DOI: 10.11992/tis.202211028
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20230731.1007.002

麻将博弈 AI 构建方法综述

李霞丽[1,2], 王昭琦[1,2], 刘博[1,2], 吴立成[1,2]

（1. 中央民族大学 信息工程学院, 北京 100081; 2. 中央民族大学 民族语言智能分析与安全治理教育部重点实验室, 北京 100081）

摘　要: 麻将及其不同变体的规则复杂, 构建高水平的麻将博弈 AI (artificial intelligence) 算法及其测试环境等面临巨大挑战。本文分析了麻将博弈的相关研究文献, 梳理出基于知识和基于数据的两大类麻将 AI 构建方法, 分析了每种类型的构建方法的优势和局限性, 重点分析了 Suphx 构建方法。指出了麻将 AI 构建面临的问题和挑战; 提出将经验回放、分层强化学习、好奇心模型、对手模型、元学习、迁移学习、课程学习等应用到麻将博弈 AI 算法优化中, 构建多元化的麻将 AI 评估指标、通用对抗平台和高质量的数据集等未来的研究重点。

关键词: 机器博弈; 非完备信息博弈; 麻将; Suphx; 知识; 对手建模; 深度学习; 强化学习
中图分类号: TP39　　**文献标志码:** A　　**文章编号:** 1673−4785(2023)06−1143−13

中文引用格式: 李霞丽, 王昭琦, 刘博, 等. 麻将博弈 AI 构建方法综述 [J]. 智能系统学报, 2023, 18(6): 1143–1155.
英文引用格式: LI Xiali, WANG Zhaoqi, LIU Bo, et al. Survey of Mahjong game AI construction methods[J]. CAAI transactions on intelligent systems, 2023, 18(6): 1143–1155.

Survey of Mahjong game AI construction methods

LI Xiali[1,2], WANG Zhaoqi[1,2], LIU Bo[1,2], WU Licheng[1,2]

(1. School of Information Engineering, Minzu University of China, Beijing 100081, China; 2. Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing 100081, China)

Abstract: Mahjong and its different variants have complex rules. Therefore, building a high-level Mahjong game artificial intelligence (AI) algorithm and its test environment is challenging. Through the analysis of relevant research literature on Mahjong game, this paper identified two types of Mahjong AI construction methods based on knowledge and data. Moreover, the advantages and disadvantages of each typical method are analyzed, emphasizing the construction method of Suphx. The problems and challenges encountered in constructing Mahjong AI are identified, suggesting the need to apply experience replay, hierarchical reinforcement learning, curiosity model, opponent model, metalearning, transfer learning, and curriculum learning to the AI algorithm optimization of Mahjong game and construct diversified Mahjong AI evaluation indicators, general confrontation platforms, and high-quality data sets. These problems are all promising research directions for the future.

Keywords: computer games; imperfect information game; Mahjong; Suphx; knowledge; opponent modeling; deep learning; reinforcement learning

机器博弈是人工智能研究领域的一个重要分支, 根据游戏参与者对他人信息的可知程度, 机器博弈分为完备信息和非完备信息博弈。复杂条件下的多智能体博弈[1-3]是当前研究的热点, 对实践和生产具有重要意义, 可以提高交通决策、优化智能生产、甚至对军事控制领域也有一定影响。麻将是典型的非完备信息博弈游戏, 其隐藏信息复杂、随机性强、参与者多, 是研究复杂条件下多智能体博弈的基础方向之一。麻将博弈 AI(artificial intelligence) 研究多以台湾麻将、日本麻将为主, 近年来也对中国的麻将展开了研究。麻将博弈 AI 构建最初大多采用基于知识的方法, 即将设计者的经验和领域专家的理解编为计算机语言, 指导 AI 的决策。随着神经网络、机器学习、深度学习、强化学习等应用于机器博弈, 麻将 AI 构建发展为基于数据的方法, 即从大量的数据中提取出特征, 利用模型的模拟能力和自学习能力, 通过不断训练得到稳定的决策模型。采用基于数据的方法训练的 AI 博弈水平越来越高, “爆打”[4] 高于人类的平均水平, Suphx [5] 超越人类高手的水平。

收稿日期: 2022−11−18.　**网络出版日期:** 2023−07−31.
基金项目: 国家自然科学基金项目（61873291, 62276285）.
通信作者: 吴立成. E-mail: wulicheng@tsinghua.edu.cn.

本文对麻将博弈的相关文献进行梳理和分析，从基于知识和数据的角度进行分类论述，还分析了当前水平最高的麻将博弈 AI Suphx 算法，以供其他复杂环境下智能体博弈的研究者参考。麻将博弈和其他复杂环境的多智能体博弈游戏一样，其 AI 构建面临奖励稀疏、算法通用性差、对手建模研究薄弱等科学问题。此外，麻将还面临博弈水平的测试环境不完善等现状。本文还指出了麻将博弈未来的研究重点所在，不仅推进麻将博弈的研究，也为解决复杂环境的多智能体博弈提供可行思路。

1　麻将博弈复杂度分析

麻将博弈具有玩家关系复杂、非完备信息庞大、博弈奖励稀疏、得分计算复杂等特点。

一盘麻将的信息集数量 I 计算公式为

$$I = \mathrm{C}_{136}^{13} \times \left(1 + 34^4 + \cdots + 34^{80}\right) \times 7^3 \approx 5.48 \times 10^{142} \quad (1)$$

一盘游戏中，第 1 轮每位玩家都拥有 13 张手牌，其未知牌的数目为 123 张，每个信息集的大小为约为1.52×10^{49}；第 2 轮，减去上一轮中公开的 4 张牌，未知牌剩余 119 张牌，每个信息集的大小约为3.23×10^{48}。以此类推，最终得到信息集的平均大小约为10^{48}。

由图 1 可以看出，国标麻将相对日本麻将有更长的出牌序列，相对信息集数目更大，但两者的信息集平均大小一致。麻将的信息集平均大小远高于德州扑克、桥牌等非完备卡牌类游戏。虽然信息集数目低于围棋，但每次决策均需考虑10^{48}种情况，因此，麻将博弈 AI 构建更具有挑战性，研究者不断尝试探索麻将博弈 AI 构建的关键技术以在游戏策略、弃牌、吃牌、碰牌、杠牌、听牌、和牌等决策中获得更好的表现。

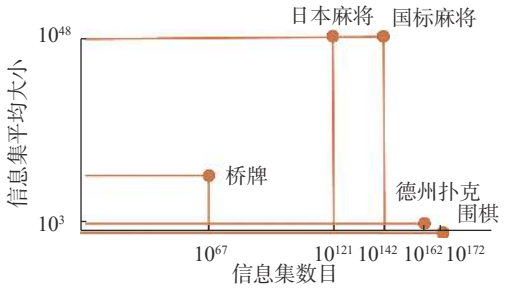

图 1　麻将与其他游戏的复杂度对比

Fig. 1　Comparison of complexity among Mahjong and other games

2　基于知识的麻将 AI 构建方法

基于知识的麻将 AI 构建方法如图 2 所示，一般根据设计者的经验对吃、碰、杠、弃等动作设计优先级和搜索算法指导 AI 的决策。能否将人类玩家的知识总结成规则并恰当的表示影响 AI 水平的高低。使用先验知识构建的 AI 虽然具备一定的水平，但不能真正地解决麻将博弈的问题。基于知识构建的 AI 多倾向于快速听牌、和牌、避免点炮等，灵活性较差，且智能普遍缺少高分牌型，在博弈中单局得分低。先验知识与蒙特卡罗模拟、缺牌数、对手建模、攻防转换等结合，在 AI 的构建中应用得较多。构建方法对比如表 1 所示。

