·“双清论坛”专题：人工智能基础理论及应用·

动态不确定条件下的人工智能

唐平中[1*] 朱 军[2] 俞 扬[3] 汤斯亮[4]

（1. 清华大学 交叉信息研究院，北京 100084；2. 清华大学 计算机科学与技术系，北京 100084；
3. 南京大学 计算机科学与技术系，南京 210023；4. 浙江大学 计算机学院，杭州 310027）

［摘 要］ 本文介绍了动态非确定条件下的人工智能最新进展，包括内部不确定性的贝叶斯人工智能、外部不确定性的非完全信息博弈、动态多回合决策以及动态开放环境决策深度强化学习等内容，并给出了今后的研究内容。

［关键词］ 贝叶斯；非完全信息博弈；深度强化学习；开放环境决策

DOI:10.16262/j.cnki.1000-8217.2018.03.006

人工智能，按照其经典的定义[1]，分为思考和行动两个方面。近年来，智能科学在这两个方面均取得了长足的进步与发展。一方面，以机器学习，计算机视觉和语言为代表的学科分支，极大的提高了机器预测与推理的准度与精度，提高了思考能力；另一方面，以搜索、博弈和强化学习为代表的学科分支，以认知与推理为输入进行建模，优化智能体的决策，从而提高了行动能力。

经典的人工智能研究，主要关注于确定条件下的决策问题，这里面经典问题如完全信息下的双人博弈，包括教科书中介绍的三子棋等；而更具挑战和具有现实意义的决策问题往往具有动态、不确定条件等特点，其需要人工智能不同方法之间相互协同，不能依赖于搜索、剪枝等单一算法解决，其给当前人工智能发展提出了更大挑战和提供了更广阔应用前景，成为当前人工智能热点前沿领域。

“动态”蕴含两个层面含义：首先该类决策问题需要智能体进行多个回合的复杂决策（如围棋）；其次，该类决策问题的环境往往随时间进行动态变化。另外，“不确定性”同样可从两个层面解读：首先在智能体内部如何表示建模知识与信息的不确定性，并对不确定性进行推理；其次如何对智能体外部，也就是对手（如多智能体系统，非完全信息博弈等）的不确定性进行建模和推理。

本文将就人工智能中动态不确定条件问题的内涵和外延进行综合性阐述，探讨该领域的科学进展和需要解决的若干重要问题。

1 内部不确定性：贝叶斯人工智能

由于物理环境的随机性、测量噪声、信息不完整等因素，如何处理不确定性是人工智能系统面临的一个基本问题[2]。在人工智能历史上，出现了很多种描述不确定性的方法，如模糊集、粗糙集、证据理论等，其中，贝叶斯方法利用概率论为不确定性建模、推断以及决策提供了严谨的数学工具，成为当前最主流的解决方案。2011 年图灵奖获得者 Judea Pearl 教授将图论与概率论相结合，提出概率图模型（如贝叶斯网络）[3]，成为人工智能领域表达和计算不确定性的基础工具。贝叶斯方法的研究包括建模、后验推断、概率编程等。

1.1 贝叶斯建模

贝叶斯建模关注如何对问题中的不确定性进行适当的概率描述和刻画。随着数据种类和规模的增加，需要更加灵活的贝叶斯模型。传统的层次化贝叶斯模型[4]或贝叶斯网络[3]一般是建立在数据独立同分布的假设上；同时，要求数据的特征表示是给定，变量之间的依赖关系是简单的。但近期深度学习的研究表明，自动学习数据的特征表示是一个提升模型表达能力和性能的有效途径。同时现实世界中的对象存在复杂相互依赖关系。

收稿日期：2018-03-16；修回日期：2018-04-08

* 通信作者，Email：kenshin@tsinghua.edu.cn

因此，需要加强以下几方面的研究：(1) 贝叶斯深度学习，有机融合贝叶斯与深度特征学习，构建灵活的建模框架；(2) 深度概率图模型，结合图论发展新型的具有特征学习能力的概率图模型；(3) 贝叶斯逻辑，将描述关系型数据的逻辑(如一阶逻辑)与不确定性推理结合；(4) 非参数化贝叶斯，自动确定模型的复杂度。

