中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

基于粒子群优化的德州扑克在线对手利用

胡振震, 陈少飞, 袁唯淋, 李鹏, 陈璟

引用本文:
胡振震,陈少飞,袁唯淋,李鹏,陈. 基于粒子群优化的德州扑克在线对手利用[J]. 控制与决策, 2024, 39(5): 1687−1696.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1790

基于粒子群优化的德州扑克在线对手利用

胡振震, 陈少飞, 袁唯淋, 李 鹏, 陈 璟†

(国防科技大学 智能科学学院, 长沙 410073)

摘 要: 德州扑克中, 相比于采用均衡策略求解的方法, 对手利用是针对存在弱点的对手以获取更大收益的更有效方法. 然而在面对一个全新对手时, 在线条件下如何高效利用对手仍然是一大难题. 现有方法常采用离线训练在线适应的方式来避开这一问题, 即利用学习、演化等方法, 通过海量离线训练来获得具有对手适应性的模型, 使其能在比赛中适应不同的对手, 而不是在比赛中针对一个新对手在线主动地优化自身策略. 对此, 以在线主动策略优化实现有效对手利用为目的, 基于时间维的粒子定义提出一种基于粒子群优化的策略优化方法, 将在线策略优化的思路引入德州扑克这种具有强随机性的博弈问题中, 开展对手利用并实现在线比赛收益最大化. 针对适应度计算受随机运气影响以及部分对手针对性策略难以优化的问题, 提出一种基于局部最优解替代、全局最优解替代的改进粒子群优化算法(BR-PSO). 实验结果表明, 对于标准PSO方法难以针对的对手, 所提出的方法能有效获得对手的针对性策略以实现最大化对手利用, 而且优化策略的收益能够媲美基于手牌预测AI的收益.

关键词: 粒子群优化；策略优化；最优解替代；对手利用；在线比赛；德州扑克

中图分类号: TP18；O225 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1790

引用格式: 胡振震, 陈少飞, 袁唯淋, 等. 基于粒子群优化的德州扑克在线对手利用 [J]. 控制与决策, 2024, 39(5): 1687-1696.

Online opponent exploitation method based on particle swarm optimization for Texas Hold'em

HU Zhen-zhen, CHEN Shao-fei, YUAN Wei-lin, LI Peng, CHEN Jing†

(College of Intelligence Science and Technology，National University of Defense Technology，Changsha 410073，China)

Abstract: In Texas Hold'em, opponent exploitation is the more effective method to obtain larger income from opponents with weakness in contrast to the Nash equilibrium searching method. However, how to effectively exploit the brand new opponent under the condition of online competitions is still a challenge. The existing methods usually use offline training and online adaptation to avoid this problem, that is, using like learning, evolution methods to obtain a model with opponent adaptability through massive offline training, so that it can adapt to different opponents in competitions, instead of actively optimizing its own policy for a new opponent in the online competition. For the purpose of online active policy optimizing to achieve effective opponent exploitation, a policy optimization method based on particle swarm optimization (PSO) is proposed to maximize the competition income, which introduces the idea of online optimization into Texas Hold'em regarded as an game problem with strong randomness. Aiming to the problems that fitness computation is affected by random luck and targeted policies for some opponents are hard to optimize with the standard PSO, a modified PSO method called BR-PSO (best replacement-PSO) is proposed based on local optimal solution replacement and global optimal solution replacement. The result of experiments indicates the proposed method can find targeted policies to maximize opponent exploitation of the opponents that are hard to counter with the standard PSO, and the income of the optimized policy is comparable to that of AI based on the hand prediction method.

Keywords: particle swarm optimization；policy optimization；optimal solution replacement；opponent exploitation；online competition；Texas Hold'em

收稿日期: 2022-10-16；**录用日期:** 2023-02-22.

基金项目: 国家自然科学基金项目(61806212, 62376280).

责任编委: 林崇.

†通讯作者. E-mail: chenjing001@vip.sina.com.

*本文附带电子附录文件, 可登录本刊官网该文“资源附件”区自行下载阅览.

0 引 言

策略游戏是具有挑战的人工智能研究领域,开发策略游戏相关的人工智能程序使得许多相关领域广泛受益[1]. 策略游戏中如井字棋、国际象棋等棋类是最先被攻克的领域,特别是围棋AI-Alpha Go-的成功,标志着“完全信息确定性两人零和博弈”策略游戏被完全解决[2]. 而以德州扑克为代表的扑克游戏,具有隐藏信息、随机发牌、潜在欺骗等特点[3],是另一类“不完全信息零和博弈”的策略游戏,最近也取得了重要突破. AI程序Deepstack、Pluribus等在两人和多人德州扑克中击败了人类选手[4-5],这标志着采用纳什均衡策略求解范式的AI算法具备了在庞大状态空间中搜索近似纳什均衡解的能力,单从博弈胜负的角度看,这类不完全信息零和博弈也已被解决.

从博弈双方采用纳什均衡策略时不会主动改变策略的概念可知,采用均衡策略等价于考虑了最强对手情况下的最优策略. 因此从另一个角度看,当双方条件一致时,采取均衡策略能够保证不会输于对手,但不能保证收益的最大化,特别是对手并非采用均衡策略时. 用石头剪刀布游戏可以很容易说明这一现象,即如果能预判对手的行动并采取针对性行动,则必然会比采用石头剪刀布各三分之一概率的均衡策略收益更大. 所以,基于近似纳什均衡求解范式的AI虽然能击败人类职业选手,但不代表在相同情况下能比人类或其他方法赢得更多. 相比之下,基于对手建模的对手利用方法,能够使用针对性的策略,是一种尽可能去获得最大收益的有效方法. 根据对手建模思路的不同,此类方法可以分为显式建模和隐式建模两类. 显式建模的对手利用方法重点在于显式地建模对手,进而基于对手模型给出最优的对抗策略,这种方法由于需要显式地给出对手模型,难度较大. 而隐式建模的方法关注于具有对手适应性的策略模型的学习,常利用神经网络等工具建立一个隐含考虑对手不同特点的决策模型[6-12]. 其核心思想是在与不同特点对手的对抗训练过程中,通过演化算法、强化学习等方法来优化参数,进而得到一个具有高适应性的决策模型,能够鲁棒地适应较弱对手,使其具备获得更高收益的可能性.

