智能体对手建模研究进展

刘婵娟， 赵天昊， 刘睿康， 张 强

(大连理工大学计算机科学与技术学院，辽宁 大连 116024)

摘 要：智能体是人工智能领域的一个核心术语。近年来，智能体技术在自动无人驾驶、机器人系统、电子商务、传感网络、智能游戏等方面得到了广泛研究与应用。随着系统复杂性的增加，关于智能体的研究重心由对单个智能体的研究转变为智能体间交互的研究。多个智能体交互场景中，智能体对其他智能体决策行为的推理能力是非常重要的一个方面，通常可以通过构建参与交互的其他智能体的模型，即对手建模来实现。对手建模有助于对其他智能体的动作、目标、信念等进行推理、分析和预测，进而实现决策优化。为此，重点关注智能体对手建模研究，展开介绍关于智能体动作预测、偏好预测、信念预测、类型预测等方面的对手建模技术，对其中的优缺点进行讨论和分析，并对手建模技术当前面临的一些开放问题进行总结，探讨未来可能的研究和发展方向。

关 键 词：决策智能；对手建模；博弈论；智能体系统；AlphaGo

中图分类号：TP 391 **DOI：**10.11996/JG.j.2095-302X.2021050703

文献标识码：A **文 章 编 号：**2095-302X(2021)05-0703-09

Research progress of opponent modeling for agent

LIU Chan-juan, ZHAO Tian-hao, LIU Rui-kang, ZHANG Qiang

(School of Computer Science and Technology, Dalian University of Technology, Dalian Liaoning 116024, China)

Abstract: Agent is a core term in the field of artificial intelligence. In recent years, agent technology has been widely studied and applied in such fields as autonomous driving, robot system, e-commerce, sensor network, and intelligent games. With the increase of system complexity, the research focus on agent technology has been shifted from single agent to interactions between agents. In scenarios with multiple interactive agents, an important direction is to reason out other agents' decisions and behaviors, which can be realized through the modeling of other agents involved in the interaction, that is, opponent modeling. Opponent modeling is conducive to reasoning, analyzing, and predicting other agents' actions, targets, and beliefs, thus optimizing one's decision-making. This paper mainly focused on the research on opponent modeling of agents, and introduced the opponent modeling technology in agent action prediction, preference prediction, belief prediction, and type prediction. In addition, their advantages and disadvantages were discussed, some current open problems were summarized, and the possible future research directions were presented.

Keywords: decision intelligence; opponent modeling; game theory; agent systems; AlphaGo

收稿日期：2021-04-01；**定稿日期：**2021-05-21

Received：1 April，2021；**Finalized：**21 May，2021

基金项目：中国科协青年人才托举工程(2018QNRC001)；国家自然科学基金项目(61702075，31370778，61425002，61772100，61751203)

Foundation items：Young Elite Scientists Sponsorship Program by CAST (2018QNRC001); National Natural Science Foundation of China (61702075, 31370778, 61425002, 61772100, 61751203)

第一作者：刘婵娟(1986–)，女，河北邯郸人，副教授，博士。主要研究方向为智能决策与智能计算。E-mail：chanjuanliu@dlut.edu.cn

First author：LIU Chan-juan (1986–), female, associate professor, Ph.D. Her main research interests cover intelligent decision and intelligent computing. E-mail：chanjuanliu@dlut.edu.cn

通信作者：张 强(1971–)，男，陕西西安人，教授，博士。主要研究方向为图形图像处理、智能计算等。E-mail：qzhangdl@163.com

Corresponding author：ZHANG Qiang (1971−), male, professor, Ph.D. His main research interests cover graphic image processing, intelligent computing, etc. E-mail：qzhangdl@163.com

在人工智能领域，智能体(agent)是一个物理的或抽象的实体，能够通过感知器感知环境，对环境做出反应并通过效应器作用于环境。近年来，智能体技术在自动无人驾驶、机器人系统、电子商务、传感网络、智能游戏等方面得到了广泛研究与应用，成为备受关注的研究方向之一。智能体技术主要分为对单个智能体和智能体间交互的研究。随着系统复杂性的增加，关于智能体的研究重点逐步由单个智能体转变为智能体之间的交互。

智能体间的交互形式主要包括竞争和合作 2 种。博弈论是研究互动决策的理论[1-2]，为刻画和分析多智能体相互之间竞争与合作决策提供了理论框架。由于现实博弈的状态动作空间太大、博弈元素不完全可知等因素，智能体难以实现完全理性，且智能体策略往往会不断学习变化。鉴于传统博弈理论中理性假设较强以及在状态转换和策略动态演化方面的建模不足，又基于有限理性的演化博弈理论及强化学习方法得到广泛研究与应用[3]，为博弈决策的建模与求解提供了新的思路。

在智能体交互环境中，智能体的收益不仅取决于环境，同时也取决于其他智能体的动作[4]。因此对其他智能体的策略、信念等特征的推理，对于智能体的有效决策至关重要。同时，参与交互的智能体在进行交互时往往会暴露出一些行为特征或策略偏向。如果一个智能体在进行交互决策时，能够通过对其他智能体的建模而对其特征加以利用，则可帮助该智能体做出更好地决策。这种对参与交互的其他智能体进行建模的方法称为对手建模[5]。经典的对手建模是根据智能体的交互历史建立一个对手模型[6]，将该模型看作一个函数，输入智能体的交互历史信息与数据，输出其动作、偏好、目标、计划等的预测。