1.2　贝叶斯推断

贝叶斯推断可对模型中未知变量后验分布进行计算。经典贝叶斯方法使用贝叶斯定理，该定理已有 250 多年历史[5]。在大数据环境下，往往要用图或者逻辑表达丰富领域知识(如知识图谱)，而经典贝叶斯框架下先验分布和似然函数在利用领域知识时会受到限制。对贝叶斯模型而言，精确贝叶斯推断通常不可行，因此需要高效、高近似算法。主流的近似推断算法包括变分贝叶斯方法和蒙特卡洛方法。

因此，贝叶斯推断要加强研究以下 3 方面：(1) 对贝叶斯深度学习、贝叶斯逻辑等复杂模型有效的推断算法；(2) 处理大数据的分布式和随机梯度推断算法；(3) 扩展贝叶斯定理，灵活考虑领域知识。

1.3　概率编程

对于描述不确定性的概率模型，使用一般的编程语言(如 C，C＋＋)通常是比较复杂的，较易犯错，而且不同用户写的特定模型和算法很难被其他用户理解。概率编程(Probabilistic Programming)的目标是设计适合概率模型的建模和推断的编程语言，将通用编程与概率建模相结合。

概率编程语言需要解决通用性和性能之间的平衡。另外，随着新硬件和新型模型的出现，需要研制高效灵活的编程语言。

因此，概率编程要加强以下方面的研究：(1) 适用于深度贝叶斯学习、贝叶斯逻辑等模型的概率编程语言；(2) 通用概率编程语言的高效推断算法；(3) 面向 GPU 等新型硬件的概率编程语言。

2　外部不确定性：非完全信息博弈

针对多人的不确定性决策场景，当前主要有两个研究方向：(1) 以玩家角度为主研究，即在多人博弈场景中如何对选手不确定性进行建模推理，以做出最优决策；(2) 以博弈设计者角度为主的研究，即针对多人场景设计博弈游戏规则以最大化设计者目标。

2.1　完全信息博弈

第一个方向的典型解决方法是将博弈场景建模成静态博弈或博弈树，并对其搜索求解。对于小规模博弈树(如三子棋)可通过暴力搜索得到最优策略；对于中等规模博弈树(如国际象棋)则需引入 alpha-beta 剪枝等策略以缩小状态空间；对于大规模博弈树(如围棋)则需要事先通过监督学习[6]或自我博弈[7]方法对状态估值进行精确评估，以制定有效启发函数，进行高效启发式搜索。

2.2　非完全信息博弈

在非完全信息博弈中，博弈树建模过程中存在所谓信息集(information set)，即玩家因为并不能完全观测自己所处状态，从而无法区分多个状态。如在扑克或桥牌中，玩家无法观测对手手牌。这类问题通常指二人零和非完全信息博弈，其典型解决方案是求解该博弈树的纳什均衡。但是，此类博弈树通常规模更大，且内部节点之间关联性也导致该类博弈树不易于被剪枝，因此现有解决方法通常包括 3 个步骤：(1) 通过博弈抽象方法(game abstraction)对博弈树进行压缩，合并等价状态[8]；(2) 通过迭代算法(如基于无悔学习的 CFR 算法)对博弈树进行自我博弈，迭代求解；(3) 残局精确求解(endgame solving)。基于上述框架的扑克机器人 Libratus(CMU Tuomas Sandholm 研究组设计)现已能在双人无限注德州博弈上击败世界顶级人类玩家。