基于隐式建模的对手利用方法常采用提前离线训练和在线适应对手结合的方式实现. 这种方式需要一个能获取海量数据或交互的训练环境,但在与一个全新对手比赛时,无法实现提前训练进而在线适应的要求,因为只能依靠在线比赛提供的数据边比赛边训练. 另一方面,提前训练得到的模型的适应性并不意味着必定能有效利用对手和收益最大化. 从适应性角度开展离线的策略模型训练,其适应性与训练所用的数据密切相关. 当训练所用对手(包括其类型或风格)不具备足够的代表性时,训练得到的模型尽管可能具有一定的泛化能力,但当面对一个新类型的对手时,对手利用很可能会面临困难. 上述两个问题实际上引出了对在线对手利用的新需求,即能否建立一种能在线进行策略优化的方法,使其能突破隐式建模在线适应的被动逻辑,在不需要提前进行海量训练的前提下,实现一种主动机制,以在面对全新对手时能更快获得针对性策略,从而实现收益最大化. 为此,本文设定对手利用的两个新目标:一是面对一个全新对手时要能通过有限的在线比赛学习到针对性策略;二是要尽快地搜索并收敛到目标策略以获得更大的收益. 更大的收益是对手利用的终极目标,而目标策略的搜索和收敛则是手段和方法. 由于不完全信息随机博弈的德州扑克游戏具有强随机性特征,策略优化过程中的适应度计算受到随机运气的影响,而且在线比赛不能通过增加比赛局数来减小其影响,优化过程必须要解决随机运气影响下解的正确收敛的问题. 基于上述考虑及粒子群优化算法在优化解搜索和收敛速度方面的优势[13-14],本文建立一种基于粒子群优化的德州扑克在线对手利用方法,以开展随机运气影响下的在线策略优化和对手利用研究.

1 背景和相关工作

德州扑克是一种不完全信息的重复博弈游戏,游戏中玩家在对手隐藏底牌情况下通过合理的行动决策赢得筹码. 研究其决策以及在线对手利用,对很多具有近似特征的现实问题中的对抗与决策具有重要的参考意义.

1.1 德州扑克规则及其行动决策

德州扑克根据游戏人数不同分为两人德州扑克和多人德州扑克,根据是否限制下注金额分为有限注德州扑克和无限注德州扑克[4,15]. 德州扑克是以局为单位的回合制扑克比赛,具有特定的游戏规则. 主要包括:一局比赛分为翻牌前轮(preflop)、翻牌轮(flop)、转牌轮(turn)和河牌轮(river)4个轮次. 比赛开始前玩家根据所处位置向底池投入指定的筹码. 翻牌前轮,庄家向各玩家发出2张隐藏的底牌,在翻牌轮、转牌轮和河牌轮各发出3、1和1张公共牌. 玩家则轮流根据自己的底牌和公共牌组合来合理地选择不同的动作,包括:弃牌(fold)、跟注(call)、过牌(check)和加注(raise),最后根据摊牌后牌组合的牌型大小来决定底池筹码的归属. 德州扑克中的决策实质是在面临

每一个行动选择时,根据底牌和公共牌、双方行动序列、对手行动特点、比赛历史记录等信息,选择最合适行动以获得最大的收益. 由于对手的底牌信息在一局比赛结束前无法获知,且发牌是随机的,不确定性和信息不完全是其显著特征. 一个好的玩家必须能处理概率推理、风险管理、潜在欺骗、对手建模等问题[3, 16]. 在人工智能研究中,人们从不同的途径来处理德州扑克中的决策问题. 基于纳什均衡求解来获得均衡策略是从博弈论角度出发的. 除博弈均衡求解外,基于强化学习[17-19]和演化算法[6,9-10]等方法从遍历、演化等角度来获得最优行动策略显然也是一条好的路径,这条路径可以视为基于隐式建模的求解路径.

1.2　基于隐式建模的对手利用

基于隐式建模的对手利用常基于演化算法[6-7]、强化学习等方法获得策略模型(或称决策模型)来实现对手利用,其基本特征是通过与不同对手的大量离线训练学习到具有对手适应性的策略模型,然后在比赛中适应对手从而获益. Barone等[6]使用一种简单的参数化模型,通过与不同对手对抗的演化训练得到最优的参数,等价于隐式地考虑了对手的游戏风格,并利用他们的弱点来最大化自身收益. 其参数化模型主要依靠加注(raise)、跟注(call)和弃牌(fold)的似然概率曲线构造. 基于该参数化模型,通过参数初始化构建一个种群,然后运用演化算法的选择、交叉、变异来获得能针对多样对手的最优个体. Nicolai等[8]使用一个35-20-3的前馈神经网络来作为策略模型,输入是关于当前状态的35个特征参数,输出是3种动作的概率值,加注的筹码数根据输出值的大小分区段确定. 在演化过程中,交叉运算时,子代个体决策网络的参数从父代网络的参数加权得到,当进化过程中出现劣化趋势时,采用遗忘策略来保留更优的前代个体. 而Li等[9-10]引入循环神经网络来作为特征提取工具,结合前馈神经网络构成一个更为复杂的决策模型. 因为该模型能根据游戏的当前状态以及从历史信息中提取的模式来选择动作,同时在利用演化算法优化网络参数过程中针对不同的对手进行整体的演化,所以它具有适应不同对手以实现有效利用的能力. 这种循环神经网络也被引入到多种不同的博弈游戏中,Van Schreven[11]将其应用到6-handed-sit-and-go等游戏中,并利用神经网络的连接与否来考虑多个对手的不同情况. 为了更快地适应对手,Wu等[12]提出了快速适应对手的框架(L2E),该框架用于优化一个基础策略网络,由于该网络是以快速适应对手的目的来训练的,在对抗时天然具有快速适应能力. 除了上述基于神经网络演化训练的方法外,采用监督学习的思想从人类高手的历史数据中学习具有适应性的决策模型也可视为一种隐式建模方法[20]. 强化学习类的方法也在与不同对手的比赛中通过探索和反馈获得优化的策略网络[21-23]. 上述这些方法都采用离线训练在线适应比赛的方式,能够获得一个具有较强适应性的策略模型,但不是一种能在线比赛中主动地显式地进行策略优化的方式.

