摘要

深度学习技术的发展，为策略学习算法在更复杂的现实场景中的应用提供了可能性。对于传统的策略学习算法，其目标往往是学习一个智能体策略，使其能够优化某一目标函数。然而现实应用场景往往会对智能体策略有更多的要求。例如，由于现实场景的复杂性，策略学习算法在学习过程中往往需要处理一个非常庞大的状态-动作空间。在这种情况下，智能体策略需要有更好的探索能力，以便更高效地对环境进行探索。再例如，由于现实场景的不确定性，策略学习算法在训练时往往难以涵盖实际应用场景中所有的可能性。在这种情况下，智能体策略需要更好的泛化能力，以保证其在意外情况下仍有较好的表现。

为了满足上述这些现实应用场景中的复杂需求，本文围绕信息论启发的策略学习算法这一主题进行了研究。在此类算法中，对智能体策略的额外要求（如探索能力、泛化能力等）可以被转化为一系列信息论指标。通过将这些描述智能体能力的信息论指标引入原有的优化过程中，策略学习算法便可以在优化原有目标函数的同时，确保所学得的智能体策略具有额外的良好性质。同时，由于通过信息论指标对原有优化目标进行了增强，本文也针对这些增强后的目标函数，对策略学习算法的学习过程进行了相应的调整，使得算法在理论上具有一定程度的保证。具体来说，本文的主要贡献包括：

1. 针对多智能体强化学习中的探索问题，本文提出了一种基于相关性的多智能体条件策略迭代算法。为了鼓励多智能体系统对环境的探索，本文引入了多智能体系统的联合策略熵对原有的目标函数进行增强。为了适应这一增强目标函数，同时也为了提高多智能体系统联合策略的表达能力，本文通过考虑智能体间的相关性，设计了多智能体条件策略迭代算法，并对其进行了理论分析。同时，本文进一步提出将具有相关性的联合策略分解为无需相关性的联合策略和考虑相关性的修正项两部分，以实现多智能体系统的分布式执行。通过多个多智能体合作场景下的实验，本文验证了这一算法可以在收敛速度或最终智能体性能方面超越其他基线算法。

2. 针对多智能体系统在动态编队下的泛化性问题，本文提出了一种基于互信息正则化的多智能体策略迭代算法。为避免智能体在决策时过度依赖编队相关信息而导致其在未见过的编队中出现性能退化，本文通过最小化智能体策略和编队相关信息之间的互信息，实现了对智能体策略的正则化。为了对互信息增强的目标函数进行优化，本文首先提出基于固定边缘策略的多智能体策略迭代，并

对其进行了理论分析。随后，本文进一步提出使用虚拟编队分布来近似动态边缘策略的方法，并得到最终的多智能体策略迭代算法。通过多种多智能体合作场景下的实验，本文验证了该算法在提升编队泛化能力方面的有效性。

3. 本文进一步将基于互信息正则化的方法应用于具有自然语言动作空间的强化学习问题中，旨在解决因自然语言组合性导致的动作空间过大问题。本文利用预训练语言模型分析动作语义和当前任务之间的相关性，并据此实现对动作空间的缩减。同时，针对预训练语言模型中先验知识和环境实际情况可能存在的偏差，本文提出了一种基于互信息正则化的策略优化和语言模型自适应算法，并对其理论性质进行了分析。通过将语言模型自适应过程引入策略迭代过程，该算法可以在策略学习的同时，根据环境实际情况对预训练语言模型进行动态调整，以防止知识偏差对智能体策略优化的干扰。通过多个文字游戏场景下的实验，本文验证了这一算法在提升训练速度和智能体最终性能方面的有效性。

4. 最后，针对视频条件策略的组合泛化问题，本文提出了一种基于信息瓶颈的视频条件策略学习算法。通过对信息瓶颈增强目标函数的优化，该算法鼓励智能体在训练集任务中隐式地学习一系列技能，并将这些技能重新组合以泛化到未出现在训练集中的任务上。在这一学习过程中，本文通过信息瓶颈指标来控制视频编码器得到的视频表征所包含的信息量，确保其尽可能仅编码与当前技能相关的信息，并从理论上对这一方法进行了验证。通过这种方法，智能体策略可以在面对未见过的任务视频时，从中提取出已学到的技能信息，从而通过重新组合这些技能，实现对未见过的任务的泛化。通过多个机器人场景下的实验，本文验证了这一算法可以在组合泛化性方面取得超越基线算法的表现。

总体而言，本文研究了一系列信息论启发的策略学习算法。在这些算法中，本文通过引入不同的信息论指标对策略学习算法原有的目标函数进行了增强，并通过优化这些增强后的目标函数来学习具有理想性质的智能体策略。同时，针对这些增强后的目标函数，本文也对策略学习算法的学习过程进行了相应的调整，以保证算法的理论性质。最后，通过不同场景下的多种实验，本文验证了此类方法的有效性，为这类方法在现实场景中的应用提供了可能。

关键词：策略学习，强化学习，多智能体强化学习，模仿学习