知识和数据协同驱动的群体智能决策方法研究综述

蒲志强 [1, 2, 3] 易建强 [1, 2, 3] 刘 振 [1, 3] 丘腾海 [1, 3] 孙金林 [1, 4] 李非墨 [1]

摘 要 群体智能 (Collectire intelligence, CI) 系统具有广泛的应用前景. 当前的群体智能决策方法主要包括知识驱动、数据驱动两大类, 但各自存在优缺点. 本文指出, 知识与数据协同驱动将为群体智能决策提供新解法. 本文系统梳理了知识与数据协同驱动可能存在的不同方法路径, 从知识与数据的架构级协同、算法级协同两个层面对典型方法进行了分类, 同时将算法级协同方法进一步划分为算法的层次化协同和组件化协同, 前者包含神经网络树、遗传模糊树、分层强化学习等层次化方法; 后者进一步总结为知识增强的数据驱动、数据调优的知识驱动、知识与数据的互补结合等方法. 最后, 从理论发展与实际应用的需求出发, 指出了知识与数据协同驱动的群体智能决策中未来几个重要的研究方向.

关键词 群体智能, 知识与数据协同, 多智能体, 决策智能

引用格式 蒲志强, 易建强, 刘振, 丘腾海, 孙金林, 李非墨. 知识和数据协同驱动的群体智能决策方法研究综述. 自动化学报, 2022, **48**(3): 627−643

DOI 10.16383/j.aas.c210118

Knowledge-based and Data-driven Integrating Methodologies for Collective Intelligence Decision Making: A Survey

PU Zhi-Qiang[1, 2, 3] YI Jian-Qiang[1, 2, 3] LIU Zhen[1, 3] QIU Teng-Hai[1, 3] SUN Jin-Lin[1, 4] LI Fei-Mo[1]

Abstract Collective intelligence (CI) shows promising application prospects. Current research methodologies of intelligent decision making for CI systems can be categorized as knowledge-based and data-driven methods, both showing inherent advantages and disadvantages. Therefore, we claim that integrating knowledge-based and data-driven paradigms offers a new and prospective research direction. In this paper, possible methods of this integration are systematically introduced, and all of these methods are classified into a framework level and an algorithm level. Specifically, the methods integrated in the algorithm level are further categorized as hierarchical and componentized methods. In the hierarchical taxonomy, neural network tree, genetic fuzzy tree, and hierarchical reinforcement learning are included. In the componentized taxonomy, knowledge enhanced data-driven, data optimized knowledge-driven, and complementary knowledge and data driven methods are introduced. Finally, several future research priorities on the knowledge-based and data-driven integrating paradigms are proposed for the considerations of theoretical development and application requirement.

Key words Collective intelligence (CI), knowledge and data integrating, multi-agent, decision intelligence

Citation Pu Zhi-Qiang, Yi Jian-Qiang, Liu Zhen, Qiu Teng-Hai, Sun Jin-Lin, Li Fei-Mo. Knowledge-based and data-driven integrating methodologies for collective intelligence decision making: A survey. *Acta Automatica Sinica*, 2022, **48**(3): 627−643

群体智能 (Collective intelligence, CI) 起源于对群居性生物及人类社会性行为的观察研究, 因其分布性、灵活性和健壮性等优势, 为很多极具挑战的复杂性问题提供了新的解决方案, 是新一代人工智能重点发展的五大智能形态之一[1]. 进一步, 由无人机、无人车等自主无人平台组成的无人集群系统获得长足发展, 在智能交通管控、区域物流调度、机器人集群控制、复杂网络同步等领域取得了一系列研究和应用成果[1−11]. 特别是在军事智能领域, 群体智能已被认为是有可能带来颠覆性变革的新技术, 国内外纷纷部署相关研究项目, 如美国的 “进攻性蜂群使能战术” (Offensive swarm-enabled tactics, OFFSET) 项目、“拒止环境中的协同作战” (Collabo-

收稿日期 2021-02-04 录用日期 2021-06-18
Manuscript received February 4, 2021; accepted June 18, 2021
科技创新 2030 “新一代人工智能”重大项目 (2020AAA0103404), 国家自然科学基金 (62073323) 资助
Supported by National Key Research and Development Program of China (2020AAA0103404) and National Natural Science Foundation of China (62073323)
本文责任编委 穆朝絮
Recommended by Associate Editor MU Chao-Xu
1. 中国科学院自动化研究所综合信息系统研究中心 北京 100190 2. 中国科学院大学人工智能学院 北京 100049 3. 泰州智能制造研究院 泰州 225321 4. 江苏大学电气信息工程学院 镇江 212013
1. Integrated Information System Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049 3. Taizhou Institute of Intelligent Manufacturing, Taizhou 225321 4. School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013

rative operations in denied environment, CODE) 项目, 印度 2019 年发布的首个无人机集群概念项目 “空射弹性资产群” (Air-launched flexible asset-swarm, ALFA-S), 国内中国电子科技集团、北航、国防科大等开展的无人机集群试飞项目等[12].

尽管群体智能已成为当前发展热点, 但现今并没有关于这一概念的统一定义[6−7]. 不同学者从生物群体智能[13]、人群智能[1]、多智能体系统[9]、复杂网络[14−15]、演化博弈论[16] 等截然不同的学科视角出发展开研究, 从不同侧面取得了丰富的研究成果. 本文统一称其为 “群体智能”, 并选择其对应英文为 Collective intelligence. 一方面因为在我国新一代人工智能中, 群体智能已显性地成为一种智能形态, 此时已有必要将不同学科下的概念加以融合; 另一方面 CI 在英文文献中的内涵也更为广泛[1- 6], 能相对更好地与 “群体智能”这一概念相对应. 特别地, 本文将融合控制论等学科进展, 较多着墨于由无人系统这类物理平台组成的群体系统. 因此, 本文在谈及统一性概念时采用 “群体智能”, 而在具体问题中则可能结合上下文称这样的系统为 “集群系统” “多智能体系统”等.

当前群体智能决策主要基于两大类方法: 知识驱动和数据驱动. 知识驱动方法[17] 可充分利用已有知识, 包括已有模型与算法知识、规则经验知识以及特定领域知识. 知识的广泛内涵便于实现多学科知识的灵活集成; 同时, 许多基于模型的知识驱动方法具有完备的理论支撑体系, 在分析算法稳定性、最优性、收敛性等方面具有天然优势; 此外, 知识驱动模型具有更好的可解释性; 而知识作为一种数据和信息高度凝练的体现, 也往往意味着更高效的算法执行效率. 但在实际应用中, 特别是大规模群体协同等复杂问题中, 群智激发汇聚的知识机理尚不完全清晰, 知识获取的代价高昂, 同时现有知识难以实现复杂群体行为庞大解空间的完备覆盖, 也难以支持集群行为的持续学习与进化. 近年来广泛兴起的深度强化学习等数据驱动方法[18] 具有无需精确建模、能实现解空间的大范围覆盖和探索、从数据中持续学习和进化、算法通用性强等特点, 同时具有海量开源模型和算法库等工具支撑. 然而, 这类方法在理论特性分析上往往存在困难, 其典型的 “黑箱”特性也带来了可解释性差等问题; 同时, 其高度依赖高质量的大数据, 而在群体智能应用中, 这类数据本身较难获取; 此外, 随着群体规模和问题复杂度的提升, 解空间维度灾难问题为学习效率带来了严峻挑战; 而其依赖庞大算力的特点也使得个人或一般性机构在开展研究时面临严重瓶颈. 知识驱动与数据驱动方法的主要优缺点总结如图 1 所示.

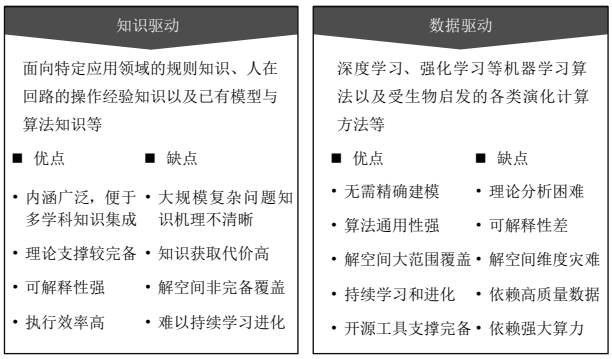


图 1 知识驱动和数据驱动各自优缺点

Fig. 1 Advantages and disadvantages of knowledge-based and data-driven methodologies

基于上述分析, 将知识驱动和数据驱动两大类方法相结合, 利用各自优势, 形成知识与数据协同驱动的新方法路径, 有望为群体智能系统研究和应用提供更为广阔的空间. 这类方法尽管在近年来逐步受到关注[19−23], 但尚未形成体系. 为此, 本文首先对知识驱动和数据驱动概念进行定性界定, 在此基础上系统梳理了知识与数据协同驱动可能存在的不同方法路径, 主要从知识与数据的架构级协同、算法级协同两个不同层面进行了方法归类, 总体框架如图 2 所示. 在架构级协同层面, 从个体架构、群体架构两方面介绍常见架构体系, 为复杂群体协同问题提供总体解决框架; 在算法级协同层面, 进一步划分为算法的层次化协同、组件化协同, 并在每类协同方法中具体选取了若干代表性方法进行介绍. 这里, 架构级协同和算法级协同间的区别和关联在于, 前者为复杂问题的解决搭建了基础框架, 这为各类知识驱动、数据驱动以及知识与数据协同驱动的算法提供了 “容器”, 体现为不同算法间的逻辑关系; 而算法级协同则主要探讨具体算法内部如何协同运用知识与数据的相关要素, 体现为某类算法内的逻辑关系. 在对上述两大类协同方法进行详细介绍后, 本文最后从群体智能理论进一步深化、应用进一步落地等实际需求出发, 指出了知识与数据协同驱动的群体智能决策中未来几个重要的研究方向. 值得说明的是, 由于知识与数据驱动的外延极其广泛, 学科交叉特点十分明显, 本文难以覆盖所有方法, 但致力于系统地为知识与数据协同驱动这类极具潜力的方法开启讨论, 并为当前群体智能以及机器学习两大热点领域各自及其交叉领域的研究提供必要借鉴.

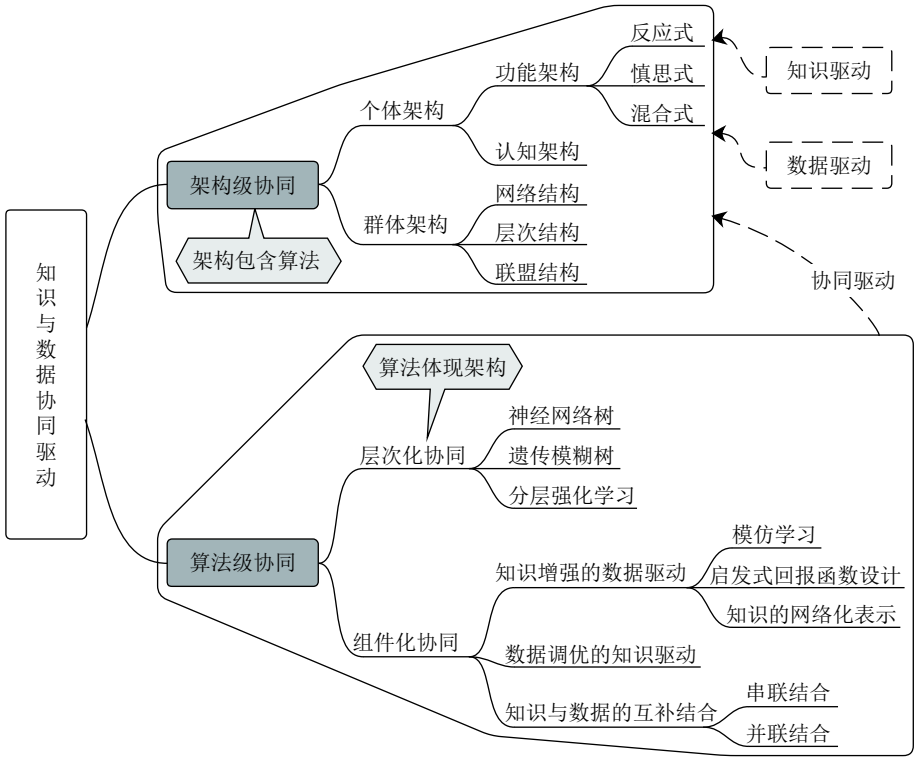


图 2　知识与数据协同驱动总体框架

Fig. 2　Overall framework of knowledge-based and data-driven methods integration

1　知识和数据驱动的概念界定

本质上来说, 任何人为设计的方法均包含 “知识”, 例如所有神经网络模型中网络结构、激活函数、超参数的选取都体现了人的经验或先验知识, 但学术界显然默认神经网络属于数据驱动方法. 从这个意义来说, 所有数据驱动方法都体现了知识和数据协同的理念. 但这样的理解却使问题变得过于 “平凡”, 失去了对方法设计的指导意义. 本文所述知识与数据协同, 体现了一种更有针对性的 “显式”协同. 以下将首先对知识驱动及数据驱动方法进行适当界定, 并简要介绍各自发展的总体情况. 值得一提的是, 这种界定本身仍停留在定性列举而非严格的概念定义层面.