1.3　粒子群优化算法及其改进途径

群体智能算法应用于现实中的各类优化任务具有独特的潜力,呈现了应用越来越广的趋势[24]. 粒子群优化算法(particle swarm optimization, PSO)作为最常用的群体智能优化算法之一[25],在函数优化、规则挖掘、特征选择、流程控制、策略优化等众多领域和无线通信、电力网络、信号处理等广泛现实问题中得到了广泛应用[14]. 粒子群算法的基本概念起源于对鸟群觅食行为的研究,它将优化问题的解视为搜索空间中运动的粒子,通过让粒子跟踪自身局部最优解(称为认知项)和全局最优解(称为社会项)来更新自己,通过迭代的更新来获得优化解. 粒子群优化算法原理和实现都较为简单,相比于遗传算法、蚁群算法等方法收敛速度更快,但容易出现优化解局部收敛等问题. 对此,人们提出了很多改进算法,改进途径主要包括: 1) 优化惯性权重、认知项权重、社会项权重和最大速度限制; 2) 引入邻居拓扑结构; 3) 引入参数维分解等协同机制; 4) 引入多子群机制; 5) 结合其他优化算法构成混合算法; 6) 引入变异等自我调整策略; 7) 优化编码方式,等等[14]. 德州扑克在线比赛要基于在线数据高效地找到优化解,可以利用粒子群优化算法较快收敛的特点,但要尽可能地搜索到更好的全局最优解而不是较差的局部最优解,还需要考虑德州扑克中策略优化适应度计算受随机性因素影响的问题,因此,将粒子群优化算法引入德州扑克在线对手利用的策略优化中,需要根据德州扑克问题本身的特点做一定的改进.

1.4　面向在线优化的参数化策略模型

开展策略优化需要对策略进行描述并构造策略模型以便实施优化. 在不完全信息博弈理论中,策略往往基于信息集给出. 考虑玩家P和对手O的博弈问题,信息集是指玩家P要行动前能获得的历史信息集合,但因为对手的信息未知却又无法分辨具体的状态. 简单来讲,状态是确定信息的集合,信息集是不确定的状态的集合. 信息集上的策略常以概率分布的

形式表示,比如混合策略形式上就是若干行动的概率分布. 从信息集上的状态信息到若干行动的概率分布的映射函数就是策略模型,这种映射函数可以是任何形式的模型,比如神经网络、概率模型等. 然而,不同于前述隐式建模离线训练策略模型的方式,在线策略优化的对手利用要求能在在线比赛过程中,优化出针对对手的策略,具有强的实时性要求,相比于基于大型神经网络的策略模型,一些非网络形式的参数化模型则更适合作为策略模型. Barone等[6]基于指数和高斯函数构造的参数化决策模型只有3个参数,能通过调整3个参数来调整不同行动的概率. Korb等[16]给出的基于S型函数的策略模型,能够通过直接获得的相对概率归一化得到行动的概率. 李翔等[26]提出的一种基于匹配区间和相对概率的策略模型,能通过数据统计获取每种动作的赢率匹配区间,然后基于匹配区间构成每种行动的相对概率模型,最后归一化获得不同行动的选择概率. Huang[27]也采用S型函数来描述策略模型,这种模型基于行动风险期望能将不同局相同信息集上的行动次数宏观统计转换为具体行动的概率,其模型参数量小,适合作为策略模型. 本文也采用类似的基于行动风险的模型来构造参数化的策略模型.

2 基于粒子群优化的在线对手利用方法

要在比赛中获取更大的收益,就要利用对手的弱点,即找到对手的针对性策略并加以运用,进而实现收益最大化. 基于粒子群优化的在线对手利用方法需要解决在有限的在线比赛中对手针对性策略的优化和利用问题. 先将德州扑克比赛转换为一个在线策略优化问题,建立可用于优化的策略模型,给出策略优化问题的数学描述并设计优化算法,再通过改进来解决随机运气影响所导致的策略优化困难问题.

2.1 基于行动风险的策略模型

要在有限的在线比赛中主动优化获得针对对手的策略,有两个关键要求:一是实时性要求,即在边比赛边策略优化的过程中,要保证优化过程的实时性,不像离线训练没有时间限制,可以采用任意大小和形式的策略模型;二是收敛性要求,在优化过程中要保证目标策略能够快速收敛,而不是等到比赛结束了才收敛. 因此,需要使用一个具有策略表现力、参数规模适中、计算量小的模型,以保证满足这两个要求.

为减少参数量以便于优化,本文将面临决策的所有信息集抽象为两类:一类是行动选择为{fold, call, raise}的A类信息集;另一类是行动选择为{check, raise}的B类信息集. 两类信息集中每个信息集均由若干状态构成,这些状态具有相同的确定信息和可能不同的未知信息,由于决策实际依赖于这些相同的确定信息,有时也可以把信息集视为一个由确定信息表征的状态,而不是由全部信息所表征的状态的集合. 一类信息集采用一个相同的策略模型M,在决策时玩家根据所处信息集是哪一类来选择策略模型,并根据当前信息集I从策略模型M中确定出行动概率$p_a(I)$从而选择行动.

两类信息集上的策略模型M(即状态到行动概率的映射函数)由基于行动风险考虑的人类玩家决策逻辑进行构造. 以一个A类信息集为例,考虑人类玩家的一般决策逻辑:玩家当前状态下牌力越强,越倾向于选择加注行动(raise);牌力越弱则越倾向于弃牌(fold);牌力适中则倾向于选择跟注(call). 反之,在相同牌力下,继续投入金额越大的行动所造成的损失越大,则行动风险也越大,即$\text{risk}(\text{raise}, I) > \text{risk}(\text{call}, I) > \text{risk}(\text{fold}, I)$. 基于不同行动的概率与不同行动风险之间的合理关联关系,并利用两条参数化曲线来估计行动的风险,则策略模型可由信息集I上的行动概率和风险拟合参数化曲线(S型曲线)构成,即

$$\begin{aligned} &p_f(I) = \text{risk}(\text{call}, I),\\ &p_c(I) = \text{risk}(\text{raise}, I) - \text{risk}(\text{call}, I),\\ &p_r(I) = 1 - \text{risk}(\text{raise}, I); \end{aligned} \tag{1}$$

$$\text{risk}(\text{call}, I) = 1 - \frac{1}{\exp(-d(I-e))+1}; \tag{2}$$

$$e(E_c) = \frac{1}{d}\log\Big(\frac{e^d - e^{dE_c}}{e^{dE_c} - 1}\Big). \tag{3}$$

其中:p是行动概率,下标f、c、r表示三类行动;d和e为在0~1范围的S型曲线的压缩因子和位移因子,d取大于1的数,e的取值范围是-1~1;各行动概率的期望为(f, c, r)时,$E_c = r + c$,该值等于1减去跟注行动风险的期望;对于加注行动的风险$\text{risk}(\text{raise}, I)$可由类似方法得到,只是计算位移因子时,从$E_c$变为$E_r = r$,该值等于1减去加注行动风险的期望. 由此,$A$类信息集上只要选择3个参数便可构成完整的策略模型:加注和跟注行动的期望概率$r + c$、加注的期望概率r和S型曲线的压缩因子d,记为x_1、x_2、x_3. 类似地,对于B类信息集,因为只有两种行动选择,也就只需考虑两种行动风险,所以仅用一条风险拟合曲线就能将两种行动的概率表达出来,即只需用B类信息集上两个参数:加注的期望概率r和S型曲线的压缩因子d便可构成策略模型,记为x_4、x_5. 因此,在忽略诸如轮次、动作历史等次要因素,考虑上

述两类抽象信息集情况下,便可获得用5个参数构成的向量($X = (x_1, x_2, x_3, x_4, x_5)$)表示的简化决策模型,将其作为AI的决策模型(即在优化算法中采用该5维向量表示解). 值得一提的是,若考虑更多的决策场景(即考虑更多类的信息集),则可以构建维数更大的向量作为策略模型的参数,这里是为了减小优化参数维度而采用了两类信息集的抽象.

