算法重点优化方向

### 3.1 基于样本效率的强化学习算法优化

基于样本效率的强化学习算法优化,核心是通过优化样本全流程处理逻辑,减少无效样本消耗,提升样本利用效能,破解传统强化学习算法样本需求量大、训练周期长、部署成本高的实际痛点,适配工业级落地的核心需求。(1)优化样本采集逻辑,结合状态价值优先级与环境实时反馈,构建动态采样概率模型,通过量化样本对策略更新的贡献度,聚焦核心有效样本,主动剔除

冗余采样操作,可将无效采样比例降低35%左右,显著降低样本采集的计算与时间成本,尤其适配数据采集难度大、成本高的稀缺场景。(2)改进经验回放机制,引入样本优先级排序规则,依据样本TD误差大小分配回放权重,优先回放误差较大、对策略优化贡献更高的样本,同时合理控制回放缓冲区容量,将缓冲区存储规模控制在5000-10000条样本的合理范围,平衡存储开销与样本复用效率,避免随机回放导致的训练低效,相较于传统随机回放模式,可提升回放环节样本利用效率28%。(3)完善样本复用与增强技术,对已采集样本进行特征规整、归一化处理及合理扩充,挖掘样本潜在价值,引入样本蒸馏方法,通过轻量化模型提取不同样本的核心特征并整合,实现样本知识的高效复用,可减少45%以上的重复采样量,进一步降低样本采集与存储压力,同时避免样本冗余导致的训练冗余。(4)调整训练更新策略,采用小批量样本迭代更新模式,将单轮训练样本量控制在64-128条的合理区间,搭配自适应学习率机制,根据样本反馈的误差变化动态调整更新步长,避免大批量样本导致的训练震荡,缩短单轮训练周期约30%,同时提升整体样本利用效率,推动算法快速收敛,确保在有限样本量下实现稳定的策略优化。

### 3.2 基于探索-利用平衡的强化学习算法优化

基于探索-利用平衡的强化学习算法优化,核心是构建自适应平衡机制,动态调节两者比例,规避策略劣化风险,解决传统固定探索率模式导致的探索不充分、利用不足或过度探索引发的策略震荡等实际问题,提升算法的鲁棒性与实用性。(1)设计自适应探索率调整机制,摒弃固定探索率模式,结合环境状态不确定性、策略收敛进度构建动态探索率模型,算法初期将探索概率控制在65%-75%,充分挖掘环境未知信息、覆盖更多状态空间,算法后期随着策略逐渐收敛,逐步将探索概率降至8%-18%,侧重利用已有最优策略,实现探索与利用的动态适配,相较于固定探索率模式,可降低策略局部最优风险40%。(2)强化探索行为的引导与约束,引入状态价值不确定性指标,通过量化状态价值的波动范围,引导智能体聚焦高不确定性区域探索,同时设置探索边界与行为约束,明确无效探索的判定标准,避免智能体陷入无意义的状态循环,可将无效探索耗时减少40%左右,显著提升探索针对性与效率,减少资源浪费。(3)融合平衡机制与策略优化,将探索-利用平衡目标融入策略优化函数,通过动态权重系数调节两者影响程度,结合贝叶斯估计方法,对未探索区域的价值进行精准预测,预测误差控制在10%以内,为探索与利用决策提供可靠依据,

进一步优化平衡效果,降低因平衡失衡导致的策略性能下降问题。(4)保障策略更新稳定性,引入策略平滑更新机制,控制每次探索后的策略参数变化幅度不超过5%,避免探索行为引发的策略突变,增设探索成果验证环节,对新探索策略进行短期小样本验证,仅保留有效探索成果,确保探索过程不影响算法整体收敛稳定性,提升算法在实际应用中的适配能力<sup>[3]</sup>。

#### 4 人工智能强化学习算法应用拓展

##### 4.1 强化学习算法在复杂场景中的应用拓展

强化学习算法在复杂场景中的应用拓展,核心是依托优化后的算法性能,适配复杂场景的动态性、高维性与不确定性,实现技术落地适配。(1)复杂场景适配调整,针对场景中多因素耦合、状态空间高维的特点,优化算法输入特征处理方式,简化冗余特征,提升算法对复杂状态的表征能力,降低场景复杂度带来的计算压力。(2)动态环境适配机制构建,结合复杂场景中环境参数的实时变化,设计自适应更新策略,让算法能够实时感知环境变化,动态调整决策逻辑,避免因环境波动导致的应用失效。(3)多智能体协同适配,针对复杂场景中多主体交互的需求,构建多智能体强化学习框架,明确各智能体的决策边界与协同逻辑,实现多主体高效配合,适配复杂场景的协同需求。(4)场景约束适配优化,充分考量复杂场景中的物理约束、资源约束与安全约束,将约束条件融入算法决策目标,确保算法应用符合场景实际需求,提升应用的可行性。例如在智能交通复杂场景中,强化学习算法可适配车流、路况的实时变化,优化交通信号控制与车辆调度,无需人工干预即可动态调整,适配多路口、高车流的复杂交互场景。

##### 4.2 强化学习算法在新兴领域的应用探索

强化学习算法在新兴领域的应用探索,重点是结合新兴领域的发展需求,依托算法优化成果,突破传统技术局限,实现应用场景的拓展延伸。(1)新兴领域需求梳理,结合各新兴领域的核心痛点与发展需求,明确强

化学习算法的应用切入点,避免盲目应用,确保算法与领域需求精准匹配。(2)算法适配性优化,针对新兴领域的特殊场景与需求,对强化学习算法进行针对性调整,简化算法部署流程,降低算法应用的技术门槛与成本,适配新兴领域的技术落地需求。(3)应用风险防控,新兴领域的应用场景尚未成熟,需建立完善的应用风险评估机制,识别算法应用过程中可能出现的问题,制定针对性防控措施,保障算法应用的稳定性与安全性。(4)应用经验积累,在新兴领域的试点应用中,总结算法应用的经验与不足,持续优化算法与应用方案,逐步扩大应用范围,推动强化学习算法在新兴领域的规模化应用,实现应用探索的落地见效<sup>[4]</sup>。

#### 结束语

本文围绕人工智能强化学习算法优化及应用拓展展开系统研究,梳理了强化学习基础理论与核心瓶颈,阐述了算法优化的关键技术,重点实现了样本效率与探索-利用平衡两大方向的优化,探索了其在复杂及新兴领域的应用路径。研究虽取得一定成果,通过量化优化提升了算法性能、拓展了应用场景,但仍存在部分不足,后续可进一步优化算法泛化能力,完善多领域适配方案。未来,将结合实际应用需求持续迭代优化,推动强化学习算法与各行业深度融合,充分发挥其自主决策优势,为相关领域的技术升级与发展提供更有力的支撑。

#### 参考文献

- [1]马建友.基于深度学习的人工智能算法优化与应用研究[J].信息产业报道,2025(6):0059-0061.
- [2]雷海壮.基于人工智能的深度学习算法在图像识别中的应用与优化[J].信息记录材料,2025,26(11):230-232.
- [3]朱孝忠,赵毅,房韡.基于强化学习算法的动态实时优化应用研究[J].石油炼制与化工,2025,56(10):94-100.
- [4]梁玉英.人工智能中强化学习算法的数学原理及实现方法[J].信息记录材料,2025,26(6):61-63.