2.3　非完全信息机制设计

针对第二个方向在微观经济学和博弈论中亦称为机制设计问题，有较长的研究历史(2007 年诺贝尔经济学曾奖颁发给三位在机制设计方面有突出贡献的经济学家)。其典型应用包括在买卖交易场景下，买卖双方在非对称信息前提下，卖家如何设计价格进行销售(如在线广告等)。基于人工智能的机制设计方法是当前人工智能与经济学交叉学科研究中热点。主要研究包括通过强化学习优化机制设计[9]、通过机器学习和行为经济学对玩家进行建模、通过实验经济学对玩家建模和机制设计以进行验证和评估。

2.4　研究内容

当前非完全信息下智能决策有如下研究方向值得探索：(1) 将二人零和博弈问题扩展到二人非零和博弈或多人博弈。在多人零和博弈中，纳什均衡

等博弈解决方案依赖于对手理性程度，从而缺乏最优性保障，因此如何设计多人博弈的策略仍然是未解决的问题。(2) 将非完全信息博弈扩展到非博弈领域。在现实世界中，大多数情形并非如博弈一样有确定性规则(如金融市场，安全博弈)，其环境难以模拟，因此单纯基于自我博弈(围棋和德州扑克均依赖这一技术)方法很难扩展到现实应用场景，解决多人非确定信息环境可模拟性是将上述技术成功应用的关键。(3) 目前在机制设计中，传统解决方案是结合行为经济学模型和机器学习模型来对玩家进行建模，从而预测玩家在机制中行为。近年来，基于数据的行为建模取得了较好进展，如何将基于数据的行为建模融入机制设计框架也是当前非完全信息决策研究热点。

3 动态多回合决策：深度强化学习

强化学习思想形成于 21 世纪初[10]，其核心概念由阿尔伯塔大学 Richard S. Sutton 整理完善，其思想假设来源于心理学中行为主义，即智能体在与环境(或对手)长期交互过程中，可通过试错(trial and error)或者搜索与记忆来优化完善自身行为。

在强化学习过程中，算法需要保证在连续变化动态环境下，让智能体可获得最大预期奖励的最佳应对策略，其中一个策略由一系列连续动作组成，分别对应于在相应环境状态下智能体对环境响应。主流强化学习算法不需对状态进行预测，也不考虑行动如何影响环境，因而几乎不需要先验知识，理论上是解决复杂多变环境中自主学习有效手段。但是，强化学习算法存在复杂度会随状态—动作空间增长而指数增长缺陷，因此在 2013 年之前，高维状态—动作空间强化学习难以突破。

3.1 基于价值的强化学习

2015 年 12 月份，DeepMind 发表深度强化学习(Deep Q-Network，DQN)[11]论文，实现了稳定的深度强化学习。DQN 是一种基于价值的强化学习方法，即直接预测某个状态下所有动作的期望价值(Q 值)，然后以每次选择期望价值最大的动作作为其行动策略。在 DQN 中，采用卷积神经网络来逼近 Q 值函数，初步解决了高维状态空间的表达与计算问题。近两年来，DQN 发展迅速，出现了一系列的改进方法，如采用两个 DQN 来解耦动作选择与评估的 Double DQN 以及分离了动作优势函数的 Duelling Network 架构等。

3.2 基于策略的强化学习

基于价值的强化学习很难求取随机策略，不适宜解决连续动作问题，因此基于策略的深度强化学习在解决动态不确定条件下实际问题中具有较好应用。基于策略的深度强化学习试图学习一个策略函数而不是每个动作价值，通过不断调整策略函数的参数来寻找最佳整体方案。由于利用神经网络直接进行策略搜索代价较高且容易陷入局部最优，往往需要对搜索空间进行限制，早期 REINFORCE 方法通过设定一些任务相关规则来进行限制，同样也适用于深度强化策略学习。另一种思路是设置一些可信区域(trust region)来保证策略学习性能单调递增(如 TRPO[12]等)。