根据交互环境的不同，对手建模可以分为竞争对手和合作对象，统称为对手[7]。象棋等对抗性博弈是对手建模研究的主要推动力之一。这类博弈的主要解决方案是基于“极大极小”原理(动态博弈中的极大极小方法如图 1 所示，其中，EV 为用启发式评估函数计算得出的博弈树叶节点的评估值；BV 为由极大极小原理得到各个状态的评估值；MAX 为选择其直接后继子节点中价值最大的节点价值作为对该节点的评估值；MIN 为选择其直接后继子节点中最小的价值)，即智能体针对最坏情况、万无一失地对手优化其决策。但通常情况下玩家对其对手并不是一无所知。与该玩家博弈交互过程中可以积累经验和知识，即具有不完美的对手信息。随着完美信息博弈的研究不断取得成功，不完美信息博弈逐渐成为研究的难点[8]。由于无法准确获取对手的全部信息，对手模型的研究成为不完美信息博弈中的一个核心方向，即在博弈过程中通过对手建模，尽可能地发现并利用对手策略中的弱点，从而获得更大的收益。处在同一环境中的多个智能体，由于资源的有限性必然面临资源竞争导致的利益冲突，因此每个智能体必须考虑其他智能体可能采取的行动及该行动对自己产生的影响。因此，如何在竞争环境中建立对手模型并不断更新，进而更好地了解和预测对手并正确制定自身对策使自己利益最大化，是智能体决策优化的一个关键问题。

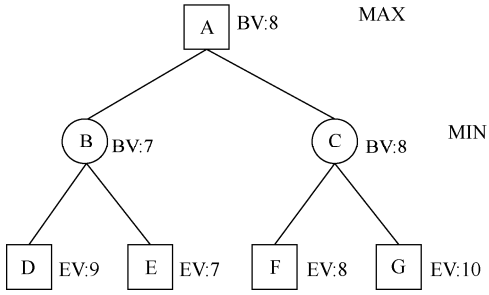

图 1 极大极小算法过程示例(在深度为 3 和分支因子为 2 的博弈树中采用极大极小策略进行决策)

Fig. 1 An illustration of minimax process (In the game tree with depth of 3 and branching factor of 2, minimax strategy is used to make decisions)

除了传统博弈论领域，在多智能体强化学习领域中对手建模也得到了一系列的研究[7]。多智能体强化学习考虑的是一群智能体彼此交互并与环境相互作用的决策场景，重点关注多智能体的协同问题。随着多智能体系统的发展，其类型也越来越多，该决策的关键是基于彼此的了解及调整自己的策略。开放多智能体环境中，智能体会遇到之前未知的对手，那么可能需要将新的对手与老对手进行对比，并用之前的成功策略作为该对手可能采用的策略，从而更好地协作。

对手建模在智能游戏、自动协商、足球机器人、人机交互等人工智能应用场景中有着重要的应用价值[7]。例如，自动驾驶汽车要保持在道路上安全稳定运行，不仅需要感知环境，还需要通过预测汽车和行人的行驶方向、运动轨迹来更好地避免事故；在谈判过程中，如果知道对方的底线，就能更快地达成理想的协议；游戏设计领域的一个重要挑战是创建自适应的可对玩家动作做出正确响应的 AI 对手，要求其能够首先识别玩家的策略。基于大

量的游戏回放，AI 对手能够更好地预测对手的策略，提升自身的能力；在商业战中，如果知道对手的商业策略，则能够利用这些信息做出更有效地打击对手的决策。

目前已有很多的智能体对手建模方法被提出。依据建模和预测的目标，对手建模方法可分为对对手动作、对手类型、对手偏好及对手信念的建模等。不同方法各具优缺点且适用于不同的情况。总体而言，理论和实践证明通过合适的对手建模方法可实现对对手属性的预测，能够在各种交互场景中帮助智能体实现决策优化。随着智能交互场景的日益丰富与完善，对手建模技术也得到不断地研究与发展，且面临新的困难与挑战。

本文基于对相关文献的研究和总结，从对智能体的动作建模、偏好建模、信念建模、类型建模以及其他建模等几个角度出发，对其主要的方法和技术进行了详细地阐述并对其优势及不足进行了分析，最后就对手建模技术存在的难点和挑战性问题进行总结，并探讨未来可能的研究和发展方向。

1　面向动作预测的对手建模研究

动作预测即通过重建对手的决策方法，显式预测对手的未来动作，是对手建模中最常见的一种方法。其基本思路是首先建立一个初始的对手模型，可以初始化一个随机模型或引用已知特定参数的一个对手模型作为初始模型；然后基于与对手的交互过程，不断对模型参数进行调整；最终拟合出与观察结果相符的对手模型。面向动作预测的对手建模方法主要包括基于动作频率和基于相似性推理2种。

1.1　基于动作频率的对手动作预测

动作频率是根据对手的历史动作行为，将智能体的可能动作构建为概率分布模型。动作频率的建模最早见于“虚拟游戏”[9]中，采用最大似然估计法将观察到的对手动作进行拟合，即计算动作的平均频率实现智能体相互建模。这种方法后被应用到矩阵博弈[10]和多智能体强化学习当中。文献[11]通过将 RL (reinforcement learning)与均衡(或协调)学习方法相结合，来学习自身及其他智能体的联合动作估值。每个智能体 i 都有一个计数器 $C^j_{a^j}$，表示其他智能体 j 在过去使用动作 a^j 的次数。该计数反映了一个智能体对于其他智能体的策略的信念。在博弈时，i 将 j 的每一个动作的相对频率作为 j 当前策略的指示，假设每一个其他的智能体均以与 i 当前关于 j 的信念来选择行动。对于每个智能体 j，i 假设 j 以概率 $Pr^j_{a^j}\Big/\left(\sum_{b^j\in A_j} C^j_{b^j}\right)$ 执行动作 a^j。这些概率集合构成一个策略集 Π_{-i}，i 则将对此做出最优反应，并对其动作 a^i 进行评价，即

$$EV(a^i)=\sum_{a^{-i}\in A^{-i}} Q\left(a^{-i}\cup\{a^i\}\prod_{j\neq i}\left\{Pr^i_{a^{-i}[j]}\right\}\right) \quad (1)$$