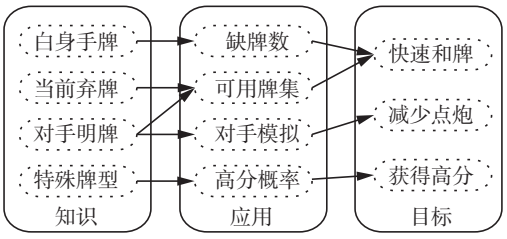

图 2　基于知识的麻将 AI 构建方法

Fig. 2　Mahjong AI construction methods based on knowledge

表 1　基于知识的麻将 AI 构建方法对比

Table 1　Comparison between the Mahjong AI construction methods based on knowledge

采用的专家知识	文献	优点	缺点
自身手牌信息	文献[6-10]	包含基本信息，常与蒙特卡罗模拟、动作优先级结合。	仅考虑自身手牌容易点炮，受发牌随机性影响大，不易获得高分。
缺牌数	文献[11-12]	对缺牌数的合理应用可实现快速和牌，常与树搜索结合。	需与局面信息相结合，避免路径不可行出现死听；前期可知信息少，失误多。
对手建模	文献[13]	可以避免点炮，善于利用对手弱点，针对单一策略的AI。	博弈树搜索开销大，限定时间内难以搜索到终局。
攻防策略划分	文献[14]	加入了对局面形势的考量，在对抗中灵活性更好。	人工设定的转换条件未能平衡攻击与防守。
高分牌型	文献[15]	选择高分牌型和牌，可以提高AI单局得分。	高分牌型往往需要等待，过久等待会失去和牌机会。
人工设定的目标函数	文献[15]	利用多目标优化达到速胜和高分之间的最优平衡。	AI未得到役牌知识，能否从单人麻将推广到四人麻将有待验证。

2.1　基于知识与决策模型方法

知识与决策模型结合是经典的方法。文献 [6] 于 2008 年提出了应用于台湾麻将的博弈 AI Long Cat, 使用“上听数”实现快速听牌的目标。

$$T = \mathrm{Min}(\pi_0,\pi_1,\pi_2,\cdots) \tag{2}$$

上听数 T 为当前手牌与几种可能听牌情况 π_0、π_1、π_2、…所缺牌数的最小值。中后期, Long Cat 根据上听数选择防守策略或且战且守的策略。为避免 AI 点炮, 使用蒙特卡罗模拟临近终局时牌局。Long Cat 结合上听数、有效牌、蒙特卡罗模拟等, 构建了快速上听、避免点炮的 AI。但专家知识不全面, 缺少可以获得高分的特殊牌型, 且策略单调, 在游戏中容易被对手钳制。

Long Cat 中缺牌数是基于经验与统计随机产生的, 对手手牌的模拟较为随机。为规避高随机性带来的误差, 文献 [7] 构建了 108 张牌麻将变体的博弈 AI, 仅对听牌者的手牌进行模拟并计算决策, 优化了 Long Cat 使用的方法。此外, 人工设定了弃牌优先级、吃牌优先级, 对吃、碰、弃、听等动作进行指导; 设置了听牌有效数对有效牌进行监测, 以解决死听问题。与 Long Cat 相比, 提高了胜率并且减少了点炮, 但也存在特殊牌型知识不足的问题。

VeryLong Cat[8] 连续多年获得 Computer Olympiad 麻将项目的冠军。其将麻将概括为取牌和弃牌两种动作交替的游戏, 通过将对手取牌限制为摸牌、删除对手的动作、将自己的动作限制为摸牌等方式对麻将动作进行简化。使用最大期望搜索算法与搜索树结合的搜索方法, 并利用麻将转换手牌进行搜索优化, 以求快速和牌, 但存在点炮次数较多的问题。LongCatMJ[9]、Mahjong DaXia[10] AI 也是在 Long Cat 基础上构建的, 不再赘述。文献 [11] 将和牌距离结合可用牌, 实现了快速和牌。和牌距离越短, 代表的线路更优。使用手牌信息集、缺牌集合、弃牌集合等对游戏路线进行剪枝。该 AI 实现了快速和牌, 但游戏前期信息集参考价值较差, AI 失误较多。文献 [12] 提出了一种在不同麻将变体中通用的、快速计算缺牌数[16] 的算法。在传统的四叉树算法基础上, 引入分块的方法, 既能利用四叉树精确地计算缺牌数, 又能通过分块算法克服四叉树算法存在的计算空间巨大导致的响应速度慢的问题, 加快了计算缺牌数的速度。将其应用于四川麻将博弈程序中, 每次动作的响应时间小于 1 s。

综上可知, 经典的基于专家知识与决策模型的麻将博弈 AI 构建方法主要专注于自身手牌, 普遍缺少高分牌型的专家知识, 使得 AI 错失获得高分的机会。目前的研究方法更关注于快速和牌与避免点炮的平衡, 对快速和牌与等待高分的平衡探究较少。虽然采用“上听数”“有效牌”“缺牌数”等指导 AI 快速和牌, 并在对弈后期模拟对手的手牌避免点炮失分, 但博弈 AI 整体水平不高。

2.2　对手建模及其他方法

文献 [13] 提出了对手建模和博弈树搜索结合的麻将博弈 AI 算法 —— KF-TREE (knowledge-based formwork tree), 获得了 2019 年 Computer Olympiad 麻将项目银牌。 KF-TREE 包含局面分析、对手建模、博弈树搜索、评估决策 4 个模块。为充分地预测对手, KF-TREE 针对上家、下家、对家分别建模, 并计算自身每张牌的风险概率:

$$P_{\mathrm{Risk}} = \alpha P_1(\mathrm{title}) + (1-\alpha)P_2(\mathrm{title}) \tag{3}$$

有人听牌时:

$$P_1(\mathrm{title}) = R_1 + R_2 + R_3 \tag{4}$$

无人听牌时:

$$P_2(\mathrm{title}) = R_1 + R_2^{\mathrm{AA}} + R_3^{\mathrm{AA}} \tag{5}$$

式中: α 表示有对手听牌的概率, R_i 表示其他三家需要这张牌的概率, R^{AA}表示手牌被对手打劫形成刻子的概率。

在节点扩展时, 除常用的出牌结点与摸牌结点探索外, 还从高分牌型和优化手牌两个角度进行扩展。在最后决策时, 对获得的搜索路径从获胜概率、风险概率、得分 3 个角度进行综合评估。KF-TREE 在快速获胜、获得高分、避免点炮之间形成了平衡, 达到了专家水平。

人类玩家在麻将游戏中会根据当前局势的好坏, 调整自身的牌风。文献 [14] 据此提出将人工提取出的攻防转换策略加入麻将博弈 AI。加入攻防转换策略的 AI 在防守阶段表现良好, 但由于博弈树搜索层数受到时间限制, AI 的表现也受到限制。文献 [15] 将麻将中快速获胜和获得高分抽象为多目标优化问题, 基于限定回合的单人麻将展开研究。提出了根据手牌计算剩余回合以及平均得分的目标函数, 并使用一种改进的模块化拓扑神经网络 MM-NEAT(modular multi-objective neuro-evolution of augmenting topologies)[17] 优化函数。但是由于输入的知识仅包含手牌信息, 未考虑役牌知识, AI 虽然能够快速和牌, 但获得的积分并不理想。此外, 多目标优化能否推广至四人麻将仍需验证。

综上可知, 加入对手建模或攻防策略转换等方法的 AI, 除己方信息外开始关注对手的情况,

博弈 AI 整体水平有所提升。但是由于 AI 构建依然主要依据人工设计专家知识，构建的 AI 普遍水平较低。

3　基于数据的麻将博弈 AI 构建方法

基于数据的麻将博弈 AI 构建方法需从大量的数据中提取出特征，并通过不断地训练得到稳定的决策模型。训练出地麻将博弈 AI 水平普遍高于基于知识的麻将博弈 AI 水平，部分 AI 甚至可以媲美人类高手。目前最强的麻将博弈 AI Suphx 由微软亚洲研究院联合京都大学、中国科学技术大学共同研发，拥有超过人类高手的优秀战力。基于数据的构建方法对比如表 2 所示。