1.1　知识驱动概念界定及简介

本文所述 “知识”包括一系列基于数学/物理模型的算法知识、规则经验知识以及面向特定应用的领域知识. 知识驱动是许多实际群体智能系统的主要研究路径, 在无人集群任务规划、博弈决策、协同控制等方方面面具有广泛的应用基础.

一是数学/物理模型知识. 以群体动力学模型为例, 典型的模型知识包括 Reynold 模型[24]、Vicsek 模型[25]、Couzin 模型[26]、Cucker-Smale 模型[27]等, 这为群体中的个体微观运动提供了动力学基础. 二是基于模型的算法知识, 包括各类基于模型推导的路径规划算法[28], 任务分配算法[29−30], 基于一阶[31−32]、二阶[33−34]、高阶[35−36] 模型的一致性控制算法等, 这类方法从解析的群体数学/物理模型出发, 基于解析求导的优化理论以及 Lyapunov 等稳定性理论实现群体问题求解. 三是规则经验知识, 包括由人们对于集群基础行为的认知构建起的集群简单行为规则, 如各类基于模糊理论[37]、知识系统[38] 构建起的规则推理方法等. 四是面向特定应用场景的领域知识, 这是群体智能系统走向实用化的重要支撑, 例如在兵棋推演系统[39] 中构建的各类实体要素模型和裁决规则知识, 这类知识为群体学习进化提出了新的约束条件, 但同时也对问题求解空间进行了极大约简.

以上基于机理模型、先验知识或规则的知识驱动方法在确定、简单、低维的单体或群体系统中表现出良好的性能, 但现实中群体系统往往难以建模, 且缺乏领域知识, 同时当集群规模扩大, 特别是集群表现出高维、复杂、强不确定性的行为特征时, 已有的模型或规则经验知识难以覆盖整个解空间, 知识驱动方法的适用性、稳定性、鲁棒性将大大降低.

1.2 数据驱动概念界定及简介

蚁群算法、粒子群优化算法以及直接对无人集群系统行为具有重要借鉴意义的狼群算法、鸽群算法等生物启发式进化计算方法在群体智能系统中具有广泛的应用[13, 40−41]. 囿于数据驱动方法广泛的外延, 本文所述 “数据驱动方法”侧重于深度学习、强化学习等近些年广泛兴起的机器学习算法, 但在某些方法分类中附带包括上述进化计算方法.

深度学习具有高维数据的 “感知”能力, 强化学习具有在与环境交互中的 “决策”能力, 因此这两种方法天然具有与大规模群体智能系统应用结合的优势, 特别是两种算法结合形成的深度强化学习 (Deep reinforcement learning, DRL) 方法. 文献 [42] 和文献 [43] 分别对深度学习和强化学习进行了综述, 而 DeepMind 团队的系列成果则为深度强化学习的研究树立起里程碑, 代表性成果为三篇发表在 *Nature*上的文章, 分别介绍了在 Atari 游戏[44]、围棋程序 AlphaGo[45] 及其进阶版 AlphaGo Zero[46] 上的应用. 针对多智能体问题, 文献 [4−5, 47−48] 系统介绍了强化学习在多智能体系统中的应用. 针对非完全信息、大规模组合空间博弈问题, DeepMind 采用模仿学习、强化学习、多智能体学习等组合方法, 训练的 AlphaStar[49] 能战胜 99.8% 的专业人类玩家, 但其 “多智能体”属性主要体现在由不同策略构成策略池从而进行联盟学习, 具体到每个策略, 仍是将所有操作算子看作一个整体进行单智能体学习. OpenAI团队提出一种多智能体深度确定性策略梯度 (Multi-agent deep deterministic policy gradient, MADDPG) 算法, 通过集中评判-分散执行方式使智能体具有自主决策能力, 在动态环境中实现智能协同合作与对抗[50], 但其端到端的学习架构在复杂问题中面临挑战. 此外, OpenAI 针对 DOTA 2 开展的多智能体研究也取得了不错的成果, 其开发的人工智能系统 OpenAI Five 于 2019 年 4 月击败 DOTA 2 人类冠军, 核心技术特点是针对 Open AI Five 这类具有上亿参数量的大规模决策系统, 设计了一种新颖的 “手术” (Surgery) 训练机制, 从而能够在模型和环境不断变化的情况下对智能体进行持续训练, 而无需从头训练获取参数, 降低了新模型设计验证的成本[51].

综上所述, 尽管 DRL 等数据驱动方法在单智能体及多智能体系统中取得了一定的研究成果, 但面对非完全信息、复杂物理约束等实际问题, 如何结合先验知识与算法模型, 提高算法效率、降低算力要求, 亟待进一步深入研究.

2 知识和数据的架构级协同

从数据驱动的角度来看, 当前一类主流的方法是端到端的机器学习算法, 即输入原始状态信息, 经黑箱模型后直接输出所需要的结果, 如感知模型中物体识别的类别、决策模型中智能体的行为动作等. 然而, 对于复杂系统和复杂任务而言, 特别是无人集群系统所面临的复杂任务, 端到端的学习模型难以奏效, 此时一个合理的智能体任务体系架构便显得尤为重要. 对群体智能系统体系架构的研究, 至少源于两方面的需求, 一是描述不同复杂任务中的通用机理和逻辑流程, 有助于挖掘问题内在的不变性机理并进行标准化建模; 二是将复杂问题分解为若干较易解决的子问题, 极大降低问题处理的复杂度. 体系架构为复杂大规模问题求解搭建起基本框架, 在此基础上, 针对架构中的不同逻辑模块 (子成员、子任务、子系统等), 确定是采用知识驱动、数据驱动还是知识与数据协同驱动等具体算法. 因此, 体系架构充当了算法容器的功能, 使得不同驱动方式的算法形成有机协同, 即实现架构级协同.

体系架构研究的内涵十分广泛, 且存在截然不同的问题研究角度和方法路径. 针对本文所讨论的群体智能系统, 大致可从两方面剖析其体系架构问题: 一是个体的体系架构, 研究个体如何自主决策; 二是群体的体系架构, 研究群体如何协同决策.

2.1 常见个体与群体体系架构

若将每个个体看作一个智能体 (Agent), 则从 Agent 建模角度来看, 个体的体系架构大致可分为 3 类: 反应式体系架构、慎思式体系架构和混合式体系架构[52]. 反应式体系架构模拟了动物反应式行为的特点, 包含多个能独立输入输出的模块, 每个模块采用反应式的 “感知—动作”结构, 对输入信息进行反应式的动作, Brooks[53] 提出的包容式体系结构便是典型的反应式体系架构, 而多智能体控制方法中基于行为的控制方法[54] 也体现了这一特点. 纯反应式架构的缺陷在于, Agent 仅基于局部信息做决策, 在大规模系统中, 这种相对 “近视”的决策机制可能难以获得理想结果. 慎思式体系架构则将对输入信息进行逻辑推理, 典型的例子为著名的信念–意图–期望 (Believe-desire-intension, BDI) 模型[55], 智能体基于所建立的信念库、意图库、期望库, 按照一定的逻辑推理规则进行推理决策. 慎思式架构的缺陷在于, 其推理过程往往较复杂, 难以很好地适应实时性要求很高的环境. 混合式体系架构兼具了反应式架构对环境的快速反应和慎思式架构的逻辑

推理特点, 采用层次化体系结构, 对于群体系统往往包含 3 层, 自上而下分别为合作层、推理层和反应层[52], 合作层处理智能体间的合作任务, 推理层完成智能体内部的慎思式推理, 反应层执行环境刺激的反应式行为和上层下达的行为指令. 混合式架构对于群体智能系统这类复杂系统具有较好的适用性. 此外, 上述 3 类体系架构主要侧重于应用导向的系统功能实现, 另一种体系架构研究思路是从认知科学出发, 致力于刻画自然或人工智能体认知、发育过程中的认知机理, 并基于此实现人类认知水平的智能行为, 著名的认知架构模型包括 "状态、算子与结果" (State, operator, and result, SOAR) 模型、基于理性思维的自适应控制 (Adaptive control of thought-rational, ACT-R) 模型等[56].

群体体系架构刻画存在于各智能体中的通讯和控制模式, 体现了集群中个体间的信息共享、存储和协作方式, 对群体系统的一致性、自主性、涌现性等特性具有直接影响[57]. 从群体中智能体的组织方式和通信、控制模式来看, 群体架构大致可分为网络结构、层次结构、联盟结构三类[52]. 网络结构中, 每个智能体的地位均等, 符合条件的智能体间均能进行信息交互, 最大限度体现了群体系统的自组织特性; 层次结构中, 智能体分为不同层次, 每层的决策和控制权来自于其上层的指令输出, 分层架构体现了问题的逐级抽象特点, 便于复杂任务的层次化分解; 联盟结构中, 智能体根据一定规则划分为不同联盟, 联盟内和联盟间分别采用不同的信息交互机制形成群体协同, 这种结构体现了一定的功能异构性.

上述个体和群体结构为复杂系统架构建模提供了基本思想和模型要素, 面向不同应用领域, 则将基于上述基础模型进行进一步设计. 以无人集群系统最为典型的应用领域 —— 军事指挥控制领域为例, 这是一个典型的多要素、巨复杂场景, 其智能指挥控制过程难以采用单一的端到端模型加以刻画, 体系架构设计便显得尤为重要. 面向多无人机任务规划等任务, 洛克希德·马丁公司提出了多态认知智能体架构 (Polymorphic cognitive agent architecture, PCCA)[58], 其核心是包含一个认知层, 并进一步自上而下分解为宏观 (Macro)、微观 (Micro)、原子 (Proto) 三层认知架构, 宏观认知层采用基于 SOAR 的知识推理模型, 微观认知层采用基于 ACT-R 的专家推理模型, 原子认知层采用基于群智分布式自组织方式实现. 面向无人机/车异构集群城市作战任务, 美国国防部高级研究计划局 (DARPA) 开展的 OFFSET 项目[59], 将复杂任务自上而下分解为集群任务层 (Swarm mission)、集群战术层 (Swarm tactics)、集群原子操作层 (Swarm primitives)、集群算法层 (Swarm algorithm), 任务层刻画宏观任务需求, 战术层描述完成任务所需的战术序列, 原子操作层表征完成某战术所需具体执行的行为, 算法层则代表为实现具体行为所需的各项技能, 每一层又进一步划分为不同功能模块, 是一个典型的层次化体系架构. 更一般地, 观察–判断–决策–执行 (Observe-orient-decide-act, OODA) 循环理论已被普遍接受为描述指挥决策过程的通用模型框架[60], 其将作战过程分解为由观察、判断、决策、执行四个环节串联形成的决策环, 并可作为一般性模型拓展到多智能体仿真[61]、应急响应[62] 等应用领域中.