2.2　德州扑克AI在线策略粒子群优化框架

2.2.1　面向在线比赛的时间维粒子设定

在群体智能优化算法中,类似粒子群这样的种群概念通常是定义在空间维度上的,比如遗传算法中种群内的个体是根据染色体的差异独立存在于空间内的,常见的优化问题大多采用这种定义. 但在在线比赛中种群无法定义在空间上,因为对手只能实时地与一个人比赛,所以种群的定义只能换一种思路. 本文采用从时间维定义种群的方法,即将一个固定时间段分配给一个粒子(个体)比赛并进行适应度评价,由全部粒子(个体)分配的时间段所构成的完整时间段作为当前种群的评价时间段. 通过这样的时间维粒子设定,使得基于群体优化的算法能在在线比赛中得以实施. 但相比于在空间维定义种群的方法,这种设定使得种群的评价只能串行进行而无法并行进行,同时因为总比赛时间(局数)的限制以及单个粒子评价受随机性影响,种群大小也会有所限制. 因此,面向在线比赛的粒子群优化问题相比于一般的粒子群优化问题存在显著差异,具体特点总结于表1中.

表1　面向在线德州扑克比赛的粒子群策略优化特点

问题	在线比赛问题	一般优化问题
优化目标	整个比赛过程收益最大	种群适应度最大
粒子维度	时间维	空间维
评价计算	串行	并行/串行
种群大小	受限	无条件限制
粒子适应度	受随机运气影响	无随机因素

2.2.2　优化问题数学描述

根据两人无限注德州扑克在线比赛整体收益最大化的目标,基于粒子群优化算法构建主动策略优化AI的目的是:在有限的总比赛局数(n)条件下,能更快更好地优化得到针对对手的策略模型以利用对手来获取最大收益. 这是一个以比赛收益最大为目标的优化问题,用数学形式描述为

$$F = \max w(n-1, \boldsymbol{X}) | \{\boldsymbol{X}(t=0), \ldots\}. \tag{4}$$

其中:w为从第0局开始到当前局的总收益,$w(n-1)$为n局比赛的总收益(即比赛中赢得的总筹码数),$\boldsymbol{X}$为策略模型,t为粒子群优化算法的迭代步,T为总迭代步数.

因为德州扑克比赛是以局为单位的重复随机博弈,所以在时间维度上可以用比赛局h为单位进行描述,即用给定局数n_f的比赛表示分配给每个粒子的时间段,而单个粒子的适应度就是该粒子在n_f局比赛中的总收益. 因此,基于时间维的粒子设定,优化问题可以进一步写为

$$F = \max \sum_{t=0}^{T-1} \sum_{i=0}^{m-1} f_i(\boldsymbol{X}(t)), \tag{5}$$

其中m是粒子的数量. 因单个粒子的策略是用一个5维的向量X表示,故粒子群的策略模型$\boldsymbol{X}$可以表示成一个$m \times 5$的矩阵. n_f为单个粒子适应度计算所需局数,粒子i的适应度为$f_i = \sum_{h}^{n_f} w_h(\boldsymbol{X}_i)$,$h$为比赛局的序号,$w_h(\boldsymbol{X})$为第$h$局的收益(即第$h$局赢得的筹码数). 显然,式(5)这样的优化问题不同于仅考虑种群适应$\left(\max \sum_{i=0}^{m-1} f_i(\boldsymbol{X}(T-1))\right)$的一般优化问题,德州扑克中的在线对手利用问题不仅要求找到最优解,而且要求整个比赛过程的收益最大.

2.2.3　适用于在线比赛的优化框架

基于前述问题定义和时间维的粒子设定,德州扑克比赛AI策略优化的粒子群优化框架如图1所示.

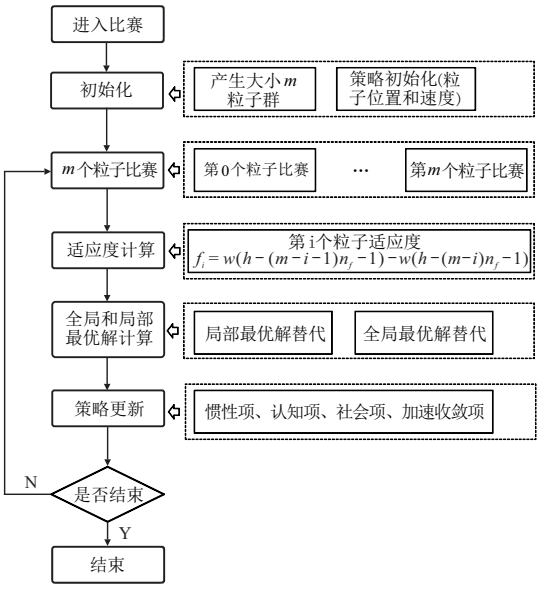


图1　德州扑克AI在线策略优化框架

基本优化过程可以理解为:在进行$n = T \times m \times n_f$局的在线比赛中,$m$个粒子共进行了$T$次迭代更新. 在开始比赛粒子初始化后,每个粒子依次进行n_f局比赛,完成$m \times n_f$局比赛后,计算所有粒子的适应度,并利用粒子更新规则开展第$t = 1$次粒子更新(一次更新表示完成了粒子群优化的一次迭代). 接着,更新后的粒子再次进行$m \times n_f$局比赛,然后开展

第$t=2$次粒子更新. 不断重复这个过程直到比赛结束,完成$t=T$次粒子更新.