一次决策交互之后，i 根据其他智能体使用的动作相应更新其计数。通过博弈实验结果可以看出，采用基于对手动作频率的联合动作估值有助于通过较少的交互次数提升多智能体系统中智能体选择最优联合动作的概率。

采用基于动作频率的对手建模方法，需要解决的关键问题是了解在建模中使用哪些历史元素。如果使用的历史信息过少或有错，就有可能做出错误或低可靠性的预测；如果使用历史信息过多，就会导致学习过慢。为解决此问题，JENSEN 等[12]提出使用历史中 n 个最近元素的可能子集来学习动作频率的方法。为解决子集组合爆炸的问题，该方法采用条件熵进行历史信息的选择。熵越高则不确定性越高，因此，最终选择的是熵最低的历史元素子集。

由此可见，基于动作频率的对手动作预测方法主要是基于历史交互信息对未来出现相同状态时的对手动作进行预测。通过此方法可以在决策开始时就建立一个相对理性的对手智能体，避免将一无所知的随机对手模型作为初始模型。但该方法主要处理历史出现过的状态，难以预测没有遇到过的未来状态。

1.2　基于相似性推理的对手动作预测

动作频率的建模方式只能预测在历史中曾经出现的状态，而对于未来未知状态没有很好的泛化能力。基于推理的方法可以解决这个问题，其基本思路是首先将历史状态归结为不同的情形，记录这些情形以及智能体遇到该情形时所采取的动作。然后，构造相似性函数来评估不同情形的相似程度。最后，当遇到新的情形时，依据相似函数判定最相近的情况[13-15]并预测相应的可能动作。

基于相似性推理方法的关键问题就在于相似度函数的设计。通常，决策情形(或案例)包含多个属性的向量表示，而相似度函数定义为向量之间的某种差异运算，且需要根据被建模智能体的特征进行自动优化。STEFFENS[16]将相似度函数定义为 2 个给定决策情形的属性差异的线性加权，可以一个

足球游戏为例，即

$$\begin{aligned}\mathrm{sim}(C_1,C_2)=&\sum_{i=1}^{22}[\omega_i\times\Delta(p(i,C_1),p(i,C_2))+\\&\omega_i'\times\Delta(v(i,C_1),v(i,C_2))]+\omega_0\times\Delta(bp(C_1),bp(C_2))+\\&\omega_0'\times\Delta(bv(C_1),bv(C_2))\end{aligned}\tag{2}$$

其中，C_1 和 C_2 为比较的 2 种决策情形(或案例)；$p(i,C_j)$ 和 $v(i,C_j)$ 为玩家 i 在 C_j 情形下的位置和速度信息；$bp(C_j)$ 和 $bv(C_j)$ 为足球在 C_j 情形下的位置和速度；$\Delta(A,B)$ 为 A 到 B 之间的欧氏距离；ω_k 和 ω_k' 为位置和速度的权重，且 $\sum_{i=0}^{22}(\omega_k+\omega_k')$。权重反映了属性的相关性，是基于对手的目标和描述子目标与情形属性间的依赖关系的“目标依赖网络”，通过学习得到。实验结果表明，这种基于对手目标而自适应调整的相似性度量增加了推理系统的预测准确度。对于大规模决策问题，需要解决的另一个问题是如何高效地存储和检索这些决策情形。为此，DENZINGER 和 HAMDAN[17]提出一种基于树搜索策略的决策情形检索方法。

上述 2 种对手动作预测方法的共同点均是基于观测到的历史信息，由此可能带来的一个问题是指数级的空间复杂度。例如，假设使用的历史信息或情形的个数为 a，每个历史信息或情形的动作或值的个数为 n，则需要存储的数据规模为 a^n。因此需要寻求其他的表示方法，例如有限自动机 DFA[18]、决策树[19]、人工神经网络[20]等方式。文献[4]将多智能体强化学习问题重新表述为贝叶斯推理，推导出最大熵目标的多智能体版本，并通过最大熵目标学习对手的最优策略，为合作博弈中的对手建模提供了新的思路。DAVIES 等[21]在多智能体强化学习算法-多智能体深度确定性策略梯(multi-agent deep deterministic policy gradient，MADDPG) (图 2)中引入双向 LSTM 网络，存储并通过学习更新对手表示来模拟对手的学习过程，作者以 Keep-Away 游戏为例说明对手建模表现出更好的效果。