2.2　知识与数据架构级协同概念模型

从知识和数据协同驱动的角度来说, 上述一般性个体架构模型、群体架构模型以及作为示例的军事指挥控制架构模型从三方面体现了知识和数据协同的特点: 一方面, 这类组织架构本身便体现了先验知识的运用, 是一类高度抽象的内嵌知识; 另一方面, 将复杂问题分解为若干子问题, 往往表现为不同问题求解子模块, 针对每个子模块, 可以进一步确定是采用数据驱动方法还是知识驱动方法加以求解, 进而便于对各类基于知识或数据驱动的方法进行灵活集成; 此外, 从数据驱动来看, 增强了数据驱动模型的可解释性, 并使数据驱动模型带来的不确定性被限定在某个子模块内.

以 OODA 循环为例, 结合 OFFSET 等采用的层次化、模块化思想, 我们可将复杂的群体决策问题描述为如图 3 所示的概念架构模型. 该模型将从原始状态输入到最终行为输出间的决策控制过程分为观察、判断、决策、执行四层, 每一层根据需要进一步分解为不同颗粒度的子模块, 知识和数据协同驱动的思想则渗透到所有层次子模块中, 即可根据每个子模块的功能特点、问题复杂度灵活选择是采用知识驱动方法 (浅灰色圆角矩形) 还是数据驱动方法 (深灰色矩形), 并进一步研究具体采用哪一种知识驱动方法, 如基于模型的解析算法 (Algorithm) 或启发式的经验知识 (Heuristic) 等, 或哪一种数据驱动方法, 如深度学习中的卷积神经网络 (Convolutional neural network, CNN) 模型、强化学习中的近端策略优化 (Proximal policy optimization, PPO) 算法、多智能体强化学习中的 MADDPG 算法等. 特别地, 涌现 (Emergence) 作为我们对群体系统重要的期待特征, 当前存在大规模复杂系统涌现机理不清晰、复杂任务涌现规则难以设计

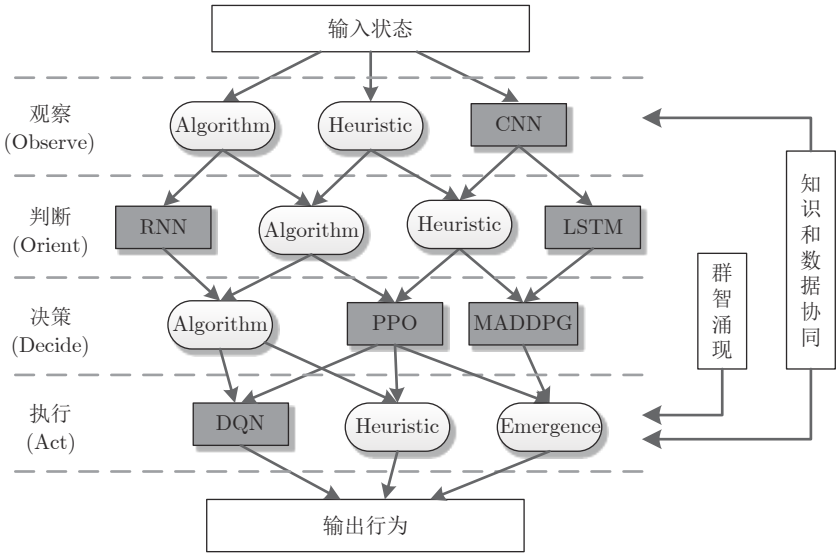


图 3 知识和数据架构级协同概念模型

Fig. 3 Conceptual model for framework-level integration of knowledge-based and data-driven methods

等问题. 为此, 结合层次化分解思想, 我们可将群智涌现行为局限在较低层次的执行层, 而非具有更高复杂度和问题抽象度的判断、决策层, 便于自组织、涌现方法在实际系统中的集成应用, 这种思想与洛克希德·马丁 PCCA 模型中的原子层设计类似.

3 知识和数据的算法级协同

前述个体或群体体系架构主要针对复杂系统、综合任务, 如图 3 所示的概念架构往往包含多种算法, 并在不同层次、不同功能模块间体现出知识与数据的协同. 与此对应, 许多算法本身便体现了知识与数据协同驱动的特点, 由此形成 “算法级”的知识和数据协同路径, 在此就几类代表性算法进行综述, 并根据算法的主要特点, 进一步分为层次化协同算法、组件化协同算法两类. 层次化协同算法与架构级协同思路类似, 算法本身体现了一种分层思想, 所不同的是, 这种分层思想被包含在一个具体的算法内部, 可以直观地理解为 “算法包含架构”, 而非架构级协同那样是 “架构包含算法”; 组件化协同则代表了其他一大类非层次化协同的方法, 我们将探讨更为 “精细”的知识与数据协同路径, 即协同不仅仅体现在分层这种单一思想上, 而是将知识驱动或数据驱动部分看作另一方的某一个算法组件, 二者紧密结合形成一个完整算法.

3.1 层次化协同算法

3.1.1 神经网络树

神经网络树是一种典型的知识与数据协同驱动模型, 其中神经网络模型代表数据驱动, 决策树结构则代表了知识驱动, 其实质是将若干神经网络模型以决策树的结构有效组织起来, 使之兼具决策树模型可解释性强、易于集成专家知识以及神经网络模型自主学习的优点. 神经网络树的研究已有数十年历史, 研究者很早便意识到将符号主义的决策树模型与联结主义的神经网络模型结合起来的优势[63], 并提出了多种结合方式, 如首先设计一个决策树, 再从中生成层次化神经网络模型[64], 或反过来从已训练好的神经网络中提取决策规则[65].

针对多机器人协同环境建模场景中的机器人异常行为检测问题, 文献 [66] 提出采用 Siamese 神经网络 (Siamese neural network, SNN)[67] 来计算两个环境信息向量 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 间的距离, 从而实现机器人异常行为的检测, 考虑到机器人群体采集到的环境信息维数十分庞大, 作者进一步将由 T 个机器人采集到的环境信息分为 T 个子向量, 并将原始的 SNN 设计为一个层次化网络结构, 由此简化了 SNN 网络的训练过程. 机器人自主导航往往包含目标搜索、避碰避障等多种任务, 各任务间的协调成为自主导航的关键, 为此, 文献 [68] 针对自主导航中的多种子任务分别设计神经网络控制器, 进一步设计一个基于神经网络的协调器来调整子任务控制器的输出权重, 子网络及协调网络间构成一个层次化体系结构. 近年来, 随着深度学习技术的兴起, 产生了基于各种深度神经网络 (Deep neural network, DNN) 的树模型. 文献 [69] 提出一种具有增量学习特点的深度神经网络树模型, 对于已经训练好的 DNN 模型, 当新数据来临后, 模型能以一种树状结构继续

层次化生长, 以学习新数据中的模式, 同时保留先前所学习到的知识, 以避免网络产生灾难性遗忘问题. 文献 [70] 提出一种层次化卷积神经网络, 用以提升分类问题结果准确率, 其核心是确定一个合理的卷积神经网络层次化结构, 为此作者采用层次化聚类方法构建一个可视化的树结构, 并定义了一个层次化聚类有效性指数来指导树结构的自动学习. 更多关于神经网络树的最新研究可参考[71−73].

3.1.2　遗传模糊树

遗传模糊树 (Genetic fuzzy tree, GFT) 除了具有像神经网络树这样的树结构外, 还代表了模糊推理这种典型知识驱动模型和遗传算法这类数据驱动模型相结合的算法, 其中模糊逻辑基于专家知识建立起推理框架, 遗传算法用以实现模糊推理中前后件规则参数的优化, 而树结构则进一步表征复杂问题中的层次化体系架构. 推而广之, 这里的模糊系统可替换为专家系统等符号逻辑系统, 遗传算法可替换为其他启发式优化算法或神经网络等数据驱动模型, 因此 GFT 具有较强代表性.

GFT 的典型应用主要体现在空战博弈对抗系统上. 针对复杂的空战博弈过程, 文献 [74] 详细阐述了 GFT 构建博弈智能体的优势. 进一步, 文献 [75] 针对多无人战斗机在复杂环境中的战术协同和行为决策问题, 利用 GFT 方法进行战术决策, 并在著名的 ALPHA 智能空战系统中, 实现了在高保真模拟环境中的无人作战飞行器空战任务. 针对多兵种异构作战问题, 文献 [76] 设计了多个级联模糊系统和遗传算法进行战术决策和优化. 这项研究中提出的 GFT, 创建了对不确定性因素具有恢复能力和自适应特性的控制器. 最终无人战斗机小组实现了在面对来自空中拦截器、地空导弹站点和电子战站点等不确定性威胁的情况下, 利用敌武器空隙穿越作战空间并成功摧毁目标的任务.

然而, 上述方法在构建模糊规则时仍需大量专业知识, 特别是当智能体数量增加时, 输入参数的增加将导致模糊规则数量指数增加. 为此, 文献 [77] 提出一种基于单一输入规则群 (Single input rule modules, SIRMs) 动态连接模糊推理模型和改进自适应遗传算法的多无人战斗机空战博弈战术决策方法. 该方法改进了传统的模糊推理方法, 基于 SIRM 模型将所有输入变量解耦, 解耦后的各模糊推理模块再通过动态权重将结果进行合并, 得到推理决策动作, 这种解耦方法解决了传统模糊规则数量随输入变量数呈指数级增长的规则爆炸问题; 同时遗传算法的优化作用使得只需建立粗略的规则框架, 而无需精确的交战规则, 大大降低了规则设计的难度.

3.1.3　分层强化学习

深度强化学习成为引领当前人工智能特别是决策智能技术发展的核心要素. 然而, 在大规模复杂问题中, 特别是在具有大量智能体的群体合作/对抗类问题中, 状态空间和动作空间指数增长带来的维数灾难问题仍然是当前强化学习面临的一大重要挑战. 分层强化学习 (Hierarchical reinforcement learning, HRL) 采用策略分层、分而治之的思想, 为解决复杂大规模问题提供了有效手段, 其本质是针对马尔科夫决策过程 (Markov decision process, MDP) 中假设每个动作都只在单个时间步内完成的问题, 采用不同的时间抽象方法将若干原子动作封装为一个个扩展动作序列 (Extended courses of action, ECA), 每个 ECA 可能包含多个时间步, 从而把微观的原子动作扩展为颗粒度更大的动作, 这样极大压缩了动作空间[78], 其理论依据则主要是半马尔科夫决策过程 (Semi-Markov decision process, SMDP)[79] 的求解理论. MDP 与 SMDP 的原理概念化对比如图 4 所示.