2.3 随机运气影响下的改进粒子群算法

德州扑克这类具有不完全信息特征的随机序列博弈问题中的策略优化,不同于一般的函数优化问题,要考虑随机因素带来的影响. 对于一个固定策略的玩家,由于发牌的随机性,在一局比赛中可能拿到好牌而获得高的收益,在下一局可能拿到差牌导致低收益,即使考虑一定局数的总收益,仍会受到运气的影响. 所以要有效地引导策略的优化,必须找到一种方法来削弱运气对于适应度计算的影响. 随机运气影响适应度计算主要导致的是错误的优化解追踪问题,表现为局部最优解的劣化追踪和全局最优解的错误追踪带来的收敛慢或无法收敛问题,表现在比赛收益上就是持续地输给对手. 为解决上述问题,本文提出一种改进的粒子群优化算法,称为最优解替代的粒子群优化算法,记为BR-PSO (best replacement PSO).

2.3.1 局部最优解(单粒子的历史最优解)替代

在粒子初始化后,通过n_f局比赛获得第1次的适应度可能是负的,表示当前粒子所代表的策略不能针对对手(若发牌运气不差但收益差,则表示策略差),若将其作为粒子历史上的局部最优解,则粒子的局部最优解追踪可能劣化. 因此,需要改进局部最优解的生成方式,当粒子的历史上所有的适应度为负时,不再使用该粒子的历史局部最优解,而从适应度为正的其他粒子中随机选择一个,作为该粒子局部最优解的替代. 首先更新粒子的历史最优适应度f_i^l,即

$$f_i^l(t)=\begin{cases}\max_t f_i(t),\ f_i(t)>0 \bigvee f_i^l(t-1)>0;\\ 0,\ \text{otherwise}.\end{cases} \tag{6}$$

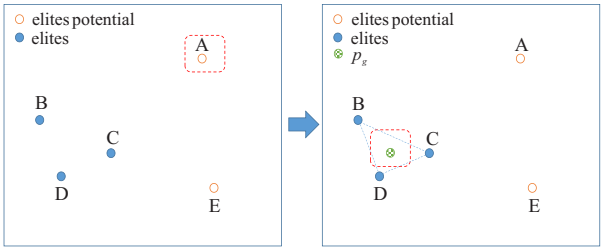
然后更新粒子的历史最优解

$$p_i^l(t)=\begin{cases}p_j^l(t),\ f_i^l(t)=0;\\ p_i^l(t-1),\ f_i^l(t-1)\geqslant f_i(t);\\ X_i(t),\ f_i^l(t-1)<f_i(t).\end{cases} \tag{7}$$

其中j表示历史最优适应度大于0的某个粒子,本文设定从历史最优适应度最好的3个粒子中随机选择一个.

2.3.2 全局最优解(群的历史最优解)替代

在标准粒子群算法中,全局最优解是所有粒子的历史中适应度最好的那个解. 然而在随机运气的影响下,有可能某个粒子在某一段时间运气特别好,即使其代表的策略不是好策略,但也可能获得最高的适应度,此时若跟踪该最优解,则导致无法获得正确的优化解. 因此,必须找到方法来引导全局最优解的产生,使其尽可能处于正确的位置上. 本文考虑两个思路:一是利用所有粒子的历史最优适应度信息;二是利用当前时刻所有粒子的适应度信息. 基于这些信息选择精英解,并将精英解的重心位置作为全局最优解的替代. 首先根据历史最优适应度为正选择对应的粒子为潜在精英粒子,并从潜在精英粒子中选择当前时刻适应度为正的粒子作为真正的精英解. 这种历史收益为正且当前收益也为正的粒子呈现了其所代表策略的正收益稳定性,而使用精英解的重心目的也是期望让全局最优解向真正的全局最优区域移动,以尽可能消除随机运气的影响. 图2给出了使用精英解重心代替标准PSO算法全局最优解的原理.



粒子的历史最优适应度排序为A>B>C>D>E,
标准PSO将会把A作为全局最优解,而本文将
精英解B、C、D的重心作为全局最优解的替代

图 2 全局最优解替代策略

在实现上考虑:当精英解的数量超过3时,选择适应度最好的3个粒子的重心作为全局最优解;当精英解数量小于3时,使用全部精英解的重心作为全局最优解. 当不存在精英解时,对潜在精英解采取相同处理来作为全局最优解,即

$$p^g(t)=\begin{cases}\bar{p}_e^l,\ |e|>0;\\ \bar{p}_{\text{ep}}^l,\ \text{otherwise}.\end{cases} \tag{8}$$

其中:$e=\{i|f_i^l(t-1)>0\bigwedge f_i(t)>0\}$表示精英解,$\text{ep}=\{i|f_i^l(t-1)>0\}$表示潜在精英解,$\bar{\cdot}$表示求均值(由于解空间上处处密度相同,重心位置可以用均值表示),$|e|$表示精英解的数量.

值得注意的是,局部最优解和全局最优解的引导作用体现在粒子运动速度和方向的修正上. 由于最大速度限制,随机系数、权重系数以及惯性项的存在,粒子的位置不会突然移动到当前最优解上,而会通过粒子的运动持续地探索解空间,这种最优解替代不会破坏粒子群算法解搜索的特性.

3 实验与分析

3.1 实验设置

实验基于开源的ACPC德州扑克平台开展两人无限注比赛(单局单玩家筹码量为20 000),以粒子群

策略优化AI与其他AI比赛的形式来考察算法对于特定对手的针对性策略的优化能力. 本文基于不同紧松和凶弱程度的决策逻辑构建4种基础风格类型[6,28]的AI对手玩家, 使用α和β两个参数, 其中α表示玩家主动玩牌(不弃牌)的最小期望赢率, β表示玩家主动加注的最小期望赢率, 期望赢率是指玩家根据当前的底牌和已经发出公共牌以及未知的对手手牌和未发出的公共牌计算的到当前局结束时能赢的期望概率. 4个对手LA (loose-aggressive 松凶)、LP (loose-passive 松弱)、TA (tight-aggressive 紧凶)、TP (tight-passive 紧弱) 的α和β取值分别为 [0.3, 0.45], [0.3, 0.85], [0.55, 0.65], [0.55, 0.85]. 我方基于粒子群优化的AI, 根据策略模型给出的行动概率和随机数来选择行动, 其中加注行动的金额也由随机数决定, 即在最小加注金额和最大加注金额中随机选择. 粒子群速度更新中的惯性权重、认知权重、社会权重根据实验条件设定.

3.2　结果与分析

3.2.1　标准PSO和BR-PSO的结果

利用前述框架考察标准PSO和BR-PSO对于不同风格对手的优化结果. 标准PSO实验中取$m = 20, n_f = 1\,000$, 惯性项权重$\omega = 0.5$, 认知项权重$c_1 = 0.1$, 社会项权重$c_2 = 0.1$.