表 2　基于数据的构建方法对比

Table 2　Comparison between the construction methods based on data

文献	网络或模型	优点	缺点	数据来源	平均水平
文献[18]	CNN	输入增加局面信息，解决了部分牌局信息未参与决策的问题。	缺少局面周边信息的输入，AI决策信息不全面。	天凤平台	普通玩家
文献[19]	CNN	输入信息全面，在局势判断的基础上会主动选择防守策略。	决策风格过于保守，攻击性较差，未能学习到合适的攻防转换条件。	天凤平台	高级玩家
文献[20]	DenseNet XGBoost	简单训练可达到初学者水平，节约前期训练数据集消耗。	XGBoost是否有助于训练高水平AI有待实验证实。	在线游戏平台	初级玩家
文献[21]	残差神经网络	可以学习到高分牌型，并能够应用。	数据集庞大，训练方法不易在其他麻将变体上应用。	在线游戏平台	高级玩家
文献[22]	PCA	学习到麻将高手对于危险局面的处理态度。	未能与实际的AI结合，实用性待考证。	天凤平台	无
文献[5] (Suphx)	DRL 自对弈	DRL提升了决策能力；自对弈的训练方法减少后期训练数据集的消耗。	采用手工制作的特征作为输入十分复杂；计算资源开销大。	天凤平台	超过顶尖高手
文献[23]	搜索算法 DQN	DRL使得博弈树在复杂状态下效果更好。	受到麻将奖励稀疏的影响，AI在对弈中和局情况居多。	对弈产生	初学者
文献[24]	搜索算法 Double DQN	改进估值函数后的Expectimax搜索算法，在非完备信息博弈中表现更好。	优化的剪枝策略需要人工设置参数，不能自适应更改。	在线游戏平台	普通玩家
文献[25]	A3C模型	结合监督学习和DRL来处理麻将，训练收敛速度快。	仅考虑自身手牌，并未将对手动作加入决策。	天凤平台	高级玩家
文献[26]	变分潜在 先知教练	鲁棒性强、通用性好，可与基于价值的DRL算法结合。	训练方法需要大量的数据集和计算资源支持。	天凤平台	顶尖高手

3.1　基于深度学习的麻将博弈 AI 构建方法

在完备信息博弈中，AlphaGo 利用深度学习与树搜索[27]首次获得了超越人类高手的成绩，随后引起了深度学习在计算机博弈中的研究热潮。在非完备信息博弈中，使用深度学习的 AI 在德州扑克[28]、斗地主[29]、六人德州扑克[30]、3D 视频游戏“Blade & Soul”[31]中均取得了超越人类高手的成绩。在麻将游戏中，深度学习也被广泛应用，并取得了不错的成绩。

图 3 给出了基于深度学习的麻将博弈 AI 构建基本模型。在输入方面通常采用编码的对战数据，若坐庄、圈风等局面信息可得时则共同输入。用于处理数据的神经网络可分两种：用于直接输出决策的单神经网络，单神经网络的输出通常为 1×38 的数组结构，表示 34 张牌与吃、碰、杠、弃 4 个动作；分别输出不同动作概率的多个神经网络，由设计者写出决策程序根据概率值进行决策。

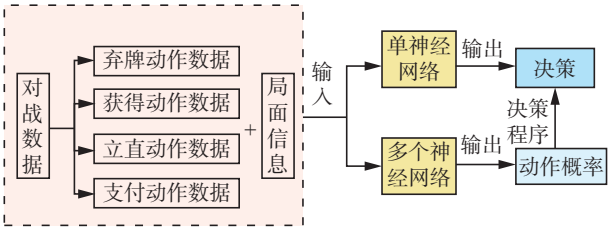

图 3　基于深度学习的麻将博弈 AI 构建方法

Fig. 3　Mahjong game AI construction methods based on deeplearning

卷积神经网络经常用于麻将博弈 AI 中。文献 [18] 将麻将游戏归纳为多分类问题，使用来自天凤平台的对战数据，利用卷积神经网络构建深度学习模型。其将模型划分为弃牌网络、动作网络（吃、碰、杠等）、立直网络，训练后的 AI 具有自主博弈能力，弃牌的准确率达到 68.8%。该 AI 是在不使用任何人类知识的前提下训练出的，其在输入数据时考虑到牌局的基本信息，部分解决了基于知识的麻将博弈 AI 中牌局信息无法参与决策的问题，但排名、轮数、分数、番数等信息仍未纳入 AI 决策模型。

文献 [19] 使用卷积神经网络进行监督学习，其其训练多个网络组合决策，并计算可能丢失的分数：

$$L_i = W \times D_i + T_i + R_i \tag{6}$$

式中：i 表示手牌序号，W 表示等待预测网络的输出，D 表示弃牌网络的输出，T 表示手牌是否为对手等待牌预测值，R 表示网络预测的支付分数。训练后的弃牌网络的准确率为 88.4%。天凤平台上的 AI 水平测试结果显示，该 AI 的攻击性较差。

文献 [20] 将卷积神经网络模型 DenseNet (densely connected convolutional networks) 与 XGBoost(extreme gradient boosting) 模型结合，训练四川麻将博弈 AI。AI 经过简单的训练即可掌握四川麻将的规则，可以节约训练前期较长的试错时间，使得 AI 迅速达到初学者水平。其实验数据来自国内在线网络血腥麻将游戏平台，由于游戏平台玩家来源复杂，可以组成的高质量数据集较小，经过充分训练后能否成为高水平的麻将博弈 AI 尚未可知。

文献 [21] 率先将残差神经网络应用于中国江西省上饶地区的麻将变体。实验所需数据来自在线的商业麻将游戏平台，为保证数据集的质量，其选用了排名靠前的大师级玩家的游戏数据；人为地加入一些对局数据以平衡数据集中各动作的数量。使用一种不平衡的残差网络，由若干 Inception+结构组成的残差块连接而成，采用非向量的原始数据作为输入，使用低级语义特征对模型学习进行引导。游戏平台的测试验证该 AI 可以学习到高分牌型并通过高分牌型赢得比赛。

文献 [22] 利用神经网络将对手策略进行分类，使己方决策选择更有针对性。文献 [32] 基于支持向量机的方法估计玩家的弃牌目的，以便对初学者进行提示。文献 [18,33] 对数据结构进行改进以提高模型训练的准确率。

综上，基于深度学习构建麻将博弈 AI 主要从网络模型、数据等方面开展研究，训练后的 AI 具有学习和决策能力，但其水平并未超越人类，且深度学习对数据集的质量、算力、测试环境等都具有较高的要求。

3.2　基于深度强化学习的麻将 AI 构建方法

深度强化学习[34] 是多智能体领域的常用的技术，结合了深度学习模型的强大模拟能力和强化学习强大的决策能力。其 Q 值也可通过网络模拟预测：

$$Q_t(s_t, a_t, \theta) = R(s_t, a_t) + \gamma \mathrm{Max}[Q_{t+1}(s_{t+1}, a_{t+1}, \theta)] \tag{7}$$

深度强化学习方法在麻将博弈中得到验证，Suphx 是典型代表。Suphx[5] 是一款应用在四人日本麻将上的 AI 系统，是基于深度强化学习训练的目前最强大的麻将 AI 系统，超过 Bakuuchi[35]、NAGA[36] 这两个当时较强的麻将 AI，在最大的日本麻将在线对战平台天凤（tenhou.net）上超过了 99% 的人类玩家。Suphx 采用深度卷积神经网络作为模型基础，利用专业玩家的日志监督学习，形成基本博弈策略；再使用策略梯度算法进行自对弈强化学习来提升博弈水平。应用了全局奖励预测、先知教练以及运行时策略适应等新技术。顶尖人类玩家与 Suphx 均可达十段，但 Suphx 的排名稳定性更强，在与人类对战时，表现出很强防守能力和低点炮率。