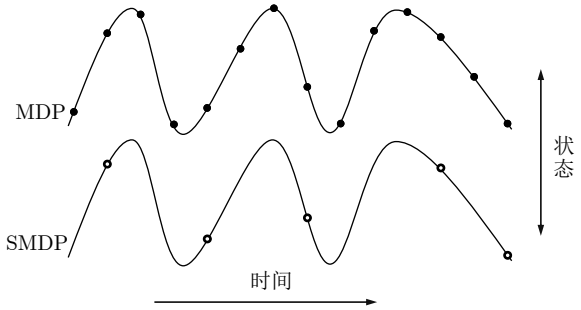


图 4　MDP 与 SMDP 比较

Fig. 4　Comparison between MDP and SMDP

最早在强化学习中提出多层次任务划分的代表性工作是 Dayan 等[80] 提出的封建强化学习 (Feudal reinforcement learning, FRL). 正如其名所示, FRL 将复杂任务在时空上分层, 当前层为 Manager, 其上层为 Super-manager, 下层为 Sub-manager, 当前层的学习目标是满足上层的任务, 并向下层下达指令, 非相邻层之间实行奖励隐藏 (Reward hiding) 和信息隐藏 (Information hiding), 实现任务解耦. 除此之外, 经典的分层强化学习还包括 Sutton 等[81] 提出的基于选项 (Option) 的强化学习、Parr 等[82] 提出的基于分层抽象机 (Hierarchies of abstract machine, HAM) 的强化学习、Dietterich[83] 提出的基于值函数分解的 MaxQ (MaxQ value function decomposition) 强化学习方法等. Option 方法定义了一系列由原子动作封装而成的 “选项”, 相对于原子动作, 选项也可看作是一种 “宏观

动作”、“抽象动作”、“子控制器”, 例如对于在多个房间内游走的移动机器人, 可以定义 “前”、“后”、“左”、“右”这样的原子动作, 也可定义 “移动到门口”这样的选项, 机器人将在原子动作和选项中进行动作选择. HAM 方法将任务定义为一个随机有限状态机, 采用 MDP 对状态机进行建模, 实现智能体在某个状态机内部的学习以及状态机间的切换调用. MaxQ 方法将一个 MDP 过程 M 分解为子任务集 $\{M_0, M_1, \cdots, M_n\}$, 对应的策略 π 也分解为策略集 $\{\pi_0, \pi_1, \cdots, \pi_n\}$, 所有子任务形成以 M_0 为根节点的分层结构, 每个子任务的动作选择既可以是原子动作, 也可以是其他子任务, 最终解决了 M_0, 即解决了完整任务.

近年来, 将分层强化学习思想应用于多智能体强化学习, 所产生的多智能体分层强化学习已成为研究热点. DeepMind 提出了一种多智能体强化学习方法, 核心是采用基于种群的训练、单个智能体内部奖励优化以及分层强化学习架构, 其在 “雷神之锤”游戏中不仅学会了如何夺旗, 还学到了一些不同于人类玩家的团队协作策略[84]. 文献 [85] 介绍了一种具有技能发现能力的双层多智能体强化学习方法: 在底层, 智能体基于独立的 Q-learning 学得特定技能; 在上层, 基于外部团队协作奖励信号并采用集中式训练方式实现多智能体间的协作. 文献 [86] 则使用多智能体分层强化学习来处理稀疏和延迟奖励问题, 作者同时研究了多种同步/异步 HRL 方法, 并提出了一种新的经验回放机制来处理多智能体学习中的非平稳性等问题. 此外, HRL 在多智能体路径规划[87]、多卫星协同任务规划[88] 等应用问题中也展现了良好的求解能力.

显然, 分层强化学习引入了大量的先验或领域知识, 如 Option 方法中如何将原子动作封装为选项并确定选项的进入、退出条件, HAM 方法中如何设计随机状态机, MaxQ 方法中如何构建子任务层次结构等. 尽管基于智能体自动任务抽象的端到端分层强化学习成为当前另一研究热点, 并出现了 Option-Critic[89]、Manager-Worker[90] 等端到端学习方法, 但在大规模复杂问题中, 特别是对系统可靠性、可解释性有着苛刻要求的物理智能体领域, 结合先验和领域知识的分层强化学习方法仍是一个有效的选择.

3.2 组件化协同算法

根据知识驱动、数据驱动方法各自所处的主次地位, 我们可大致将组件化协同算法分为知识增强的数据驱动方法、数据调优的知识驱动方法、知识和数据互补结合三类方法. 其中, 知识增强的数据驱动方法以数据驱动方法构成算法的主体框架, 算法的部分组件或某个操作步骤采用现有知识加以辅助或增强设计, 目的是相较纯数据驱动方法获得性能提升; 数据调优的知识驱动方法则以知识驱动方法构成算法主体框架, 同样算法的部分组件或某些操作步骤采用数据驱动方法、特别是数据驱动强大的寻优能力来实现相对于纯知识驱动方法的性能改善; 在知识和数据互补结合方法中, 知识驱动、数据驱动两类方法的主次关系相对不明显, 二者将以互补方式构成集成算法.

3.2.1 知识增强的数据驱动

如图 5 所示, 在此主要介绍强化学习中的模仿学习、启发式回报函数设计以及深度学习中的网络化知识表示三种知识增强的数据驱动方法, 每种方法的不同组件将基于先验知识进行辅助增强设计, 如直接模仿学习中的行为策略、逆强化学习及启发式回报函数设计方法中的回报函数, 以及网络化知识表示中的网络结构、参数和学习策略等.

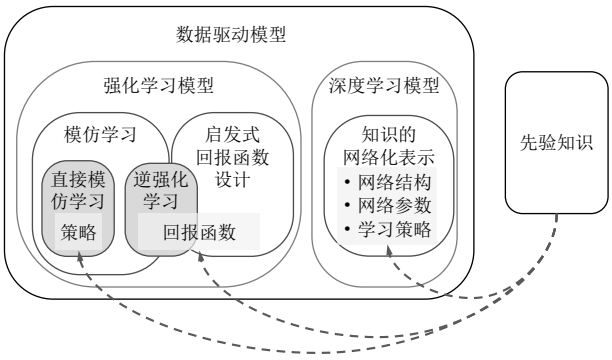


图 5 知识增强的数据驱动方法

Fig. 5 Knowledge enhanced data-driven methods

1) 模仿学习

多智能体强化学习中搜索状态空间和策略空间巨大, 且由于稀疏奖励、延迟回报等问题, 基于累积奖赏来学习多步之前的决策非常困难, 而在现实任务中, 我们往往能够获得一批专家的决策过程示例, 由此可使强化学习模型直接模仿专家的示例轨迹来缓解前述困难, 这一方法即为模仿学习. 根据在强化学习框架下所 “模仿”的对象, 可进一步将模仿学习划分为直接模仿学习、逆强化学习两类[20, 91–93].

直接模仿学习中, 首先获取到专家的 “状态–动作对”示例数据, 然后采用监督学习方式来学得符合专家决策轨迹的策略模型. DeepMind 团队的 AlphaStar[49] 首先针对人类玩家中排名前 22% 的玩家获取到百万规模的对战数据集, 采用监督学习方式

对策略网络进行预训练, 此后再采用强化学习和联盟学习方式进行策略提升和进化. 文献 [94] 采用层次化学习架构来研究 5V5 的多玩家在线对战竞技 (Multiplayer online battle arena, MOBA) 游戏, 定义了 “对战阶段”、“注意力”两层宏观策略和 “行为执行”一层微观操作, 并采用监督学习方式分别学习宏观策略和微观操作. 前述针对电竞游戏的研究能较便捷地获取到大规模先验数据集, 与此不同, 实际物理环境下的无人集群应用场景往往缺乏人类经验或先验数据, 但可能存在许多基于先验模型或解析算法的知识类模型. 为此, 文献 [95] 针对多智能体编队和避碰问题, 分别采用一致性编队协议和最优互补避碰 (Optimal reciprocal collision avoidance, ORCA) 算法设计知识驱动型编队和避碰算法, 并利用该算法产生示例数据, 进一步基于该示例数据采用模仿学习方式训练初始值网络, 为后续强化学习提供初始网络参数, 这种由 “模仿人类”改为 “模仿算法”的思想很有借鉴意义.

与直接模仿学习从示例数据中直接学习行为策略不同, 逆强化学习[96] 的思想是从专家示例中学习回报函数, 这在专家示例数据较少时表现出更好的问题抽象能力和泛化性能. 文献 [97−98] 对逆强化学习进行了综述, 根据是否人为指定回报函数的形式, 将逆强化学习分为两类: 一类是人为指定回报函数形式的传统方法, 主要包括学徒学习方法、最大边际规划算法、结构化分类方法以及基于最大熵、交叉熵等概率模型形式化表达方法; 另一类方法为深度逆强化学习方法, 即为了克服大规模问题中人为指定特征函数表现能力不足、只能覆盖部分回报函数解空间等问题, 采用深度神经网络来设计回报函数学习模型[99−100]. 与前述完全从专家正向示例样本中学习不同, 文献 [101] 介绍了一种能同时学习正向样本和负向样本数据的机器人自主导航学习框架, 正向样本告诉机器人应该怎么做, 而负向样本教会机器人不应该怎么做, 与单纯采用正向样本的方法相比, 在机器人避碰成功率等方面得到了提升. 在多智能体场景中, 平衡解的非唯一性意味着同一个平衡策略可能对应多个逆模型, 这为多智能体逆强化学习的研究带来了挑战. 文献 [102] 将单智能体逆强化学习[96] 拓展到多智能体领域, 并将环境建模为一个一般和随机博弈过程, 以分布式方式来求取智能体各自的策略; 文献 [103] 则针对双人零和博弈问题, 采用贝叶斯方法来建模回报函数, 即首先为回报函数分配一个先验分布, 再基于观察到的策略从后验分布中生成回报函数的点估计.

2) 启发式回报函数设计

在强化学习中, 许多问题存在奖励稀疏或延迟等问题, 恰当的回报函数设计是算法优异表现的关键. 鉴于回报函数设计复杂, 利用各种先验知识来优化奖励信号的启发式回报函数设计方法[104−105] 成为一大类重要的知识与数据协同驱动方法. 事实上, 前述逆强化学习正是一种启发式回报函数设计的特殊形式, 其特别之处在于是从专家示例数据中去学得回报函数, 因此, 本部分介绍除逆强化学习之外的启发式回报函数设计方法.

启发式回报函数设计的第 1 种通用方法是直接利用经验或先验知识来设计回报函数. 例如, 文献 [106] 针对多智能体协同区域覆盖与网络连通保持这一复合任务, 在回报函数设计中充分运用了先验知识: 在区域覆盖子任务中计算覆盖率作为奖惩因素, 在网络连通保持子任务中计算代数连通度来作为连通性奖惩因素, 最终实现了复杂任务的知识引导学习. 文献 [107] 针对无人车车道变换问题设计了基于深度 Q 网络 (Deep Q-network, DQN) 的自主决策模型, 在回报函数中综合考虑了车道变换的安全性和驾驶速度等因素. 文献 [108] 则基于控制论思想, 采用被控量误差绝对值的累加和作为回报函数来调节基于 DRL 的控制器.

启发式回报函数设计的第 2 种方法是引入附加回报函数. 为表述清晰, 在此对一个 MDP 问题 M 进行五元组定义表示, 即 $\langle S, A, R, T, \gamma\rangle$, 五个变量分别表示环境状态集合、动作集合、奖赏函数、状态转移函数和折扣因子. 在附加回报函数设计中, 为了对决策过程进行引导, 在原 MDP 问题 M 的回报函数 R 上叠加一个附加回报函数 F, 构成新的 MDP 问题 M', 其回报函数为 $R' = R + F$. 特别地, Ng 等[109] 证明可将附加回报函数设计为某个势函数关于相邻两个状态的差分形式而不是仅与当前状态相关, 即

$$F(s, s') = \gamma\phi(s') - \phi(s) \tag{1}$$

其中, $s, s' \in S$ 表示当前及下一时刻状态, $\phi(\cdot)$ 为需要设计的势函数, 从而有利于维持从 M 到 M' 的策略不变性. 文献 [110] 进一步从理论上证明了这一策略不变性结论. 基于上述势函数, 可将附加回报函数 F 的设计转化为势函数 $\phi(s)$ 的设计, 而势函数则可基于先验知识进行设计, 例如选为状态 s 与目标或者子目标之间广义距离的相反数, 进而产生一个 “势场”的吸引作用[111]. 进一步, 文献 [112] 将附加回报函数从单纯依赖状态空间拓展到依赖状态-动作联合空间, 即

$$F(s, a, s', a') = \gamma\phi(s', a') - \phi(s, a) \tag{2}$$

其中, $a, a' \in A$ 表示当前时刻及下一时刻选取的动作, 这样构成基于势函数的建议, 即鼓励智能体在

某一状态下采取某一特定动作; 文献 [113] 则将文献 [109] 中的原始势函数推广为动态势函数, 即在势函数中显式增加了时间变量, 并证明仍然能保持策略的不变性.

$$F(s, t, s', t') = \gamma\phi(s', t') - \phi(s, t) \tag{3}$$

结合上述基于势函数的建议和动态势函数, 文献 [114] 证明可将任意奖励函数转化为基于势函数的动态建议.