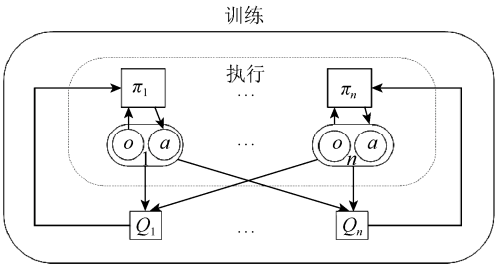

图 2 多智能体强化学习 MADDPG 算法框架

Fig. 2 Framework of a multi-agent reinforcement learning algorithm MADDPG

2 面向偏好预测的对手建模研究

对手的偏好建模是通过对对手策略的根本原因的建模实现对策略的预测与分析，主要包括对手意图、收益函数、目标计划等方面。

2.1 基于意图识别的偏好预测

对手的行为意图识别是通过观察到的对手前期动作信息，预测对手当前动作的意图进而对其完整动作进行间接预测的方式。薛方正等[22]研究足球机器人中的对手意图识别。贝叶斯网络是一种有向无环图，除了图的节点 V，弧 A，还有概率注释 P 表示节点的条件概率密度。作者将贝叶斯网络引入对手目标识别中，其中节点包括表示对手目标意图的目标节点以及对抗态势节点，弧表示目标节点间的依赖关系的目标-目标弧，每个目标节点 T_i 有一个置信度 $Bel(T_i)=\sum_{n=1}^{N}Bel(T_i^n)$，其中 T_i^n 为 T_i 的子节点；表示目标节点和态势节点间因果关系的目标-态势弧；以及由概率矩阵来表示态势节点间的推理关系的态势-态势弧。依据贝叶斯定理和检测到的新信息对目标节点的置信度进行更新。建立了由对手规划网络和对手识别网络组成的混合贝叶斯网络，即先建立对手规划图表明对手的目标分区节点及其关系，再为每个目标节点建立对手识别网络从而判断对手在每个分区内的意图，以解决多机器人对抗系统中对手意图的预测问题。通过该方法能够分析和判断对手的意图是将球踢向哪个分区，因而在足球机器人应用中相对于无对手建模的策略具得更好的表现。

基于行为意图的对手建模方法对算法实时性的要求较高，即需要在实时交互的过程中，基于已做出的部分动作信息和历史相同态势下的决策，快速判断其正在采取的行动的真实意图，这一点对于一些场景(如人机交互中机器人对于人的意图识别)等非常重要。

2.2 基于收益函数重构的偏好预测

收益函数重构的建模方法是通过估计对手的效用函数实现对手可能策略的预测。假设对手在博弈过程中总是尽量最大化其收益，基于建模出的对手收益函数模型，就可以根据收益函数最优化的原则推断对手即将采取的策略。文献[5]在博弈游戏中应用了该方法。其将对手模型定义为对手的估值函数和搜索深度。一个玩家定义为(S_{player}，S_{model})，其中 S_{player}=(f_{player}, d_{player})为该玩家的策略，包括其估值

函数 f_{player} 及其搜索深度 d_{player}，而 $S_{\text{model}}=(f_{\text{model}}, d_{\text{model}})$为对手的策略，包括对手的估值函数 f_{model} 及其搜索深度 d_{model}。以对手博弈的历史数据，即博弈局面及相应的对手决策，作为算法的输入，最终学习到对手的搜索深度和估值函数。假设估值函数是博弈状态特征的线性组合，即 $f(b)=\bar{\omega}\bullet\bar{h}(b)=\sum_i \omega_i h_i(b)$，其中 b 为博弈局面，$h_i(b)$函数为返回局面的第 i 个特征。假设局面特征是公共知识，则对手估值函数的学习归结为对手的估值函数权重的学习。本文使用爬山搜索法迭代学习各个深度 d 下的权重向量 $\bar{\omega}_d$，使得相应的策略 $(\bar{\omega}_d\bullet\bar{h},d)$ 与对手历史信息最相符，实验结果表明基于对手建模方法的玩家比其他玩家的获胜率更高。

基于收益函数重构对智能体的偏好进行预测的方法往往需要基于对手是理性的假设，并且需要对对手模型的部分信息(如收益函数的影响因素)有一定了解，同时借助大量的历史信息才能学到更准确的参数。

2.3　基于计划识别的偏好预测

计划识别是根据所观察的对手动作确定对手的最终目标及其实现该目标的计划[23]。目标对象的描述通常采用基于层次表示法的“计划库”形式[24]，从顶层的最终目标子任务到最终不可分解的动作，逐层分解和细化。计划库中除了计划分解关系，还包含各个步骤之间的时序等依赖关系以及步骤的先决环境条件信息。计划库的表示形式可以使用有向无环图。基于计划库以及一组观察到的历史动作，计划识别的方法可以生成满足计划库规则的可能计划。将对手的规划问题视为马尔可夫决策过程，通过求解最优随机策略实现对手行为序列的预测。

基于计划识别的偏好预测方法，可以预知相对较长的一段交互中对手的策略，因此更适用于长期或持续决策场景。例如，智能个性化推荐系统可以基于对用户历史目标和计划的分析，为其推荐所需要的信息[25]；一个智能的入侵检测系统可通过检测攻击者的目标和计划，持续地采取相应的防御措施[26]。但是该方法需要有设计良好的计划库，使其能够准确表达各种可能的计划，甚至非常复杂的计划。

3　面向信念建模的对手建模研究

除了环境因素，其他自主智能体的“心智状态”也会影响智能体的决策。智能体对于当前的环境以及其他智能体持有自己的信念，与此同时，其他智能体对环境也持有自己的信念，这样的信念层层嵌套就形成了无限的递归信念推理。信念建模就是通过“模拟”其他智能体的推理过程来预测其行为策略。

3.1　定量信念建模

信念建模首先在博弈论中的不完全信息博弈中得到研究[8,27]。为解决信念推理的无限递归问题，文献[5]将信念嵌套终止于一个固定的递归深度，从而实现信念的近似模拟。在递归底部的预测被传递到上层以选择该层的最佳操作，然后再传递到更高层，依此类推，直到智能体可以在递归开始时做出选择。通常假设每个智能体自认比其他智能体拥有更深的信念，且假设其他智能体是理性的，即将根据自己的信念选择最优的行动。GALLEGO 等[28]将 *k*-层的信念建模方法用在多智能体强化学习中，基于模型平均算法来更新对手信念中最有可能的模型。BAKER 等[29]利用贝叶斯对手信念来构建对手模型，并在扑克游戏中展示出贝叶斯对手模型的有效性。