3.2.1　Suphx 优点与局限性分析

通过分析微软发布的 Suphx 论文[5]，发现 Suphx 的成功是两方面相辅相成的结果：一方面是对麻将的建模贴合实际，另一方面是深度强化学习的深入应用。贴合现实的决策流程建模使 Suphx 的决策过程流畅且简化，高度分工决策确保了每个决策的专业性。深度强化学习中将麻将的决策判断全部交给神经网络模型：决策模型、先知教练、全局奖励预测。两方面的创新使 Suphx 得到超越人类战绩，也成为 Suphx 的枷锁。

Suphx 具体决策流程如图 4 所示，其将麻将的决策判断分为 6 类，覆盖了现实麻将博弈的各类决策，如黄色菱形表示。除和牌模型使用规则，其他模型使用神经网络实现。多模型分工，

降低了单模型决策时不同种类判断之间的干扰。例如在吃碰杠等鸣牌决策时，当前弃牌对于自己手牌的价值越大，越倾向于鸣牌操作，而弃牌决策则是考虑当前 14 张手牌中哪一张的价值最小。

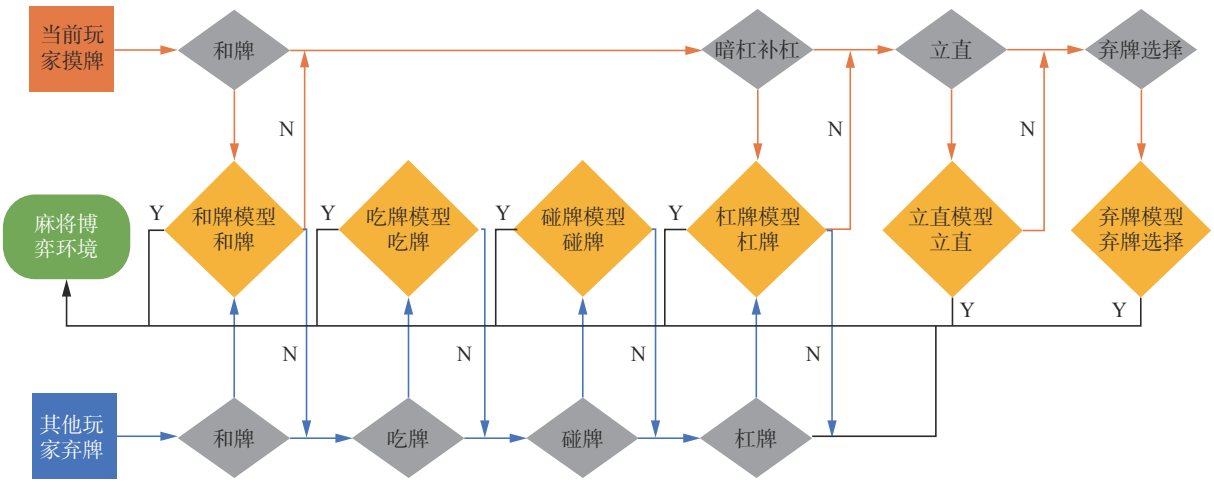

图 4　Suphx 决策流程

Fig. 4　Suphx decision flow

固定的决策流程，也造成 Suphx 的局限性。其一，决策之间无法进行权衡比较。由于和牌、立直、吃、碰、杠的决策判断是分开进行的，且只有是与否两种结果，并非一个决策的评分。在决策时，一旦靠前的决策成立，靠后的决策即使收益更高也无法实施。其二，Suphx 的和牌模型基于规则实现，并未采用深度强化学习方法，全局奖励预测器在训练阶段，并没有成为和牌模型的一部分。如何训练神经网络使得 AI 能够根据当前牌局合理追求高分，是留待其他研究者探索的课题。

Suphx 的深度强化学习训练原理如图 5 所示。训练分为 3 个步骤：通过监督学习掌握基础游戏规则，再使用自对弈强化学习对弃牌模型的参数进行优化以增强鲁棒性，最后使用改进的蒙特卡罗树搜索算法增强运行适应能力。

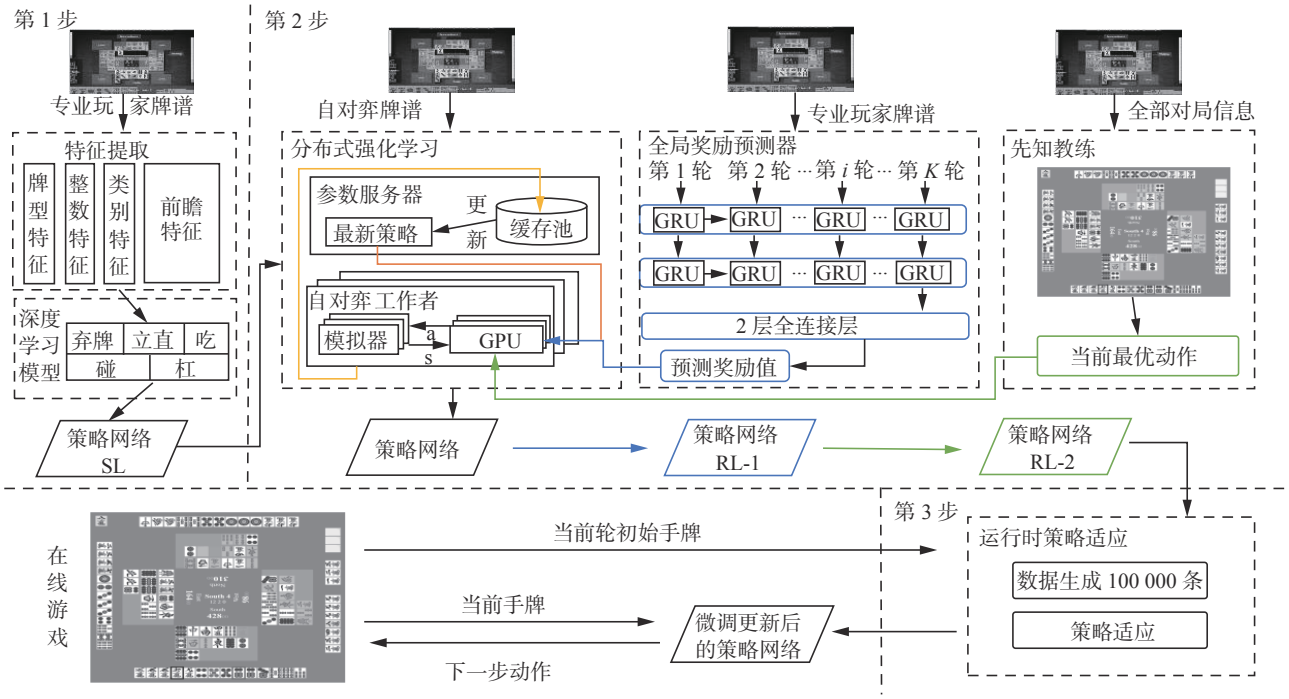

图 5　Suphx 原理

Fig. 5　Suphx schematic

麻将对局信息的手工特征提取是 Suphx 成功的第一大保障，但也意味着模型在高概率获胜牌型的推理方面存在一定不足。训练采用天凤平台上顶级人类玩家对局数据，提取到的特征除常规

的牌型特征、整数特征、类别特征, 还加入基于树搜索产生的前瞻特征(look-ahead features), 实际上是将规则的专家知识输入麻将博弈 AI。模型只需要根据专家的特征信息做出判断, 缺少了理解、推理的过程。不同模块的输入输出特征维度如表 3。

表 3　各模型输入输出维度

Table 3　Input and output dimensions of each model

模型	输入维度	输出维度
弃牌模型	34×838	34
立直模型	34×838	2
吃模型	34×958	2
碰模型	34×958	2
杠模型	34×958	2

在强化学习阶段, 通过自对弈对模型进行改进, 并引入全局奖励预测和先知教练, 增加 AI 的全局观与鲁棒性。先知教练可以获知其他三人手牌、牌墙等非公开信息。在掌控全局信息的训练下, Suphx 形成独特牌风: 高染手率与门清防守。这也是区别于其他 AI 和顶尖人类的地方。一方面染手的两番弥补役牌副露打点不足的缺点, 大量字牌又给手牌提供足够的防守能力。另一方面门清防守使 Suphx 在严密防守中兜牌又高效进攻, 直至高分和牌。Suphx 根据自家手牌的向听数确定安全牌数量, 手牌优势大则不保留安全牌进攻, 手牌优势小就留下安全牌防守。有别于其他 AI 和顶尖高手, Suphx 不轻易放牌给其他玩家副露。