大部分强化学习的奖励信号都是通过环境给定的外在奖励, 事实上学习的收益还有可能来源于内在奖励 (Intrinsic reward), 例如智能体的好奇心以及对于内部信息的反应[115]. 文献 [116] 即给出了一个形象的例子说明, 单纯依赖外部奖励可能会遗漏智能体内部的重要信息, 而增加内部奖励则可能提升智能体的性能表现; 在大量稀疏奖励问题中, 如何使智能体经过有效探索以最快速度获得外部奖励, 是强化学习研究的热点问题, 为此, 文献 [117] 提出了一种新的基于内在奖励的强化学习探索准则: BeBold, 能够使智能体在不知道具体环境语义的情况下以一种普适准则快速地适应各种环境, 训练出有效策略; 更进一步, 文献 [118] 研究如何在完全没有外部奖励的环境下通过内在奖励实现智能体的训练, 并在 54 个基准环境下进行测试, 验证了这一方法的有效性. 在知识与数据协同驱动的框架内, 上述内在奖励可以通过知识引导的方式设计, 也可以通过数据驱动的方式来自动寻优[116, 119].

3) 知识的网络化表示

知识和数据协同驱动的另一种方法是将知识展开成数据化表示, 特别是采用神经网络来进行表示, 从而形成一种特殊形式的知识嵌套网络, 该网络的结构、参数等将体现领域或专家知识的特点, 进一步可将该网络嵌入到更大的神经网络中进行统一训练学习, 概念模型如图 6 所示. 例如, Xu等[19] 提出一种将知识驱动和数据驱动相结合的框架, 该框架首先根据问题物理机理、先验知识等建立一个具有若干未知参数的模型族, 然后基于数据驱动算法设计算法族, 对模型族中的未知参数寻优, 最后将整个模型展开为深度网络以实施深度学习, 该架构对知识与数据的深度集成具有很好的启发意义. 事实上, 这种将某一模型算法展开成神经网络进行统一训练的思想很早便得到关注. 例如, 模糊神经网络[120–121] 便是将模糊推理的隶属度函数计算、模糊规则推理等过程展开成神经网络表示, 随后采用训练的方式实现模糊推理前后件参数规则的寻优; 又如, PID 神经网络[122] 将控制中应用最广泛的 PID 控制器展开成神经网络表示, 随后采用网络训练方式来寻优

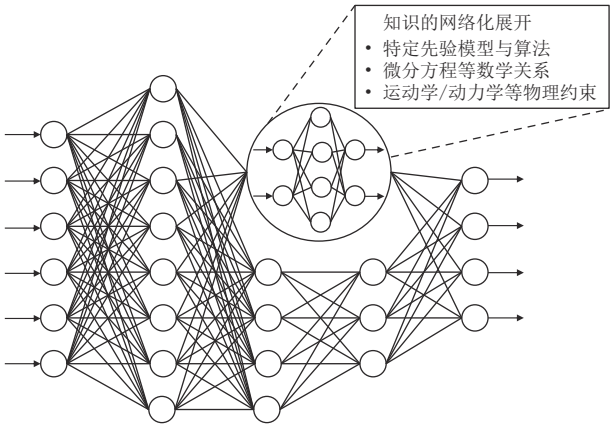


图 6 知识的网络化展开概念模型

Fig. 6 Conceptual networking expansion of knowledge

控制参数. 除了将具体的模型或算法展开为神经网络表示外, 还可以将某些数学方程展开为网络表达, 例如利用神经网络来表示非线性偏微分方程约束[123] 或直接求解偏微分方程[124].

除了将解析模型/算法或数学关系进行神经网络展开外, 针对某些实际物理系统, 还可将物理约束进行网络化展开. 例如, 针对真实机器人所受的动力学等物理约束, 文献 [123] 提出一种新颖的深度拉格朗日网络 (Deep Lagrangian networks, DeLaN), 即将物理对象的拉格朗日动力学模型表示成神经网络形式, 进一步采用深度网络的训练方式实现学习, 从而在利用深度学习高效计算的同时保证物理约束. 文献 [125] 也提出采用神经网络来表示机器人机理模型, 并验证了该模型在表示 7 自由度机械臂逆向动力学模型时, 具有比传统前馈神经网络和循环神经网络更好的表示精度和泛化性能. 文献 [126] 提出将复杂、动态系统采用图神经网络来表示, 例如机器人的身体和关节可分别用图模型中的节点和边来表示, 从而采用一种统一的网络方式实现模型的表征. 而图神经网络[127] 在表征多智能体系统时具有更加直观的意义, 结合注意力机制, 图注意力网络[128] 可有效地提取智能体之间的隐藏时空特征关系, 从而为多智能体协同决策提供特征输入.

除了上述三种方法外, 知识增强的数据驱动还有许多路径选择. 例如, 基于模型的强化学习便是一大类方法, 其本质是对 MDP 模型 M 中状态转移函数 T 的处理和运用, 通常是采用神经网络等模型对环境 (即状态转移概率) 进行建模, 然后基于该模型来生成用于后期策略训练的数据, 或是直接产生基于优化的预测控制器. 文献 [129] 便采用这样的思路, 基于元学习来使得智能体能够在线自适应地

学到动态变化的环境模型, 从而提升策略的鲁棒性, 在实际物理环境下的验证表明, 算法能使多足机器人在变化的地形条件、姿态估计存在偏差、负载变化、甚至是缺失一条腿的复杂情况下表现出良好的适应性. 此外, 若 T 已知, 另一类通用方法是动态规划[130−131], 由于其内涵过于广泛, 本文不做更进一步展开介绍.

3.2.2　数据调优的知识驱动

数据调优的知识驱动方法总体思想是利用数据驱动方法强大的寻优能力来实现知识驱动方法中结构或参数的优化, 这类方法在感知、决策、控制等领域已几乎无处不在. 例如, 前述的遗传模糊方法, 即是采用进化计算这类数据驱动方法来优化模糊推理这类知识驱动方法中的规则前后件; 控制领域中的自适应控制、优化控制等方法群也大量采用数据驱动方法来实现参数调优. 又如, 文献 [132] 设计了模糊 Q 学习控制器, 采用强化学习方法对模糊控制器参数进行优化. 在集群编队方面, 文献 [133−134] 以基于模型的一致性控制器为主控制器, 采用径向基神经网络方法估计集群编队中的不确定性, 设计了最小参数学习自适应控制算法. 类似地, 文献 [135] 在考虑全状态约束和指定性能的基础上提出了一种事件触发自适应控制算法, 采用反步法构建控制框架, 采用径向基神经网络处理多智能体模型中的非线性函数. 这类方法在基于模型的规划、控制、决策等研究中已经得到广泛关注, 故在此不做展开介绍.

3.2.3　知识与数据的互补结合

在这类方法中, 知识驱动和数据驱动方法没有明显的主次关系, 二者通过不同形式紧密集成. 文献 [21] 系统总结了基于模型的知识驱动方法和基于神经网络的数据驱动方法的不同结合形式, 从架构上主要分为二者并联结合、串联结合两类: 在并联结合中, 知识驱动和数据驱动模型采用相同的输入, 在输出端将二者输出结果进行并联; 在串联结合中, 可将知识驱动模型的输出作为数据驱动模型的输入, 或反过来将数据驱动模型的输出作为知识驱动模型的输入, 文章还框架性地给出了这些结合形式在系统建模、预测、控制等不同问题中的应用. 以控制系统设计为例, 两种结合方式衍生出 3 种常见的系统框架, 如图 7 所示[21].

在框架 A 中, 控制律 u 为

$$u = \mathcal{K}\left(y,\, \mathcal{D},\, \mathcal{M},\, p\right) + \mathcal{N}\left(y,\, w\left(P\left(y,\, u\right)\right)\right) \tag{4}$$

其中, $\mathcal{K}$ 表示知识驱动控制器, 输出为 u_k, $\mathcal{N}$ 表示神经网络, 输出为 u_n, $y = [y_m,\, y_{sp}]$, 其中 y_{sp} 为被控量设定值, y_m 为其测量值, $\mathcal{D}$, $\mathcal{M}$ 分别表示先验知识中的状态模型和输出模型, p 为先验模型参数,

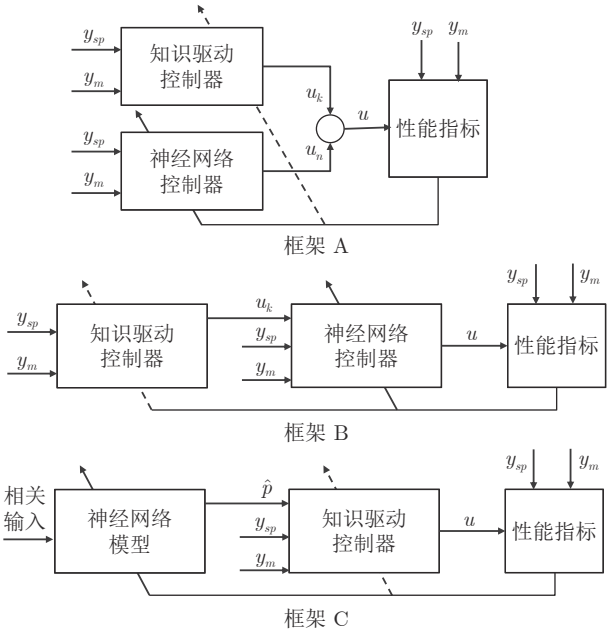


图 7　知识驱动与神经网络互补结合控制框架

Fig. 7　Control diagrams of complementary knowledge-driven and neural network methods

w 为神经网络权重, 其根据性能指标函数 P 调整; 同时, 知识驱动控制器中的参数 p 也可根据 P 调整.

类似地, 框架 B 中的控制律可表示为

$$u = \mathcal{N}\left(\mathcal{K}\left(y,\, \mathcal{D},\, \mathcal{M},\, p\right),\, y,\, w\left(P\left(y,\, u\right)\right)\right) \tag{5}$$

框架 C 中的控制律可表示为

$$u = \mathcal{K}\left(y,\, \mathcal{D},\, \mathcal{M},\, \mathcal{N}\left(I,\, w\left(P\left(y,\, u\right)\right)\right)\right) \tag{6}$$

其中, I 为神经网络模型的相关输入. 这些不同的结合形式体现出不同的实际意义, 例如, 在框架 A 中, 往往采用数据驱动模型构建不确定性补偿模型, 从而实现算法的优化和鲁棒增强[136]; 在框架 B 中, 可采用神经网络估计系统逆向动力学模型, 然后采用知识驱动模型加以控制[137]; 在框架 C 中, 神经网络的作用则是估计知识驱动控制器中的参数 p [134].

除了神经网络外, 强化学习也被用于与知识驱动方法形成互补结合. 例如, 文献 [138] 采用 Q-learning 构成补偿控制器, 与基于模型的基准控制器一起工作, 实现了四旋翼无人机的稳定控制; 类似地, 文献 [108] 采用二型模糊方法构成基准控制器, 采用基于深度确定性策略梯度 (Deep deterministic policy gradient, DDPG) 的强化学习方法构成互补控制器, 实现了电网调节频率的控制. 在串联结合方式中, 文献 [139] 在策略学习框架中增加了一个盾牌 (Shield), 用来监督所学习的动作是否安全合理, 具体结合方式有两种, 一是智能体做决策时, 直接从盾牌中获取一个安全行为, 二是监督智能体的

学习, 一旦出现非安全行为时盾牌将加以动作修正; 文献 [140] 在 MOBA 类游戏中也采取了类似的思想, 采用一个动作掩码 (Mask) 来对强化学习的探索过程进行剪枝, 而掩码的设计则继承了有经验的人类玩家的先验知识. 当然, 无论是盾牌法还是动作掩码法, 其知识驱动部分仅作为数据驱动部分的一个组件, 仍体现出一定主次性, 应归为前述知识增强的数据驱动方法一类, 在此介绍主要是体现其串联结合的特性.