3.3 策略与价值结合强化学习

将策略与价值结合的 Actor-Critic 方法也受到广泛关注，2015 年 DDPG[13]首次将 Actor-Critic 算法与 DQN 结合，Actor 与 Critic 分别用一个神经网络来逼近，即 Actor (策略网络) 根据状态采取行动，Critic (价值网络) 给该行动打分，其中 Actor 的学习基于策略梯度，Critic 则可通过蒙特卡洛采用直接从回报(reward)中学习给出状态与行动的 Q 值。Actor-Critic 较好的平衡了策略与价值两种方法，实现了策略梯度的单步更新，这种方法在 AlphaGo 中也得到了应用。后续的 Asynchronous Advantage Actor-Critic(A3C)算法采用并行方式来同时执行多个 Actor 与环境之间的交互，并将异步交互结果汇聚到一个全局 Critic 网络中予以策略更新，该方法消除了单个智能体所学到策略之间相关性，解决了 Actor-Critic 收敛难题。

3.4 研究内容

尽管深度强化学习的理论与应用取得了长足的进步，但基于试错的学习方式并不是人类获得智能最主要的手段，人的学习过程是外部知识与强化学习的完美融合。需要在现有的强化学习过程中融入知识，因此要加强如下研究：1) 融合知识的强化学习：在不确定条件下，如能实现对环境的有效建模，将极大的提高强化学习算法的性能。目前面向环境建模的强化学习算法主要基于模型(model-based)来实现，即利用现有状态与动作来预测下一个状态，这在状态空间高维境况下较难实现，可考虑将变分自编码器(VAE)或生成对抗网络(GAN)等生成模型与 model-based

方法结合来生成下一可能状态。另一方面也可考虑将包含丰富视觉和文本信息的跨媒体知识图谱等结构化知识直接引入到深度强化学习环境中，即作为深度强化学习神经网络的输入，或在状态空间表达与动作选择的后验概率上融入逻辑规则与知识，也可作为条件约束加入到生成对抗网络中来丰富基于模型的强化学习方法。2）群智融合的强化学习：如何在任务学习过程中，更好的结合人类专家的智慧也是强化学习需要进一步探索领域。目前在这一领域主要方法为模仿学习，即在强化学习过程中利用人类专家决策轨迹来快速准确辅助策略学习，如 AlphaGo 利用人类历史棋局即是一种非常初步的模仿学习。另一种模仿学习方法被称为逆向强化学习，即根据标注样例来反推出回报函数。在不确定条件下强化学习中需接受更为自然的输入，如以自然语言方式对强化学习算法施加影响，算法也可以自然语言形式描述习得策略。3）任务可迁移强化学习：任务迁移是一种重要能力，这是在不确定或者信息不完全环境中实现快速响应关键，目前 DQN 虽然能够做到任务适应，但其算法仍不具备迁移能力。任务迁移对象既可是抽象知识，也可是获取知识方法，在这方面研究面临极大挑战。

4　动态开放环境决策

以机器学习为主要推动力的人工智能技术发展，形成了从数据中提炼有价值信息有效手段，在商业、经济、军事上显露出重要价值，例如在智能制造领域生产线监控、供应链调配、市场管理预测等。智能制造等领域所涉物理过程天然构成了开放动态环境，常表现分布偏移、类别增加、属性变动、目标多样等特点。经典机器学习理论方法主要面向预先设定封闭静态环境。在开放动态环境中，以往假设与条件不再满足而导致学习性能严重下降。

开放动态环境弱化了人工智能技术鲁棒性，近年来受到越来越多重视。国际机器学习学会创始主席 T. Dietterich 在对鲁棒人工智能研究总结和展望中指出[14]，当人工智能技术用于关键应用时，将面临对人类用户错误鲁棒、对网络攻击鲁棒、对错误目标鲁棒、对不正确模型鲁棒、对未建模现象鲁棒等挑战，而这些不鲁棒因素根源正是开放动态真实环境所致。