在麻将游戏中以多盘累计排名为目标, 单盘得分无法完整评估每盘游戏的优劣。Suphx 引入全局奖励预测器解决单盘与整局奖励之间的问题, 根据一盘的分数、当前累计的分数、庄家位置、连庄和立直赌注等信息, 使用门控循环网络、最小均方误差来拟合并预测最终的游戏奖励。全局奖励预测器的训练中, 通过学习人类在非完备信息的条件下的自主判断, 模拟出一个合理的预测奖励。使得 Suphx 形成综合考虑平和、七对、染手和防守的平衡打法。在与先知教练结合后, Suphx 变得更有进攻性。

在线游戏阶段采用运行时策略适应, 利用初始手牌快速进行对手建模。Suphx 提出了一种新的参数蒙特卡罗策略自适应方法 (pMCPA), 作为初始化手牌时的有限前瞻, 调整适应对局策略。由于搜索模拟时间较长, 决策时间有限, 对搜索性能要求苛刻。亟待在有限硬件条件下减少运行时间, 高效建模, 实现根据每次手牌动态调整策略的构想。

Suphx 将麻将与深度强化学习的方法结合, 但其手工制作特征、庞大的数据结构、复杂的架构、较高的计算资源消耗是许多研究人员与实验室望而却步的原因。文献 [37] 提出一维数组的数据结构结合基于注意力的模型架构, 解决了数据结构庞大, 不利于在小规模硬件上重复实验的问题。文献 [38] 采用 Ray 分布式训练架构, 优化特征工程以提高训练效果。文献 [39] 通过观测公共信息与私密信息获得替代特征, 在更小的网络结构训练出相同的效果。目前在 Suphx 基础上的研究, 目标是以更小的消耗或更简便的方法使得模型的实验室效果达到 Suphx 水平, 在线对抗中后续的 AI 还未超越 Suphx。

3.2.2　其他基于深度强化学习的研究

除 Suphx 之外, 有研究者利用深度强化学习训练麻将博弈 AI。将 Expectimax 搜索与 PER DQN[28] 或 Double DQN[29] 算法结合, 既保留了 Expectimax 算法的高随机性优势, 也增加了决策的准确性。虽然 AI 的水平不高, 但实验采用的由德州农工大学实验室开发的、支持多种非完备卡牌游戏进行深度强化学习的博弈环境 RLCard[40] 十分便捷。文献 [30] 通过改进 A3C 网络模型, 实现竞争策略, 由于其未充分考虑对手的情况, 模型仅达到中等水平。微软亚洲研究院在先知教练的研究[31] 中, 基于贝叶斯理论提出了一种新的目标函数, 并提出了一种适用于麻将游戏的通用强化学习框架, 在特定情况下训练出的 AI 胜率超过了 Suphx。

综上, 深度强化学习训练的麻将博弈 AI 平均水平高, 并能产生超越人类高手的高水平 AI, 但对数据集、算力的需求更高。

4　麻将博弈 AI 构建面临的挑战

本文对麻将模型构建的相关文献进行梳理和分析, 重点分析了 Suphx 的原理。麻将博弈 AI 的构建方法经历了以专家知识到基于数据的转变, 大部分麻将变体被研究, 但众多麻将变体 AI 的水平还有很大上升空间, AI 构建算法、麻将博弈 AI 对战水平的测试平台等研究存在一些问题和挑战。

4.1　麻将博弈 AI 构建算法面临的主要问题

1) 人工设计的专家知识灵活性差。麻将规则复杂、博弈空间大, 无法人工设计出完整的专家

知识，知识的缺失会使 AI 在某些时刻做出完全随机的、错误的决定。此外，基于知识的 AI 决策流程固定，决策模式单一容易被对手欺骗，决策流程过于复杂则速度较慢。

2）马尔可夫决策过程建模受麻将博弈奖励延时影响。在麻将博弈中约 100 手之后才能获得一次奖励，终局时得到的奖励并不能表示该局的每个动作都是正确（错误）的。这些使得马尔可夫决策过程长时间难以收敛。

3）对手建模未被充分利用。目前，麻将博弈中对手建模多将三位对手视为单智能体统一建模，但三位对手的水平、决策风格、手牌状态、动作意图均有差异，统一建模不仅缺少针对性、也忽略了三家之间竞争与合作的关系，不能充分利用对手建模进行对手剥夺[41-42]。

4）麻将博弈 AI 构建算法通用性差。麻将变体众多，但核心规则如吃、碰、杠的条件、和牌规则等较为相似。当前的麻将博弈 AI 构建多是针对于某一种麻将变体的专用算法，尚缺少能够应用于不同麻将变体的通用算法。AI 在训练过程如何将某一麻将中学习的知识和策略应用在其他麻将变体的游戏中，是面临的一个挑战。

5）麻将博弈 AI 的决策逻辑可解释性差[43]，消耗计算资源多。现阶段构建麻将 AI 使用深度学习作为骨干网络，导致研究者只能通过对战最终结果，以及现有理论进行贴合，试图理解 AI 的决策思路。无法做到对 AI 直观理解、控制、优化。

4.2　麻将博弈 AI 对战水平测试存在的问题和挑战

1）麻将博弈的评估标准不统一。欢乐麻将、天凤平台、深圳快乐麻将等在线游戏平台均有自己的分级评估标准。但用于科学研究的麻将博弈 AI 测试平台的评估标准不统一。文献 [6,8] 以获胜次数和点炮次数来评价 AI 的水平，文献 [23-24] 以神经网络在某一实验的验证集上的准确率作为评估标准，文献 [5,33] 以某个在线的分级结果为标准。

2）适合学术研究的麻将博弈平台缺乏。麻将游戏的商业化不能满足博弈学术研究的需要，研究采用的平台需要具备以下特点：①有大量的高水平用户来保证对弈数据的质量②提供统一的 API 接口，以便接入 AI 程序进行训练和测试③极高的平台稳定性和安全性，保证研究的顺利进行。满足这些要求的、仅适用于单一麻将变体的平台少之又少，麻将通用的大型专用平台更是尚未出现。

3）缺乏高质量数据集。通过基于数据方法训练麻将博弈 AI，需要高质量对弈数据集支持。目前实验所用对弈数据一般来自天凤平台[5,23-24,31]、在线麻将游戏平台[25-26]，缺少公开的、免费的、质量较高的对弈数据集来支持研究。日本麻将是当前在线数据最多的麻将变种，可以从天凤平台上下载到高手的对弈数据，但并未有经过预处理的测试数据集发布；其他变种尤其是国内的大众麻将、四川麻将、各地的小众麻将均未有大型的可供研究使用的数据集。

5　展望

麻将作为典型的非完备博弈，是复杂环境下多智能体博弈的简单体现，解决麻将博弈智能体问题，对推进复杂环境多智能体的知识获取、模型构建、决策研究等具有重要意义。麻将博弈 AI 构建面临的诸多挑战，在其他复杂环境博弈中依旧存在[44-45]，优化 AI 构建算法和构建 AI 对战水平的测试环境是未来的重点研究所在。

5.1　优化麻将博弈 AI 构建算法

利用人类的专家知识构建的麻将博弈 AI 平均水平较低，灵活性差。使用深度学习和强化学习构建的麻将博弈 AI 整体水平较高，甚至超越了人类高手，但是麻将博弈 AI 构建算法研究仍然存在很大的提升空间。