4 几个重要的研究方向

无论从群体智能系统这一应用主体还是深度学习、强化学习这类方法主体来看, 当前都已逐步走向应用问题具象化、多领域概念深度融合的发展阶段, 从理论进一步深化、应用进一步落地等角度来看, 以下几个方面将是未来重要的发展方向.

1) 多学科融合视角下的群体智能机理研究. 如前所述, 当前, "群体智能"这一概念尚未形成统一认识, 不同学者从不同的学科视角出发展开了丰富的研究. 未来的重点方向之一势必是打破这样的学科壁垒, 建立多学科融合的群体智能统一话语体系, 汲取不同学科所包含的理论工具、研究路径等知识内核, 形成更高层次和水平、具有更丰富路径选择的知识与数据协同体系. 这方面已逐步引起关注, 如[141−142] 从博弈论和人工智能等不同角度探讨了多智能体学习的问题, 但仍未形成完善的理论方法体系.

2) 知识与数据协同框架的理论分析. 传统基于数学/物理模型的知识驱动方法往往具有理论支撑较完备的特点, 但当融合数据驱动模式后, 如何开展整个协同框架的理论分析, 是实现安全、可信任人工智能的关键. 例如, 在融入实际物理模型稳定性、正定性等特性以及等式、不等式、动力学等约束后, 如何设计能表征上述特性和约束的神经网络模型 (网络结构、激活函数形式等) 以及如何开展受限网络的学习律设计和理论分析, 是值得研究的重要理论方向.

3) 群体系统智能决策的可解释性研究. 对于无人集群系统这样的实际物理系统, 可解释性显得尤为重要. 在机器学习领域, 可解释性描述一个算法模型输出结果能为人们所理解的程度[143]. 传统机器学习的可解释性研究主要包括两条路径: 一是建立本身易于解释的模型; 二是对建立好的数据驱动模型采用可解释性方法进行解释, 即模型无关的可解释性. 但针对群体系统, 这里的可解释性多了另一层含义, 即群体由于自组织特性所产生的涌现行为可解释性. 因此, 如何统筹考虑数据驱动模型的黑箱可解释性和群智行为的涌现可解释性, 是群体智能系统走向实用化的关键.

4) 知识与数据的迭代进化. 以知识来引导产生数据模型, 从数据模型中归纳生成新的知识, 形成知识与数据的交替迭代, 是实现智能系统自主进化的重要路径, 也是实现能被人所理解却又超越人类知识体系的人工智能系统的重要范式. 从知识到数据的方法包括模仿学习以及各种启发式的数据驱动方法, 从数据到知识则包括各种规则学习、对手建模[144] 等方法, 但在决策智能这一当前最具挑战性的问题下, 尤其是针对群体智能系统的智能决策行为, 如何结合实际应用背景形成知识与数据的迭代进化范式, 是极具吸引力的研究方向.

5 结束语

群体智能理论和应用发展方兴未艾, 是新一代人工智能的一个热点研究领域, 但当前存在群智激发汇聚机理不清、对群体智能系统认知有限、高质量训练数据缺乏等问题, 无论对知识驱动还是数据驱动方法都提出了严峻挑战, 因此知识与数据协同驱动将是推进群体智能特别是群智决策研究的重要方法, 也将为实现可引导、可信任、可学习、可进化的群体智能系统提供方法支撑. 本文系统梳理了知识与数据协同驱动的多种方法路径, 并从架构级协同、算法级协同等不同层面进行了方法归类, 最后从理论和应用等发展需求角度提出了几个未来重点发展方向, 以期为相关领域的研究提供必要借鉴.

References

1 Li W, Wu W J, Wang H M, Cheng X Q, Chen H J, Zhou Z H, et al. Crowd intelligence in AI 2.0 era. *Frontiers of Information Technology and Electronic Engineering*, 2017, **18**(1): 15−43

2 Chung S J, Paranjape A A, Dames P, Shen S J, Kumar V. A survey on aerial swarm robotics. *IEEE Transactions on Robotics*, 2018, **34**(4): 837−855

3 Du Yong-Hao, Xing Li-Ning, Cai Zhao-Quan. Survey on intelligent scheduling technologies for unmanned flying craft clusters. *Acta Automatica Sinica*, 2020, **46**(2): 222−241
(杜永浩, 邢立宁, 蔡昭权. 无人飞行器集群智能调度技术综述. 自动化学报, 2020, **46**(2): 222−241)

4 Nguyen T T, Nguyen N D, Nahavandi S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Transactions on Cybernetics*, 2020, **50**(9): 3826−3839

5 Sun Chang-Yin, Mu Chao-Xu. Important scientific problems of multi-agent deep reinforcement learning. *Acta Automatica Sinica*, 2020, **46**(7): 1301−1312
(孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题. 自动化学报, 2020, **46**(7): 1301−1312)

6 He F J, Pan Y D, Lin Q K, Miao X L, Chen Z G. Collective intelligence: A taxonomy and survey. *IEEE Access*, 2019, **7**: 170213−170225

7 Krause J, Ruxton G D, Krause S. Swarm intelligence in animals and humans. *Trends in Ecology and Evolution*, 2010, **25**(1): 28−34

8 Wu T, Zhou P, Liu K, Yuan Y L, Wang X M, Huang H W, et al. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. *IEEE Transactions on Vehicular Technology*, 2020, **69**(8): 8243−8256

9 Chen Jie, Fang Hao, Xin Bin. *Cooperative Flocking Control of Multi-Agent Systems*. Beijing: Science Press, 2017.
(陈杰, 方浩, 辛斌. 多智能体系统的协同群集运动控制. 北京: 科学出版社, 2017.)

10 Zhu B, Zaini A H B, Xie L H. Distributed guidance for interception by using multiple rotary-wing unmanned aerial vehicles. *IEEE Transactions on Industrial Electronics*, 2017, **64**(7): 5648−5656

11 Qin J H, Gao H J, Zheng W X. Exponential synchronization of complex networks of linear systems and nonlinear oscillators: A unified analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(3): 510−521

12 Wang Xiang-Ke, Liu Zhi-Hong, Cong Yi-Rui, Li Jie, Chen Hao. Miniature fixed-wing UAV swarms: Review and outlook. *Acta Aeronautica et Astronautica Sinica*, 2020, **41**(4): 023732
(王祥科, 刘志宏, 丛一睿, 李杰, 陈浩. 小型固定翼无人机集群综述和未来发展. 航空学报, 2020, **41**(4): 023732)

13 Duan Hai-Bin, Qiu Hua-Xin. *Unmanned Aerial Vehicle Swarm Autonomous Control Based on Swarm Intelligence*. Beijing: Science Press, 2018.
(段海滨, 邱华鑫. 基于群体智能的无人机集群自主控制. 北京: 科学出版社, 2018.)

14 Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. *Nature*, 1998, **393**(6684): 440−442

15 Barabasi A L, Albert R. Emergence of scaling in random networks. *Science*, 1999, **286**(5439): 509−512

16 Su Q, McAvoy A, Wang L, Nowak M A. Evolutionary dynamics with game transitions. *Proceedings of the National Academy of Sciences of the United States of America*, 2019, **116**(51): 25398−25404

17 Xing Li-Ning, Chen Ying-Wu. Research progress on intelligent optimization guidance approaches using knowledge. *Acta Automatica Sinica*, 2011, **37**(11): 1285−1289
(邢立宁, 陈英武. 基于知识的智能优化引导方法研究进展. 自动化学报, 2011, **37**(11): 1285−1289)

18 Xu J X, Hou Z S. Notes on data-driven system approaches. *Acta Automatica Sinica*, 2009, **35**(6): 668−675

19 Xu Z B, Sun J. Model-driven deep-learning. *National Science Review*, 2018, **5**(1): 22−24

20 Li Chen-Xi, Cao Lei, Zhang Yong-Liang, Chen Xi-Liang, Zhou Yu-Huan, Duan Li-Wen. Knowledge-based deep reinforcement learning: A review. *Systems Engineering and Electronics*, 2017, **39**(11): 2603−2613
(李晨溪, 曹雷, 张永亮, 陈希亮, 周宇欢, 段理文. 基于知识的深度强化学习研究综述. 系统工程与电子技术, 2017, **39**(11): 2603−2613)

21 Agarwal M. Combining neural and conventional paradigms for modelling, prediction and control. *International Journal of Systems Science*, 1997, **28**(1): 65−81

22 Hsiao Y T, Lee W P, Yang W, Muller S, Flamm C, Hofacker I, et al. Practical guidelines for incorporating knowledge-based and data-driven strategies into the inference of gene regulatory networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016, **13**(1): 64−75

23 Zhang J, Xiao W D, Li Y J. Data and knowledge twin driven integration for large-scale device-free localization. *IEEE Internet of Things Journal*, 2021, **8**(1): 320−331

24 Reynolds C W. Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH Computer Graphics*, 1987, **21**(4): 25−34

25 Vicsek T, Czirok A, Ben-Jacob E, Cohen I, Shochet O. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 1995, **75**(6): 1226−1229

26 Couzin I D, Krause J, James R, Ruxton G D, Franks N R. Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology*, 2002, **218**(1): 1−11

27 Cucker F, Smale S. Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 2007, **52**(5): 852−862

28 Frazzoli E, Dahleh M A, Feron E. Real-time motion planning for agile autonomous vehicles. *Journal of Guidance, Control, and Dynamics*, 2002, **25**(1): 116−129

29 Choi H L, Brunet L, How J P. Consensus-based decentralized auctions for robust task allocation. *IEEE Transactions on Robotics*, 2009, **25**(4): 912−926

30 Sui Z Z, Pu Z Q, Yi J Q. Optimal UAVs formation transformation strategy based on task assignment and particle swarm optimization. In: Proceedings of the 2017 IEEE International Conference on Mechatronics and Automation (ICMA). Takamatsu, Japan: IEEE, 2017. 1804−1809

31 Huang J. The cooperative output regulation problem of discrete-time linear multi-agent systems by the adaptive distributed observer. *IEEE Transactions on Automatic Control*, 2017, **62**(4): 1979−1984

32 Jiang H, He H B. Data-driven distributed output consensus control for partially observable multiagent systems. *IEEE Transactions on Cybernetics*, 2019, **49**(3): 848−858

33 Tian B L, Lu H C, Zuo Z Y, Yang W. Fixed-time leader-follower output feedback consensus for second-order multiagent systems. *IEEE Transactions on Cybernetics*, 2019, **49**(4): 1545−1550

34 Gao F, Chen W S, Li Z W, Li J, Xu B. Neural network-based distributed cooperative learning control for multiagent systems via event-triggered communication. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(2): 407−419

35 Dong X W, Shi Z Y, Lu G, Zhong Y S. Output containment analysis and design for high-order linear time-invariant swarm systems. *International Journal of Robust and Nonlinear Control*, 2015, **25**(6): 900−913

36 Zhang Y H, Sun J, Liang H J, Li H Y. Event-triggered adaptive tracking control for multiagent systems with unknown disturbances. *IEEE Transactions on Cybernetics*, 2020, **50**(3): 890−901

37 Lee M, Tarokh M, Cross M. Fuzzy logic decision making for multi-robot security systems. *Artificial Intelligence Review*, 2010, **34**(2): 177−194

38 Burgin G H, Sidor L B. Rule-Based Air Combat Simulation, NASA Contractor Report 4160, Titan Systems Inc., USA, 1988.

39 Liu Yuan. *War Game and War-Game Deduction*. Beijing: National Defense University Press, 2013.
(刘源. 兵棋与兵棋推演. 北京: 国防大学出版社, 2013.)