3.2　定性信念建模

在形式逻辑领域，研究学者们对智能体的信念建立了定性的描述与推理方法。通常用处理不确定性的信念算子“B”来表示智能体相信。信念算子与知识算子“K”都属于认知逻辑算子。与定量信念类似，定性信念的研究涉及单层的信念，以及多层的信念嵌套[30]；同时还要考虑信念是否动态变化[31]。基于信念逻辑方法，可以借助模型检测技术实现多智能体互动环境下的策略推理与自动验证[32]。

面向信念建模的方法充分体现了智能体之间的相互建模，智能体在对对手预测的同时也将对手对于该智能体的预测考虑在内。然而由于这种层层嵌套的递归推理或计算，导致计算代价非常大，且在递归的交互推理中往往需要加入一些理性假设，因此建模准确性的保证也是一个难点。

4　面向类型建模的对手建模

4.1　基于模型选择的类型建模

上述对手建模方法都需要基于对手的交互信息从零开始学起。显然，该方式构建模型的速度比较慢。基于类型的建模是假设建模对象属于已知的几种类型之一，因而可以直接重用先前的模型。

每一个类模型可以对智能体的完整参数进行描述，以交互历史作为输入，以被建模对象的动作概率分布作为输出。对于当前的交互，在已有模型中找出最为相似的模型即可。类模型的参数可以由领域内的专家指定或根据任务领域合理假设[33]，也可以通过已有的交互学习信息或历史数据生成[19]。罗键和武鹤[34]利用交互式动态影响图(interactive dynamic influence diagrams，I-DID)对未知对手进行建模。交互式动态影响图以图形形式表达多智能体交互决策过程并在复杂的交互中对其他智能体的动作行为进行预测。图 3 为包含 2 个智能体(位于 l 层的主观智能体 i 和位于 l–1 层的智能体 j)的交互式动态影响图。其中的节点包括：①决策节点，表示智能体 i 的决策行为 A_i；②随机节点，包括客观存在的环境状态 S 以及智能体 i 的观察函数 O_i，智能体 j 动作的概率分布 A_j；③值节点 R_i 为智能体 i 的值函数；④模型节点 $m_{j,l\text{-}1}$ 为 l–1 层的智能体 j 所有可能存在的模型，即由上层智能体 i 描述的下层智能体 j 的所有模型，$m_{j,l-1}=\left\langle b_{j,l-1},\hat{\theta}_j\right\rangle$，其中 $b_{j,l-1}$ 为该模型的信度，$\hat{\theta}_j$ 为智能体 j 的框架(包括决策节点、观察节点和值节点)。除此之外，还有策略链，用随机节点 A_j 与模型节点之间的虚线表示，代表模型的解即智能体 j 的策略。一个 l 层的交互式动态影响图的求解是采用自底向上的方法，即求解智能体 i 的第 l 层的 I-DID，需要先解出所有低一层即 l–1 层的智能体 j 的模型 $m_{j,l-1}$。该工作将对手的候选模型保存在模型节点并随时间更新其信度，通过智能体 i 观察到 j 的动作及 j 的观察，基于最大效用理论，在模型空间中使用“观察-动作”序列逐步排除候选模型，最终判定对手的真实模型。该方法大大提高了建模的速度，但需要基于智能体都是理性的假设。

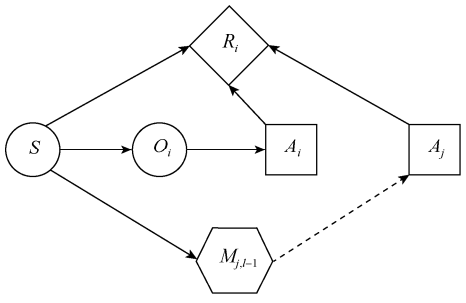

图 3 l 层上的交互式动态影响图示例

Fig. 3 An illustration of a l-level I-DID

4.2 基于特征分类的类型建模

除了建立预选类模型之外，也可以通过特征分类来预测对手的行为风格，例如是“攻击型”还是“防守型”[35]，以揭示对手在何时会采取相应的动作。该方法将有限的特征标签中的一种分配给对手，因此是典型的分类问题。在策略游戏领域，该方法得到了广泛地应用。WEBER 和 MATEAS[36]提出预测星际争霸游戏中玩家策略的方法。在网络中收集大量玩家的游戏回放，再将游戏回放转换为游戏动作日志，每个玩家 P 的行动都被编码成一个单一的特征向量 f，并使用基于专家游戏玩法分析的规则标记出特定的策略。将策略预测问题表示成分类问题，利用这些数据，使用一些经典的分类及回归算法如 C4.5 算法，k-NN 等机器学习算法检测对手的策略并估计对手执行动作的时间。SYNNAEVE 和 BESSIÈRE[37]利用收集到的相同回放数据，基于期望最大化和 k-means 算法，提出了将星际争霸玩家的开放策略从有限的策略集中分类的方法。文献[35]提出特定的分类器来预测玩家在游戏 Spring 中的游戏风格。DAVIDSON 等[38]使用深度神经网络表达对手模型，通过只对专家进行特征选择，减轻了特征选择的计算负担并在德州扑克游戏中应用该方法取得了较好的分类效果。

因此，基于类型的对手建模通过分类模型的建立，很好地避免了对对手从零学起的问题。然而，该方法的有效性依赖于分类模型的准确性和完整性，错误的或覆盖面不足的类型空间会直接导致错误的预测。