麻将博弈状态空间巨大，环境奖励稀疏。经验回放[46-47]、分层强化学习[48-50]、好奇心模型[51-53]能够充分利用现有的数据信息来解决奖励稀疏的问题。还可以引入认知行为模型[54]，将先验知识描述为人和 AI 均能理解的格式，指导 AI 选择，加快前期训练速度、减少失误。

采用离线学习与在线学习结合的方法，通过在线对弈减少训练后期的数据集需求。也可以利用小样本机器学习中常用的元学习[55-57]、迁移学习[58-63] 等方法，在数据集有限的情况下，训练出更加强大的麻将博弈 AI。此外，广泛借鉴参考德州扑克中在线训练过程采用演化学习与深度神经网络结合的方法[64]，调整麻将博弈算法训练的架构，提高样本利用率，最终达到提升麻将博弈 AI 学习效率的目的。

针对麻将对手建模研究仍然很薄弱的现状，从对手剥削[65] 的角度着手，在实时在线对抗过程中，通过对手模型预测对手状态及可能采取的行动，发掘可利用空间，增加己方收益。也可以采用分阶段课程学习[66]、多样性自主课程学习[67] 等

方法, 通过种群课程训练和演化选择复杂规则之间的权重调配。使用集成学习[68]将多个对手模型纳入强化学习过程, 学习鲁棒的策略。可以针对不同对手的缺陷, 利用元学习[69]与不同风格的对手进行训练。

对深度学习网络进行优化。分布式的训练架构是缓解高算力硬件需求的直接方法, 采用轻量化的模型则是缓解硬件需求的有效方法。轻量化的模型在保持性能的条件下尽可能地减小网络结构, 甚至以一定的精度换取网络的精简。另外, 引入擅长求解复杂优化问题的进化算法, 优化神经网络架构搜索[70]和深度学习超参数选取[71]等, 也是加快训练的可行方法。随着神经网络应用更深入, 神经网络可解释性差的问题也随之而来, 如能对“黑盒”进行分析, 也将对模型的优化起到相当大的作用。

5.2　构建麻将博弈 AI 对战水平的测试环境

建立统一的多元化的麻将博弈 AI 评估指标。如何为麻将游戏制定一个多元化、标准化、统一的评估标准是一个重要而开放的问题。目前评估麻将 AI 多是从胜率[6,8-9]、动作(吃、碰、杠、弃)准确率[30-31]、排名[5,39]等角度进行分析, 评估方法不全面。可以考虑为在线博弈的 AI 构建包含胜率、排名、响应速度、点炮概率的集成评估体系。

对于麻将等竞技类游戏, 一般借助技术等级分段与专业分来判断。去年 8 月, 国际麻将联盟和中国棋院杭州分院共同成立了麻将运动技术等级评定中心, 同时出台了《麻将运动技术技能等级评定管理办法(修订版)》的通知。对麻将运动员分级评定提出具体规则。借鉴此技能等级管理办法, 可以设置对应的麻将博弈 AI 的等级管理办法。

另外在分段的基础上, 为了更好地评估、了解个人风格, 可采用雷达图的方式统计个人的历史表现。以国标麻将为例, 一局为 16 盘, 统计 AI 每盘的表现分可以简要设计如下:

$$\text{表现分} = \frac{\text{和牌得分} + \text{荒牌上听数}}{\text{点炮失分} + \text{对手自摸失分}}$$

此处借鉴多人在线战术竞技 (MOBA) 类游戏 KDA(kill dead aid) 的计算思想, 荒牌牌局的上听数较小, 则可能意味着错失得分机会, 但如果上听数仍然较大, 则认为避免了他人和牌, 保全了自己的分数。

将表现分、累计和牌番数、累计搭子牌数(每盘终局时, 已经成为搭子的牌数目)、累计未点炮数、累计扣牌数 (终局时, 手牌中包含其他玩家和牌所需牌的数目) 组成雷达图, 如图 6 所示, 集合 5 个方面综合表现一个麻将博弈 AI 的实力。

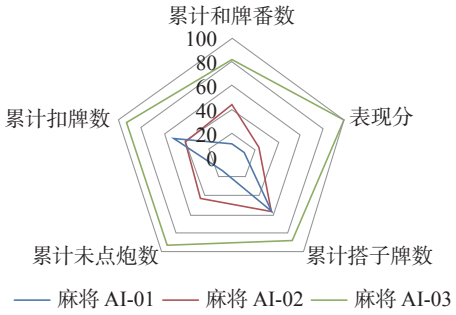

图 6　麻将博弈 AI 实力评估雷达图示例

Fig. 6　Mahjong game AI strength assessment radar graph

搭建通用对抗博弈环境。对抗博弈环境可分为线下麻将程序和线上公开的麻将服务器。线下麻将程序可安装至本地, 通过与标准程序比赛来评估 AI 的能力。目前, 开源的线下麻将程序有德州农工大学开发的 RLCard[40], 国内尚未发布开源且安装便捷的麻将程序。线上的麻将服务器, 人类和 AI 均可注册并参与比赛、获得评分。麻将服务器通过互联网提供水平更高、竞技性更强的比赛, 从而更易收集到高质量数据。线上服务器还能提供 AI 在线训练、在线权威评估等功能。线上麻将服务器的发展略好, 天凤平台是针对日本麻将的成熟的国际博弈平台。国内发展较好的平台有北京大学创立的 Botzone[72], 该网站从 2020 年开始与 IJCAI 会议合作举办 Mahjong AI Competition 比赛。另外, 竞技世界公司的大众麻将平台, 与中国计算机博弈大赛组委会合作提供麻将比赛平台。麻将变体众多, 发布适用于某种变体的对抗环境, 或开发适用较多变体的集成对抗环境, 不仅利于统一麻将博弈 AI 水平的评估标准, 也是构建高质量数据集的有效途径。

构建高质量数据集。目前, 除日本麻将拥有大量的数据可以供研究人员使用外, 其余的麻将变体数据较少且质量差, 如何低成本且高效地构建高质量的数据集是也是未来待研究的方向。

参考文献:

[1]　陆升阳, 赵怀林, 刘华平. 场景图谱驱动目标搜索的多智能体强化学习 [J]. 智能系统学报, 2023, 18(1): 207–215.
LU Shengyang, ZHAO Huailin, LIU Huaping. Multi-agent reinforcement learning for scene graph-driven tar-

get search[J]. CAAI transactions on intelligent systems, 2023, 18(1): 207–215.

[2] 欧阳勇平, 魏长赟, 蔡帛良. 动态环境下分布式异构多机器人避障方法研究 [J]. 智能系统学报, 2022, 17(4): 752–763.
OUYANG Yongping, WEI Changyun, CAI Boliang. Collision avoidance approach for distributed heterogeneous multirobot systems in dynamic environments[J]. CAAI transactions on intelligent systems, 2022, 17(4): 752–763.

[3] 齐小刚, 陈春绮, 熊伟, 等. 基于博弈论的预警卫星系统抗毁性研究 [J]. 智能系统学报, 2021, 16(2): 338–345.
QI Xiaogang, CHEN Chunqi, XIONG Wei, et al. Research on the invulnerability of an early warning satellite system based on game theory[J]. CAAI transactions on intelligent systems, 2021, 16(2): 338–345.

[4] MIZUKAMI N, NAKAHARI R, URA A, et al. Realizing a four-player computer mahjong program by supervised learning with isolated multi-player aspects[J]. Transactions of information processing society of Japan, 2014, 55(11): 1–11.

[5] LI Junjie, KOYAMADA S, YE Qiwei, et al. Suphx: mastering mahjong with deep reinforcement learning[EB/OL]. (2020−03−30)[2022−11−18].https://arxiv.org/abs/2003.13590

[6] 乔继林. 麻将机器博弈方法研究 [D].沈阳: 沈阳航空航天大学, 2022.
QIAO Jilin. Research on the Mahjong machine game method [D]. Shenyang: Shenyang Aerospace University, 2022.