40 Gao K Z, Cao Z G, Zhang L, Chen Z H, Han Y Y, Pan Q K. A review on swarm intelligence and evolutionary algorithms for solving flexible job shop scheduling problems. *IEEE/CAA Journal of Automatica Sinica*, 2019, **6**(4): 904−916

41 Huang Gang, Li Jun-Hua. Multi-UAV cooperative target allocation based on AC-DSDE evolutionary algorithm. *Acta Automatica Sinica*, 2021, **47**(1): 173−184
(黄刚, 李军华. 基于 AC-DSDE 进化算法多 UAVs 协同目标分配. 自动化学报, 2021, **47**(1): 173−184)

42 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436−444

43 Littman M L. Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 2015, **521**(7553): 445−451

44 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

45 Silver D, Huang A, Maddison C J, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, **529**(7587): 484−489

46 Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. *Nature*, 2017, **550**(7676): 354−359

47 Schwartz H M. *Multi-agent Machine Learning: A Reinforcement Approach*. Hoboken: Wiley, 2014.

48 Liang Xing-Xing, Feng Yang-He, Ma Yang, Cheng Guang-Quan, Huang Jin-Cai, Wang Qi, et al. Deep multi-agent reinforcement learning: A survey. *Acta Automatica Sinica*, 2020, **46**(12): 2537−2557
(梁星星, 冯旸赫, 马扬, 程光权, 黄金才, 王琦, 等. 多 Agent 深度强化学习综述. 自动化学报, 2020, **46**(12): 2537−2557)

49 Vinyals O, Babuschkin I, Czarnecki W M, Mathieu M, Dudzik A, Chung J, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019, **575**(7782): 350−354

50 Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: MIT Press, 2017. 6382−6393

51 Berner C, Brockman G, Chan B, Cheung V, Debiak P, Dennison C, et al. Dota 2 with large scale deep reinforcement learning. arXiv: 1912.06680, 2019.

52 Yan Yue-Jin, Li Zhou-Jun, Chen Yue-Xin. Multi-agent system architecture. *Computer Science*, 2001, **28**(5): 77−80
(颜跃进, 李舟军, 陈跃新. 多 Agent 系统体系结构. 计算机科学, 2001, **28**(5): 77−80)

53 Brooks R. A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, 1986, **2**(1): 14−23

54 Lawton J R T, Beard R W, Young B J. A decentralized approach to formation maneuvers. *IEEE Transactions on Robotics and Automation*, 2003, **19**(6): 933−941

55 Bratman M E, Israel D J, Pollack M E. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 1988, **4**(3): 349−355

56 Lieto A, Bhatt M, Oltramari A, Vernon D. The role of cognitive architectures in general artificial intelligence. *Cognitive Systems Research*, 2018, **48**: 1−3

57 Li D Y, Ge S S, He W, Li C J, Ma G F. Distributed formation control of multiple Euler-Lagrange systems: A multilayer framework. *IEEE Transactions on Cybernetics*, 2020, DOI: 10.1109/TCYB.2020.3022535

58 Amduka M, Russo J, Jha K, DeHon A, Lethin R, Springer J, et al. The Design of A Polymorphous Cognitive Agent Architecture (PCCA), Final Technical Report AFRL-RI-RS-TR-2008-137, Lockheed Martin Advanced Technology Laboratories, USA, 2008.

59 Keller J. DARPA to develop swarming unmanned vehicles for better military reconnaissance. *Military and Aerospace Electronics*, 2017, **28**(2): 4−6

60 Sun Rui, Wang Zhi-Xue, Jiang Zhi-Ping, Jiang Xin. Analysis of the foreign military command and control process model. *Ship Electronic Engineering*, 2012, **32**(5): 12−14
(孙瑞, 王智学, 姜志平, 蒋鑫. 外军指挥控制过程模型剖析. 舰船电子工程, 2012, **32**(5): 12−14)

61 Fusano A, Sato H, Namatame A. Multi-agent based combat simulation from OODA and network perspective. In: Proceedings of the UkSim 13rd International Conference on Computer Modelling and Simulation. Cambridge, UK: IEEE, 2011. 249−254

62 Huang Y Y. Modeling and simulation method of the emergency response systems based on OODA. *Knowledge-Based Systems*, 2015, **89**: 527−540

63 Zhao Q. Training and retraining of neural network trees. In: Proceedings of the 2001 International Joint Conference on Neural Networks. Washington, USA: IEEE, 2001. 726−731

64 Brent R P. Fast training algorithms for multilayer neural nets. *IEEE Transactions on Neural Networks*, 1991, **2**(3): 346−354

65 Schmitz G P J, Aldrich C, Gouws F S. ANN-DT: An algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks*, 1999, **10**(6): 1392−1401

66 Utkin L V, Zhuk Y A, Zaborovsky V S. An anomalous behavior detection of a robot system by using a hierarchical Siamese neural network. In: Proceedings of the 21st International Conference on Soft Computing and Measurements (SCM). St. Petersburg, Russia: IEEE, 2017. 630−634

67 Bromley J, Bentz J W, Bottou L, Guyon I, Lecun Y, Moore C, et al. Signature verification using a "Siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 1993, **7**(4): 669−688

68 Calvo R, Figueiredo M. Reinforcement learning for hierarchical and modular neural network in autonomous robot navigation. In: Proceedings of the 2003 International Joint Conference on Neural Networks. Portland, USA: IEEE, 2003. 1340−1345

69 Roy D, Panda P, Roy K. Tree-CNN: A hierarchical deep convolutional neural network for incremental learning. *Neural Networks*, 2020, **121**: 148−160

70 Zheng Y, Chen Q Y, Fan J P, Gao X B. Hierarchical convolutional neural network via hierarchical cluster validity based visual tree learning. *Neurocomputing*, 2020, **409**: 408−419

71 Yang Y X, Morillo I G, Hospedales T M. Deep neural decision trees. In: Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning. Stockholm, Sweden: ACM, 2018. 34−40

72 Fei H, Ren Y F, Ji D H. A tree-based neural network model for biomedical event trigger detection. *Information Sciences*, 2020, **512**: 175−185

73 Ren X M, Gu H X, Wei W T. Tree-RNN: Tree structural recurrent neural network for network traffic classification. *Expert Systems with Applications*, 2021, **167**: 114363

74 Ernest N D. Genetic Fuzzy Trees for Intelligent Control of Unmanned Combat Aerial Vehicles [Ph.D. dissertation], University of Cincinnati, USA, 2015.

75 Ernest N, Carroll D, Schumacher C, Clark M, Cohen K, Lee G. Genetic fuzzy based artificial intelligence for unmanned combat aerial vehicle control in simulated air combat missions. *Journal of Defense Management*, 2016, **6**(1): 1000144

76 Ernest N, Cohen K, Kivelevitch E, Schumacher C, Casbeer D. Genetic fuzzy trees and their application towards autonomous training and control of a squadron of unmanned combat aerial vehicles. *Unmanned Systems*, 2015, **3**(3): 185−204

77 Kang Y M, Pu Z Q, Liu Z, Li G, Niu R Y, Yi J Q. Air-to-air combat tactical decision method based on SIRMs fuzzy logic and improved genetic algorithm. In: Proceedings of the 2020 International Conference on Guidance, Navigation and Control. Tianjin, China: Springer, 2020. 3699−3709

78 Botvinick M M. Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, 2012, **22**(6): 956−962

79 Bradtke S J, Duff M O. Reinforcement learning methods for

continuous-time Markov decision problems. In: Proceedings of the 7th International Conference on Neural Information Processing Systems. Denver, USA: ACM, 1994. 393−400

80 Dayan P, Hinton G E. Feudal reinforcement learning. In: Proceedings of the 5th International Conference on Neural Information Processing Systems. Denver, USA: ACM, 1993. 271−278

81 Sutton R S, Precup D, Singh S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 1999, **112**(1–2): 181−211

82 Parr R, Russell S. Reinforcement learning with hierarchies of machines. In: Proceedings of the 10th International Conference on Neural Information Processing Systems. Denver, USA: ACM, 1997. 1043−1049

83 Dietterich T G. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, 2000, **13**: 227−303

84 Jaderberg M, Czarnecki W M, Dunning I, Marris L, Lever G, Castaneda A G, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 2019, **364**(6443): 859−865

85 Yang J C, Borovikov I, Zha H Y. Hierarchical cooperative multi-agent reinforcement learning with skill discovery. In: Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems. Auckland, New Zealand: ACM, 2020. 1566−1574

86 Tang H Y, Hao J Y, Lv T J, Chen Y F, Zhang Z Z, Jia H T, et al. Hierarchical deep multiagent reinforcement learning with temporal abstraction. arXiv: 1809.09332, 2019.

87 Zheng Yan-Bin, Li Bo, An De-Yu, Li Na. Multi-agent path planning algorithm based on hierarchical reinforcement learning and artificial potential field. *Journal of Computer Applications*, 2015, **35**(12): 3491−3496
(郑延斌, 李波, 安德宇, 李娜. 基于分层强化学习及人工势场的多 Agent 路径规划方法. 计算机应用, 2015, **35**(12): 3491−3496)

88 Wang Chong, Jing Ning, Li Jun, Wang Jun, Chen Hao. An algorithm of cooperative multiple satellites mission planning based on multi-agent reinforcement learning. *Journal of National University of Defense Technology*, 2011, **33**(1): 53−58
(王冲, 景宁, 李军, 王钧, 陈浩. 一种基于多 Agent 强化学习的多星协同任务规划算法. 国防科技大学学报, 2011, **33**(1): 53−58)

89 Pierre B L, Harb J, Precup D. The option-critic architecture. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2017. 1726−1734

90 Vezhnevets A S, Osindero S, Schaul T, Heess N, Jaderberg M, Silver D, et al. Feudal networks for hierarchical reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 3540−3549

91 Piot B, Geist M, Pietquin O. Bridging the Gap between imitation learning and inverse reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**(8): 1814−1826

92 Zhou Zhi-Hua. Machine Learning. Beijing: Tsinghua University Press, 2016. 390–393
(周志华. 机器学习. 北京: 清华大学出版社, 2016. 390-393)

93 Argall B D, Chernova S, Veloso M, Browning B. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 2009, **57**(5): 469−483

94 Wu B. Hierarchical macro strategy model for MOBA game AI. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI, 2019. 1206−1213

95 Sui Z Z, Pu Z Q, Yi J Q, Wu S G. Formation control with collision avoidance through deep reinforcement learning using model-guided demonstration. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, **32**(6): 2358−2372

96 Ng A Y, Russell S J. Algorithms for inverse reinforcement learning. In: Proceedings of the 17th International Conference on Machine Learning. Stanford, USA: ACM, 2000. 663−670

97 Shao Z F, Er M J. A review of inverse reinforcement learning theory and recent advances. In: Proceedings of the 2012 IEEE Congress on Evolutionary Computation. Brisbane, Australia: IEEE, 2012. 1−8

98 Chen Xi-Liang, Cao Lei, He Ming, Li Chen-Xi, Xu Zhi-Xiong. Overview of deep inverse reinforcement learning. *Computer Engineering and Applications*, 2018, **54**(5): 24−35
(陈希亮, 曹雷, 何明, 李晨溪, 徐志雄. 深度逆向强化学习研究综述. 计算机工程与应用, 2018, **54**(5): 24−35)

99 Finn C, Levine S, Abbeel P. Guided cost learning: Deep inverse optimal control via policy optimization. In: Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR, 2016. 49−58

100 Wulfmeier M, Ondruska P, Posner I. Maximum entropy deep inverse reinforcement learning. arXiv: 1507.04888, 2015.

101 Choi S, Kim E, Lee K, Oh S. Leveraged non-stationary Gaussian process regression for autonomous robot navigation. In: Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA). Seattle, USA: IEEE, 2015. 473−478

102 Reddy T S, Gopikrishna V, Zaruba G, Huber M. Inverse reinforcement learning for decentralized non-cooperative multiagent systems. In: Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Seoul, Korea (South): IEEE, 2012. 1930−1935

103 Lin X M, Beling P A, Cogill R. Multiagent inverse reinforcement learning for two-person zero-sum games. *IEEE Transactions on Games*, 2018, **10**(1): 56−68

104 Wang Xue-Song, Zhu Mei-Qiang, Cheng Yu-Hu. *Principle and Applications of Reinforcement Learning*. Beijing: Science Press, 2014. 15−16
(王雪松, 朱美强, 程玉虎. 强化学习原理及其应用. 北京: 科学出版社, 2014. 15-16)

105 Laud A D. Theory and Application of Reward Shaping in Reinforcement Learning [Ph.D. dissertation], University of Illinois, USA, 2004.