5 其他建模方式

对手建模方法大多都采用显示建模的方式，即建立一个显示的对手模型(如决策树、神经网络、贝叶斯网络)等来预测对手策略。同时，对手建模较多的是从主智能体的角度对交互中的其他智能体进行建模，因此，对个体智能体的建模是一个主流模式。

5.1 隐式建模方法

隐式建模不再是显式构建一个对手模型，而是将对手的某些特征隐含地编码到其他结构或推理过程中。专家算法[39]能够从一组给定的策略中选择最佳的专家策略，基于专家算法的隐式建模在扑克游戏中展示出良好的效果[40]。HERNANDEZ-LEAL 等[41]将对手建模为 MDP 动态转换的一部分。文献[7]采用深度神经网络 DRON 对对手的历史交互进行建模，并将对手模型隐含地嵌套在 DQN 网络中训练智能体(图 4)，该方法将强化学习中状态转移概率 $T(s,a,s')=Pr(s'|s,a)$ 加入对手因素，变为 $T(s,a,o,s')=Pr(s'|s,a,o)$，回报函数由 $R(s,a,s')$ 变

为 $R(s,a,o,s')$，其中 s 和 s' 为系统状态；a 为主智能体(所控制的智能体)的动作；o 为所有其他智能体的联合动作。同时，不同于 DQN 中的动作价值函数 $q_\pi(s,a)$，DRON 将对手视作环境的一部分，重新定义相对于对手联合策略的最优动作价值函数为 $Q^{(*|\pi^o)}=\max\limits_{\pi} Q^{\pi|\pi^o}(s,a)$，其中 π^o 为其他智能体的联合策略。DRON 模型中有一个预测状态 Q 值的策略学习模块(N_Q)和一个推断对手策略的对手学习模块(N_O)，其未有明确预测对手的属性，而是根据过去的观察学习对手的隐藏表示。模拟足球游戏和问答游戏中的表现说明，通过对对手行为的隐式预测，DRON 在稳定性和性能方面都优于未进行对手建模的 DQN。

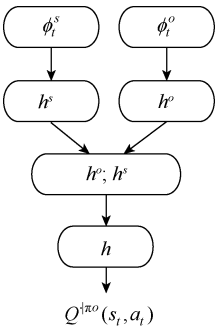

图 4　基于 DQN 的隐式对手建模网络

Fig. 4　Implicit DQN-based opponent modeling

吴天栋和石英[42]基于贝叶斯统计方法提出策略自扩展算法，改进了显式建模的效率；同时基于对手博弈行为在不同信息集中的关联性提出了子策略发现算法，提高了隐式模型的准确率。

相对于显示建模，隐式建模的优点在于合并了建模与规划这 2 个过程，可以更自然地利用二者之间的协同作用。但显式建模更易于直接观察建模结果，并且通过模型与规划的解耦，对手模型可以由不同的规划算法使用。

5.2　群体建模方法

上述的模型大都是预测单个智能体的行为，而在多智能体的环境中，智能体之间可能具有显著的相关性，对每个智能体单独建模则将由于无法捕捉到这样的关联特征而使预测效果大打折扣。联盟博弈中，联盟内部的智能体尽力地使联盟的总体收益最大化，不同联盟之间也应该将其他联盟中的成员作为一个整体来进行对手建模。FOERSTER 等[43]在深度强化学习中使智能体在更新自己策略的同时，考虑到他人的学习过程，即基于其他智能体的行动来预测其参数。将所有参与交互的其他智能体建模视为其对手，主体不是根据当前的环境参数优化策略，而是考虑到下一步对手更新策略之后的环境，并依此环境来更新策略。实验表明该方法在重复囚徒困境和投币游戏中都取得了很好的效果。AUMANN[44]提出了推广的纳什均衡概念且定义为个体行为的联合分布。将智能体分组，通过利用群体中的额外结构(如团队内成员角色信息、通信协议[45-46]等)对群体进行建模往往更加高效和准确[29]。

相对于个体对手建模，群体建模方式可以刻画群组内部成员间的策略依赖，因此在以群体形式决策的场景中有望取得更好的预测效果。同时由于不需要单独考虑每个对手和主智能体的决策交互，能一定程度上提高建模的效率。但由于群组内部具有各种相互依赖的关系，群体模型相对于个体模型而言往往更加复杂。

6　未来发展方向

对手建模理论及方法的研究已经取得了长足的进展。然而面对复杂开放环境，智能体需应对动态不确定性变化并自适应地做出相应调整，这也对对手建模提出了更高的要求。

6.1　部分可观测条件下的对手建模

目前的研究往往假设能够获取完整的环境和对手信息。然而，许多领域应用场景具有环境部分可观测的特性，智能体的交互性与自适应性更为环境增加了不确定性。智能体只能获取关于环境或其他智能体不完整或不确定的观察，这种部分可观测的特性使对手建模变得愈发困难：一方面建模可能面临信息错误或遗漏的问题；另一方面，对手的行为模式与其持有信息之间的对应关系不能明确。目前，在扩展式博弈领域，已经提出一些符号和概率方法来解决这一问题[47]。如何实现更广泛的智能交互场景中部分可观测条件下的对手建模是一个亟待解决的问题。

6.2　对手动态策略的建模

现有的建模方法中，多数方法都假设智能体在交互过程中的策略是固定的。但在自适应系统中，智能体也在不断学习与动态适应调整中，因此这种假设是不现实的，对手模型的构建需要根据对手的实时策略做出适应性变化。但对手策略可能变化的情况过多，使得动态建模变得非常困难。当前已经有一些方法针对这一问题提出了相应的解决方案，从而在不同程度上实现了动态的对手建模。例如，