图3给出了针对4类不同对手的优化结果, 其中ptct指应用了粒子群优化的AI. 从收益历史可知: 对于LA和LP对手, 采用标准粒子群优化很快就获得了

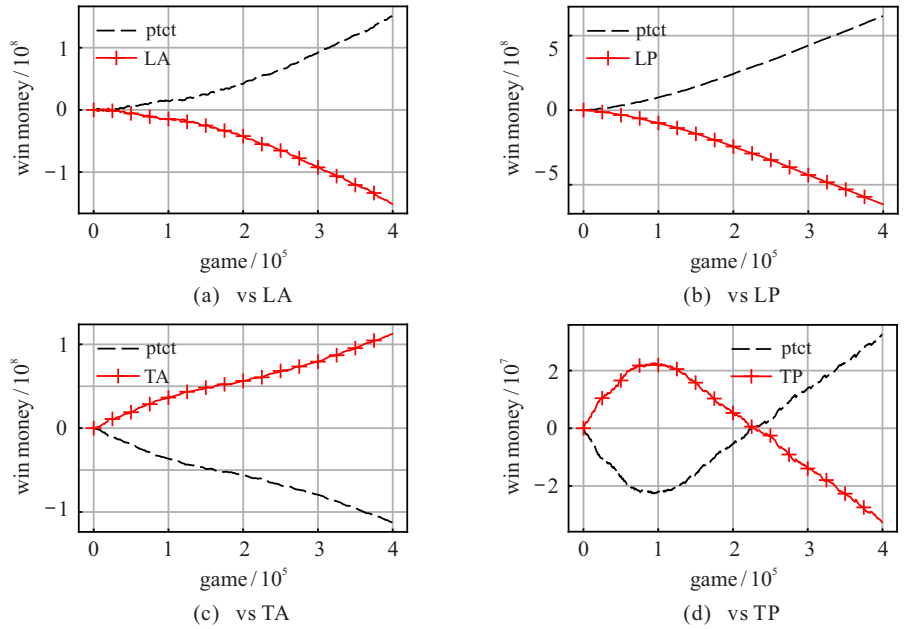


图 3　面对不同对手标准PSO优化的收益历史

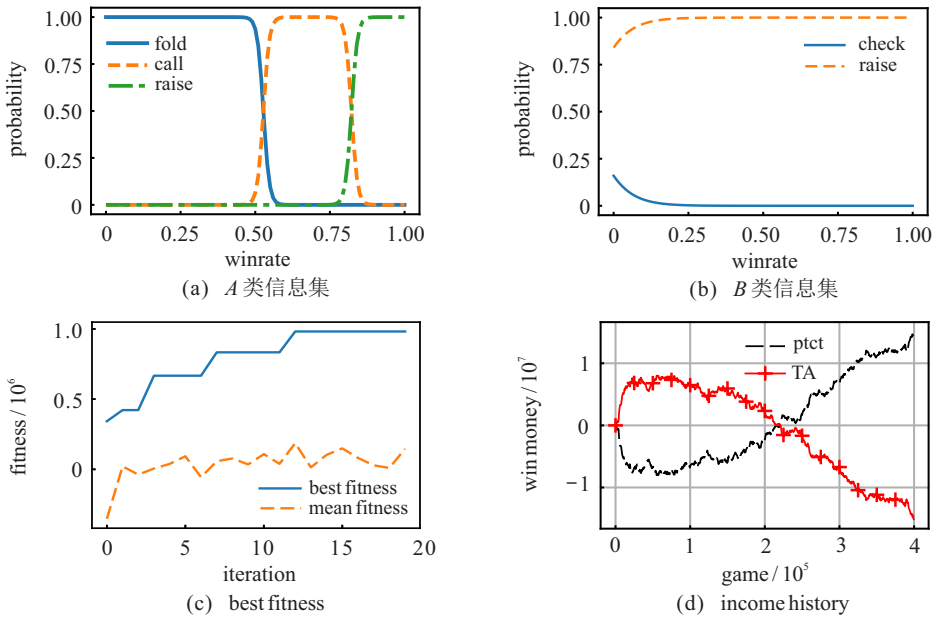


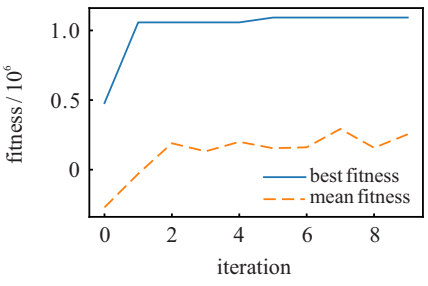
图 4　TA对手的BR-PSO优化结果

具有针对性的强策略;对于TP对手,在经过80 000局比赛后逐渐找到了针对性策略;而对于TA对手,经过400 000局比赛仍然无法优化出针对性策略. 从总收益上看,针对LP和LA对手的最大收益在10的8次方量级,针对TP对手的最大收益在10的7次方量级,也说明松的玩家相比紧的玩家更容易被针对.

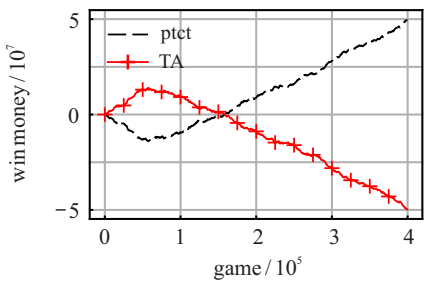
BR-PSO主要针对标准PSO难以针对的TA对手开展实验. 图4给出了考虑$m = 20, n_f = 1\,000$,权重系数根据文献方法基于sin/cos函数动态变化[29],ω从0.95到0.675,c_1从2.5到0.5,c_2从0.5到2.5,且不限制最大速度仅限制参数取值范围的结果. 图4(d)的收益历史表明优化方法逐渐找到了TA对手的针对性策略. 从图4(c)看,尽管全局最优适应度呈现明显上升趋势,但平均适应度的增长并不明显,这充分反映了随机运气的影响,即全局最优适应度的获得是由运气较好的粒子获得的. 图4(a)和图4(b)是根据原始的全局最优解$p^g = [0.41, 0.14, 97.57, 0.99, 19.61]$给出$A$类和$B$类信息集上的策略模型,大体反映了对紧凶(TA)对手的一种应对策略:在对手不加注情况下,我方加强加注会有利;在对手加注情况下,我方在牌力很强时(赢率大于0.8)加注;牌力较强时(赢率大于0.5)跟注会有利.

