摘要

强化学习中，智能体通过与环境交互并利用交互所得的经验数据提升策略，从而完成相应的决策任务。近些年，深度学习的兴起使得利用数据进行反向传播训练得到有效的特征表达成为可能，克服了传统强化学习中特征表达难以设计的问题，从而将强化学习推广到高维复杂决策任务中，例如围棋、机器人控制、大模型价值对齐等。然而很多决策任务不能仅依赖一个智能体独立完成，而是需要多个智能体协同决策，例如多无人机控制、电力分配系统等。多智能体强化学习旨在训练多个智能体使其能合作地完成决策任务，其对于这类涉及合作的实际应用场景有着广阔的前景。

人类社会是一种更高级的合作形式，其很多特征可以为多智能体学习算法提供启发。本文研究思路来源于人类社会中两个重要特征：社会参与个体的去中心化和个体之间的社会关系所对应的社会化行为。社会参与个体往往仅利用自己的局部信息独立地优化策略，这一去中心化的学习范式对于很多受限于隐私、通信或智能体数量规模而无法获得其他智能体信息的实际场景有重要帮助。在去中心化学习中，每个智能体不能获得其他智能体动作等信息，因此只能将其他智能体视为环境的一部分。由于其他智能体的策略在同步更新，从每个智能体的视角来看，所交互的环境的转移概率，即智能体执行决策后环境所转移到各个状态的概率，是不稳定的，这使得绝大部分已有算法在去中心化学习下缺少收敛性和最优性的理论保证。如何在不稳定的转移概率下使智能体去中心化地收敛至全局最优策略，是多智能体强化学习领域的根本性问题之一。目前学术界对于去中心化学习的研究非常有限，已有的工作均不能从理论上解决这一难题。由于其重要的应用场景和有限的已有研究两方面原因，本文将去中心化学习作为研究重点。

本文将从在线学习和离线学习两个方面分析研究去中心化学习。在线学习方面，为解决转移概率不稳定所带来的收敛性与最优性缺失问题，本文分别针对确定性环境和随机环境进行研究。通过去中心化地建模理想转移概率并使智能体在理想转移概率上独立更新值函数，提出一个确定性环境下能够收敛至最优策略的去中心化算法。对于更一般化的随机环境，提出一种新的去中心化更新值函数的算子：最优可能算子，从而得到第一个随机环境下能够收敛至最优策略的去中心化算法。离线学习方面，转移概率不稳定问题转化为转移概率偏差问题。本文首先分析了去中心化学习中离线转移概率和在线转移概率之间的偏差会导致较大的值估计误差。为了缓解转移概率偏差，本文引入两个权重因子修正离线转移概率，得到第一个解决转移概率偏差的离线去中心化学习方法。本文进而研究了对离线预训练策略的去中心化微调，根据少量在线交互

得到的数据，提出一种基于转移相似度的优先级采样来修正离线转移概率使其接近在线转移概率，从而实现高效的离线预训练策略的去中心化微调。

尽管社会参与者往往是去中心化的个体，但个体之间存在广泛的社会关系并对应于不同的社会化行为。人类在发展过程中形成了多种社会化行为，如分工合作、利他主义、公平、层次化管理等等。这些社会化行为提升了社会生产力，促进了社会稳定。类比地，多智能体系统也是一个微型的社会结构，如果智能体能够在学习过程中不依赖先验知识自发形成这样的符合人类认知的社会化行为，将有利于提升多智能体系统的效率与稳定。本文选取了两种常见的社会化行为：个性和公平性，用以探究如何促进社会化行为在多智能体学习过程中产生以及其对多智能体系统的帮助。在个性方面，本文从信息论的角度提出将预测智能体所获的环境局部观察来自于该智能体的概率作为个性化奖励给予该智能体，以鼓励智能体逐步分化出个性化策略，从而克服智能体策略同质化的问题。在公平性方面，本文提出一种公平-效率奖励函数，并证明了若每个智能体独立提升自己的公平-效率奖励可以实现系统的帕累托最优和资源的公平分配，从而帮助智能体在一些资源共享的应用场景中实现公平与效率的兼顾。这两个方法均通过额外的奖励函数诱导相应的社会化行为产生，因此可以方便地施加到所提出的去中心化算法之上。

概括来讲，本文研究了去中心化多智能体强化学习和多智能体系统中社会化行为的产生。在去中心化学习方面进行了一系列开创性的研究，包括提出首个在随机环境中收敛至全局最优策略的去中心化算法、首次探究去中心化离线学习与策略微调中的转移概率偏差。在社会化行为方面通过促进个性化策略和智能体公平性两种行为模式的产生，有助于解决策略同质化和资源分配不均衡两种多智能体学习中常见的难题。这些研究对于存在智能体数目较多、智能体通信受限、智能体需要与人类社会合作等限制要求的现实应用具有一定的应用价值。

关键词：机器学习，强化学习，多智能体