106 Wu S G, Pu Z Q, Liu Z, Qiu T H, Yi J Q, Zhang T L. Multi-target coverage with connectivity maintenance using knowledge-incorporated policy framework. In: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA). Xi′an, China: IEEE, 2021. 8772−8778

107 Wang J J, Zhang Q C, Zhao D B, Chen Y R. Lane change decision-making through deep reinforcement learning with rule-based constraints. In: Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN). Budapest, Hungary: IEEE, 2019. 1−6

108 Khooban M H, Gheisarnejad M. A novel deep reinforcement learning controller based type-II fuzzy system: Frequency regulation in microgrids. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021, **5**(4): 689−699

109 Ng A Y, Harada D, Russell S J. Policy invariance under reward transformations: Theory and application to reward shaping. In: Proceedings of the 16th International Conference on Machine Learning (ICML). Bled, Slovenia: ACM, 1999. 278−287

110 Wiewiora E. Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 2003, **19**: 205−208

111 Hussein A, Elyan E, Gaber M M, Jayne C. Deep reward shaping from demonstrations. In: Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN). Anchorage, USA: IEEE, 2017. 510−517

112 Wiewiora E, Cottrell G W, Elkan C. Principled methods for

advising reinforcement learning agents. In: Proceedings of the 20th International Conference on Machine Learning (ICML). Washington, USA: AAAI, 2003. 792−799

113 Devlin S, Kudenko D. Dynamic potential-based reward shaping. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems. Valencia, Spain: ACM, 2012. 433−440

114 Harutyunyan A, Devlin S, Vrancx P, Nowe A. Expressing arbitrary reward functions as potential-based advice. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA: AAAI, 2015. 2652−2658

115 Singh S, Barto A G, Chentanez N. Intrinsically motivated reinforcement learning. In: Proceedings of the 17th International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM, 2004. 1281−1288

116 Singh S, Lewis R L, Barto A. Where do rewards come from? In: Proceedings of the 31st Annual Conference of the Cognitive Science Society. Amsterdam, the Netherlands: Cognitive Science Society, 2009. 2601−2606

117 Zhang T J, Xu H Z, Wang X L, Wu Y, Keutzer K, Gonzalez J E, et al. BeBold: Exploration beyond the boundary of explored regions. arXiv: 2012.08621, 2020.

118 Burda Y, Edwards H, Pathak D, Storkey A J, Darrell T, Efros A A. Large-scale study of curiosity-driven learning. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ACM, 2019. 1−17

119 Yang D, Tang Y H. Adaptive inner-reward shaping in sparse reward games. In: Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, United Kingdom: IEEE, 2020. 1−8

120 Kasabov N. Evolving fuzzy neural networks for supervised/unsupervised online knowledge-based learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2001, **31**(6): 902−918

121 Zhao D B, Yi J Q. GA-based control to swing up an acrobot with limited torque. *Transactions of the Institute of Measurement and Control*, 2006, **28**(1): 3−13

122 Zhou M L, Zhang Q. Hysteresis model of magnetically controlled shape memory alloy based on a PID neural network. *IEEE Transactions on Magnetics*, 2015, **51**(11): 7301504

123 Lutter M, Ritter C, Peters J. Deep Lagrangian networks: Using physics as model prior for deep learning. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ACM, 2019.

124 Raissi M, Perdikaris P, Karniadakis G E. Physics informed deep learning (Part I): Data-driven solutions of nonlinear partial differential equations. arXiv: 1711.10561, 2017.

125 Ledezma F D, Haddadin S. First-order-principles-based constructive network topologies: An application to robot inverse dynamics. In: Proceedings of the 17th International Conference on Humanoid Robotics (Humanoids). Birmingham, UK: IEEE, 2017. 438−445

126 Sanchez-Gonzalez A, Heess N, Springenberg J T, Merel J, Riedmiller M A, Hadsell R, et al. Graph networks as learnable physics engines for inference and control. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 4467−4476

127 Jiang J C, Dun C, Huang T J, Lu Z Q. Graph convolutional reinforcement learning. In: Proceedings of the 2020 International Conference on Learning Representation (ICLR). 2020.

128 Velickovic P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ACM, 2018. 1−12

129 Nagabandi A, Clavera I, Liu S M, Fearing R S, Abbeel P, Levine S, et al. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ACM, 2019. 1−17

130 Liu D R, Xue S, Zhao B, Luo B, Wei Q L. Adaptive dynamic programming for control: A survey and recent advances. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, **51**(1): 142−160

131 Chen Xin, Wei Hai-Jun, Wu Min, Cao Wei-Hua. Tracking learning based on Gaussian regression for multi-agent systems in continuous space. *Acta Automatica Sinica*, 2013, **39**(12): 2021−2031
(陈鑫, 魏海军, 吴敏, 曹卫华. 基于高斯回归的连续空间多智能体跟踪学习. 自动化学报, 2013, **39**(12): 2021−2031)

132 Desouky S F, Schwartz H M. $Q(\lambda)$-learning fuzzy logic controller for a multi-robot system. In: Proceedings of the 2010 IEEE International Conference on Systems, Man and Cybernetics. Istanbul, Turkey: IEEE, 2010. 4075−4080

133 Xiong T Y, Pu Z Q, Yi J Q, Sui Z Z. Adaptive neural network time-varying formation tracking control for multi-agent systems via minimal learning parameter approach. In: Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN). Budapest, Hungary: IEEE, 2019. 1−8

134 Xiong T Y, Pu Z Q, Yi J Q, Tao X L. Fixed-time observer based adaptive neural network time-varying formation tracking control for multi-agent systems via minimal learning parameter approach. *IET Control Theory & Applications*, 2020, **14**(9): 1147−1157

135 Yang Bin, Zhou Qi, Cao Liang, Lu Ren-Quan. Event- triggered control for multi-agent systems with prescribed performance and full state constraints. *Acta Automatica Sinica*, 2019, **45**(8): 1527−1535
(杨彬, 周琪, 曹亮, 鲁仁全. 具有指定性能和全状态约束的多智能体系统事件触发控制. 自动化学报, 2019, **45**(8): 1527−1535)

136 Yu J L, Dong X W, Li Q D, Ren Z. Practical time-varying formation tracking for second-order nonlinear multiagent systems with multiple leaders using adaptive neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(12): 6015−6025

137 Patino D, Carelli R, Kuchen B. Stability analysis of neural networks based adaptive controllers for robot manipulators. In: Proceedings of the 1994 American Control Conference. Baltimore, USA: IEEE, 1994. 609−613

138 Lin X B, Yu Y, Sun C Y. Supplementary reinforcement learning controller designed for quadrotor UAVs. *IEEE Access*, 2019, **7**: 26422−26431

139 Alshiekh M, Bloem R, Ehlers R, Konighofer B, Niekum S, Topcu U. Safe reinforcement learning via shielding. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 2669−2678

140 Ye D H, Liu Z, Sun M F, Shi B, Zhao P L, Wu H, et al. Mastering complex control in MOBA games with deep reinforcement learning. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 6672−6679

141 Shoham Y, Powers R, Grenager T. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 2007, **171**(7): 365−377

142 Tuyls K, Parsons S. What evolutionary game theory tells us about multiagent learning. *Artificial Intelligence*, 2007, **171**(7): 406−416

143 Molnar C. Interpretable machine learning — a guide for making black box models explainable [Online], available: https://christophm.github.io/interpretable-ml-book/, June 24, 2021

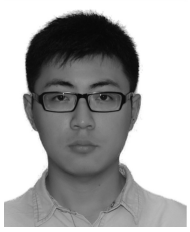
144 Albrecht S V, Stone P. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 2018, **258**: 66−95

蒲志强 中国科学院自动化研究所综合信息系统研究中心副研究员. 2014 年获得中国科学院大学控制理论与控制工程博士学位. 主要研究方向为群体智能, 多智能体强化学习, 无人系统鲁棒自适应控制. 本文通信作者.
E-mail: zhiqiang.pu@ia.ac.cn
(**PU Zhi-Qiang** Associate professor at the Integrated Information System Research Center, Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree in control theory and control engineering from University of Chinese Academy of Sciences, in 2014. His research interest covers collective intelligence, multi-agent reinforcement learning, and robust adaptive control of unmanned systems. Corresponding author of this paper.)

易建强 中国科学院自动化研究所综合信息系统研究中心研究员. 1992 年获得日本九州工业大学自动控制博士学位. 主要研究方向为智能控制, 智能机器人, 自主无人系统.
E-mail: jianqiang.yi@ia.ac.cn
(**YI Jian-Qiang** Professor at the Integrated Information System Research Center, Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree in automation control from the Kyushu Institute of Technology, Kitakyushu, Japan, in 1992. His research interest covers intelligent control, intelligent robotics, and autonomous unmanned systems.)

刘 振 中国科学院自动化研究所综合信息系统研究中心副研究员. 2015 年获得中国科学院大学控制理论与控制工程博士学位. 主要研究方向为飞行控制, 鲁棒自适应控制, 多智能体强化学习.
E-mail: liuzhen@ia.ac.cn
(**LIU Zhen** Associate professor at the Integrated Information System Research Center, Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree in control theory and control engineering from University of Chinese Academy of Sciences, in 2015. His research interest covers flight control, robust adaptive control, and multi-agent reinforcement learning.)

丘腾海 中国科学院自动化研究所综合信息系统研究中心助理研究员. 2016 年获得北京航空航天大学控制理论与控制工程硕士学位. 主要研究方向为智能决策, 多智能体, 自主无人系统应用.
E-mail: tenghai.qiu@ia.ac.cn
(**QIU Teng-Hai** Research assistant at the Integrated Information System Research Center, Institute of Automation, Chinese Academy of Sciences. He received his master degree in control theory and control engineering from Beihang University, in 2016. His research interest covers intelligence decision making, multi-agent, and the applications of unmanned autonomous systems.)



孙金林 江苏大学电气信息工程学院讲师. 主要研究方向为鲁棒与自适应控制, 计算智能, 抗干扰控制.
E-mail: jinlinsun@outlook.com
(**SUN Jin-Lin** Lecturer at the School of Electrical and Information Engineering, Jiangsu University. His research interest covers robust and adaptive control, computational intelligence, and anti-disturbance control.)

李非墨 中国科学院自动化研究所综合信息系统研究中心助理研究员. 2017 年获得中国科学院大学计算机应用技术博士学位. 主要研究方向为遥感图像处理, 计算机视觉, 智能感知. E-mail: lifeimo2012@ia.ac.cn
(**LI Fei-Mo** Research assistant at the Integrated Information System Research Center, Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree in computer applied technology from University of Chinese Academy of Sciences in 2017. His research interest covers remote sensing image processing, computer vision, and intelligent perception.)