3.2.2 种群大小等参数的影响

对于TA对手采用BR-PSO算法并增加粒子数来考察种群大小的影响. 图5给出了种群大小$m = 40$时的优化结果. 从收益历史看,种群增大后同样获得了TA的针对性策略,而且相比于种群大小$m = 20$时收益更好,一定程度上反映了其搜索到的优化策略更好. 得到的代表性原始全局最优解为$p^g = [0.86, 0.85, 90.96, 0.99, 15.11]$,表明在TA对手加注情况下采用较大的加注概率也是一种针对性策略,这个全局最优解与种群数为20时搜索到的全局最优解是不同的,说明对于一个相同对手存在不同的针对性策略.



(a) best fitness



(b) income history

图 5 TA对手的BR-PSO优化结果($m = 40$)

因为每个粒子适应度计算所用的局数n_f是有限的,所以适应度计算不可避免地受到随机运气的影响. 在前面$n_f = 1\,000$的基础上增加一倍,即$n_f = 2\,000$,来考察n_f、权重及速度限制对改进算法(BR-PSO)的影响. 在粒子数$m = 20$时考虑如下情况:

case 1: $n_f = 2\,000, \omega \in [0.95, 0.675], c_1 \in [2.5, 0.5], c_2 \in [0.5, 2.5]$,最大速度不做限制;

case 2: $n_f = 2\,000, \omega = 0.5, c_1 = 1.5, c_2 = 1.5$,最大速度不做限制;

case 3: $n_f = 2\,000$,速度权重同case1,最大速度限制为$[0.1, 0.1, 5, 0.1, 5]$;

case 4: $n_f = 2\,000$,速度权重同case2,最大速度限制为$[0.3, 0.3, 5, 0.3, 5]$.

图6给出了4种情况的收益结果. case 1的收益表明优化算法在第2次位置更新后(80 000局)基本找到了针对对手的策略. case 2的收益表明优化算法在第1次位置更新后(40 000局)就基本找到了针对性策略,最终收益也达到了7次方量级. 从case 2与case 1的比较看,固定速度权重比动态调整权重反而更快地收敛到针对性策略,说明相比于动态权重调整尝试更广泛的策略搜索,固定权重倾向于快速收敛到当前搜索到的好的解. case 3的收益表明优化算法在10次位置更新后(400 000局)才基本找到针对性策略,相比于前面的方法收敛速度过慢,显然这是由于最大速度限制所导致的,x_4等参数的最大调整速度0.1限制了解的快速收敛. case 4的收益表明优化算法在4次位置更新前(160 000局)呈现了一段收益波动的区间,说明粒子群的策略尚未达到整体较优,但在160 000局之后明显找到针对性策略,收益呈现快速增长趋势,其代表性的最优解为$[0.97, 0.22, 61.78, 0.99, 64.79]$. 这也说明固定权重和放宽速度限制能够加速针对性策略的收敛,尽管这种针对性策略可能并不是最优的,但对于博弈而言,要尽可能地获得更大的收益,加速收敛到一个能够较好地利用对手的策略,不失为一种好的选择.

此外,考虑到针对TA存在多个针对性策略,将前述3个代表性策略固定后与TA对手比赛,并与基于手牌预测方法的简单规则决策AI(名为ph)的比赛

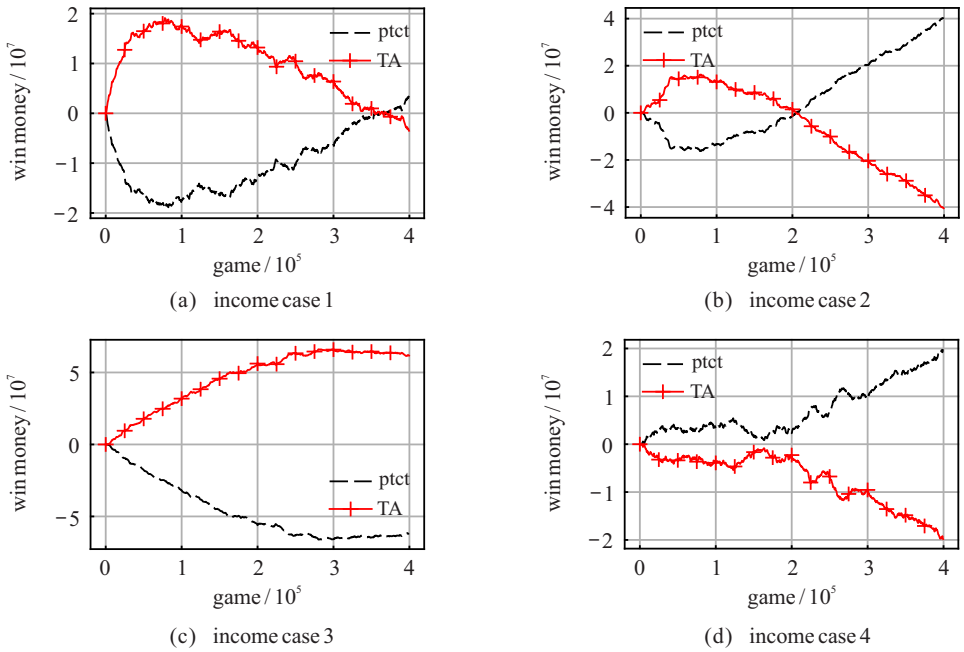


图 6　不同参数下 TA 对手的 BR-PSO 优化结果

结果做比较, 其中策略 $X = [0.86, 0.85, 90.96, 0.99, 15.11]$ 获得的总收益 (93 453 887) 与 ph 的收益 (95 420 539) 接近, 说明通过粒子群实现的策略优化能在一定程度上弥补未采用手牌预测等措施所带来的局限.

4　结　论

本文针对德州扑克面对全新对手时的在线对手利用问题, 开展了基于粒子群优化方法的策略优化方法研究. 以获取比赛最大收益为目的, 给出了问题的数学模型和优化框架, 并针对适应度受随机运气影响和部分对手针对性策略较难收敛的问题, 提出了基于最优解替代的改进粒子群优化方法 (BR-PSO). 实验表明: 1) 粒子群优化算法可以应用于德州扑克这种具有强随机性特征的问题中, 能够仅通过在线比赛就优化得到对手的针对性策略; 2) 在以收益为目标的比赛中, 要权衡解的优化与收益的最大化, 若单以解的优化为目标, 则很可能达不到利用对手以最大化收益的目的; 3) 对于一个相同的对手, 存在不同的针对性策略, 这些策略的收益程度会存在差异, 为了快速针对对手以获得更大收益, 加速收敛到非最优的针对性策略以避免过度搜索浪费对手利用时间也是一种好的选择.