文献[48]规定动作必须在极限内收敛，文献[41]要求在多种固定策略中周期性选择。然而，要真正处理变化的对手策略，构建能够动态预测对手行为的智能体仍然是面临的一个困难。

6.3 开放多智能体系统中的对手建模

目前，对于多智能体场景下智能体的建模均存在系统环境封闭的假设，即在整个交互的过程中，系统中的智能体数目是不变的。而在一些开放的多智能体系统中[49]，智能体可能随时进入或离开系统而不通知其他智能体，这种特性在网络环境当中是经常出现的[50]。当下，已经有一些针对开放多智能体系统的研究，但是为此类系统开发高效的对手建模方法，例如开放环境下合作智能体间信任与信誉模型的建立[51]，仍然是一个挑战。

6.4 可迁移的对手建模方法

基于特定的决策场景和特定的对手信息建立对手模型是比较直接的思路。然而某一场景下表现良好的方法换到其他场景或其他对手时则难以有效应用。足够智能的智能体应具备面对未知环境和对手做出自适应调整的能力。目前的对手建模大都是针对特定决策过程下的特定对手，不具有一般性。具有一定通用性的对手建模方法具有很大的研究吸引力，同时也带来更大的挑战。迁移学习可以通过重用过去的经验来改进新场景和未知对手的学习过程[19]，可能是研究此类问题的解决方案之一。

7 总 结

本文主要针对当前存在的智能体对手建模技术进行了归纳和总结，详细阐述了对手建模的不同类型。另外，本文还列举了各种建模方法中的典型算法，分析了各类技术的优缺点及技术特点。最后对对手建模技术研究中存在的难点问题及未来发展方向进行了探讨。

参考文献 (References)

[1] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[2] OSBORNE M J, RUBINSTEIN A. A course in game theory[M]. Cambridge: MIT Press, 1994: 1-368.

[3] LOWE R, WU Y I, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//The 31st International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 6382-6393.

[4] TIAN Z, WEN Y, GONG Z, et al. A regularized opponent model with maximum entropy objective[C]//The 28th International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 602-608.

[5] CARMEL D, MARKOVITCH S. Learning models of opponent's strategy game playing[C]//Proceedings of the AAAI Fall Symposium on Games: Learning and Planning. Palo Alto: AAAI Press, 1993: 140-147.

[6] LOCKETT A J, CHEN C L, MIIKKULAINEN R. Evolving explicit opponent models in game playing[C]//Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation - GECCO'07. New York: ACM Press, 2007: 2106-2113.

[7] HE H, BOYD-GRABER J, KWOK K, et al. Opponent modeling in deep reinforcement learning[C]//The 33rd International Conference on Machine Learning. New York: ACM Press, 2016: 1804-1813.

[8] HARSANYI J C. Games with incomplete information played by "Bayesian"[J]. Management Science, 1967, 14(3): 159-182.

[9] BROWN G. Iterative solution of games by fictitious play[M]. New York: John Wiley & Sons, 1951: 374-376.

[10] FUDENBERG D, LEVINE D K. The theory of learning in games[M]. Cambridge: MIT Press, 1998: 1-292.

[11] CLAUS C. The dynamics of reinforcement learning in cooperative multiagent systems[C]//The 15th National/10th Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence. Palo Alto: AAAI Press, 1998: 746-752.

[12] JENSEN S, BOLEY D, GINI M, et al. Rapid on-line temporal sequence prediction by an adaptive agent[C]//Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems-AAMAS'05. New York: ACM Press, 2005: 67-73.

[13] ALBRECHT S V, RAMAMOORTHY S. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems[EB/OL]. [2021-06-18]. https://arxiv.org/abs/1506.01170.

[14] HSIEH J L, SUN C T. Building a player strategy model by analyzing replays of real-time strategy games[C]//2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). New York: IEEE Press, 2008: 3106-3111.

[15] BORCK H, KARNEEB J, ALFORD R, et al. Case-based behavior recognition in beyond visual range air combat[C]//The 28th International Florida Artificial Intelligence Research Society Conference. Palo Alto: AAAI Press, 2015: 379-384.

[16] STEFFENS T. Adapting similarity measures to agent types in opponent modelling[C]//AAMAS'04 Workshop on Modeling Other Agents from Observations. Heidelberg: Springer, 2004: 125-128.

[17] DENZINGER J, HAMDAN J. Improving modeling of other agents using tentative stereotypes and compactification of observations[C]//Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology. New York: IEEE Press, 2004: 106-112.

[18] CARMEL D, MARKOVITCH S. Model-based learning of interaction strategies in multi-agent systems[J]. Journal of Experimental & Theoretical Artificial Intelligence, 1998, 10(3): 309-332.

[19] BARRETT S, STONE P, KRAUS S, et al. Teamwork with limited knowledge of teammates[C]//Proceedings of the 27th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2013:102-108.

[20] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[21] DAVIES I, TIAN Z, WANG J. Learning to model opponent learning (student abstract)[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(10): 13771-13772.

[22] 薛方正, 方帅, 徐心和. 多机器人对抗系统仿真中的对手建模[J]. 系统仿真学报, 2005, 17(9): 2138-2141.
XUE F Z, FANG S, XU X H. Opponent modeling in adversarial multi-robot system simulation[J]. Acta Simulata Systematica Sinica, 2005, 17(9): 2138-2141 (in Chinese).

[23] CARBERRY S. Techniques for plan recognition[J]. User Modeling and User-Adapted Interaction, 2001, 11(1-2): 31-48.