针对开放动态环境不同特点，近年已提出了一些应对方法。在克服分布偏移方面，通过“查询扩展”等方式获取先验偏移趋势[15]，对特定领域中的分布偏移产生了良好作用；对复杂对象的多示例表示可缓解将对象压缩在单示例所造成分布偏移。在类别增加方面，近期引入了样本生成的思想[16]，通过生成已知类别边界样本，学习鉴别已知与未知数据，从而能够识别未知类别样本；[17]则进一步通过生成新类样本，在发现一个新类样本时就可以学习识别新类别。在属性更替方面，已有研究提出共享属性和共享子空间思路，能有效将旧属性信息迁移到新属性，使得对于属性变化数据的学习变得可行。与此同时，感知与表示的研究集中在设计有效模型结构方面，学习与推理研究开始呈现融合，决策与控制研究在围棋一类简单规则封闭环境中取得突破[7]；系统性的处理开放动态环境问题的方法还尚缺乏。

为此，需要加强如下研究：对感知与表示进行研究，弥补感知渠道独立、缺乏渠道协作、数据源孤立、模型可理解性差、语义层次低等缺陷。对学习与推理进行研究，有效处理环境开放动态、样本数量巨大、标记稀缺低质等普遍特性，增强可理解性，整合知识获取与知识推理能力。对决策与控制技术进行研究，克服专家知识依赖严重、环境适应能力差、样本利用率低等不足。

5　总　结

对动态非确定条件下人工智能问题的研究，是经典人工智能的自然延伸，是人工智能理论与实践链接的关键桥梁，是当前人工智能研究的前沿热点课题。

参　考　文　献

[1] Russell SJ, Norvig P. Artificial Intelligence, A modern approach. Third Edition. Pearson. 2010.

[2] Ghahramani G. Probabilistic machine learning and artificial intelligence. Nature, 2015, 521:452—459.

[3] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann; 1 edition (September 15, 1988).

[4] Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. Boca Raton, FL: CRC press, 2014.

[5] Efron B. Bayes' Theorem in the 21st Century. Science, 2013, 340: 1177—1178.

[6] Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587): 484—489.

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

[8] Sandholm T. Solving imperfect-information games. Science, 2015, 347(6218): 122—123.

[9] Tang P. Reinforcement mechanism design. Early Carrer Highlights at Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017: 5146—5150.

[10] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Cambridge: MIT Press 1998.

[11] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529.

[12] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization. International Conference on Machine Learning. 2015: 1889—1897.

[13] Lillicrap TP, Hunt JJ, Pritzel A, et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv: 1509.02971, 2015.

[14] Dietterich T. Steps Toward Robust Artificial Intelligence. AI Magazine, 2017, 38(3): 3—24.

[15] Carpineto C, Romano G. A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR), 2012, 44(1): 1.

[16] Yu Y, Qu WY, Li N, et al. Open category classification by adversarial sample generation. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17), Melbourne, Australia, 2017: 3357—3363.

[17] Zhu Y, Ting KM, Zhou ZH. New Class Adaptation via Instance Generation in One-Pass Class Incremental Learning. Data Mining (ICDM), 2017 IEEE International Conference on. IEEE, 2017: 1207—1212.

AI under dynamics and uncertainties

Tang Pingzhong[1] Zhu Jun[2] Yu Yang[3] Tang Siliang[4]

(1. *Institute of Interdisciplinary Information Sciences, Tsinghua University, Beijing*, 100084;
2. *Department of Computer Science and Technology, Tsinghua University, Beijing*, 100084;
3. *Department of Computer Science and Technology, Nanjing University, Nanjing* 210023;
4. *School of Computer Science, Zhejiang University, Hangzhou* 310027)

Abstract This article presents an overview of AI under dynamics and uncertainties, including important topics such as Bayesian AI under uncertainty, Games under incomplete information, dynamic multi-rounds decision making, as well as decision making under dynamic open environments. For each topic above, directions for future research are described.

Key words Bayes AI; games with incomplete information; deep reinforcement learning; decision making under open environments.