[7] 王亚杰, 乔继林, 梁凯, 等. 结合先验知识与蒙特卡罗模拟的麻将博弈研究 [J]. 智能系统学报, 2022, 17(1): 69–78.
WANG Yajie, QIAO Jilin, LIANG Kai, et al. Research on Mahjong game based on prior knowledge and Monte Carlo simulation[J]. CAAI transactions on intelligent systems, 2022, 17(1): 69–78.

[8] 王松. 基于深度学习的非完备信息博弈对手建模的研究 [D]. 南昌: 南昌大学, 2023.
WANG Song. Research on incomplete information game opponent model based on deep learning [D]. Nanchang: Nanchang University, 2023.

[9] 赵海璐. 大众麻将计算机博弈智能搜索算法的应用研究 [D]. 重庆: 重庆理工大学, 2023.
ZHAO Hailu. Application research on intelligent search algorithm of popular Mahjong computer game [D]. Chongqing: Chongqing University of Technology, 2023.

[10] 任航. 基于知识与树搜索的非完备信息博弈决策的研究与应用 [D]. 南昌: 南昌大学, 2020.
REN Hang. Research and application of imperfect information game decision based on knowledge and game-tree search[D]. Nanchang: Nanchang University, 2020.

[11] 彭丽蓉, 赵海璐, 甘春晏, 等. 一种大众麻将计算机博弈的胡牌方法研究 [J]. 重庆理工大学学报(自然科学版), 2021, 35(12): 127–133.
PENG Lirong, ZHAO Hailu, GAN Chunyan, et al. Research on the hu method of a popular mahjong computer game[J]. Journal of Chongqing University of Technology (natural science edition), 2021, 35(12): 127–133.

[12] YAN Xueqing, LI Yongming, LI Sanjiang. A fast algorithm for computing the deficiency number of a mahjong hand[EB/OL]. (2021−08−15)[2022−11−13]. https://arxiv.org/abs/2108.06832.

[13] WANG Mingyan, REN Hang, HUANG Wei, et al. An efficient AI-based method to play the Mahjong game with the knowledge and game-tree searching strategy[J]. ICGA journal, 2021, 43(1): 2–25.

[14] XU D. Mahjong AI/analyzer[D]. Los Angeles: California State University Northridge, 2015.

[15] IHARA K, KATO S. Neuro-evolutionary approach to multi-objective optimization in one-player mahjong[C]// International Conference on Network-Based Information Systems. Cham: Springer, 2018: 492−503.

[16] LI Sanjiang, YAN Xueqing. Let's play Mahjong![EB/OL]. (2019−03−08)[2022−11−13].https://arxiv.org/abs/1903.03294.

[17] SCHRUM J, MIIKKULAINEN R. Evolving multimodal behavior with modular neural networks in Ms. Pac-Man[C]//Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation. New York: ACM, 2014: 325−332.

[18] GAO Shiqi, FUMINORI O, YOSHIHIRO K, et al. Supervised learning of imperfect information data in the game of mahjong via deep convolutional neural networks[J]. Information processing society of Japan, 2018: 43−50.

[19] ZHENG Y, YOKOYAMA S, YAMASHITA T, et al. Study on evaluation function design of Mahjong using supervised learning[J]. SIG-SAI, 2019, 34(5): 1–9.

[20] GAO Shijing, LI Shuqin. Bloody Mahjong playing strategy based on the integration of deep learning and XGBoost[J]. CAAI transactions on intelligence technology, 2022, 7(1): 95–106.

[21] WANG Mingyan, YAN Tianwei, LUO Mingyuan, et al. A novel deep residual network-based incomplete information competition strategy for four-players Mahjong games[J]. Multimedia tools and applications, 2019, 78(16): 23443–23467.

[22] SATO H, SHIRAKAWA T, HAGIHARA A, et al. An analysis of play style of advanced mahjong players toward the implementation of strong AI player[J]. International journal of parallel, emergent and distributed systems, 2017, 32(2): 195–205.

[23] 孙一铃. 基于 Expectimax 搜索的非完备信息博弈算法的研究 [D]. 北京: 北京交通大学, 2021.
SUN Yiling. Research on incomplete information game algorithm based on Expectimax search[D]. Beijing: Beijing Jiaotong University, 2021.

[24] 雷捷维, 王嘉旸, 任航, 等. 基于 Expectimax 搜索与 Double DQN 的非完备信息博弈算法 [J]. 计算机工程, 2021, 47(3): 304–310,320.
LEI Jiewei, WANG Jiayang, REN Hang, et al. Incomplete information game algorithm based on expectimax search and double DQN[J]. Computer engineering, 2021, 47(3): 304–310,320.

[25] ZHAO Cong, XIAO Bing, ZHA Lin. Incomplete information competition strategy based on improved asynchronous advantage actor critical model[C]//Proceedings of the 2020 4th International Conference on Deep Learning Technologies. New York: ACM, 2020: 32−37.

[26] HAN D, KOZUNO T, LUO X, et al. Variational oracle guiding for reinforcement learning[C]//International Conference on Learning Representations. Vienna: ICLR, 2021: 1−22.

[27] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.

[28] MICHAEL B, NEIL B, MICHAEL J, et al. Heads-up limit hold'em poker is solved[J]. Science, 2015, 347(6218): 145–149.

[29] ZHA Daochen, XIE Jingru, MA Wenye, et al. DouZero: mastering DouDizhu with self-play deep reinforcement learning[EB/OL]. (2021−06−11)[2022−11−13].https://arxiv.org/abs/2106.06135.

[30] BROWN N, SANDHOLM T. Superhuman AI for multiplayer poker[J]. Science, 2019, 365(6456): 885–890.

[31] OH I, RHO S, MOON S, et al. Creating pro-level AI for a real-time fighting game using deep reinforcement learning[J]. IEEE transactions on games, 2022, 14(2): 212–220.

[32] UENO M, HAYAKAWA D, ISAHARA H. Estimating the purpose of discard in mahjong to support learning for beginners[C]//International Symposium on Distributed Computing and Artificial Intelligence. Cham: Springer, 2019: 155−163.

[33] Long H, Tomoyuki K. Improving Mahjong Agent by Predicting Types of Yaku[C]//Proceedings game programming workshop. Venue: IPSJ, 2019: 206−212.

[34] 龚慧雯, 王桐, 陈立伟, 等. 基于深度强化学习的多智能体对抗策略算法 [J]. 应用科技, 2022, 49(5): 1–7.
GONG Huiwen, WANG Tong, CHEN Liwei, et al. A multi-agent adversarial strategy algorithm based on deep reinforcement learning[J]. Applied science and technology, 2022, 49(5): 1–7.

[35] KURITA M, HOKI K. Method for constructing artificial intelligence player with abstractions to Markov decision processes in multiplayer game of mahjong[J]. IEEE transactions on games, 2021, 13(1): 99–110.

[36] VILLAGE D M. NAGA: deep learning Mahjong AI[EB/OL]. (2022−07−29)[2022−11−13].https://dmv.nico/ja/articles/mahjong_ai_naga/.

[37] TRUONG T D. A supervised attention-based multiclass classifier for tile discarding in Japanese Mahjong[D]. Grimstad : University of Agder, 2021.

[38] LIN J. Phoenix: an open-source, reproducible and interpretable Mahjong agent[EB/OL]. (2021−05−05)[2022−11−13].https://csci527-phoenix.github.io/documents.html.

[39] Long H, Tomoyuki K. Training japanese mahjong agent with two dimension feature representation[C]//Proceedings game programming workshop. online: IPSJ, 2020: 125-130.

[40] ZHA Daochen, LAI K H, HUANG Songyi, et al. RLCard: a platform for reinforcement learning in card games[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2020: 5264−5266.

[41] LOCKHART E, LANCTOT M, PÉROLAT J, et al. Computing approximate equilibria in sequential adversarial games by exploitability descent[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao: International Joint Conferences on Artificial Intelligence Organization, 2019: 464−470.