本文将在线策略优化的思路引入德州扑克这种随机博弈问题中, 基于粒子群优化特点, 结合改进的粒子更新策略实现了有效的在线对手利用, 证明了在线比赛策略优化中改造并应用高效优化算法的可行性, 为一些现实对抗问题中的对手针对性策略在线获取提供了有益借鉴. 值得一提的是, 本文仅讨论了两人对抗情况下的对手利用. 在多人对抗情况下, 由于对手人数的增加和对手之间的差异使得问题变得更为复杂, 是针对所有对手开展策略优化, 还是只针对较弱对手开展利用等问题值得讨论; 而与之相对应的引导策略优化的适应度计算等方面也可能需要进行新的调整, 这些都有待于下一步的深入研究.

参考文献 (References)

[1] Russel S J, Norvig P. Artificial intelligence a modern approach[M]. The 4th edition. Harlow: Pearson Education, 2022.

[2] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[3] Billings D, Davidson A, Schaeffer J, et al. The challenge of poker[J]. Artificial Intelligence, 2002, 134(1/2): 201-240.

[4] Moravčík M, Schmid M, Burch N, et al. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker[J]. Science, 2017, 356(6337): 508-513.

[5] Brown N, Sandholm T. Superhuman AI for multiplayer poker[J]. Science, 2019, 365(6456): 885-890.

[6] Barone L, While L. An adaptive learning model for simplified poker using evolutionary algorithms[C]. Proceedings of the Congress on Evolutionary Computation. Washington, 2002: 153-160.

[7] Sipper M, Azaria Y, Hauptman A, et al. Designing an evolutionary strategizing machine for game playing and beyond[J]. IEEE Transactions on Systems Man & Cybernetics, Part C — Applications & Reviews, 2007, 37(4): 583-593.

[8] Nicolai G, Hilderman R. Countering evolutionary for getting in no-limit Texas Hold'em poker agents[C]. Computational Intelligence. Berlin: Springer, 2012: 31-48.

[9] Li X, Miikkulainen R. Evolving adaptive LSTM poker players for effective opponent exploitation[C]. AAAI Workshops. San Francisco, 2017: 1-13.

[10] Li X, Miikkulainen R. Opponent modeling and exploitation in poker using evolved recurrent neural networks[C]. Proceedings of the Genetic and Evolutionary Computation Conference. Kyoto, 2018: 189-196.

[11] Van Schreven C. Deepbot-neuroevolutionary agent for opponent exploitation in 6-handed sit-and-go poker[D]. Geneva: Distributed Information Systems Laboratory. 2019: 1-49.

[12] Wu Z, Li K, Xu H, et al. L2E: Learning to exploit your opponent[C]. 2022 International Joint Conference on Neural Networks. Padua, 2022: 1-8.

[13] Bassimir B, Rab A, Schmitt M. Theory of particle swarm optimization: A survey of the power of the swarm's potential[J]. IT: Information Technology, 2019, 61(4): 169-176.

[14] Shami T M, El-Saleh A A, Alswaitti M, et al. Particle swarm optimization: A comprehensive survey[J]. IEEE Access, 2022, 10: 10031-10061.

[15] Rubin J, Watson I. Computer poker: A review[J]. Artificial Intelligence, 2011, 175(5/6): 958-987.

[16] Korb K B, Nicholson A E, Jitnah N. Bayesian poker[C]. Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Stockholm, 1999: 343-350.

[17] Dahl F A. A reinforcement learning algorithm applied to simplified two-player Texas Hold'em poker[C]. Machine Learning: ECML 2001. Berlin: Springer, 2001: 85-96.

[18] Brown N, Bakhtin A, Lerer A, et al. Combining deep reinforcement learning and search for imperfect-information games[J/OL]. 2020, arXiv: 2007.13544.

[19] Zhao E M, Yan R Y, Li J Q, et al. AlphaHoldem: High-performance artificial intelligence for heads-up No-limit poker via end-to-end reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(4): 4689-4697.

[20] Teofilo L, Reis L P. Building a no limit Texas Hold'em poker agent based on game logs using supervised learning[C]. International Conference on Autonomous and Intelligent Systems. Burnaby, 2011: 1-10.

[21] Hoehn B. The effectiveness of opponent modelling in a small imperfect information game[D]. Edmonton: University of Alberta, 2006.

[22] Bard N, Johanson M, Burch N, et al. Online implicit agent modelling[C]. AAMAS'13: Proceedings of the 2013 International Conference on Autonomous Agents and Multiagent Systems. St. Paul: International Foundation for Autonomous Agents, 2013: 255-262.

[23] Foerster N J, Chen Y R, Al-Shedivat M, et al. Learning with opponent-learning awareness[C]. Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'18), 2018: 122-130.

[24] Tang J, Liu G, Pan Q T. A review on representative swarm intelligence algorithms for solving optimization problems: Applications and trends[J]. IEEE/CAA Journal of Automatica Sinica, 2021, 8(10): 1627-1643.

[25] Kennedy J, Eberhart R. Particle swarm optimization[C]. Proceedings of International Conference on Neural Networks. Perth, 2002: 1942-1948.

[26] 李翔, 姜晓红, 陈英芝, 等. 基于手牌预测的多人无限注德州扑克博弈方法[J]. 计算机学报, 2018, 41(1): 47-64.
(Li X, Jiang X H, Chen Y Z, et al. Game in multiplayer no-limit Texas Hold'em based on hands prediction[J]. Chinese Journal of Computers, 2018, 41(1): 47-64.)

[27] Huang J. Building a computer poker agent with emphasis on opponent modeling[D]. Massachusetts: Massachusetts Institute of Technology, 2011: 1-54.

[28] Mandziuk J. Knowledge-free and learning-based methods in intelligent game playing[M]. Berlin: Springer-Verlag, 2010: 169-180.

[29] Chen K, Zhou F Y, Yin L, et al. A hybrid particle swarm optimizer with sine cosine acceleration coefficients[J]. Information Sciences, 2018, 422: 218-241.

作者简介

胡振震 (1984–), 男, 工程师, 博士生, 从事人工智能、认知决策博弈等研究, E-mail: hzzmail@163.com;

陈少飞 (1987–), 男, 副教授, 博士, 从事多智能体系统、机器学习等研究, E-mail: chenshaofei01@nudt.edu.cn;

袁唯淋 (1994–), 男, 助理研究员, 博士, 从事人工智能、信息通信等研究, E-mail: yuanweilin@nudt. edu.cn;

李鹏 (1997–), 男, 博士生, 从事强化学习、多智能体协作、模式识别等研究, E-mail: lipeng@nudt.edu.cn;

陈璟 (1972–), 男, 教授, 博士生导师, 从事人工智能、智能规划等研究, E-mail: chenjing001@vip.sina.com.