摘要

近十年来，深度学习的快速发展，极大地推动了人工智能领域的技术革新。深度学习也在图像识别、文本检测、语音识别等各个领域大放异彩，实现了前所未有的突破。深度学习同时也推动了对抗博弈领域深度强化学习的发展，著名的Alpha Go、Libratus等均是在对抗博弈领域新兴的深度强化学习算法模型，在围棋和德州扑克各自取得了傲人的成绩。此外，在电子竞技、自动驾驶、量化投资，甚至在军事对抗等领域也都相继出现了深度强化学习的身影。深度强化学习广泛的应用场景意味着深度强化学习的研究发展关系着整个科技领域竞争。

对抗博弈长期客观存在于人类的社会生活之中，如何使机器智能化地进行博弈决策一直以来都受到研究人员的关注。对抗博弈中的对手策略的不确定性是这类问题棘手的根本原因，归根到底是因为其他同样有自主决策能力的智能体的策略、目的不可知。对手建模就是这类问题的解决方式之一，即通过观察对手的历史行为动作，并对其进行建模，用以预测未来对手的行为动作或者推断对手的目的意图。传统对手建模方法关注于客观环境对对手行为产生的影响而忽视了对手策略与智能体策略之间的依赖关系。

本文受到人类博弈中递归推理的启发，结合对抗博弈场景中的特点，设计了基于环境模型的对手建模算法。该方法将对手建模模块分为两个部分：递归想象和贝叶斯混合。递归想象是使用自身策略和对手策略的嵌套推理关系，在环境模型中模拟递归推理过程，形成一系列不同层次的想象对手策略。贝叶斯混合利用真实对手的行为动作当做贝叶斯后验信息，在想象对手策略的基础上进行混合，不断逼近真实对手策略。基于环境模型的对手建模先从递归想象中形成对手可能产生的变化策略，再以贝叶斯混合拟合对手的学习模式，从而达到建模不同学习类型对手的目的。

首先，本文算法在三角博弈、足球一对一、捕猎者三个环境上，分别与固定策略对手、学习型对手、思考型对手三类对手进行对抗测试。实验结果相较于对手策略适应性领域其他先进算法有明显的效果提升，特别是在对抗学习型对手和思考型对手时，算法体现出基于递归推理的对手建模的优势，能够提前预判对手将要变化的策略，并加以利用，在对抗中形成利于自身的策略。其次，本文还对算法进行了消融研究，分别测试了递归想象和贝叶斯混合的模块性能，结果表明单独运行递归想象某一层想象对手策略并不能够适配真实对手的策略，同样地将随意生成的想象对手策略进行混合也无法达到递归想象模块所表现出的性能。再次，本文进行了超参数的敏感度分析，从规划长度和递归想象层数两个参数分别进行了实验比较，综合来看规划长度由于受到

计算性能和误差影响的限制，不宜取过大；递归想象层数在通常情况下使用较小的值也可取得不错的效果。综合来看，本文算法结构耦合强，从理论和直觉来看也符合现实逻辑，最终算法性能表现超过了现有算法的水平，能够实现适应不同学习类型对手并形成有利于自身策略的目的。

关键词：强化学习，多智能体，对手建模

Model-based Opponent Modeling

Xiaopeng Yu (Computer Application Technology)
Directed by Zongqing Lu

ABSTRACT

In the past decade, the rapid development of deep learning has greatly promoted technological innovation in artificial intelligence. Deep learning has made unprecedented breakthroughs in various domains such as image recognition, text detection, and speech recognition. It has also boosted the evolution of reinforcement learning in competitive scenarios. AlphaGo and Libratus are famous algorithms in deep reinforcement learning, which have gained remarkable achievements in Go and Texas Hold'em, respectively. In addition, deep reinforcement learning has also emerged in the fields of eSports, autonomous driving, quantitative investment, and even military confrontation. Deep reinforcement learning has a wide range of application scenarios, which is critical to advance science and technology.

Competitive scenarios have long existed objectively in human social life, and how to enable the agent autonomously make decisions has long been of interest. The uncertainty of the opponent's policy in adversarial games is the fundamental reason why such problems are tricky, because the policies and goals of the opponent with autonomous decision-making capabilities are not known. One of the solutions to such problems is opponent modeling, in which the historical actions of the opponent are observed and modeled to predict the future actions of the opponent or to infer the goal of the opponent. Traditional opponent modeling methods focus on the effect of the environment on the opponent's action, but ignore the effect between the opponent's policy and the agent's policy.

Inspired by the recursive reasoning and considering the features of competitive scenarios, the paper proposes an algorithm, model-based opponent modeling. The opponent modeling module is composed of two parts: recursive imagination and Bayesian mixing. Recursive imagination uses the nested reasoning relationship between the agent's policy and the opponent's policy to simulate the recursive reasoning in the environment model, forming a series of different levels of imagined opponent policies. Bayesian mixing uses the true actions of the opponent as Bayesian posterior information to mix imagined opponent policies in order to approximate the real opponent's policy. The model-based opponent modeling first forms the possible change of the opponent's policy from recursive imagination, and then fits the learning patterns of the opponent with Bayesian mixing, so as to achieve the purpose of modeling different learning types of opponents.

In experiments, the algorithm is tested against three types of opponents: fixed-policy opponents, naïve learners, and reasoning learners. The test environments are Triangle Game, One-on-One, and Predator-Prey. First, the experimental results show significant improvement compared with other state-of-art methods in this domain. Especially when fighting against learning and thinking opponents, the algorithm shows the advantage of recursive-reasoning-based opponent modeling, which predicts the opponent's future policy and exploits it. Second, the paper performed ablation study to test the performance of the recursive imagination and Bayesian mixing respectively. The results show that an imagined opponent policy does not adapt to the real opponent policy, and mixing randomly generated imagined opponent policies does not achieve good performance. Thirdly, the paper performed the sensitivity analysis of hyperparameters, planning length and the number of recursive imagination layers. In general, the planning length should not be taken too large due to the limitation of computational cost and environment model error, and the number of recursive imagination layers can also achieve good results by using smaller values. Finally, the structure of the algorithm is tightly coupled, it also conforms to realistic logic, both from a theoretical and intuitive point of view. The algorithm performance outperforms other methods and achieves the purpose of adapting to different learning types of opponents.

KEY WORDS: reinforcement learning, multi-agent, opponent modeling