[42] WANG Zhikun, BOULARIAS A, MÜLLING K, et al. Balancing safety and exploitability in opponent modeling[J]. Proceedings of the AAAI conference on artificial intelligence, 2011, 25(1): 1515–1520.

[43] 董胤蓬, 苏航, 朱军. 面向对抗样本的深度神经网络可解释性分析 [J]. 自动化学报, 2022, 48(1): 75–86.
DONG Yinpeng, SU Hang, ZHU Jun. Interpretability analysis of deep neural networks with adversarial examples[J]. Acta automatica sinica, 2022, 48(1): 75–86.

[44] 刘佳, 陈增强, 刘忠信. 多智能体系统及其协同控制研究进展 [J]. 智能系统学报, 2010, 5(1): 1–9.

LIU Jia, CHEN Zengqiang, LIU Zhongxin. Advances in multi-Agent systems and cooperative control[J]. CAAI transactions on intelligent systems, 2010, 5(1): 1–9.

[45] 殷昌盛, 杨若鹏, 朱巍, 等. 多智能体分层强化学习综述 [J]. 智能系统学报, 2020, 15(4): 646–655.
YIN Changsheng, YANG Ruopeng, ZHU Wei, et al. A survey on multi-agent hierarchical reinforcement learning[J]. CAAI transactions on intelligent systems, 2020, 15(4): 646–655.

[46] LIN Longji. Self-improving reactive agents based on reinforcement learning, planning and teaching[J]. Machine learning, 1992, 8(3): 293–321.

[47] ANDRYCHOWICZ M, WOLSKI F, RAY A, et al. Hindsight experience replay[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 5055−5065.

[48] PATERIA S, SUBAGDJA B, TAN A H, et al. Hierarchical reinforcement learning: a comprehensive survey[J]. ACM computing surveys, 54(5): 109.

[49] DAYAN P, HINTON G E. Feudal reinforcement learning[J]. Advances in neural information processing systems, 1992, 5: 271–278.

[50] VEZHNEVETS A S, OSINDERO S, SCHAUL T, et al. FeUdal networks for hierarchical reinforcement learning[EB/OL]. (2017−03−03)[2022−11−13].https://arxiv.org/abs/1703.01161.

[51] CAMERON J, PIERCE W D. Reinforcement, reward, and intrinsic motivation: a meta-analysis[J]. Review of educational research, 1994, 64(3): 363–423.

[52] OSTROVSKI G, BELLEMARE M G, VAN DEN OORD A, et al. Count-based exploration with neural density models[EB/OL]. (2017−03−03)[2022−11−13].https://arxiv.org/abs/1703.01310.

[53] PATHAK D, AGRAWAL P, EFROS A A, et al. Curiosity-driven exploration by self-supervised prediction[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017: 488-489.

[54] 陈浩, 李嘉祥, 黄健, 等. 融合认知行为模型的深度强化学习框架及算法 [J/OL]. 控制与决策: 1−9. [2023−10−06]. https://doi.org/10.13195/j.kzyjc.2022.0281.
CHEN Hao, LI Jiaxiang, HUANG Jian, et al. Deep reinforcement learning framework and algorithm integrating cognitive behaviour model[J/OL]. Control and decision: 1−9. [2023−10−06]. https://doi.org/10.13195/j.kzyjc.2022.0281.

[55] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. New York: ACM, 2017: 1126−1135.

[56] 谭晓阳, 张哲. 元强化学习综述 [J]. 南京航空航天大学学报, 2021, 53(5): 653–663.
TAN Xiaoyang, ZHANG Zhe. Review on meta reinforcement learning[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021, 53(5): 653–663.

[57] 王方伟, 柴国芳, 李青茹, 等. 基于参数优化元学习和困难样本挖掘的小样本恶意软件分类方法 [J]. 武汉大学学报(理学版), 2022, 68(1): 17–25.
WANG Fangwei, CHAI Guofang, LI Qingru, et al. Classification of few-sample malware based on parameter-optimized meta-learning and hard example mining[J]. Journal of Wuhan University (natural science edition), 2022, 68(1): 17–25.

[58] 宋佳蓉, 杨忠, 张天翼, 等. 基于卷积神经网络和多类 SVM 的交通标志识别 [J]. 应用科技, 2018, 45(5): 71–75, 81.
SONG Jiarong, YANG Zhong, ZHANG Tianyi, et al. Traffic sign identification based on convolutional neural network and multiclass SVM[J]. Applied science and technology, 2018, 45(5): 71–75, 81.

[59] WEISS K, KHOSHGOFTAAR T M, WANG Dingding. A survey of transfer learning[J]. Journal of big data, 2016, 3(1): 1–40.

[60] RUDER S, PETERS M E, SWAYAMDIPTA S, et al. Transfer learning in natural language processing[C]//Proceedings of the 2019 Conference of the North. Minneapolis, Minnesota. Stroudsburg: Association for Computational Linguistics, 2019: 15−18.

[61] SHAO Kun, ZHU Yuanheng, ZHAO Dongbin. StarCraft micromanagement with reinforcement learning and curriculum transfer learning[J]. IEEE transactions on emerging topics in computational intelligence, 2019, 3(1): 73–84.

[62] SUN Qianru, LIU Yaoyao, CHUA T S, et al. Meta-transfer learning for few-shot learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 403−412.

[63] OLIVAS E S, GUERRERO J D M, MARTINEZ-SOBER M, et al. Handbook of research on machine learning applications and trends: algorithms, methods, and techniques[M]. IGI global, 2009.

[64] 张蒙, 李凯, 吴哲, 等. 一种针对德州扑克 AI 的对手建模与策略集成框架 [J]. 自动化学报, 2022, 48(4): 1004–1017.
ZHANG Meng, LI Kai, WU Zhe, et al. An opponent modeling and strategy integration framework for texas hold'em[J]. Acta automatica sinica, 2022, 48(4):

1004–1017.

[65] GANZFRIED S, SANDHOLM T. Safe opponent exploitation[J]. ACM transactions on economics and computation, 2015, 3(2): 1–28.

[66] LONG Qian, ZHOU Zihan, GUPTA A, et al. Evolutionary population curriculum for scaling multi-agent reinforcement learning[EB/OL]. (2020−03−23)[2022−11−13]. https://arxiv.org/abs/2003.10423.

[67] YANG Y, LUO J, WEN Y, et al. Diverse auto-curriculum is critical for successful real-world multiagent learning systems[C]//Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems. Richland: ACM, 2021: 51−56.

[68] WU Zhe, LI Kai, XU Hang, et al. L2E: learning to exploit your opponent[C]//2022 International Joint Conference on Neural Networks. Padua: IEEE, 2022: 1−8.

[69] SHEN Macheng, HOW J P. Robust opponent modeling via adversarial ensemble reinforcement learning[J]. Proceedings of the international conference on automated planning and scheduling, 2021, 31: 578–587.

[70] 苏子美, 董红斌. 面向无人机路径规划的多目标粒子群优化算法 [J]. 应用科技, 2021, 48(3): 12–20, 26.
SU Zimei, DONG Hongbin. Multi-objective particle swarm optimization algorithm for UAV path planning[J]. Applied science and technology, 2021, 48(3): 12–20, 26.

[71] JADERBERG M, DALIBARD V, OSINDERO S, et al. Population based training of neural networks[EB/OL]. (2017−11−27)[2022−11−13].https://arxiv.org/abs/1711.09846.

[72] LI Wenxin, ZHOU Haoyu, WANG C, et al. Teaching AI algorithms with games including Mahjong and fight the landlord on the botzone online platform[C]//Proceedings of the ACM Conference on Global Computing Education. New York: ACM, 2019: 129−135.

作者简介:

李霞丽, 教授, 主要研究方向为计算机博弈。

王昭琦, 硕士研究生, 主要研究方向为计算机博弈。

吴立成, 教授, 中国人工智能学会机器博弈专委会副主任, 主要研究方向为智能系统及机器人、计算机博弈。主持国家自然科学基金等项目 10 余项, 发表学术论文 80 余篇。