[24] OH J, F MENEGUZZI, SYCARA K P, et al. An agent architecture for prognostic reasoning assistance[C]//The 22nd International Joint Conference on Artificial Intelligence, Palo Alto: AAAI Press, 2011: 2513-2518.

[25] MCTEAR M F. User modelling for adaptive computer systems: a survey of recent developments[J]. Artificial Intelligence Review, 1993, 7(3-4): 157-184.

[26] GEIB C W, GOLDMAN R P. Plan recognition in intrusion detection systems[C]//Proceedings DARPA Information Survivability Conference and Exposition II. DISCEX'01. New York: IEEE Press, 2001: 46-55.

[27] HARSANYI J C. Bargaining in ignorance of the opponent's utility function[J]. Journal of Conflict Resolution, 1962, 6(1): 29-38.

[28] GALLEGO V, NAVEIRO R, INSUA D R, et al. Opponent aware reinforcement learning[EB/OL]. [2021-04-18]. www.researchgate.net/publication/335276661_Opponent_Aware_Reinforcement_Learning.

[29] BAKER R J S, COWLING P I, RANDALL T W G, et al. Can opponent models aid poker player evolution?[C]//2008 IEEE Symposium On Computational Intelligence and Games. New York: IEEE Press, 2008: 23-30.

[30] VAN DITMARSCH H, VAN DER HOEK W, KOOI B. Dynamic epistemic logic[M]. Heidelberg: Springer, 2008: 1-282.

[31] MARQUIS P, SCHWIND N. Lost in translation: Language independence in propositional logic - application to belief change[J]. Artificial Intelligence, 2014, 206: 1-24.

[32] VAN BENTHEM J, VAN EIJCK J, GATTINGER M, et al. Symbolic model checking for Dynamic Epistemic Logic—S5 and beyond[J]. Journal of Logic and Computation, 2018, 28(2): 367-402.

[33] ALBRECHT S V, CRANDALL J W, RAMAMOORTHY S. An empirical study on the practical impact of prior beliefs over policy types[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 1988-1994.

[34] 罗键, 武鹤. 基于交互式动态影响图的对手建模[J]. 控制与决策, 2016, 31(4): 635-639.
LUO J, WU H. Opponent modeling based on interactive dynamic influence diagram[J]. Control and Decision, 2016, 31(4): 635-639 (in Chinese).

[35] SCHADD F, BAKKES S, SPRONCK P. Opponent modeling in real-time strategy games[EB/OL]. [2021-04-18]. www.researchgate.net/publication/335276661_Opponent_Aware_Reinforcement_Learning.

[36] WEBER B G, MATEAS M. A data mining approach to strategy prediction[C]//2009 IEEE Symposium on Computational Intelligence and Games. New York: IEEE Press, 2009: 140-147.

[37] SYNNAEVE G, BESSIÈRE P. A Bayesian model for opening prediction in RTS games with application to StarCraft[C]// 2011 IEEE Conference on Computational Intelligence and Game. New York: IEEE Press, 2011: 281-288.

[38] DAVIDSON A, BILLINGS D, SCHAEFFER J, et al, Improved opponent modeling in poker[C]//Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI 2000), Palo Alto: AAAI Press, 2000: 493-499.

[39] CRANDALL J W. Towards minimizing disappointment in repeated games[J]. Journal of Artificial Intelligence Research, 2014, 49: 111-142.

[40] BARD N, JOHANSON M, BURCH N, et al. Online implicit agent modelling[C]//2013 International Conference on Autonomous Agents and Multiagent Systems. New York: ACM Press, 2013: 255-262.

[41] HERNANDEZ-LEAL P, ZHAN Y S, TAYLOR M E, et al. Efficiently detecting switches against non-stationary opponents[J]. Autonomous Agents and Multi-Agent Systems, 2017, 31(4): 767-789.

[42] 吴天栋, 石英. 不完美信息博弈中对手模型的研究[J]. 河南科技大学学报: 自然科学版, 2019, 40(1): 54-59, 7.
WU T D, SHI Y. Research on opponent modeling in imperfect information games[J]. Journal of Henan University of Science and Technology: Natural Science, 2019, 40(1): 54-59, 7 (in Chinese).

[43] FOERSTER J N, CHEN R Y, AL-SHEDIVAT M, et al. Learning with opponent-learning awareness[C]//AAMAS'18: Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems. New York: ACM Press, 2018: 122-130.

[44] AUMANN R J. Subjectivity and correlation in randomized strategies[J]. Journal of Mathematical Economics, 1974, 1(1): 67-96.

[45] STONE P, VELOSO M. Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork[J]. Artificial Intelligence, 1999, 110(2): 241-273.

[46] TAMBE M. Towards flexible teamwork[J]. Journal of Artificial Intelligence Research, 1997, 7: 83-124.

[47] PANELLA A, GMYTRASIEWICZ P. Interactive POMDPs with finite-state models of other agents[J]. Autonomous Agents and Multi-Agent Systems, 2017, 31(4): 861-904.

[48] CONITZER V, SANDHOLM T. AWESOME: a general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents[J]. Machine Learning, 2007, 67(1-2): 23-43.

[49] PINYOL I, SABATER-MIR J. Computational trust and reputation models for open multi-agent systems: a review[J]. Artificial Intelligence Review, 2013, 40(1): 1-25.

[50] VARMA V S, MORĂRESCU I C, NEŠIĆ D. Open multi-agent systems with discrete states and stochastic interactions[J]. IEEE Control Systems Letters, 2018, 2(3): 375-380.

[51] HUYNH T D, JENNINGS N R, SHADBOLT N R. An integrated trust and reputation model for open multi-agent systems[J]. Autonomous Agents and Multi-Agent Systems, 2006, 13(2): 119-154.